Title
Metagenomic and Cultivation-Based Analysis of Novel Microorganisms and Functions in Metal-Contaminated Environments

Author
Yelton, Alexis Pepper

Publication Date
2012

Peer reviewed|Thesis/dissertation

Metagenomic and Cultivation-Based Analysis of Novel Microorganisms and Functions in Metal-Contaminated Environments

By

Alexis Pepper Yelton

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Science, Policy, and Management

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jillian F. Banfield, Chair
Professor Mary K. Firestone
Professor Lisa Alvarez-Cohen

Fall 2012

Abstract

Metagenomic and Cultivation-Based Analysis of Novel Microorganisms and Functions in Metal-Contaminated Environments

by

Alexis Pepper Yelton

Doctor of Philosophy in Environmental Science, Policy, and Management

University of California, Berkeley

Professor Jillian F. Banfield, Chair

Some bacteria and archaea have evolved metabolic strategies that enable them to live in environments contaminated by toxic metals. In fact, many bacteria and archaea take advantage of the redox sensitivity of these very same metals to gain energy via anaerobic respiration. Here, metagenomic techniques were developed and applied along with conventional physiological and ecological methods to elucidate multiple modes of adaptation of bacteria and archaea in metal-contaminated acid mine drainage and aquifer environments contaminated by mine tailings. These approaches provided insight into how these organisms survive and thrive in these environments and how they differentiate themselves from each other.

Many of the microbial species in acid mine drainage and mine tailings-contaminated aquifer environments are difficult to culture in the laboratory. Thus, a focus of the research was metabolic analysis of these organisms via analysis of genes and genomes recovered from microbial communities and isolates. Many of the genes are novel, and likely required for specific environmental adaptation, but they are difficult or impossible to functionally characterize based on conventional homology methods. A new method was developed to deal with the challenge of identifying poorly annotated or hypothetical genes of importance in adaptation to metal-contaminated environments. This probabilistic approach is based on conserved gene order between the genomes of interest with distant relatives.

The annotation method was used in conjunction with traditional comparative genomics to differentiate a group of co-occurring archaea based on their genetic metabolic potential. These microorganisms grow in biofilm communities in an acid mine drainage system within the Richmond Mine, near Redding in Northern California, USA. Microbial biofilms growing at the air-solution interface were sampled from solutions with pH values of < 1.2, temperatures of up to 48 °C, and mM concentrations of zinc, copper, arsenic, and sub-molar concentrations of dissolved iron. We used a metagenomic approach in which DNA was extracted from biofilm samples, sequenced, and analyzed in order to evaluate differences in the metabolic potential of five closely related *Thermoplasmatales* archaea and one distant relative. A subset of these organisms appears to be capable of iron oxidation, whereas others appear to live primarily heterotrophically. Another subset is potentially capable of CO oxidation. There are also major differences in motility within the group.

A metal-contaminated aquifer adjacent to the Colorado River in Colorado, USA, was studied to investigate microorganisms adapted to high vanadium concentrations. A vanadium-reducing *Betaproteobacterium* of the genus *Simplicispira* was isolated (strain BDI). This organism's genome encodes a large number of toxic metal resistance, chemotaxis, motility, and conjugation-related genes that likely allow it to detoxify, avoid contaminants and rapidly adapt in a changing environment. Physiological characterization in the laboratory shows that it is a facultative anaerobic nitrate-reducer capable of reduction of up to 3 mM vanadate.

In order to determine the effect of vanadium contamination on the aquifer community structure, soluble vanadium was added to an in-well, flow-through sediment column. Reduction of dissolved vanadate was documented, along with an increase in the number of cells capable of vanadium reduction, and an increase in the relative abundance of strain BDI. An increase in the relative abundance of three families of known vanadium reducing bacteria (*Commomonadaceae*, *Geobacteraceae*, and *Pseudomonadaceae*) was also noted. This experiment confirmed the environmental importance of BDI, and microbial vanadium reduction in response to acetate addition. Following short-term acetate addition to the aquifer, vanadium remained immobile for at least two years. Because the organisms stimulated by amendment were resident in the aquifer and removal of vanadium from solution persists, the acetate addition approach has significant potential for bioremediation of vanadium.

In summary, this research used culture-based and culture-independent techniques to elucidate microbial metabolisms that allow organisms to colonize metal-contaminated environments. Vanadium reduction was linked to specific subsurface bacteria, one of which was isolated and characterized. The findings have significance in the fields of genomics, metagenomics, microbial ecology and biogeochemistry, and have potential application for bioremediation.

To my parents Carol Chandler and Robert Yelton

For the countless ways they have helped me get here.

And to my friend Corwin Hardham (1974-2012)

May I live my life like he lived his, always trying to make the world a better place with his heart and his work.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I have many people to thank for supporting me in throughout my Ph.D. studies. Firstly, the person who is primarily responsible for guiding me in this process was my advisor, Dr. Jill Banfield. She has not only guided me in my learning and research, but she has also taught me to write, how to interact with editors, and how to apply for grants. She is a wonderful mentor and also a great friend. I hope that I continue to see Jill and share ideas in the years to come.

I would also like to express my gratitude to a number of my colleagues who have collaborated with me or mentored me. Dr. Kenneth Williams, Dr. Kim Handley, Dr. Kelly Wrighton, and Dr. Mike Wilkins were my primary mentors in the laboratory. They taught me much of the details of experimental design in the field and in the laboratory. They took time to listen to me and went through protocols with me in the lab. Dr. Williams is an amazing manager for all of the field experiments going on at the Rifle site. He is always there to help solve problems and somehow manages to coordinate research teams from all over the country. I am particularly indebted to Dr. Wrighton, who taught me most of what I know about bacterial physiology research. Dr. Handley and Dr. Wilkins both gave me wonderful ideas for research directions and helped me design experiments. I also need to thank Brian Thomas and Chris Miller who taught me much of what I know about linux and scripting. Sue Spaulding and Brett Baker provided wonderful laboratory support. Birgit Luef and Luis Comolli did all of the cryogenic electron microscopy and provided me with some wonderful images.

Outside of current and former Banfield lab members, I need to thank Dr. Mary Firestone and Dr. John Coates, both of whom were like second advisors to me. They both allowed the use of equipment in their laboratories for this research and they also lent their ears to hear about my results and difficulties. They offered sound advice when I needed an outside perspective the most.

I also have some people to thank for their personal involvement in my life over the past five years. Without my friends and partners I could not have done this. Specifically I want to thank Samuel Collier, Patrick Oliver, Arianna Vaewsorn, and Anjuli Mahendra. They have been there for me through thick and thin.

Finally I want to thank my family. I would never have made it to UC Berkeley without the support and motivation of my parents. They have been pushing me to do the best I can and encouraging my intellectual interests since I was in elementary school. My sisters Ellis Dillon and Lourie Yelton are rocks. We will always be good friends and I feel so lucky to have such a kind, moderating force in my life.

INTRODUCTION

Metal contamination is a serious problem in industrial and industrializing societies and at locations impacted by weathering of sulfide-rich rocks. Mining of metals can contaminate the environment through production of metal-rich solutions formed by dissolution of minerals in ore and tailings piles (detritus remaining after mining or milling operations). Tailings are usually stored on site and exposed to the air, water and microorganisms. When rainwater permeates tailings piles, it can leach toxic metals into the soils and sediments in the surrounding area, potentially contaminating drinking water and irrigation water.

The acidic, metal-rich solutions generated following mining-related exposure of sulfide-containing rocks to water are referred to as acid mine drainage (AMD). These solutions are enriched in toxic elements such as copper, chromium, zinc, selenium, and arsenic. Microorganisms are essential to generation of acid mine drainage. They oxidize ferrous iron, creating ferric iron that reacts with sulfide mineral surfaces, promoting metal release into solution [1, 2]. This ferric iron can then be reduced in a reaction coupled to sulfide oxidation to sulfuric acid [3]. The sulfuric acid lowers the pH of the mine drainage and results in the flow of toxic metals out of the mine and contamination of surrounding ecosystems.

To better understand these metal-contaminated environments, it is important to understand how the microorganisms that live there interact with the metals and other abiotic factors in these ecosystems. For example, by changing the oxidation state of metals, bacteria and archaea can cause minerals to dissolve or precipitate. This can increase contamination, in the case of acid mine drainage, or result in removal of contaminants from groundwater, in the case of metal bioremediation.

This body of work focuses on two metal-contaminated systems that exemplify these two ways in which microbes interact with contaminants. The first environment is an acid mine drainage system in a former iron, silver, gold, copper, zinc, and pyrite mine (the Richmond Mine) in Northern California, USA. The second system is a shallow aquifer contaminated by tailings from a former uranium and vanadium mill (the Old Rifle Mill) in Northwestern Colorado, USA. At this site, on-going bioremediation experiments have demonstrated the ability of the native microbial community to remove both uranium and vanadium from solution.

The Richmond Mine has previously been studied as a model system for microbial ecology ([4] and references therein) because the microbial communities are typically dominated by a few species [3]. Other features that make it tractable for integrated molecular analyses include high biomass, and the ability to sample well-defined spatial structure over time. Most biofilm samples are dominated by bacteria and contain some archaea [3], but archaea dominate the community under some conditions [5]. Bacteria in general have been more extensively studied than archaea. However, bacteria in AMD and bioleaching systems remained genomically and metabolically poorly understood until cultivation-independent metagenomic methods were introduced. In the case of one important bacterial species, *Leptospirillum ferrodiazotrophum*, metagenomic analysis guided development of a cultivation strategy that led to isolation [6]. Similarly there are few isolated AMD archaea and thus their physiology and ecosystem functions remain unknown, with the notable exception of *Ferroplasma acidarmanus*, which was isolated from the Richmond Mine and

subsequently characterized extensively in the laboratory [7-12]. As with *Leptospirillum*, metagenomics provides a route to investigate the functions of the AMD archaea. However, the lack of well-characterized isolate archaea results in a large number of poorly annotated genes and hypothetical proteins, making metabolic prediction challenging.

We sought to address the problem of prediction of gene function by developing a new method for functional annotation based on synteny (conserved gene order) in addition to homology. The new method was then applied to the newly assembled genomes of four uncultivated archaea as well as two previously sequenced genomes from the Richmond Mine.

The application of this method to compare and contrast AMD archaea gives insight into the ways in which these organisms interact with their environment, how they deal with high concentrations of dissolved metals and how they function in general. It also provides information on how these organisms differentiate themselves enough to co-occur in the same environment. This is of particular interest for the Richmond Mine AMD archaea because many of them are close relatives in the order *Thermoplasmatales*, comprised of mesophilic and moderately thermophilic acidophiles. They are named "*plasmatales*" because they do not have cell walls, though some have surface layer proteins, notably the species of *Picrophilus*. Many of them are capable of the iron oxidation that contributes to AMD generation, but also grow organotrophically. Because of the ability of these organisms to free iron from pyrite and thus contribute to mineral dissolution, there is potential for bioleaching applications using *Thermoplasmatales* and related archaea. Bioleaching is a technology that enhances extraction of metals from minerals through dissolution by living organisms. In fact, several *Thermoplasmatales* organisms have been isolated from bioleaching operations [13].

In contrast to bioleaching, bioremediation of mining wastes generally relies on biotic immobilization of toxic metals. Extensive research has been carried out on the potential for the bioremediation of uranium via biostimulation with organic carbon [14, 15]. This type of stimulation can lead to microbial reduction of U(VI) to U(IV), which is less soluble and precipitates as uraninite [14]. Research on uranium bioreduction indicates that direct enzymatic reduction is possible [15] via metal reductases. Biostimulation with ethanol, glucose, or acetate results in U(V) reduction [16, 17] by stimulating iron reducers, notably *Geobacter*, and sulfate-reducing bacteria. Biostimulation with acetate to induce U reduction has also been shown to remove vanadium (V) from solution [18], suggesting potential for vanadium bioremediation applications. Only four V-reducing subsurface bacteria have been isolated [18-22]. However, it is not known which V-reducers contribute to V-removal in subsurface systems. Finding these organisms and assessing their environmental ubiquity would aid in the evaluation of V-bioremediation potential along with the estimation of vanadium reduction rates.

In order to address these questions both culturing and culture-independent approaches were employed. A V-reducer was isolated from a vanadium-contaminated aquifer, the Old Rifle Mill site, in Rifle, Colorado, USA. The organism's physiology was studied in the laboratory (Chapter 3). Its genome was sequenced and its genome was examined for clues as to its physiology. The V-reducer's 16S ribosomal RNA sequence was compared to the NCBI nr database to determine the ubiquity of the isolate at the Rifle site and in other contaminated areas. An *in situ* experiment was carried out to evaluate changes in the geochemistry and bacterial community that result from vanadium and acetate addition (Chapter 4).

Overall, this dissertation sheds light on both how bacteria and archaea interact and survive in contaminated mine environments, and how they contribute to or mitigate contamination problems. It presents a new method for gene function annotation that can be used to help understand novel organisms in these environments. This method is then applied

to the interpretation of both AMD archaeal genomes and the genome of a newly isolated subsurface V-reducer. The subsurface organism was studied in detail in the laboratory, and field experiments indicated that it is both ubiquitous and relevant to vanadium removal during bioremediation. Finally, the response of the bacterial community to vanadium and acetate amendment was studied. The results indicate both vanadium removal from groundwater and selection for specific taxa closely related to known vanadium reducers under high vanadium and acetate conditions.

CHAPTER 1.

A Semi-Quantitative, Synteny-Based Method to Improve Functional Predictions for Hypothetical and Poorly Annotated Bacterial and Archaeal Genes

Authors: Alexis P. Yelton[1], Brian C. Thomas[1], Sheri L. Simmons[2,+], Paul Wilmes[2,*], Adam Zemla[3], Michael P. Thelen[3], Nicholas Justice[4], Jillian F. Banfield[1,2]

[1]Department of Environmental Science, Policy, and Management, University of California, Berkeley, California 94720, USA

[2] Department of Earth and Planetary Sciences, University of California, Berkeley, California 94720, USA

[+] Current address: Josephine Bay Paul Center for Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole MA 02543, USA

[*] Current address: Department of Environment and Agro-Biotechnologies, Centre de Recherche Public – Gabriel Lippmann, Belvaux, Grand-Duchy of Luxembourg

[3] Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California, USA

[4] Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA

This chapter was previously published in PLoS Computational Biology with the following reference:

Yelton AP, Thomas BC, Simmons SL, Wilmes P, Zemla A, *et al*. (2011) A Semi-Quantitative, Synteny-Based Method to Improve Functional Predictions for Hypothetical and Poorly Annotated Bacterial and Archaeal Genes. Plos Computational Biology 7.

Abstract

During microbial evolution, genome rearrangement increases with increasing sequence divergence. If the relationship between synteny and sequence divergence can be modeled, gene clusters in genomes of distantly related organisms exhibiting anomalous synteny can be identified and used to infer functional conservation. We applied the phylogenetic pairwise comparison method to establish and model a strong correlation between synteny and sequence divergence in all 634 available archaeal and bacterial genomes from the NCBI database and four newly assembled genomes of uncultivated archaea from an acid mine drainage (AMD) community. In parallel, we established and modeled the trend between synteny and functional relatedness in the 118 genomes available in the STRING database. By combining these models, we developed a gene functional annotation method that weights evolutionary distance to estimate the probability of functional associations of syntenous proteins between genome pairs. The method was applied to the hypothetical proteins and poorly annotated genes in newly assembled acid mine drainage archaeal genomes to add or improve gene annotations. This is the first method to assign possible functions to poorly annotated genes through quantification of the probability of gene functional relationships based on synteny at a significant evolutionary distance, and has the potential for broad application.

Introduction

Gene function prediction is currently one of the fundamental problems in microbiology [23]. The improvement in DNA sequencing technologies has allowed for the sequencing of hundreds of full bacterial and archaeal genomes. However, in the dataset of full bacterial and archaeal genomes from NCBI, 874,583 genes out of 2,668,809 (~ 33%) are annotated as hypothetical proteins, and 25% of the protein families in the PFAM database have unknown functions [23]. In addition to these un-annotated genes, many of the genes in these databases only have general function predictions or may have incorrect function predictions. Thus, improved protein functional prediction methods are urgently needed.

It has been proposed that correlations between synteny (the conservation of gene order between genomes) and evolutionary distance, in concert with homology, can be used for predicting protein function [24-27]. Synteny has been used to predict the functional interaction of proteins, where interaction is defined as direct physical interaction, the regulation of one protein by the other, membership in a protein complex, or the sharing of a metabolic (or non-metabolic) pathway [26-28]. Various protein function prediction methods make use of synteny, as reviewed by Rogozin *et al*. in 2004 [25, 29-33], but do not consider evolutionary distance between genomes in their predictions. Preservation of synteny over large evolutionary distances should be weighted strongly in gene function prediction because it is likely the result of selection against rearrangements. Huynen and Snel noted the importance of finding the evolutionary distance at which gene order conservation becomes significant [27, 34]. Snel *et al*. simulated random genome shuffling to determine the probability of conserved gene order in a specific number of genomes [35], and Von Mering *et al*. assessed the likelihood of protein relatedness based on the number of times gene order is conserved in the STRING database of genomes [26]. Here, we link the probability of syntenous protein relatedness and evolutionary distance so that we can determine which genomes are distant enough to accurately utilize synteny-based gene annotation. An overview of our method is provided in Figure S1. Our analyses included genomes of coexisting archaea reconstructed from metagenomic sequence from biofilms growing in an extreme acid mine drainage (AMD) environment as well as published genomes. The inclusion of AMD archaea allowed us to apply the method to newly assembled genomes from uncultivated organisms, and to show the utility of the method for comparative genomics and for improving annotations of proteins of unknown function.

Results

Genome rearrangement and genome reconstruction: We reconstructed four new genomes of uncultivated archaea: A-, E-, G-, and Iplasma (archaea; Euryarchaeota; Thermoplasmata; Thermoplasmatales). *Ferroplasma acidarmanus* (Ferroplasma Type I, Fer1) and *Ferroplasma* Type II (Fer2) have previously been described [12, 36, 37]. Only Fer1 has been isolated [12]. The phylogenetic placement of these organisms based on 16S rRNA gene sequences is shown in Figure 1. E- and Gplasma are most closely related, whereas Iplasma is distantly related and may not actually belong to the *Thermoplasmatales* lineage. Data describing the manually curated and binned composite genomes of these archaea are listed in Table 1. Note that the estimates of the sizes of all genomes are similar. We used standard measures to evaluate genome completeness: a full suite of tRNAs, rRNAs, and orthologous marker genes in all genomes [38]. All of the genomes of the AMD *Thermoplasmatales* organisms except for Aplasma are near complete, according to our analysis (Table S1).

Evolution: In order to carry out regression analysis on genome rearrangements and evolutionary distance, we used gene order conservation (GOC) as a measure of whole genome rearrangement. This metric is described by Rocha [39]. Figure 2 shows the relationship between GOC and evolutionary distance as measured by average normalized BLASTP bit score, a proxy for evolutionary distance.

Figure 2 includes results for genomes reconstructed for uncultivated AMD archaea from metagenomic data. These genomes are incomplete (Table 1), and remaining gaps may affect our analyses. Thus, we investigated the effect of a limited amount of fragmentation on trends by shearing the genome of the *Ferroplasma acidarmanus* (Fer1) isolate into fragments that corresponded to the lengths of the fragments from our environmental datasets. The fragmented Fer1 pairwise comparisons followed the trend defined by all other genomes with a slight downward shift (Figure S2).

Functional prediction of hypothetical and poorly annotated genes using synteny: Based on the clear relationship between evolutionary distance and synteny, we explored an improved neighborhood approach to protein functional prediction. We developed a method that involves an evolutionary distance-weighting for each pairwise comparison and incorporates the high probability of synteny due to chance in closely related organisms. We assumed that genes that remain syntenous in organisms separated by large evolutionary distances do so because of selective pressure to maintain function. Genes in predicted operons have previously been shown to rearrange at a slower rate than genes that are never found in operons [39]. We quantified the statistical significance of the difference between the populations for operon and non-operon genes using the phylogenetic pairwise comparison method [40] and the Wilcoxon signed-rank test. We used the phylogenetic pairwise comparison method to choose independent pairs of genomes for comparison and the Wilcoxon signed-rank test to test the hypothesis that there is a significant difference between the values of GOC for populations of pairwise comparisons that include operon genes versus those that include genes not in operons in the two genomes that were compared. The Wilcoxon test indicated a significant difference with a p-value of 1.017×10^{-13}. We posit that this difference is due to stronger selection against the rearrangement of genes in operons because of co-regulation and functional linkage. As an approximation, we also assumed that genes that are not in operons and retain synteny do so solely by chance, that is, selection against rearrangements on non-operon genes is negligible for the purposes of our analysis.

We used the trend between gene order conservation (GOC) and gene sequence divergence in genes not found in operons between the two genomes being compared (non-operon genes) to determine the degree of evolutionary divergence necessary to ensure that genes that retain synteny do *not* do so by chance. Because a GOC value approximates the probability of that any two genes retain synteny in a pairwise comparison at a given evolutionary distance, to estimate the probability that genes retained synteny due *solely to chance*, P_{GOCn}, we modeled the relationship between the gene order conservation of non-operon genes (GOC_n) and evolutionary distance (Figure 3). We calculated a measure of goodness of fit with the sum of squared errors (SSE) and the total sum of squares (TSS); 1-SSE/TSS = 0.9282. P_{GOCn} values were then compared to the percentage of syntenous genes that were functionally related in genomes included in the STRING database. We modeled this relationship as well, and interpreted the response variable as the probability that any two syntenous genes are functionally related, P_r, (1-SSE/TSS = 0.7648, Figure 4). Both models were chosen from a set of models, using Akaike's information criterion (Table 2). We combined the models to predict P_r from measurements of evolutionary distance (Figure S1). Thus, for pairwise comparisons below a certain evolutionary distance threshold (a bit score value of 0.3129), P_r was statistically significant; syntenous genes have a 95% or greater probability of being functionally related ($P_r > 0.95$). A gene of unknown function in one such comparison is likely functionally related to its syntenous orthologs. In these cases, functional information for syntenous orthologs that would otherwise be disregarded due to low sequence similarity was used to improve annotations of genes of unknown function. Alternatively, if functional information was available for neighboring genes in a block for which synteny was preserved, the gene of unknown function was annotated as related to its neighbor.

We applied this evolutionary distance-weighted method to improve protein functional annotation in AMD archaea for genes involved in the following pathways and processes i) cobalamin biosynthesis ii) molybdopterin guanine dinucleotide (MOB-DN) synthesis and MOB-DN-utilizing enzymes iii) ether lipid biosynthesis and iv) CRISPR-related proteins. We improved the annotation of 25 genes involved in cobalamin salvage in A, G, and Iplasma as well as Fer1 and Fer2 (Figure 5 and Table S2). An additional 34 genes were annotated with our method as part of the *de novo* cobalamin synthesis pathway or as cobalamin-related, including several cobalamin-binding proteins. We inferred a cobalamin-related function for two genes with very general annotations due to their synteny-based annotations (Table S2). The near complete *de novo* cobalamin synthesis pathway was found only in the Ferroplasma genomes, indicating a possible difference in these organisms' growth requirements.

The synteny-based annotation of molybdopterin synthesis genes also differentiates the various AMD archaea. Our synteny-based method improved annotations or provided annotations for seventeen genes in Aplasma, eleven genes in Iplasma, ten genes in Fer1, and six genes in Fer2 that were involved in molybdopterin synthesis, utilization or molybdate uptake (Figure 6 and Table S3). The A, I, Fer1, and Fer2 genomes have full pathways for the synthesis of molybdopterin guanine dinucleotide (MOB-DN), a molybdopterin cofactor that is used by proteins involved in anaerobic energy metabolism, while E and Gplasma have very few annotated molybdopterin synthesis genes of any kind. Formate dehydrogenase subunit genes are found in Aplasma, Iplasma, Fer1, and Fer2 genomes within molybdopterin synthesis gene clusters. Formate dehydrogenase is a MOB-DN-utilizing enzyme. *In silico* protein modeling provided additional evidence for the formate dehydrogenase annotation of these genes (Table S3).

Ether lipid biosynthesis genes were found in all of the AMD *Thermoplasmatales* archaea, as expected. Synteny-based annotation improved or provided annotations for a number of genes involved in ether lipid biosynthesis and its feeder pathway, the mevalonate

pathway (Figure 7 and Table S4). This included five genes in Aplasma, seven genes in Eplasma and in Gplasma, ten genes in Iplasma, eight genes in Fer1 and eight genes in Fer2. A hypothetical protein was identified in all of the AMD archaea studied that appeared to be associated with the mevalonate pathway based on synteny. Manual curation indicated that it may encode a nucleic-acid binding protein.

All of the AMD archaeal genomes except for that of Aplasma contained most or all of the genes involved in the mevalonate and ether lipid biosynthesis pathways. Aplasma is missing key genes in the mevalonate pathway likely due to its incomplete genome assembly. Aplasma is also the only genome missing genes involved in the ether lipid synthesis pathway found in *Archaeoglobus fulgidus* [41]. Two genes that maintained synteny with the ether lipid synthesis genes were investigated for possible involvement in the final steps of ether lipid biosynthesis (e.g., polar head group attachment and side chain modifications). However, BLASTs against all available NCBI bacterial genomes, indicated that these genes were also found in a number of bacteria and were thus unlikely to be involved in archaeal ether lipid synthesis pathways.

All of the AMD *Thermoplasmatales* archaeal genomes contain some CRISPR-associated proteins that occur in gene clusters with CRISPR spacer regions. A number of the CRISPR proteins in the AMD archaea are syntenous with distant relatives, allowing us to improve annotations and annotate hypothetical proteins at these loci for twenty-seven CRISPR-associated proteins (Figure 8 and Table S5). All of the archaeal genomes contained Cas1 genes, which are generally thought to be in all Cas systems as well as Cas2 genes that are found in most Cas systems [42].

Method validation: In order to test this new synteny-based method, we compared four well characterized, very distantly related bacteria and archaea to one another. We made two comparisons, one between the two bacteria and one between the two archaea. We examined a total of 175 unique genes and their syntenous orthologs in the four organisms. Of these 175 genes we found that our method correctly annotated the genes in one organism (we chose the better characterized one in both cases) 97.1% of the time (Table S6). In five cases, the annotation appeared to be correct, but one organism had only the general annotation of ABC transporter with a likely substrate specificity instead of a specific ABC transporter protein. In three other cases, the annotations in the well characterized organism did not concur with our manual curation of the gene's function. In only two cases was the annotation method clearly incorrect, in this case substituting two very closely related protein functions that are sometimes found in the same bidirectional enzyme, fumarate reductase and succinate dehydrogenase subunits A and B.

The method was also able to reconstruct parts of the Trp operon for *E. coli* and *H. volcanii*. This is significant because not only are the functions of the genes in this operon well characterized, but their associations and regulatory systems are also well understood. In the case of *E. coli*, the method correctly predicted the functions of TrpA and TrpB (Table S8), while in the case of *H. volcanii*, the method correctly predicted the functions of TrpD, TrpE, and TrpG (Table S8).

Discussion

We reconstructed four new genomes for acidophilic archaea from environmental samples and compared them with the genomes of cultivated organisms. This investigation allowed us to develop a new, generally applicable, synteny-based method for improving annotations of poorly annotated genes and genes of unknown function in bacteria and archaea. We used this method to annotate a number of important genes in uncultivated *Thermoplasmatales* archaeal genomes and therefore to better understand the functioning of these organisms in the AMD environment.

Regression analysis and model selection: The trends reported here between synteny and sequence divergence and between protein functional relatedness and synteny were determined based on the phylogenetic pairwise comparison method. This method takes into account phylogeny in order to assign pairs of genomes for comparison that do not share recent phylogenetic history with other pairs. This produces phylogenetically independent data points and allows regression analysis to be carried out without pseudoreplication. We found high measures of goodness of fit for synteny and sequence divergence, as well as for synteny and percent of syntenous genes with related functions. Because there is no known mechanistic link between point mutations and gene rearrangements, these results indicate similar selective pressures on rearrangements and mutations.

Despite the advantages of the phylogenetic pairwise comparison method, we recognize that it has inherent biases. Specifically, picking the maximal number of pairs for analysis results in the choice of many closely related pairs. Pairs that are clustered in one portion of the tree may have similar levels of synteny and sequence divergence, but this correlation may be due to some third unknown trait that is also present in this clade. We chose to use all available data as opposed to more evenly spaced taxa in order to obtain enough information for regression analysis. For our analysis, we are interested in more distantly related organisms, thus partially resolving the problem of bias in close relatives. We also recognize that the use of all of the bacterial and archaeal genomes available in the NCBI and STRING databases has resulted in a bias in our data towards certain clades and organism types that are overrepresented (e.g., pathogens). Inclusion of genomes reconstructed from metagenomic sequence data from the natural environment slightly reduces this bias. However, this method could be greatly improved in the future when more fully sequenced genomes are available.

In a few cases, two unrelated blocks of syntenous genes were conserved adjacent to one another at significant evolutionary distances. This problem can be avoided by enforcing a stricter evolutionary distance cutoff in the cases where it can be observed that two syntenous blocks are sometimes, but not always conserved next to one another. The mechanism resulting in this type of synteny conservation is unknown.

It is important to note that the model we developed for synteny-based annotation assumes that all genes in operons are rearranged slowly compared to those that are not. This is consistent with data analysis shown in Figure 3 and with Rocha's analyses [39]. We also assumed that genes that are rearranged rapidly are not under significant selective pressure so that the trend for the non-operon genes could be used to estimate the probability that any two genes stay together due to chance. Deviation from this assumption in a subset of cases could contribute to scatter in the trend (thus poorer regressions and weaker correlations) and lead to a higher value of GOC for significance of synteny for annotation purposes. Thus, our method

provides a conservative estimate of the evolutionary distance necessary for functional predictions and the probability of functional relatedness is higher than stated.

Protein function prediction: Understanding the relationship between gene order and evolutionary distance is essential for accurate synteny-based gene functional annotation. In the case of the AMD archaea, the weighting of conservation of gene order at large evolutionary distances resulted in improved annotations for genes involved in a number of processes, including cobalamin biosynthesis, molybdopterin guanine dinucleotide (MOB-DN) synthesis and MOB-DN-utilizing enzymes, ether lipid biosynthesis, and CRISPR-based immunity.

Synteny-based annotation of cobalamin biosynthesis genes indicated a clear difference between the nutritional requirements of A, E, G, and Iplasma versus those of the *Ferroplasma* species. Both of the Fer1 and Fer2 genomes contained full *de novo* anaerobic cobalamin synthesis pathways, while the other archaeal genomes contained nearly complete cobalamin salvage pathways [43]. This difference may be important in differentiating the niches of the various types of AMD archaea. It may allow the *Ferroplasma* spp. to compete better with other archaea in low nutrient conditions, i.e., in early growth stage biofilms.

The synteny-based annotation of molybdopterin biosynthesis and molybdopterin-binding proteins in the AMD *Thermoplasmatales* archaea also helped to differentiate their respective physiologies. The molybdopterin guanine dinucleotide synthesis protein (MobA) in Aplasma, Iplasma, Fer1, and Fer2 makes a specific type of molybdenum cofactor that is only used by dimethyl sulfoxide (DMSO) reductase family enzymes. These genomes also include a gene for a formate dehydrogenase protein (a member of the DMSO reductase family) in their molybdopterin synthesis gene clusters, indicating that they may be able to use this enzyme for nitrate reduction, mixed acid fermentation, or anaerobic carbon fixation. Previously published proteomic data demonstrate that some of these MobA genes are expressed and suggests that some AMD archaea use one of these anaerobic energy or carbon metabolisms. E- and Gplasma's genomes contain only one molybtoperin-related gene (moeB), which may be a misannotation, and thus likely do not use a MOB-DN cofactor.

In silico protein structure modeling supported the functional annotation of certain molybdopterin synthesis genes of interest (Table S3). Specifically, structural modeling suggested that the potential MobA genes in Aplasma and Iplasma do in fact make MobA. Interestingly, the Iplasma homolog for MoaB fits a protein model for MogA. This is intriguing because no archaea to date have been shown to have true MogA homologs, but MoaB is thought to play the same role in molybdopterin biosynthesis for archaea as MogA does for *E. coli* [44]. Structural modeling also supported the functional annotation of the FdhF alpha subunit genes found in the Aplasma, Iplasma, Fer1, and Fer2 genomes (Figure S3). These proteins fit the FdhF of the hydrogenase-linked formate dehydrogenase model from *Escherichia coli*, suggesting a potential involvement of these genes in a formate hydrogen lyase complex and mixed acid fermentation.

The synteny-based method identified two new genes that may be involved in MOB-DN synthesis. These are a thioredoxin family gene and a SurE: 5'/3'-nucleotidase. SurE is of particular interest, as it functions in *E. coli* to remove a phosphate group from nucleotides [45]. SurE has the highest affinity for AMP among nucleotides tested by Proudfoot *et al*. [45]. An intermediate in molydopterin biosynthesis, adenylated molybdopterin, contains a covalently-bound AMP. Thus, this SurE homolog is potentially involved in dephosphorylation related to molybdopterin biosynthesis or modification.

Ether lipid biosynthesis is a pathway common to all archaea. The mevalonate pathway precedes ether lipid biosynthesis [41]. Thus, we looked for mevalonate pathway genes as well

as ether lipid biosynthesis genes. Of the twenty-five mevalonate pathway genes annotated via our synteny-based approach, one hypothetical protein has orthologs in all of the AMD archaea analyzed. This gene contains the PFAM domain of unknown function 35 that is hypothesized to bind nucleic acids. A possible ortholog of this gene is found in all archaeal genomes available on NCBI, further supporting some role of this gene in the mevalonate pathway.

The CRISPR-related proteins annotated with synteny-based annotation included a number of genes previously annotated as hypothetical proteins. Iplasma and Fer1 included the typical operon configuration of Cas module family I [46], while the other genomes included novel Cas system arrangements. These annotations provide a starting point for further investigation of the biochemistry of the CRISPR/Cas system.

Availability of computational tools: The Ruby scripts used for our analyses are open source and are available at https://github.com/pyelton/Synteny-based-annotator.

Materials and methods

Sampling and genome reconstruction: For a detailed explanation of sampling, DNA extraction, sequencing and assembly protocol see Text S1. The completeness of the archaeal genomes was evaluated based on the number of tRNA, rRNA, and other orthologous marker genes [38]. Binning accuracy was also evaluated by analysis of fragment clustering in emergent self-organizing maps (ESOM) created based on tetranucleotide frequencies of consensus contig sequences [47]. Genes missing from the pathways that we analyzed were searched for in the overall AMD DNA dataset. BLAST hits to these genes were then binned via tetranucleotide frequency, using ESOM, and assigned to organisms if possible.

Orthology, synteny, and measures of evolutionary distance: Because the objective of this work was to analyze lineage divergence and develop a gene functional prediction method applicable to all archaea and bacteria, our analyses included all publicly available archaeal genome sequences downloaded from NCBI. All available published bacterial sequences were also added to the analysis for a more comprehensive comparison of genome rearrangements. Note that this consisted of 634 genomes because we used only the full genomes published on the NCBI website that also had full 16S rRNA sequences available on NCBI. These genomes were selected from across all major lineages of archaea and bacteria.

We identified orthologs and syntenous genes using pairwise comparisons between 638 organisms (Table S7) for Figures 2 and 3 and using pairwise comparisons between all 118 prokaryotic organisms from the STRING database for Figure 4 (Table S8). Orthologs were operationally defined as those genes that were reciprocal best BLASTP hits that shared 30% or greater amino acid identity over 70% or more of the gene length or BLASTP hits that shared 20% or greater amino acid identity over 50% or more of the gene length and maintained synteny. Synteny was initially defined as conservation of two or more adjacent genes in two genomes. Subsequent analyses defined synteny as the conservation of genes separated by no more than one intervening gene. Trends in synteny versus evolutionary distance did not differ substantially between these two definitions (data not shown). Thus, we generally refer to synteny in this paper as conservation of a gene pair with no more than one intervening gene.

We used an established measure of synteny, the fraction of orthologous genes that are syntenous based on at least one shared neighbor (allowing for a specified number of gene insertions) in the two genomes compared (gene order conservation; GOC) as described by Rocha [39]. For our measure of genome sequence divergence we chose average normalized BLASTP bit score normalized to the maximum possible bit score between two genes. Normalization consisted of dividing the bit score of the alignment by the average of the two maximum possible bit scores of the alignments of self to self for each respective gene (for details see Text S1). We chose this measure for two reasons. Firstly, previous work has shown that whole genome amino acid identity is a robust measure of evolutionary distance even between close relatives [48], while sequence insertions and deletions are important in sequence divergence for distant relatives [49]. Average normalized bit score is a measure that captures both insertions/deletions and amino acid identity. 16S rRNA gene sequence divergence was also considered in the analysis because it is a standard measure and for comparison to previous studies. Trends between GOC and 16S rRNA divergence were similar to those using average normalized bit score as a measure of evolutionary distance, but were more variable (data not shown).

Manual curation: All genes used as examples in this analysis were manually curated according to the following criteria: Genes were aligned against the interpro and nr databases with a BLASTP algorithm. Genes were then annotated if they had a TIGR or Pfam domain hit that predicted a specific function with greater than 70% amino acid identity, an e-value of at least 1×10^{-10} and coverage of more than 70% of the protein. Genes were given a "putative" annotation if they met the previous criteria except they had an amino acid identity of 30-70%, an e-value between 1×10^{-4} and 1×10^{-10}, and matched 50-70% of the protein, or if their domain-based hits provided only general functional information. In these cases, additional evidence from hits from the nr database was used if possible to provide a specific functional annotation. Genes were given a "probable" annotation if they had annotated hits in the nr database with greater than 30% amino acid identity over 70% of the length of the gene.

Comparative method for correlation analysis: In order to determine the rates of synteny loss over different evolutionary distances, we looked for correlations and trends between average normalized bit score, GOC, and the percentage of syntenous genes that are known to be functionally related. Our initial regressions compared genome pairs from NCBI and our dataset and regressed GOC on average normalized bit score. The regression of percent syntenous genes that are related on GOC used genome pairs from the STRING database. Genes were considered related if they had a predicted association in STRING based on fusion events, experimental evidence, co-expression, database information (involvement in the same pathway or complex), and text mining information (co-occurrence in multiple papers). To avoid circularity in our method genome context was not used in predicting functional relatedness, that is, neither co-occurrence in genomes nor synteny was used to predict protein functional relatedness. Because of the inherent non-independence of pairwise comparisons between different taxa, we made use of a method to select phylogenetically independent pairs [40, 50]. For details on this method see Text S1.

Operon prediction: Genes were predicted to be in operons when they had the same transcription direction and no more than thirty bases between the two. We compared genes that were in predicted operons in one or both of the two genomes in a pairwise comparison, "operon genes", and genes that were not found in predicted operons in either genome of the pairwise comparison, "non-operon genes". For a more detailed explanation of the operon prediction method see Text S1.

Wilcoxon signed-rank test: In order to show that the GOC of pairwise comparisons of operon genes were significantly different from comparisons of non-operon genes, we chose to use a non-parametric test because of the unknown distribution of the data.

Applicability to near-complete environmental genomes: In order to test the validity of this method to near-complete environmental genomes, the *Ferroplasma acidarmanus* (Fer1) isolate genome was sheared into fragments. For information on genome shearing see Text S1.

Gene annotation: Open reading frames for the archaeal genomes were identified using the Prodigal software [51]. Annotations were automatically generated through a pipeline that includes homology searches against KEGG and Uniref90, and domain/motif homology searches using InterProScan. Annotations were ranked in order of increasing confidence of a match: Rank A annotations are the most confident and Rank G annotations represent gene predictions with no functional assignment. For an explanation of rankings, see Text S1. All annotations specifically mentioned in this paper were manually curated based on conserved

domains in InterProScan and similarity to the nr sequence database from NCBI. The remainder of the functional annotation and physiological inferences for the genomes of the AMD archaea is reported separately (Chapter 2).

Our weighted synteny-based annotation approach is related to a previously published approach [39]. Rocha noted that GOC is an estimate of the probability that genes remain unshuffled over a certain evolutionary distance t. He also noted that genes in operons in either organism experience much slower rates of gene rearrangements than other orthologs [39]. We calculated the probability that genes are syntenous due solely to chance at a given evolutionary distance (P_{GOCn}) by assuming that the GOC for rapidly shuffling genes (those not in operons; GOC_n) was due entirely to chance. GOC_n was plotted against evolutionary distance and was fitted to the data by nonlinear regression.

The regressions were based on the following functions:

Average normalized bit score regression:

$GOC_n = ae^{be^{ct}}$

Where a, b, and c are constants, e is Euler's number, and t is the average normalized BLASTP bit score between two genomes.

Percent of syntenous genes that are related regression on P_{chance}:

$P_r = e^{cP_{chance}}$

Where c is constant, e is Euler's number, and P_{chance} is the value of GOC calculated from the bit score of the comparison based on the GOC_n regression.

These functions were chosen for each regression based on comparison of the following types of models using Akaike's information criterion: for the GOC on average bit score regression we looked at linear models, log models, exponential models, and sigmoidal models. AIC indicated that a Gompertz model fit the GOC and average bit score data the best (Table 2). This was not surprising because the data appears to be sigmoidal and asymmetrical. For the percent of functionally related syntenous genes on GOC regression we considered linear models, exponential decay models, and quadratic models. These models were forced through the point (0,1) because at a P_{chance} of zero, where the probability that two genes retain synteny due to chance is zero, the probability that syntenous genes are functionally related must be equal to one. AIC indicated that the exponential model fit the data the best in this case (Table 2).

We found the t at $GOC_n = P_{GOCn} = 0.05$, an average normalized bit score of 0.3129, the evolutionary distance at which there is a 95% probability that syntenous genes are functionally related according to the STRING database information (Figure S1). At this evolutionary distance, there is at least a 95% probability that genes that retain synteny have done so for a reason (presumably selective pressures). In fact, the probability that syntenous genes have related function is likely higher than 95% because the STRING database does not have exhaustive protein interaction data.

Based on the derived values of t, we chose genomes that were sufficiently distant relatives that genes are not likely to be syntenous by chance so that synteny could be used to annotate genes in the AMD archaea. For genome comparisons with a bit score of less than 0.3129, we assigned or improved annotations of genes that are found in syntenous blocks in AMD *Thermoplasmatales* archaea. Each gene was then annotated with the annotation of its ortholog if that gene had an annotation, or as "functionally related to gene X" where gene X

is its syntenous neighbor gene. If the orthologous genes in these pairs had the same annotation but one was poorly annotated, the poorly annotated gene was given an additional score that indicated a synteny-based annotation improvement.

Method validation: Our annotation method was tested against 175 genes of known function in four genomes, two bacterial genomes and two archaeal genomes. The organisms used were *Escherichia coli* K12 MG1655, *Chlamydia trachomatis* D/UW-3/CX, *Haloferax volcanii*, and *Sulfolobus solfataricus* P2. These organisms were chosen for three reasons: 1. They are all very well experimentally characterized and there are more than 600 articles on each of them in the ISI Web of Science database 2. They are sufficiently distant relatives that they pass the significance threshold for using our synteny-based method. It was particularly hard to find a well-characterized Bacterium that was sufficiently distant to *E. coli* K12. 3. With the exception of *Chlamydia trachomatis*, they all have genetic systems that have been used for a number of years, allowing for genetic confirmation of gene function. We chose *Chlamydia trachomatis* because it is very distant from *E. coli* and there have been recent advances in the development of a genetic system for this organism [52] that may lead to future confirmation of our findings.

The method was tested in the following manner. Syntenous orthologs were found between *Escherichia coli* K12 MG1655 and *Chlamydia trachomatis* D/UW-3/CX, and between *Haloferax volcanii* and *Sulfolobus solfataricus* P2. 88 syntenous orthologs were found between the two bacteria and 117 syntenous orthologs were found between the two archaea. Of these, we determined that 145 were unique to one or the other pairwise comparison based on KEGG identifiers and E.C. numbers. 30 genes were potentially shared between the two pairwise comparisons. We were able to analyze a total of 145 unique syntenous orthologs and 30 shared syntenous orthologs, thus 175 genes overall.

For these 175 syntenous orthologs, we chose to mask their function in one of the organisms in each pairwise comparison, reannotating the genes as “hypothetical proteins”. We chose to hide the functions of the genes in *E. coli* K12 in the first comparison and in *S. solfataricus* in the second comparison. We chose these organisms because they are better characterized than their pair in each case. We then took these “hypothetical proteins” and applied our synteny-based annotation method to them, determining their function solely based on the function of their counterpart in the given comparison. Then we compared the new function attributed to the “hypothetical protein” by our method to the original annotation of the protein. We considered the functions the same if they had the same KEGG identifier [53] or gene name and E.C. number in the cases where the gene did not have a KEGG identifier.

Acknowledgements

Mr. Ted Arman (President, Iron Mountain Mines), Mr. Rudy Carver, and Mr. Richard Sugarek are thanked for site access and on site assistance. L. Hauser is thanked for helpful comments on the manuscript.

Tables

Table 1: Genome Information for AMD *Thermoplasmatales* archaea. Genome completeness was estimated based on number of tRNAs, rRNAs, and orthologous marker genes. * Reflects tRNAs in several closely related Fer2 strains sampled independently.

Genome	Length (Mb)	Coverage	GC content (%)	Number of tRNAs	rRNAs	Completeness
Aplasma	1.94	8X	46	48	16S, 23S, 5S	71.4%
Eplasma	1.66	9X	38	41	16S, 23S, 5S	100.0%
Gplasma	1.84	8X	38	44	16S, 23S, 5S	94.3%
Iplasma	1.69	20X	44	46	16S, 23S, 5S	100.0%
Fplasma 1	1.46	4.5X	36	44	16S, 23S, 5S	NA
Fplasma 2	1.82	10X	37	63*	16S, 23S, 5S	100.0%
Fplasma 1 isolate	1.94	13X	38	44	16S, 23S, 5S	100.0%

Table 2: AIC comparison of regression models for NCBI and STRING data.

X variable	Y variable	Model type	AIC
Average normalized bit score	GOC	Linear	-455.4074
Average normalized bit score	GOC	Logarithmic	-493.2857
Average normalized bit score	GOC	Cubic	-573.4456
Average normalized bit score	GOC	Gompertz	-585.8681
GOC	Percent of syntenous genes that are related	Exponential	-69.750892
GOC	Percent of syntenous genes that are related	Linear	-7.646633
GOC	Percent of syntenous genes that are related	Quadratic	-51.991032

Figures

Figure 1: 16S rRNA gene tree of AMD *Thermoplasmatales* archaea. Bootstrapping values are indicated as percentages out of 1000 random samples.

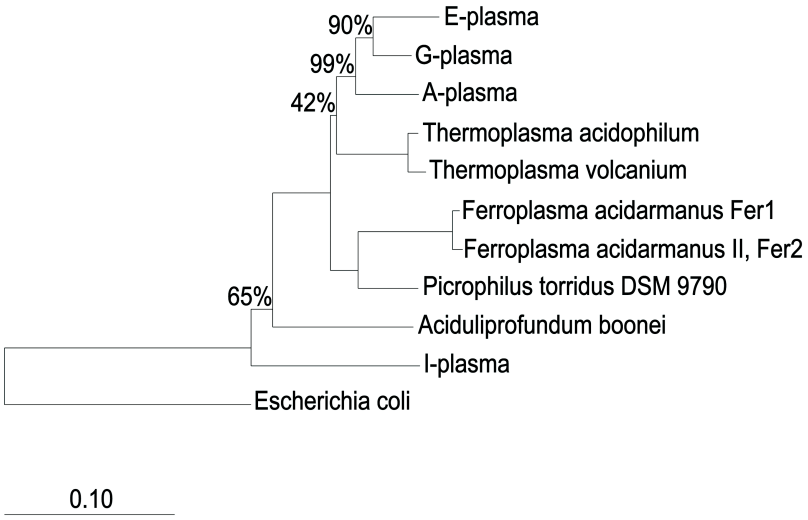


Figure 2: Synteny (gene order conservation) versus sequence divergence (average normalized bit score) in pairwise comparisons of 638 bacterial and archaeal genomes. The phylogenetic pairwise comparison method was used to carry out 319 comparisons. See Text S1 for details.

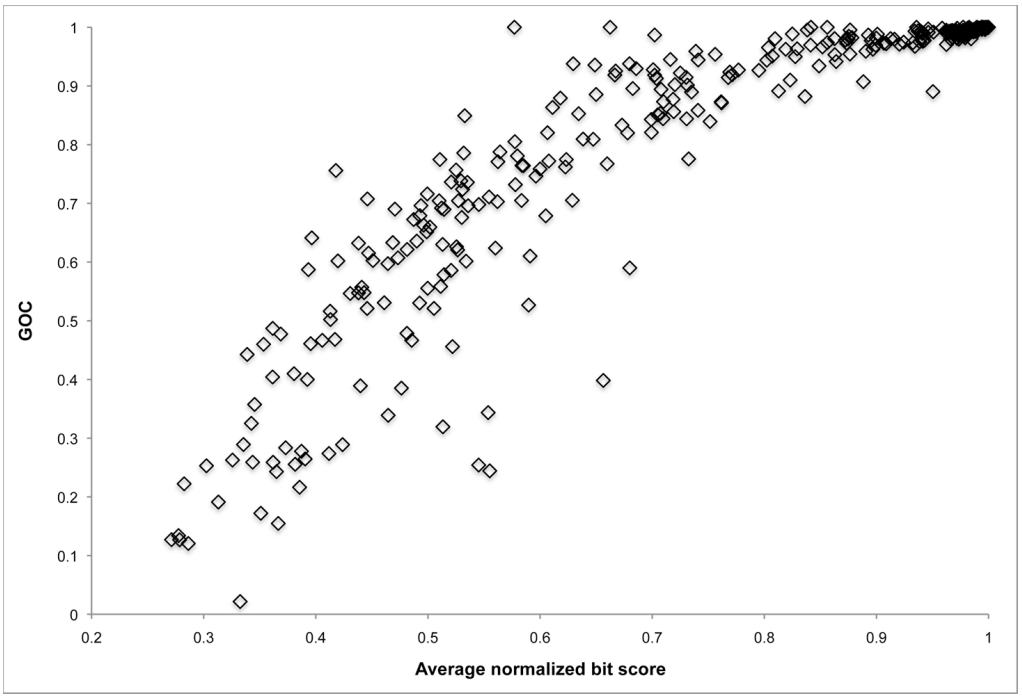


Figure 3: Synteny (gene order conservation) versus sequence divergence (average normalized bit score) in pairwise comparisons of 638 bacterial and archaeal genomes. Orthologs that are sometimes in predicted operons (operon genes) are compared separately from those that are never in operons (non-operon genes). The circled outliers come from comparisons of endosymbiont genomes, which have very small genomes and greater than expected conserved gene order in non-operon genes.

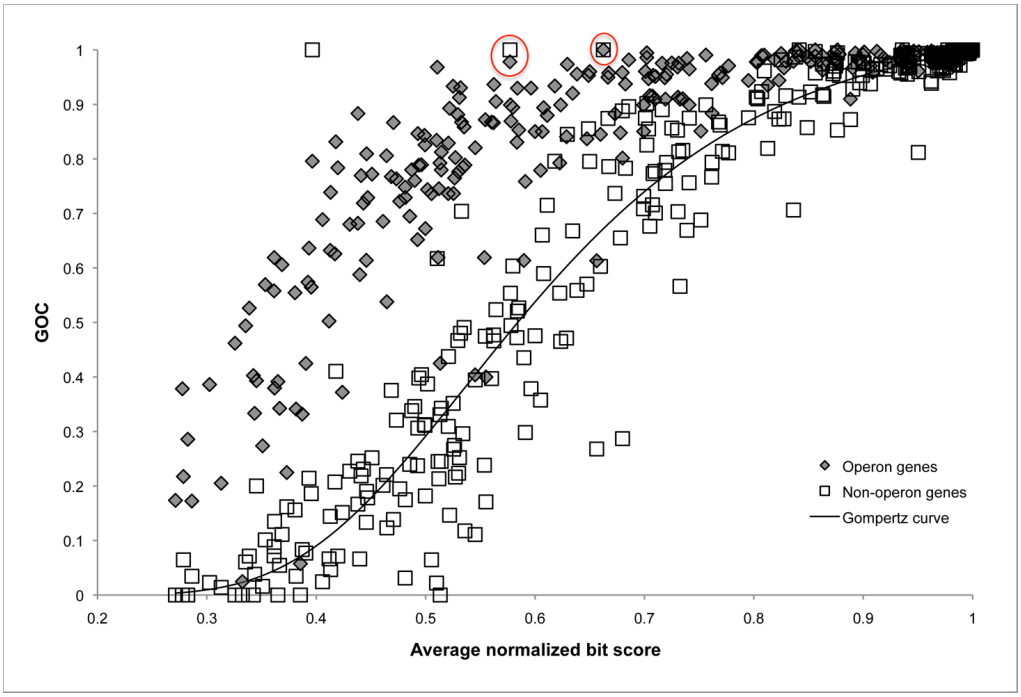


Figure 4: P_{chance} versus percent of syntenous genes that are related. P_{chance} is estimated from average normalized bit score, based on the model of GOC versus average normalized bit score. The percentage of syntenous genes that are related is based on data from the STRING protein database.

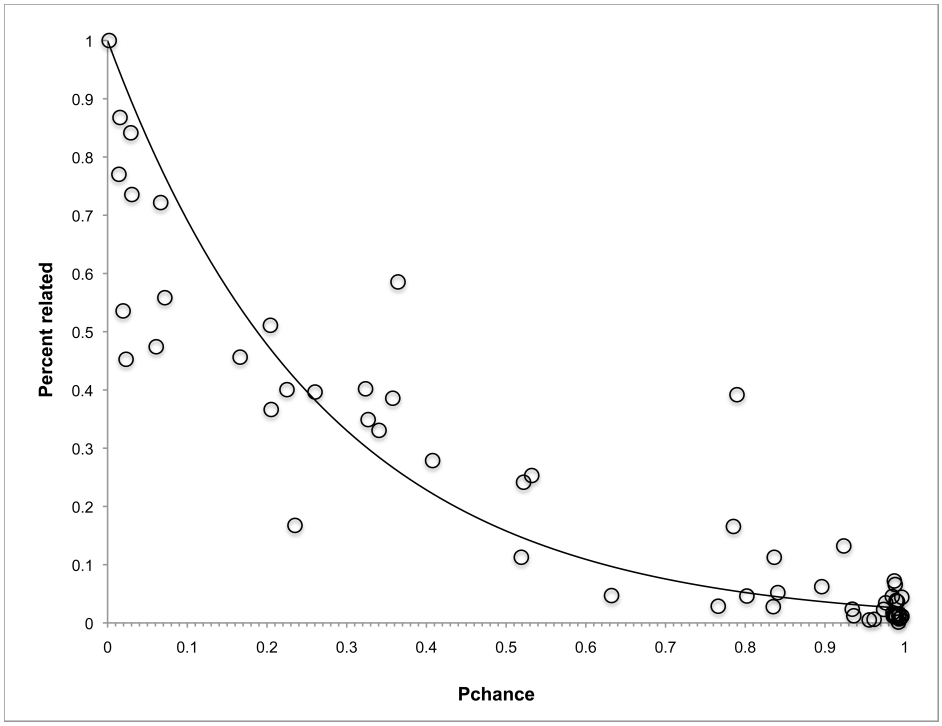


Figure 5: Cobalamin related genes. Arrow lengths are proportional to gene lengths. However, intergenic distances are not shown to scale. The colors indicate syntenous genes that we annotated with our synteny-based method. Genes of the same color are in the same syntenous block. Text in blue indicates de novo cobalamin synthesis genes. Text in red indicates cobalamin salvage pathway genes. A) Cobalamin salvage pathway genes. B) *De novo* cobalamin synthesis pathway genes.

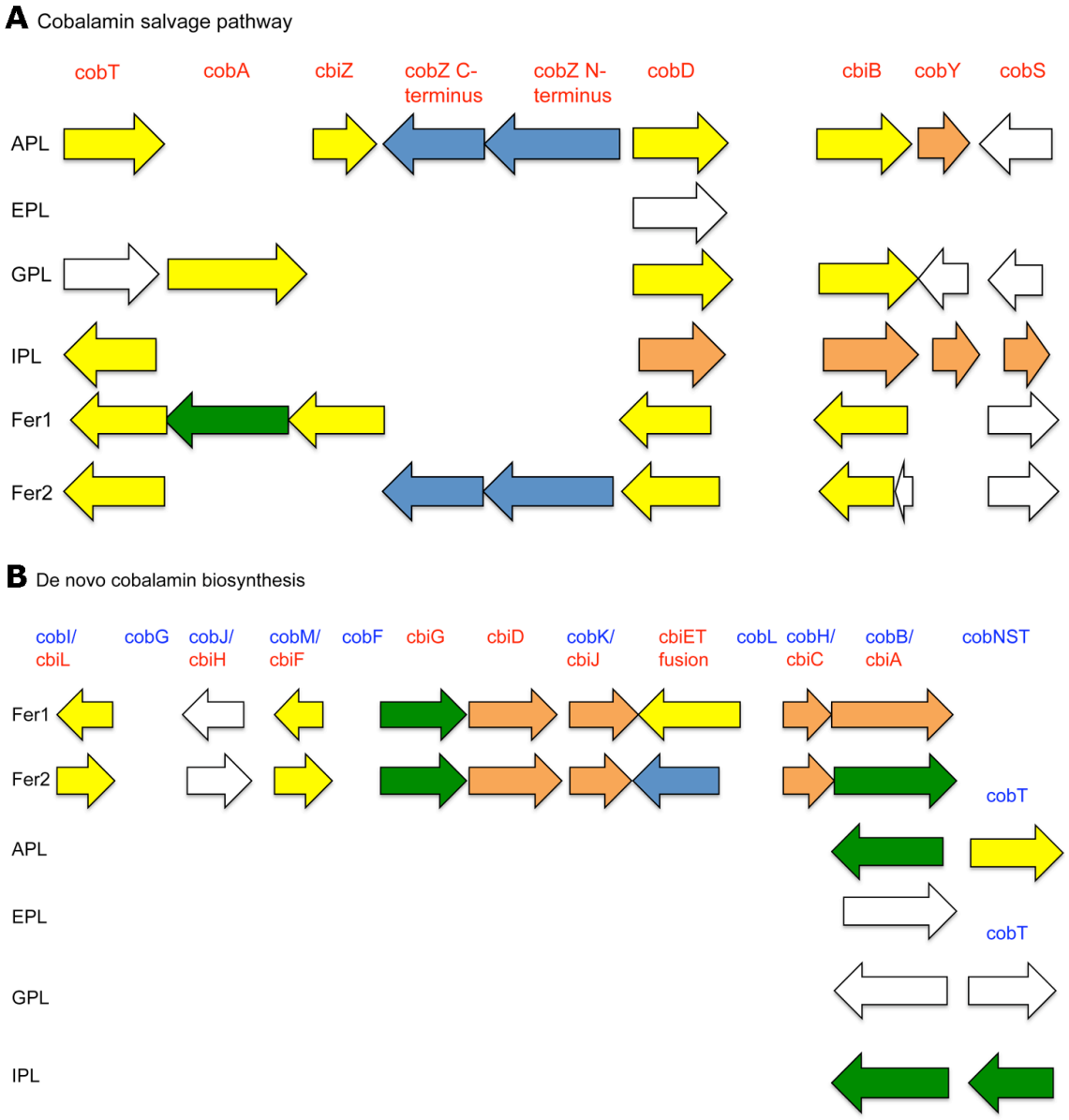


Figure 6: Molybdopterin synthesis clusters. Arrow lengths are proportional to gene lengths. However, intergenic distances are not shown to scale. * indicates annotations that are putative. ** indicates genes that are found outside of the cluster. Color indicates syntenous genes that we annotated with our synteny-based method and genes of the same color are in the same syntenous block. A) Cluster 1 B) Cluster 2 C) Clusters 3 and 4.

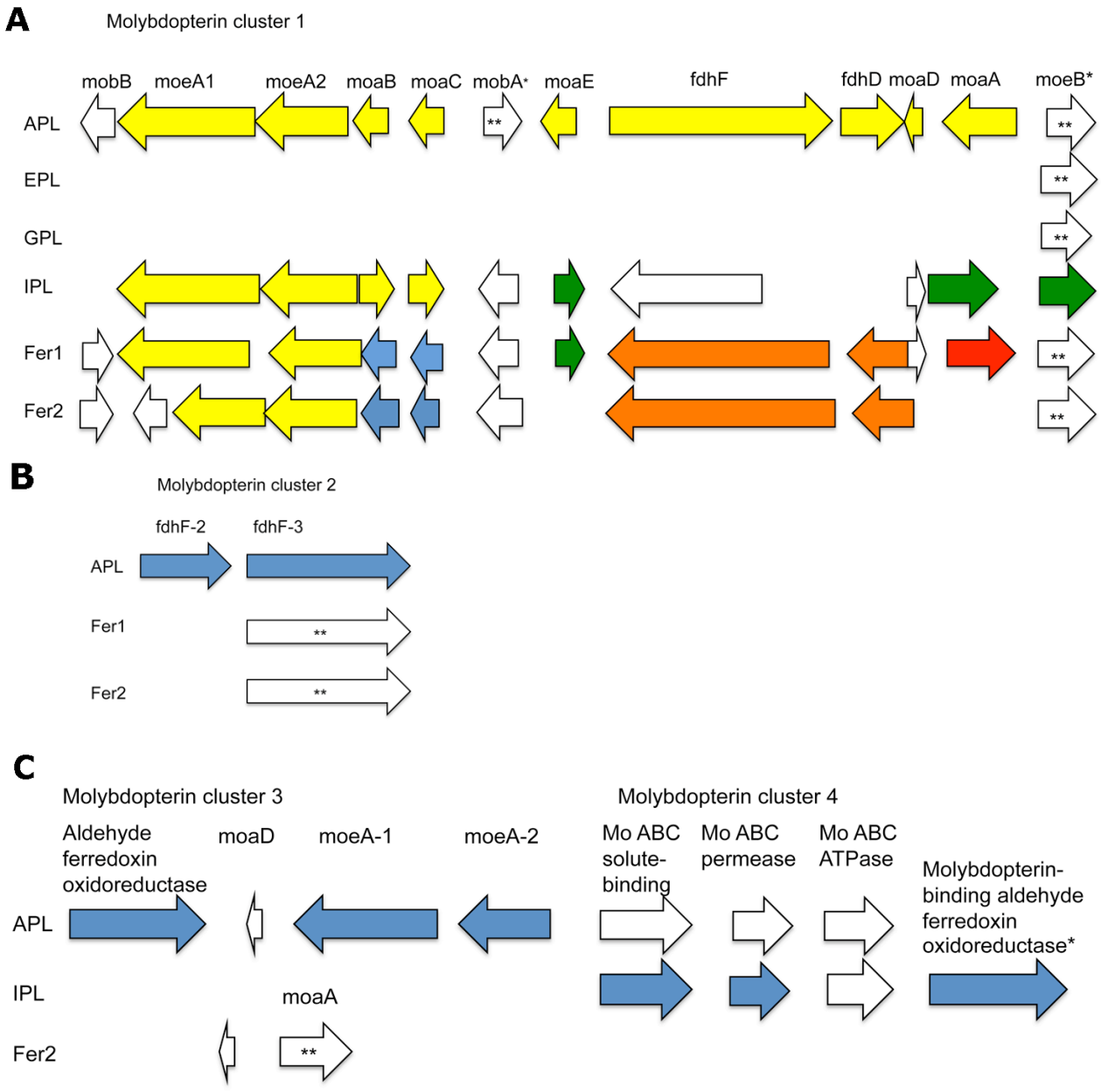


Figure 7: Genes related to ether lipid biosynthesis. Arrow lengths are proportional to gene lengths. However, intergenic distances are not shown to scale. The colors indicate syntenous genes that we annotated with our synteny-based method. Genes of the same color are in the same syntenous block. A) Mevalonate pathway genes B) Ether lipid biosynthesis pathway genes.

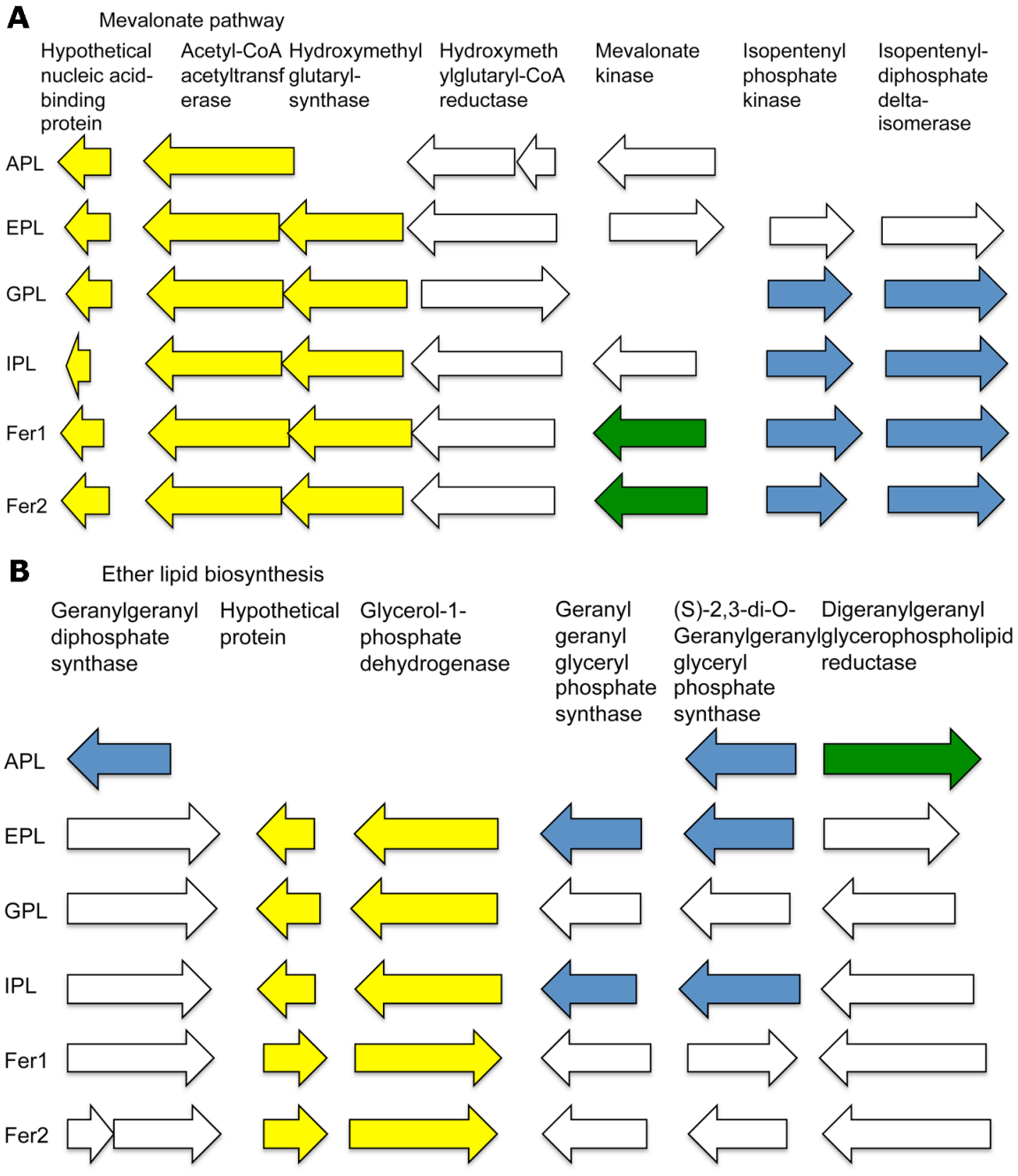
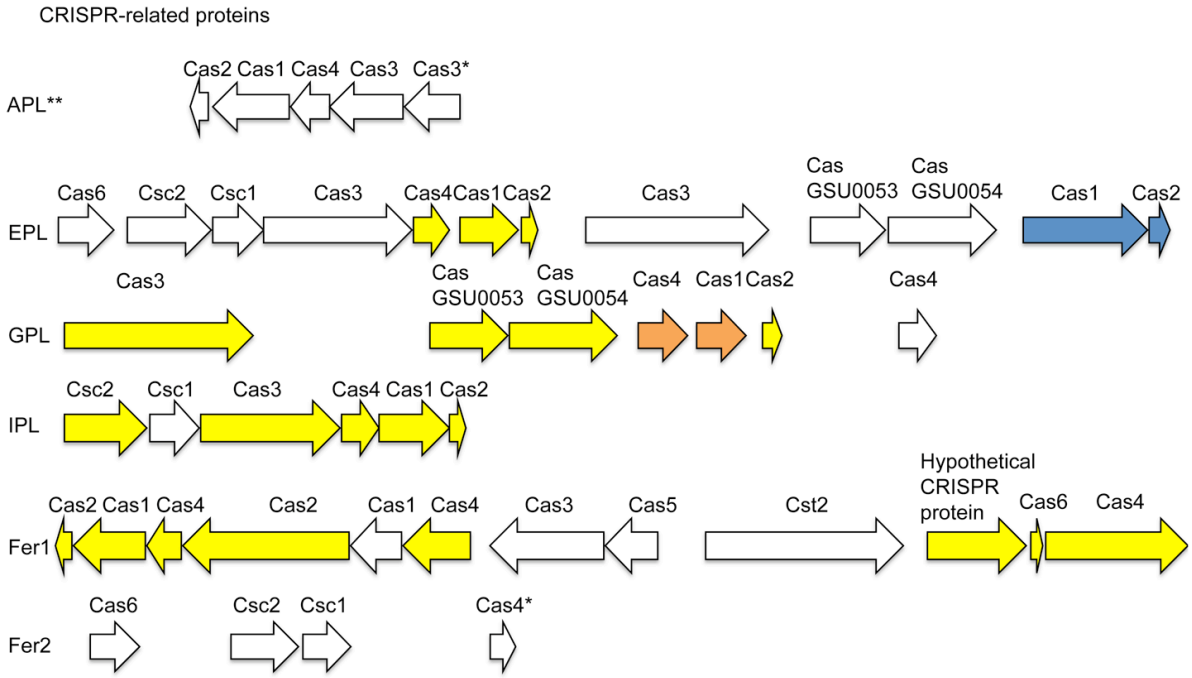


Figure 8: CRISPR-related protein genes. Arrow lengths are proportional to gene lengths. However, intergenic distances are not shown to scale. * indicates annotations that are putative. ** indicates a locus that is unbinned and potentially belongs in the Aplasma genome. The colors indicate syntenous genes that we annotated with our synteny-based method. Genes of the same color are in the same syntenous block.



Supplementary Materials

Figure S1: Overview of the synteny-based annotation method. 1. Comparison of GOC and average normalized bit score for NCBI database genomes. 2. Comparison of GOC and average normalized bit score for NCBI genomes split into groupins of genes found in predicted operons and non-operon genes. 3. Comparison of the percent of syntenous genes that are related and the probability that syntenous genes remain together due to chance for STRING database genomes. The green line illustrates where 95% of the syntenous genes are related. The blue line indicates the Pchance at 95% related. 4. This value is substituted into the model for evolutionary distance from the NCBI genomes to yield the average normalized bit score where 95% of syntenous genes have related functions (in red ~0.31). 5 and 6. Based on the comparison of genomes more distantly related than this value, annotations for poorly-annotated genes are improved.

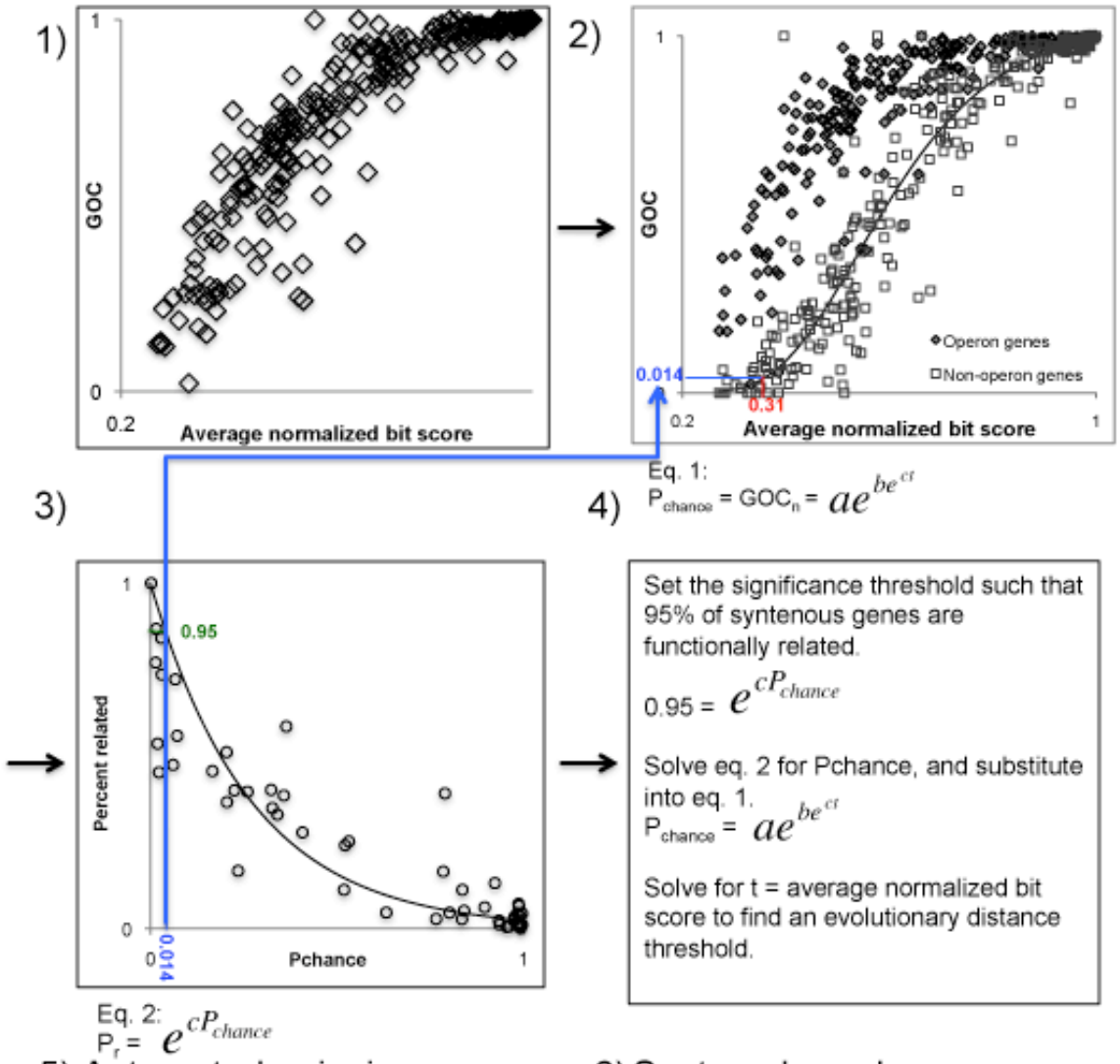

1)
2)
3)
4)
GOC
Average normalized bit score
Operon genes
Non-operon genes
0.014
0.31
Eq. 1:
$P_{chance} = GOC_n = ae^{be^{ct}}$
Percent related
Pchance
0.95
Eq. 2:
$P_r = e^{cP_{chance}}$
Set the significance threshold such that 95% of syntenous genes are functionally related.
$0.95 = e^{cP_{chance}}$
Solve eq. 2 for Pchance, and substitute into eq. 1.
$P_{chance} = ae^{be^{ct}}$
Solve for t = average normalized bit score to find an evolutionary distance threshold.
5) Automated pairwise comparison:
Looks for synteny in distant relatives (Bit score < 0.3202)
Increasing evolutionary distance
Evo distance threshold
G-plasma
EPL
P torridus
Fer1
Fer2
T volcanium
E. coli
GENE 020
GENE 021
GENE 022
GENE 023
GENE 024
GENE 025
GENE 026
6) Synteny-based annotations:
Annotates hypothetical proteins with the annotation of their syntenous orthologs or as "related to" their syntenous gene.
G-plasma
E. coli
G-plasma new annotation
Gene_020 hypothetical protein
ribosomal protein
Gene_020 ribosomal protein
Gene_021 translation elongation factor
translation elongation factor
Gene_021 translation elongation factor
GENE_022
GENE_023
GENE_024
GENE_025
GENE_026

Figure S2: GOC versus sequence divergence (average normalized bit score) in pairwise comparisons of genomes, including sheared Fer1 isolate genome. The filled blue diamond indicates the comparison between Fer1 and Fer2 in the overall dataset. The open blue diamonds indicate the comparison between the fragmented Fer1 genome and the full Fer2 genome.

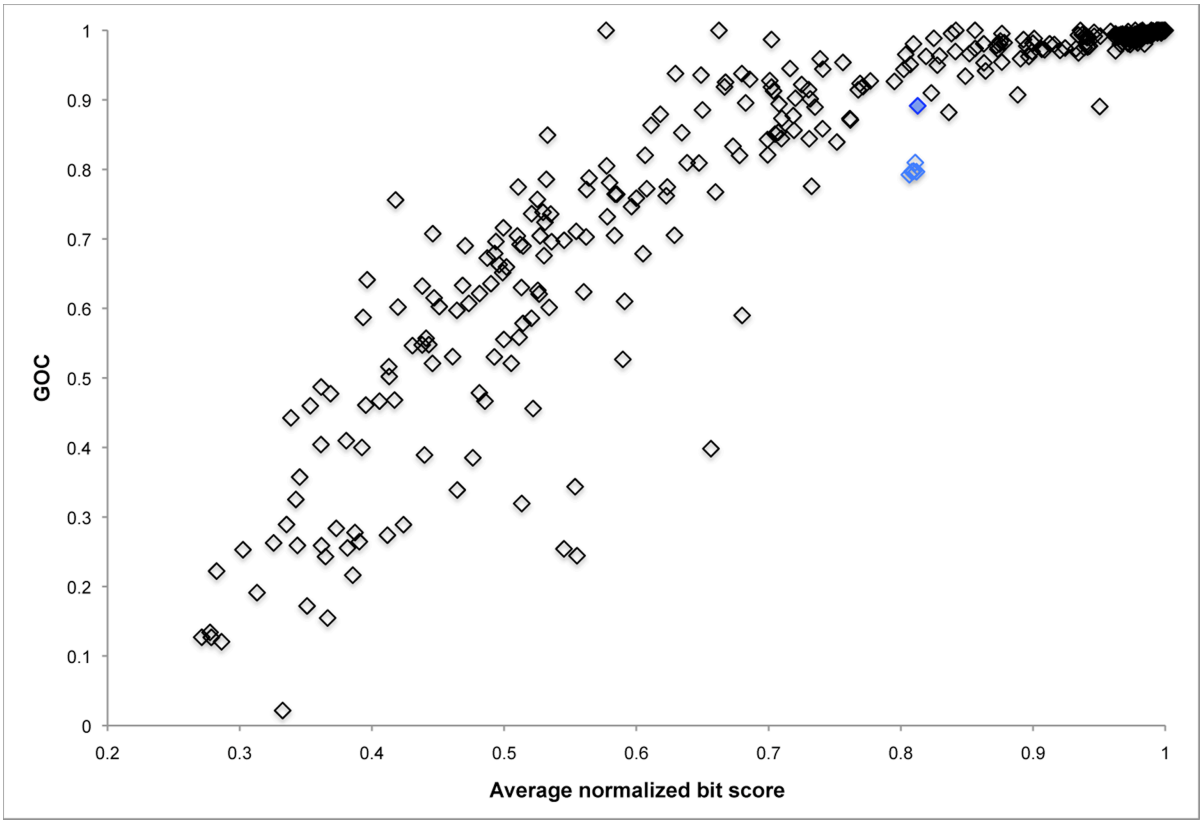


Figure S3: Protein model of FdhF alpha subunit in Iplasma on the *E. coli* hydrogenase-linked formate dehydrogenase alpha subunit protein. Conserved residues from active site are highlighted in red. Conserved residues from molybdenum coordinating site are highlighted in purple.

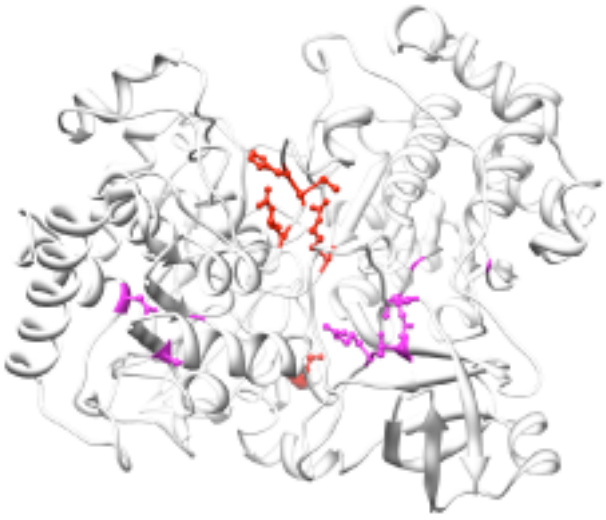


Table S1: Estimation of genome completeness based on orthologous marker genes. COG numbers and annotations for orthologous marker genes found in one copy in all prokaryotic genomes are given on the left. The number of occurrences and the gene number for each marker gene in the *Thermoplasmatales* AMD archaea are shown on the right.

35 Orthologous group markers	Apl	APL	Gpl	GPL	Epl	EPL	Ipl	IPL	Fer1	Fer1	Fer2	Fer2
COG0012 Predicted GTPase, probable translation factor	1	00004_0007	1	13318_280	1	15243_573	1	13624_54	1	1281	1	scaffold_56_1
COG0016 Phenylalanyl-tRNA synthetase alpha subunit	1	17445_0016	1	12302_32	1	15243_601	1	15911_423	1	1504	1	scaffold_1_20
COG0048 Ribosomal protein S12	1	5769_0003	1	13318_183	1	15243_193	1	13624_357	1	1280	1	scaffold_163_5 (probable)
COG0049 Ribosomal protein S7	1	5769_0004	1	13318_182	1	15243_194	1	13624_358	1	1279	1	scaffold_40_45
COG0052 Ribosomal protein S2	1	17087_0034	0		1	17965_528	1	13606_464	1	889	1	scaffold_11_49
COG0080 Ribosomal protein L11	1	17068_0016 , 17068_0017	1	13318_236	1	15243_111	1	13606_498	1	720	1	scaffold_31_10
COG0081 Ribosomal protein L1	1	17068_0015	1	13318_235	1	15243_112	1	13606_497	1	719	2	scaffold_31_9
												scaffold_730_6 (probable)

COG0085 DNA-directed RNA polymerase, beta subunit/140 kD subunit	1	17357_0065	1	13296_48	1	15243_499	1	15911_252	1	1178	1	scaffold_15_42
COG0087 Ribosomal protein L3	1	17357_0005	1	13296_96	1	15243_450	1	15911_349	1	1133	1	scaffold_60_8
COG0088 Ribosomal protein L4	1	17357_0006	1	13296_95	1	15243_451	1	15911_348	1	1134	1	scaffold_60_7
COG0090 Ribosomal protein L2	1	17357_0008	1	13296_93	1	15243_453	1	15911_346	1	1136	1	scaffold_60_5
COG0091 Ribosomal protein L22	1	17357_0010	1	13296_91	1	15243_455	1	15911_344	1	1138	1	scaffold_60_3
COG0092 Ribosomal protein S3	1	17357_0011	1	13296_90	1	15243_456	1	15911_343	1	1139	2	scaffold_15_1
												scaffold_60_2
COG0093 Ribosomal protein L14	1	17357_0015	1	13296_85	1	15243_461	1	15911_338	1	1144	1	scaffold_15_6
COG0094 Ribosomal protein L5	1	17357_0018	1	13296_82	1	15243_464	1	15911_334 (split)	1	1147	1	scaffold_15_9
								15911_335 (split)				
COG0096 Ribosomal protein S8	1	17357_0020	1	13296_80	1	15243_465	1	15911_332	1	1148	1	scaffold_15_10
COG0097 Ribosomal	1	17357_0021	1	13296_79	1	15243_466	1	15911_331	1	1149	1	scaffold_15_11

protein L6P/L9E												
COG0098 Ribosomal protein S5	1	17357_0025	1	13296_75	1	15243_470	1	15911_327	1	1153	1	scaffold_15_15
COG0099 Ribosomal protein S13	1	17357_0081	1	13296_35	1	15243_517	1	13606_290	1	1254	1	scaffold_29_48
COG0100 Ribosomal protein S11	1	17357_0083	1	13296_33	1	15243_519	1	13606_287	1	1256	1	scaffold_29_46
COG0102 Ribosomal protein L13	1	17498_0005	1	13296_5	1	15243_533	1	15911_238	1	699	1	scaffold_12_18
COG0103 Ribosomal protein S9	1	17498_0004	1	13296_4	1	15243_534	1	15911_239	1	698	1	scaffold_12_19
COG0124 Histidyl-tRNA synthetase	1	17087_0042	0		1	15243_316	1	15911_589	1	884	2	scaffold_463_9
												scaffold_11_53
COG0184 Ribosomal protein S15P/S13E	1	17298_0054	1	13302_150 (putative)	1	17965_271	1	13624_103	1	1237	1	scaffold_29_27
COG0185 Ribosomal protein S19	1	17357_0009	1	13296_92	1	15243_454	1	15911_345	1	1137	1	scaffold_60_4
COG0186 Ribosomal protein S17	1	17357_0014	1	13296_86	1	15243_460	1	15911_339	1	1143	1	scaffold_15_5
COG0197 Ribosomal protein L16/L10E	1	13214_0046	1	13302_29	1	17965_211	1	13624_39	1	1206	1	scaffold_85_5

COG0200 Ribosomal protein L15	1	17498_0006	1	13296_6	1	15243_532	1	15911_237	1	1155 (probable)	1	scaffold_15_17 (probable)
COG0201 Preprotein translocase subunit SecY	1	17357_0028	1	13296_72	1	15243_473	1	15911_0324	1	1156	1	scaffold_15_18
COG0256 Ribosomal protein L18	1	17357_0024	1	13296_76	1	15243_469	1	15911_328	1	1152	1	scaffold_15_14
COG0495 Leucyl-tRNA synthetase	1	17442_0015	1	13302_138	1	17965_267	1	13606_182	1	1303	1	scaffold_9_15
			1	13280_38								
COG0522 Ribosomal protein S4 and related proteins	1	17357_0082	1	13296_34	1	15243_518	1	13606_288	1	1255	1	scaffold_29_47
COG0525 Valyl-tRNA synthetase	1	14887_0061			1	15243_007	1	13624_74	1	959	1	scaffold_24_52
COG0533 Metal-dependent proteases with possible chaperone activity (TIGR gcp: metalloendopeptidase) (PFAM Peptidase M22, glycoprotease)	1	17306_0016	1	12263_29	1	15243_429	1	13249_151	1	1107	1	scaffold_93_6

COG0541 Signal recognition particle GTPase (Ffh in bacteria, SRP54 in archaea)	1	17481_0001	2	13318_114	1	17965_500	2	13606_285	1	161	2	scaffold_104_23 (split)
				13318_17				15911_444				scaffold_104_22 (split)
												scaffold_18_19
Percentage of genome assembled		100.0%		94.3%		100.0%		100.0%		100.0 %		100.0%

Table S2: Cobalamin synthesis genes and gene synteny conservation at $P_{related} > 0.95$.
Synteny conservation at $P_{related} > 0.95$ is indicated in yellow. Red indicates genes involved in the cobalamin salvage pathway whereas blue indicates genes involved in cobalamin biosynthesis.

Gene name	APL	GPL	EPL	IPL	FER1	FER2
CysG						
CbiK						
CobI/CbiL					fer1_1324	fer2_scaff_72_0002
CobG						
CobJ/CbiH					fer1_0205	fer2_scaff_557_0006
CobM/CbiF					fer1_1323	fer2_scaff_72_0003
CobF						
CbiG					fer1_1827	fer2_scaff_18_0028
CbiD					fer1_0365	fer2_scaff_17_0018
CobK/CbiJ					fer1_0363	fer2_scaff_17_0016
CbiE					fer1_1325	fer2_scaff_618_0005
CobL						
CbiT					fer1_1325	fer2_scaff_618_0005
CobH/CbiC					fer1_0364	fer2_scaff_17_0017
CbiA/CobB	APL_17087_0043	GPL_13334_0133	EPL_15243_251	IPL_15911_0376	fer1_1828	fer2_scaff_18_0029

CbiA/CobB					fer1_1327	
CbiA N-terminus				IPL_13624_0310		
CbiA C-terminus				IPL_13624_0311		
BluB					fer1_0489	
CobNST					fer1_1606	fer2_scaff_83_0014
CobT	APL_13077_0015	GPL_13477_0053			fer1_1057	fer2_scaff_47_0011
CobT					fer1_1328	
CobN				IPL_15911_0378		
CobA* or CobO or BtuR					fer1_1326	
CobQ or CbiP	APL_13077_0011	GPL_13477_0051	EPL_15243_80	IPL_15911_0382	fer1_1059	fer2_scaff_47_0014
CobD1	APL_13077_0014	GPL_13477_0052	EPL_15243_683	IPL_15911_0383	fer1_1058	fer2_scaff_47_0015
CobC	APL_13077_0013			IPL_15911_0381		
CobU or CobP						
CobS/CobS	APL_13077_0016	GPL_13477_0049			fer1_0558	fer2_scaff_83_0014
CobV				IPL_15911_0379		
CbiZ	APL_13077_0012				fer1_1896	
PduO/EutT	APL_12068_0004	GPL_13334_0122	EPL_15243_46	IPL_13624_0131	fer1_0871	fer2_scaff_11_0057
CobD2/CbiB	APL_17325_0019	GPL_13477_0050		IPL_13624_0131	fer1_1061	fer2_scaff_47_0013
CobD2/CbiB						fer2_scaff_47_0012
CobY	APL_17325_0023	GPL_13374_0175		IPL_15911_0038		
CobZ C-terminus	APL_13077_0020			IPL_15911_0380		fer2_scaff_37_0018
CobZ N-terminus	APL_13077_0021					
thioredoxin peroxidase				IPL_15911_0385		

methylmalonyl-CoA mutase, alpha subunit N-terminus	APL_13077_0002						

Table S3: Molybdopterin biosynthesis, utilization, and transport genes. Synteny conservation at $P_{related} > 0.95$ is indicated in yellow. Annotations indicated in blue are for genes necessary for molybdopterin guanine dinucleotide biosynthesis. Red text indicates formate dehydrogenase genes. AS2TS model score indicates the protein structural modeling score assigned to these proteins. Only reasonable fits were included.

Manual curation	**APL**	**AS2TS model score**	**EPL**	**GPL**	**IPL**	**AS2TS model score**	**FER1**	**AS2TS model score**	**FER2**	**AS2TS model score**
Cluster 1										
MobB	17445_0021						1918		17_0057	
MoeA-1	17445_0022	**B**			13606_0508	**C1**	1920		17_0059	
MoeA-2	17445_0023	**B**			13606_0509	**B**	1921		17_0060 17_0061	
moaB	17445_0024				13606_0507	**B** (MogA model)	1786		64_0018	
MoaC	17445_0025	**A**			13606_0506		1787		64_0019	
Possible MobA	17445_0026	**C1**			13606_0517	**B**	1919		64_0029	
MoaE	17445_0027	**B**			15911_0168		212			
fdhA1	17445_0030	**C1**			13606_0516	**B**	1916	**C1**	17_0055	**B**
fdhD	17445_0032						1914		17_0053	
moaD	17445_0033				15911_0170		228		216_0008	
moaA	17445_0034		17965_270		15911_0169		227		216_0009	
moaA									9_0006	
Cluster 3										

fdhA2	17306_0015						549		5_0011	**B**
formate dehydrogenase (based on synteny)	17306_0014									
Cluster 4										
Aldehyde ferredoxin oxidoreductase with molybdopterin cofactor (based on synteny)	17112_0007									
MoaD	17112_0008	**A**							217_0004	
MoeA-1	17112_0010	**C1**								
MoeA-2	17112_0011	**B**								
MoaA									217_0005	
tungstate/molybdate binding ABC transporter solute-binding component	17112_0004				13606_0512					
molybdate ABC-transporter permease component	17112_0005				13606_0511	**B**				

molybdate ABC-transporter ATPase component	17112_0006		17965_470		13606_0510					
tungstate/molybdopterin-binding aldehyde ferredoxin oxidoreductase	17112_0007				13606_0513					
Possible MoeB	17454_0009			13290_0005	15911_0171		994		24_0011	
Possible MoeB	17452_0003								863_0003	
thioredoxin related to molybdopterin synthesis			15243_750				210			
surE: 5'/3'-nucleotidase SurE							224			

Table S4: The mevalonate pathway and ether lipid biosynthesis genes. Synteny conservation at $P_{related} > 0.95$ is indicated in yellow.

Annotation	**E.C.**	**APL**	**EPL**	**GPL**	**FER1**	**FER2**	**IPL**
Mevalonate pathway							
hypothetical nucleic acid-binding protein likely involved in mevalonate pathway		17433_0048	15243_104	13459_0238	1029	20_0023	13624_0355
Acetyl-CoA acetyltransferase	2.3.1.9	17433_0047	15243_105	13459_0237	1028	20_0024	13624_0356

hydroxymethylglutaryl-CoA synthase	2.3.3.10		15243_106	13459_0236	1027	20_0025	13624_0357
hydroxymethylglutaryl-CoA reductase	1.1.1.88	17387_0042	17965_86	13477_0039		82_0017	13624_0213
hydroxymethylglutaryl-CoA reductase	1.1.1.88 or	17387_0043					
hydroxymethylglutaryl-CoA reductase	1.1.1.34				271		
Mevalonate kinase	2.7.1.36	12885_0014	15243_764		294	107_0008	13606_0039
isopentenyl phosphate kinase			12876_19*	13334_0003	945	97_0016	13249_0046
isopentenyl-diphosphate delta-isomerase	5.3.3.2		12876_20	13334_0004	944	97_0015	13249_0045
To ether lipid biosynthesis (A. fulgidus pathway)							
geranylgeranyl diphosphate synthase	2.5.1.29	17387_0030	17965_136	13459_0158	1283	56_0006 56_0007	13606_0087
hypothetical protein likely involved in ether lipid synthesis			15243_0189	13459_0274	1278	56_0002	13624_0051
glycerol-1-phosphate dehydrogenase	1.1.1.261		15243_188	13459_0275	1280	56_0003*	13624_0050
geranylgeranylglyceryl phosphate synthase	2.5.1.41		17965_136	13327_0027	508	9_0060	13606_0280
(S)-2,3-di-O-geranylgeranylglyceryl phosphate synthase	2.5.1.42	17387_0029	17965_135	13327_0026	303	107_0020	13606_0281
Digeranylgeranylglycerophospholipid reductase	1.3.1.-	17087_0052		13459_0113	166	18_0018	13624_0101

Table S5: CRISPR-associated protein genes. Synteny conservation at $P_{related} > 0.95$ is indicated in yellow. Gene numbers indicated in red have the same order as Cas system type 1.

Cas gene	APL	EPL	GPL	IPL	FER1	FER2
Cas6	00190_7	15243_144			1216	362_0002
Csc2	00190_10	15243_146		15911_0446		480_0004
Csc1	00190_11	15243_147		15911_0445	1223	480_0003

Cas3	00190_12	15243_148	13455_0101		1220	
Cas4	00190_13	15243_149	13455_0105		1221	480_0001
Cas1	00190_14	15243_151	13455_0106		0003	
hypothetical CRISPR-protein					0011	
Cst2					0010	
Cas5		17965_118			0009	
Cas3	0004	15243_155		15911_0444	0008	
Cas4	0003	17965_341	13290_0050	15911_0443	0007	
Cas1	0002	15243_159		15911_0442	0006	
Cas2	0001 partial	15243_160 partial		15911_0441 partial	0005 full	
Cas GSU0053		15243_157	13455_0103			
Cas GSU0054		15243_158	13455_0104			
Cas4					0013	
Cas4					0004	
Csh2					1218	
Cas5					1219	
Cas3					1220	
Cas2	00190_15 partial				0002 full	
Csh1					1217	

Table S6: Method validation with well-characterized genomes. The query species' genes were annotated via the synteny-based method with the annotations of its ortholog in the subject species. White indicates correct annotations. Red indicates incorrect annotations. Gray indicates ambiguous cases. Bold indicates gene with shared function between the two comparisons. We only included the shared genes from the *E. coli* and *T. maritima* comparison.

Query species	Subject species	Query gene	Subject gene	Original annotation	Synteny-based annotation
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341639	YP_003536791	rpoN DNA-directed RNA polymerase subunit N (EC:2.7.7.6) K03058 DNA-directed RNA polymerase subunit N [EC:2.7.7.6] (db=KEGG evalue=6.0e-32 bit_score=139.0 identity=100.0 coverage=98.4848484848485) (BLAST)	rpoN; DNA-directed RNA polymerase subunit N (EC:2.7.7.6); K03058 DNA-directed RNA polymerase subunit N [EC:2.7.7.6] (db=KEGG evalue=2.0e-30 bit_score=134.0 identity=100.0 coverage=98.4375) (BLAST)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341642	YP_003536794	rpl18E 50S ribosomal protein L18e K02883 large subunit ribosomal protein L18e (db=KEGG evalue=1.0e-62 bit_score=241.0 identity=100.0 coverage=99.1596638655462) (BLAST)	rpl18R; ribosomal protein L18.eR; K02883 large subunit ribosomal protein L18e (db=KEGG evalue=2.0e-60 bit_score=233.0 identity=100.0 coverage=99.1379310344828) (BLAST)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341643	YP_003536795	sso:SSO0071 rpoD DNA-directed RNA polymerase, subunit D (RpoD) (EC:2.7.7.6) K03047 DNA-directed RNA polymerase subunit D [EC:2.7.7.6] (db=KEGG) (RBH)	hvo:HVO_2781 rpoD; DNA-directed RNA polymerase subunit D (EC:2.7.7.6); K03047 DNA-directed RNA polymerase subunit D [EC:2.7.7.6] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341645	YP_003536797	sso:SSO0073 rps4p 30S ribosomal protein S4P K02986 small subunit ribosomal protein S4 (db=KEGG) (RBH)	hvo:HVO_2783 rps4; ribosomal protein S4; K02986 small subunit ribosomal protein S4 (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341667	YP_003536952	sso:SSO0101 pheT phenylalanyl-tRNA synthetase subunit beta (EC:6.1.1.20) K01890 phenylalanyl-tRNA synthetase beta chain [EC:6.1.1.20] (db=KEGG) (RBH)	hvo:HVO_2947 pheT; phenylalanyl-tRNA synthetase beta subunit (EC:6.1.1.20); K01890 phenylalanyl-tRNA synthetase beta chain [EC:6.1.1.20] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341711	YP_003534119	sso:SSO0155 argC N-acetyl-gamma-glutamyl-phosphate reductase (EC:1.2.1.38) K00145 N-acetyl-gamma-glutamyl-phosphate/N-acetyl-gamma-aminoadipyl-phosphate reductase [EC:1.2.1.38 1.2.1.-] (db=KEGG) (RBH)	hvo:HVO_0045 argC; N-acetyl-gamma-glutamyl-phosphate reductase (EC:1.2.1.38); K00145 N-acetyl-gamma-glutamyl-phosphate/N-acetyl-gamma-aminoadipyl-phosphate reductase [EC:1.2.1.38 1.2.1.-] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341712	YP_003534118	sso:SSO0156 argB acetylglutamate/acetylaminoadipate kinase (EC:2.7.2.8) K00930 acetylglutamate/acetylaminoadipate kinase [EC:2.7.2.8 2.7.2.-] (db=KEGG) (RBH)	hvo:HVO_0044 argB; acetylglutamate kinase (EC:2.7.2.8); K00930 acetylglutamate/acetylaminoadipate kinase [EC:2.7.2.8 2.7.2.-] (db=KEGG) (RBH)

Sulfolobus solfataricus P2	Haloferax volcanii	NP_341716	YP_003534117	sso:SSO0160 argD acetylornithine aminotransferase ArgD (EC:2.6.1.11) K05830 acetylornithine/acetyl-lysine aminotransferase [EC:2.6.1.11 2.6.1.-] (db=KEGG) (RBH)	hvo:HVO_0043 argD; acetylornithine aminotransferase (EC:2.6.1.11); K00821 acetylornithine/N-succinyldiaminopimelate aminotransferase [EC:2.6.1.11 2.6.1.17] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341717	YP_003534116	sso:SSO0162 acetyl-lysine deacetylase K05831 acetyl-lysine deacetylase [EC:3.5.1.-] (db=KEGG) (RBH)	hvo:HVO_0042 argE; glutamate carboxypeptidase (EC:3.5.1.-); K05831 acetyl-lysine deacetylase [EC:3.5.1.-] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341731	YP_003534720	peptidyl-tRNA hydrolase (EC:3.1.1.29) K04794 peptidyl-tRNA hydrolase, PTH2 family [EC:3.1.1.29] (db=KEGG evalue=1.0e-62 bit_score=241.0 identity=100.0 coverage=99.1666666666667) (BLAST)	conserved hypothetical protein TIGR00283; K04794 peptidyl-tRNA hydrolase, PTH2 family [EC:3.1.1.29] (db=KEGG evalue=6.0e-58 bit_score=225.0 identity=100.0 coverage=99.1071428571429) (BLAST)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341738	YP_003534155	sso:SSO0182 hemL glutamate-1-semialdehyde aminotransferase (EC:5.4.3.8) K01845 glutamate-1-semialdehyde 2,1-aminomutase [EC:5.4.3.8] (db=KEGG) (RBH)	hvo:HVO_0081 hemL; glutamate-1-semialdehyde 2,1-aminomutase (EC:5.4.3.8); K01845 glutamate-1-semialdehyde 2,1-aminomutase [EC:5.4.3.8] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341739	YP_003534152	sso:SSO0183 hemC porphobilinogen deaminase K01749 hydroxymethylbilane synthase [EC:2.5.1.61] (db=KEGG) (RBH)	hvo:HVO 0078 hemC; porphobilinogen deaminase (EC:2.5.1.61); K01749 hydroxymethylbilane synthase [EC:2.5.1.61] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341772	YP_003534428	rps12P 30S ribosomal protein S12P K02950 small subunit ribosomal protein S12 (db=KEGG evalue=2.0e-60 bit_score=234.0 identity=100.0 coverage=99.3333333333333) (BLAST)	rps12; ribosomal protein S12; K02950 small subunit ribosomal protein S12 (db=KEGG evalue=2.0e-76 bit_score=286.0 identity=100.0 coverage=99.2957746478873) (BLAST)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341781	YP_003536755	ndk nucleoside diphosphate kinase (EC:2.7.4.6) K00940 nucleoside-diphosphate kinase [EC:2.7.4.6] (db=KEGG evalue=9.0e-75 bit_score=281.0 identity=100.0 coverage=99.2753623188406) (BLAST)	hvo:HVO_2740 ndk; nucleoside diphosphate kinase (EC:2.7.4.6); K00940 nucleoside-diphosphate kinase [EC:2.7.4.6] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341783	YP_003536753	rps28e 30S ribosomal protein S28e K02979 small subunit ribosomal protein S28e (db=KEGG evalue=2.0e-47 bit_score=191.0 identity=100.0 coverage=98.9795918367347) (BLAST)	rps28R; ribosomal protein S28.eR; K02979 small subunit ribosomal protein S28e (db=KEGG evalue=6.0e-35 bit_score=149.0 identity=100.0 coverage=98.6486486486486) (BLAST)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341803	YP_003534841	sso:SSO0254 hypothetical protein K09134 hypothetical protein (db=KEGG) (RBH)	hvo:HVO_0781 hypothetical protein; K09134 hypothetical protein (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341804	YP_003534842	nicotinamide mononucleotide adenylyltransferase, putative (EC:2.7.7.1) K00952 nicotinamide-nucleotide	hvo:HVO_0782 nicotinamide-nucleotide adenylyltransferase (EC:2.7.7.1); K00952 nicotinamide-nucleotide adenylyltransferase

				adenylyltransferase [EC:2.7.7.1] (db=KEGG evalue=2.0e-61 bit_score=237.0 identity=100.0 coverage=99.1452991452991) (BLAST)	[EC:2.7.7.1] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341815	YP_003535225	sso:SSO0266 tfe transcription factor E K03136 transcription initiation factor TFIIE alpha subunit (db=KEGG) (RBH)	hvo:HVO_1174 TFIIE alpha subunit; K03136 transcription initiation factor TFIIE alpha subunit (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341816	YP_003535224	sso:SSO0267 hypothetical protein K07254 hypothetical protein (db=KEGG) (RBH)	hvo:HVO_1173 SpoU-like RNA methylase; K07254 hypothetical protein (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341870	YP_003535038	sso:SSO0325 nuoH NADH dehydrogenase subunit H (EC:1.6.5.3) K00337 NADH dehydrogenase I subunit H [EC:1.6.5.3] (db=KEGG) (RBH)	hvo:HVO_0981 nuoH; NADH dehydrogenase-like complex subunit H (EC:1.6.5.-); K00337 NADH dehydrogenase I subunit H [EC:1.6.5.3] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341871	YP_003535039	sso:SSO0326 nuoI NADH dehydrogenase subunit I (EC:1.6.5.3) K00338 NADH dehydrogenase I subunit I [EC:1.6.5.3] (db=KEGG) (RBH)	hvo:HVO_0982 nuoI; NADH dehydrogenase-like complex subunit I (EC:1.6.5.-); K00338 NADH dehydrogenase I subunit I [EC:1.6.5.3] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341883	YP_003536771	sso:SSO0344 rplP0 acidic ribosomal protein P0 K02864 large subunit ribosomal protein L10 (db=KEGG) (RBH)	hvo:HVO_2756 rpl10; ribosomal protein L10; K02864 large subunit ribosomal protein L10 (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341945	YP_003536758	sso:SSO0406 metE-1 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase (EC:2.1.1.14) K00549 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase [EC:2.1.1.14] (db=KEGG) (RBH)	hvo:HVO_2743 methionine synthase vitamin-B12 independent; K00549 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase [EC:2.1.1.14] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341946	YP_003536757	sso:SSO0407 metE-2 methionine synthase (EC:2.1.1.14) K00549 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase [EC:2.1.1.14] (db=KEGG) (RBH)	hvo:HVO_2742 metE; 5-methyltetrahydropteroyltriglutamate-homocysteine methyltransferase (EC:2.1.1.-); K00549 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase [EC:2.1.1.14] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341950	YP_003535931	sso:SSO0412 eif2G translation initiation factor IF-2 subunit gamma K03242 translation initiation factor eIF-2 gamma subunit (db=KEGG) (RBH)	hvo:HVO_1901 tif2g; translation initiation factor aIF-2 gamma subunit; K03242 translation initiation factor eIF-2 gamma subunit (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341952	YP_003535929	sso:SSO0415 rpoE1 DNA-directed RNA polymerase subunit E' (EC:2.7.7.6) K03049 DNA-directed RNA polymerase subunit E' [EC:2.7.7.6] (db=KEGG) (RBH)	rpoE1; DNA-directed RNA polymerase subunit E (EC:2.7.7.6); K03049 DNA-directed RNA polymerase subunit E' [EC:2.7.7.6] (db=KEGG evalue=1.0e-79 bit_score=298.0 identity=100.0 coverage=99.4736842105263) (BLAST)
Sulfolobus	Haloferax	NP_341954	YP_003535927	sso:SSO0416 hypothetical protein K09735	hypothetical protein; K09735 hypothetical protein

solfataricus P2	volcanii			hypothetical protein (db=KEGG) (RBH)	(db=KEGG evalue=1.0e-66 bit_score=254.0 identity=100.0 coverage=99.4413407821229) (BLAST)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342006	YP_003535645	sis:LS215_1758 type III restriction protein res subunit (db=KEGG) (RBH)	hvo:HVO_1598 rad25c; DNA repair helicase RAD25 (EC:3.6.1.-) (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342007	YP_003535643	siy:YG5714_1721 protein of unknown function DUF790 K09744 hypothetical protein (db=KEGG) (RBH)	hvo:HVO_1596 protein of unknown function (DUF790) superfamily; K09744 hypothetical protein (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342019	YP_003536402	siy:YG5714_1709 phosphate ABC transporter ATP-binding protein K02036 phosphate transport system ATP-binding protein [EC:3.6.3.27] (db=KEGG) (RBH)	hvo:HVO_2378 pstB1; ABC-type transport system ATP-binding protein (probable substrate phosphate) (EC:3.6.3.27); K02036 phosphate transport system ATP-binding protein [EC:3.6.3.27] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342021	YP_003536400	sso:SSO0490 pstC phosphate ABC transporter permease K02037 phosphate transport system permease protein (db=KEGG) (RBH)	hvo:HVO_2376 pstC1; ABC-type transport system permease protein (probable substrate phosphate); K02037 phosphate transport system permease protein (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342057	YP_003534553	sso:SSO0527 pgk phosphoglycerate kinase (EC:2.7.2.3) K00927 phosphoglycerate kinase [EC:2.7.2.3] (db=KEGG) (RBH)	hvo:HVO_0480 pgk; phosphoglycerate kinase (EC:2.7.2.3); K00927 phosphoglycerate kinase [EC:2.7.2.3] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342058	YP_003534551	sso:SSO0528 gap glyceraldehyde-3-phosphate dehydrogenase (EC:1.2.1.59) K00150 glyceraldehyde-3-phosphate dehydrogenase (NAD(P)) [EC:1.2.1.59] (db=KEGG) (RBH)	hvo:HVO_0478 glyceraldehyde-3-phosphate dehydrogenase, type II (EC:1.2.1.59); K00150 glyceraldehyde-3-phosphate dehydrogenase (NAD(P)) [EC:1.2.1.59] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342089	YP_003534393	sso:SSO0563 atpA V-type ATP synthase subunit A (EC:3.6.3.14) K02117 V-type H+-transporting ATPase subunit A [EC:3.6.3.14] (db=KEGG) (RBH)	hvo:HVO_0316 atpA; A-type ATP synthase subunit A (EC:3.6.3.14); K02117 V-type H+-transporting ATPase subunit A [EC:3.6.3.14] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342090	YP_003534394	sso:SSO0564 atpB V-type ATP synthase subunit B (EC:3.6.3.14) K02118 V-type H+-transporting ATPase subunit B [EC:3.6.3.14] (db=KEGG) (RBH)	hvo:HVO_0317 atpB; A-type ATP synthase subunit B (EC:3.6.3.14); K02118 V-type H+-transporting ATPase subunit B [EC:3.6.3.14] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342100	YP_003535555	sso:SSO0576 ilvC-1 ketol-acid reductoisomerase (EC:1.1.1.86) K00053 ketol-acid reductoisomerase [EC:1.1.1.86] (db=KEGG) (RBH)	hvo:HVO_1506 ilvC; ketol-acid reductoisomerase (EC:1.1.1.86); K00053 ketol-acid reductoisomerase [EC:1.1.1.86] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342102	YP_003535557	sso:SSO0579 ilvB-2 acetolactate synthase catalytic subunit (EC:2.2.1.6) K01652 acetolactate synthase I/II/III large subunit [EC:2.2.1.6] (db=KEGG) (RBH)	hvo:HVO_1508 ilvB1; acetolactate synthase large subunit (EC:2.2.1.6); K01652 acetolactate synthase I/II/III large subunit [EC:2.2.1.6] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342116	YP_003536993	sso:SSO0594 hisA 1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino]	hvo:HVO_2988 hisA; phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase

				imidazole-4-carboxamide isomerase (EC:5.3.1.16) K01814 phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase [EC:5.3.1.16] (db=KEGG) (RBH)	(EC:5.3.1.16); K01814 phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase [EC:5.3.1.16] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342117	YP_003536991	sso:SSO0596 hisB imidazoleglycerol-phosphate dehydratase (EC:4.2.1.19) K01693 imidazoleglycerol-phosphate dehydratase [EC:4.2.1.19] (db=KEGG) (RBH)	hvo:HVO_2986 hisB; imidazoleglycerol-phosphate dehydratase (EC:4.2.1.19); K01693 imidazoleglycerol-phosphate dehydratase [EC:4.2.1.19] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342134	YP_003535504	sso:SSO0613 pyrI aspartate carbamoyltransferase regulatory subunit (EC:2.1.3.2) K00610 aspartate carbamoyltransferase regulatory subunit (db=KEGG) (RBH)	hvo:HVO_1455 pyrI; aspartate carbamoyltransferase, regulatory subunit; K00610 aspartate carbamoyltransferase regulatory subunit (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342135	YP_003535503	sso:SSO0614 pyrB aspartate carbamoyltransferase catalytic subunit (EC:2.1.3.2) K00609 aspartate carbamoyltransferase catalytic subunit [EC:2.1.3.2] (db=KEGG) (RBH)	hvo:HVO_1454 pyrB; aspartate carbamoyltransferase (EC:2.1.3.2); K00609 aspartate carbamoyltransferase catalytic subunit [EC:2.1.3.2] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342156	YP_003534123	sso:SSO0638 argG argininosuccinate synthase (EC:6.3.4.5) K01940 argininosuccinate synthase [EC:6.3.4.5] (db=KEGG) (RBH)	hvo:HVO_0049 argG; argininosuccinate synthase (EC:6.3.4.5); K01940 argininosuccinate synthase [EC:6.3.4.5] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342157	YP_003534122	sso:SSO0639 argH argininosuccinate lyase (EC:4.3.2.1) K01755 argininosuccinate lyase [EC:4.3.2.1] (db=KEGG) (RBH)	hvo:HVO_0048 argH; argininosuccinate lyase (EC:4.3.2.1); K01755 argininosuccinate lyase [EC:4.3.2.1] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342208	YP_003536561	sso:SSO0697 rpl30p 50S ribosomal protein L30P K02907 large subunit ribosomal protein L30 (db=KEGG) (RBH)	hvo:HVO_2543 rpl30; ribosomal protein L30; K02907 large subunit ribosomal protein L30 (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342212	YP_003536565	rpl32e 50S ribosomal protein L32e K02912 large subunit ribosomal protein L32e (db=KEGG evalue=5.0e-73 bit_score=275.0 identity=100.0 coverage=99.2753623188406) (BLAST)	hvo:HVO_2547 rpl32R; ribosomal protein L32.eR; K02912 large subunit ribosomal protein L32e (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342217	YP_003536570	sso:SSO0705 rps4E 30S ribosomal protein S4e K02987 small subunit ribosomal protein S4e (db=KEGG) (RBH)	hvo:HVO_2552 rps4R; ribosomal protein S4.eR; K02987 small subunit ribosomal protein S4e (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342218	YP_003536571	rpl24p 50S ribosomal protein L24P K02895 large subunit ribosomal protein L24 (db=KEGG evalue=1.0e-60 bit_score=234.0 identity=100.0 coverage=99.1596638655462) (BLAST)	rpl24; ribosomal protein L24; K02895 large subunit ribosomal protein L24 (db=KEGG evalue=8.0e-60 bit_score=231.0 identity=100.0 coverage=99.1525423728814) (BLAST)

Sulfolobus solfataricus P2	Haloferax volcanii	NP_342227	YP_003536581	sso:SSO0718 rpl4lp 50S ribosomal protein L4P K02930 large subunit ribosomal protein L4e (db=KEGG) (RBH)	hvo:HVO_2563 rpl4R; ribosomal protein L4.eR; K02930 large subunit ribosomal protein L4e (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342255	YP_003536762	sso:SSO0750 hypothetical protein K07572 putative nucleotide binding protein (db=KEGG) (RBH)	hvo:HVO_2747 predicted RNA-binding protein; K07572 putative nucleotide binding protein (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342257	YP_003536764	rpl21E 50S ribosomal protein L21e K02889 large subunit ribosomal protein L21e (db=KEGG evalue=2.0e-42 bit_score=174.0 identity=100.0 coverage=99.009900990099) (BLAST)	rpl21R; ribosomal protein L21.eR; K02889 large subunit ribosomal protein L21e (db=KEGG evalue=8.0e-49 bit_score=195.0 identity=100.0 coverage=98.9583333333333) (BLAST)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342335	YP_003536084	sso:SSO0830 rfbB-1 dTDP-glucose 4,6-dehydratase (RfbB-1) (EC:4.2.1.46) K01710 dTDP-glucose 4,6-dehydratase [EC:4.2.1.46] (db=KEGG) (RBH)	hvo:HVO_2059 galE5; UDP-glucose 4-epimerase (EC:5.1.3.2); K01710 dTDP-glucose 4,6-dehydratase [EC:4.2.1.46] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342336	YP_003536082	sso:SSO0831 sugar phosphate nucleotydyl transferase (EC:2.7.7.-) K00973 glucose-1-phosphate thymidylyltransferase [EC:2.7.7.24] (db=KEGG) (RBH)	hvo:HVO_2057 graD2; sugar nucleotidyltransferase (EC:2.7.7.-); K00973 glucose-1-phosphate thymidylyltransferase [EC:2.7.7.24] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342337	YP_003536083	sso:SSO0832 rfbD-1 dTDP-4-dehydrorhamnose reductase (RfbD-1) (EC:1.1.1.133) K00067 dTDP-4-dehydrorhamnose reductase [EC:1.1.1.133] (db=KEGG) (RBH)	hvo:HVO 2058 RmlD substrate binding domain superfamily; K00067 dTDP-4-dehydrorhamnose reductase [EC:1.1.1.133] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342383	YP_003536477	sso:SSO0890 trpD anthranilate phosphoribosyltransferase (EC:2.4.2.18) K00766 anthranilate phosphoribosyltransferase [EC:2.4.2.18] (db=KEGG) (RBH)	hvo:HVO_2456 trpD1; anthranilate phosphoribosyltransferase (EC:2.4.2.18); K00766 anthranilate phosphoribosyltransferase [EC:2.4.2.18] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342385	YP_003536475	sso:SSO0893 trpE anthranilate synthase component I K01657 anthranilate synthase component I [EC:4.1.3.27] (db=KEGG) (RBH)	hvo:HVO_2454 trpE; anthranilate synthase component I (EC:4.1.3.27); K01657 anthranilate synthase component I [EC:4.1.3.27] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342386	YP_003536474	sso:SSO0894 trpGD anthranilate synthase component II (EC:4.1.3.27) K01658 anthranilate synthase component II [EC:4.1.3.27] (db=KEGG) (RBH)	hvo:HVO_2453 trpG; anthranilate synthase component II (EC:4.1.3.27); K01658 anthranilate synthase component II [EC:4.1.3.27] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342409	YP_003536424	sso:SSO0917 glycine dehydrogenase subunit 2 (EC:1.4.4.2) K00283 glycine dehydrogenase subunit 2 [EC:1.4.4.2] (db=KEGG) (RBH)	hvo:HVO_2401 glycine cleavage system P-protein (EC:1.4.4.2); K00283 glycine dehydrogenase subunit 2 [EC:1.4.4.2] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342410	YP_003536425	sso:SSO0918 glycine dehydrogenase subunit 1 (EC:1.4.4.2) K00282 glycine	hvo:HVO_2402 gcvP; glycine dehydrogenase (decarboxylating) (glycine cleavage system protein

				dehydrogenase subunit 1 [EC:1.4.4.2] (db=KEGG) (RBH)	P-1) (EC:1.4.4.2); K00282 glycine dehydrogenase subunit 1 [EC:1.4.4.2] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342411	YP_003536427	sso:SSO0919 gcvT glycine cleavage system aminomethyltransferase T (EC:2.1.2.10) K00605 aminomethyltransferase [EC:2.1.2.10] (db=KEGG) (RBH)	hvo:HVO_2404 gcvT; aminomethyltransferase (glycine cleavage system protein T) (EC:2.1.2.10); K00605 aminomethyltransferase [EC:2.1.2.10] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342412	YP_003536426	gcvH glycine cleavage system protein H K02437 glycine cleavage system H protein (db=KEGG evalue=1.0e-59 bit_score=231.0 identity=100.0 coverage=99.2700729927007) (BLAST)	gcvH; glycine cleavage system protein H; K02437 glycine cleavage system H protein (db=KEGG evalue=3.0e-65 bit_score=249.0 identity=100.0 coverage=99.2063492063492) (BLAST)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342425	YP_003535715	sso:SSO0939 C/D box methylation guide ribonucleoprotein complex aNOP56 subunit (db=KEGG) (RBH)	hvo:HVO_1670 nop56; archaeal nucleolar protein-like protein (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342426	YP_003535714	sso:SSO0940 fibrillarin K04795 fibrillarin-like pre-rRNA processing protein (db=KEGG) (RBH)	hvo:HVO_1669 fib; fibrillarin-like pre-rRNA processing protein; K04795 fibrillarin-like pre-rRNA processing protein (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342477	YP_003536595	sso:SSO0996 nicotinate-nucleotide pyrophosphorylase (EC:2.4.2.19) K00767 nicotinate-nucleotide pyrophosphorylase (carboxylating) [EC:2.4.2.19] (db=KEGG) (RBH)	hvo:HVO_2579 nadC; nicotinate-nucleotide pyrophosphorylase (carboxylating) (EC:2.4.2.19); K00767 nicotinate-nucleotide pyrophosphorylase (carboxylating) [EC:2.4.2.19] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342478	YP_003536596	sso:SSO0997 nadB aspartate oxidase (NadB) (EC:1.4.3.16) K00278 L-aspartate oxidase [EC:1.4.3.16] (db=KEGG) (RBH)	hvo:HVO_2580 FAD binding domain, putative; K00278 L-aspartate oxidase [EC:1.4.3.16] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342479	YP_003536597	sso:SSO0998 nadA quinolinate synthetase K03517 quinolinate synthase [EC:2.5.1.72] (db=KEGG) (RBH)	hvo:HVO_2581 nadA; quinolinate synthetase complex, A subunit; K03517 quinolinate synthase [EC:2.5.1.72] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342537	YP_003535034	purE phosphoribosylaminoimidazole carboxylase catalytic subunit (PurE) (EC:4.1.1.21) K01588 5-(carboxyamino)imidazole ribonucleotide mutase [EC:5.4.99.18] (db=KEGG evalue=1.0e-77 bit_score=290.0 identity=100.0 coverage=99.3670886075949) (BLAST)	hvo:HVO_0977 purE; phosphoribosylaminoimidazole carboxylase, catalytic subunit (EC:4.1.1.21); K01588 5-(carboxyamino)imidazole ribonucleotide mutase [EC:5.4.99.18] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342538	YP_003535033	sso:SSO1065 purK phosphoribosylaminoimidazole carboxylase ATPase subunit (EC:4.1.1.21) K01589 5-(carboxyamino)imidazole ribonucleotide synthase [EC:6.3.4.18] (db=KEGG) (RBH)	hvo:HVO_0976 purK; phosphoribosylaminoimidazole carboxylase, ATPase subunit (EC:4.1.1.21); K01589 5-(carboxyamino)imidazole ribonucleotide synthase [EC:6.3.4.18] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342626	YP_003534628	sso:SSO1168 sugar ABC transporter K02023 multiple sugar transport system	hvo:HVO_0565 malK; ABC-type transport system ATP-binding protein (probable substrate maltose);

				ATP-binding protein (db=KEGG) (RBH)	K10112 maltose/maltodextrin transport system ATP-binding protein (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342628	YP_003534626	sso:SSO1170 sugar transport protein K02025 multiple sugar transport system permease protein (db=KEGG) (RBH)	hvo:HVO_0563 malF; ABC-type transport system permease protein (probable substrate maltose); K10114 maltooligosaccharide transport system permease protein (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342813	YP_003536611	sso:SSO1369 pdhA-1 pyruvate dehydrogenase alpha subunit (lipoamide) (EC:1.2.4.1) K00161 pyruvate dehydrogenase E1 component subunit alpha [EC:1.2.4.1] (db=KEGG) (RBH)	hvo:HVO_2595 oadhA2; 2-oxoacid dehydrogenase E1 component alpha subunit (EC:1.2.4.-); K00161 pyruvate dehydrogenase E1 component subunit alpha [EC:1.2.4.1] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342814	YP_003536612	sso:SSO1370 pdhB-1 pyruvate dehydrogenase beta subunit (lipoamide) (EC:1.2.4.1) K00162 pyruvate dehydrogenase E1 component subunit beta [EC:1.2.4.1] (db=KEGG) (RBH)	hvo:HVO_2596 oadhB2; 2-oxoacid dehydrogenase E1 component beta subunit (EC:1.2.4.-); K00162 pyruvate dehydrogenase E1 component subunit beta [EC:1.2.4.1] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342958	YP_003534730	sso:SSO1525 pdhA-2 pyruvate dehydrogenase alpha subunit (lipoamide) (EC:1.2.4.1) K00161 pyruvate dehydrogenase E1 component subunit alpha [EC:1.2.4.1] (db=KEGG) (RBH)	hvo:HVO_0669 oadhA3; 2-oxoacid dehydrogenase E1 component alpha subunit (EC:1.2.4.-); K00161 pyruvate dehydrogenase E1 component subunit alpha [EC:1.2.4.1] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_342960	YP_003534728	sso:SSO1527 acoX acetoin catabolism protein AcoX (db=KEGG) (RBH)	hvo:HVO_0667 ATP-NAD kinase (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_343639	YP_003534891	sso:SSO2265 hypothetical protein K07588 LAO/AO transport system kinase [EC:2.7.-.-] (db=KEGG) (RBH)	hvo:HVO_0831 argK; ArgK-type transport ATPase (EC:2.7.-.-); K07588 LAO/AO transport system kinase [EC:2.7.-.-] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_343640	YP_003534890	mcmA2 methylmalonyl-CoA mutase, alpha-subunit, chain B (mcmA2) (EC:5.4.99.2) K01849 methylmalonyl-CoA mutase, C-terminal domain [EC:5.4.99.2] (db=KEGG evalue=3.0e-63 bit_score=243.0 identity=100.0 coverage=99.290780141844) (BLAST)	methylmalonyl-CoA mutase , subunit B; K01849 methylmalonyl-CoA mutase, C-terminal domain [EC:5.4.99.2] (db=KEGG evalue=8.0e-74 bit_score=278.0 identity=100.0 coverage=99.2805755395683) (BLAST)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_343703	YP_003534755	sso:SSO2342 gpT-1 purine phosphoribosyltransferase (gpT-1) (EC:2.4.2.-) K07101 (db=KEGG) (RBH)	hvo:HVO_0694 gptA; purine phosphoribosyltransferase (EC:2.4.2.22); K07101 (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_343704	YP_003534753	sso:SSO2343 mtaP 5'-methylthioadenosine phosphorylase II (EC:2.4.2.28) K00772 5'-methylthioadenosine phosphorylase [EC:2.4.2.28] (db=KEGG) (RBH)	hvo:HVO_0692 mtaP; methylthioadenosine phosphorylase (EC:2.4.2.28); K00772 5'-methylthioadenosine phosphorylase [EC:2.4.2.28] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_343734	YP_003534209	sso:SSO2373 putative RNA-processing protein K06961 (db=KEGG) (RBH)	hvo:HVO_0134 RNA-binding Pno1 homolog; K06961 (db=KEGG) (RBH)
Sulfolobus	Haloferax	NP_343735	YP_003534210	sso:SSO2374 hypothetical protein K07178	hvo:HVO_0135 rio1; atypical protein kinase;

solfataricus P2	volcanii			RIO kinase 1 [EC:2.7.11.1] (db=KEGG) (RBH)	K07178 RIO kinase 1 [EC:2.7.11.1] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_343736	YP_003534211	eiF1A translation initiation factor IF-1A K03236 translation initiation factor eIF-1A (db=KEGG evalue=5.0e-56 bit_score=219.0 identity=100.0 coverage=99.0740740740741) (BLAST)	tif1A1; translation initiation factor aIF-1A; K03236 translation initiation factor eIF-1A (db=KEGG evalue=3.0e-48 bit_score=193.0 identity=100.0 coverage=98.9583333333333) (BLAST)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_344136	YP_003534948	sso:SSO2815 2-oxoacid--ferredoxin oxidoreductase, alpha chain (EC:1.2.7.-) K00174 2-oxoglutarate ferredoxin oxidoreductase subunit alpha [EC:1.2.7.3] (db=KEGG) (RBH)	hvo:HVO_0888 korA; oxoglutarate--ferredoxin oxidoreductase alpha subunit (EC:1.2.7.3); K00174 2-oxoglutarate ferredoxin oxidoreductase subunit alpha [EC:1.2.7.3] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_344137	YP_003534947	sso:SSO2816 2-oxoglutarate ferredoxin oxidoreductase subunit beta (EC:1.2.7.-) K00175 2-oxoglutarate ferredoxin oxidoreductase subunit beta [EC:1.2.7.3] (db=KEGG) (RBH)	hvo:HVO_0887 korB; oxoglutarate--ferredoxin oxidoreductase beta subunit (EC:1.2.7.3); K00175 2-oxoglutarate ferredoxin oxidoreductase subunit beta [EC:1.2.7.3] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_344168	YP_003534597	sso:SSO2848 ABC transporter, permease (glucose) K10197 glucose/arabinose transport system permease protein (db=KEGG) (RBH)	hvo:HVO_0531 tsgB1; ABC-type transport system permease protein (probable substrate sugar); K02025 multiple sugar transport system permease protein (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_344169	YP_003534598	sso:SSO2849 ABC transporter, permease (glucose) K10198 glucose/arabinose transport system permease protein (db=KEGG) (RBH)	hvo:HVO_0532 tsgC1; ABC-type transport system permease protein (probable substrate sugar); K02026 multiple sugar transport system permease protein (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341854	YP_003535358	sso:SSO0307 aroC; chorismate synthase (EC:4.2.3.5); K01736 chorismate synthase [EC:4.2.3.5] (db=KEGG) (RBH)	hvo:HVO_1306 aroC; chorismate synthase (EC:4.2.3.5); K01736 chorismate synthase [EC:4.2.3.5] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_341856	YP_003535360	sso:SSO0309 aroA; 3-phosphoshikimate 1-carboxyvinyltransferase (EC:2.5.1.19); K00800 3-phosphoshikimate 1-carboxyvinyltransferase [EC:2.5.1.19] (db=KEGG) (RBH)	hvo:HVO_1308 aroA; 3-phosphoshikimate 1-carboxyvinyltransferase (EC:2.5.1.19); K00800 3-phosphoshikimate 1-carboxyvinyltransferase [EC:2.5.1.19] (db=KEGG) (RBH)
Sulfolobus solfataricus P2	Haloferax volcanii	NP_343945	YP_003536146	sso:SSO2615 dppF-3; peptide ABC transporter ATP-binding protein; K02032 peptide/nickel transport system ATP-binding protein (db=KEGG) (RBH)	hvo:HVO_2122 dppF4; ABC-type transport system ATP-binding protein (probable substrate dipeptides/oligopeptides); K02032 peptide/nickel transport system ATP-binding protein (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414624	NP_219777	ecv:APECO1_1904 mraW; S-adenosyl-methyltransferase MraW; K03438 S-adenosyl-methyltransferase [EC:2.1.1.-] (db=KEGG) (RBH)	cta:CTA_0294 mraW S-adenosyl-methyltransferase MraW (EC:2.1.1.-) K03438 S-adenosyl-methyltransferase [EC:2.1.1.-] (db=KEGG) (RBH)
Escherichia	Chlamydia	NP_414626	NP_219775	sbo:SBO_0072 ftsI; penicillin-binding	cta:CTA_0292 pbp3 penicillin-binding protein

coli K12	trachomatis D/UW-3/CX			protein 3; K03587 cell division protein FtsI (penicillin-binding protein 3) [EC:2.4.1.129] (db=KEGG) (RBH)	K03587 cell division protein FtsI (penicillin-binding protein 3) [EC:2.4.1.129] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414627	NP_219774	eco:b0085 murE, ECK0086, JW0083; UDP-N-acetylmuramoyl-L-alanyl-D-glutamate:meso-diaminopimelate ligase (EC:6.3.2.13); K01928 UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase [EC:6.3.2.13] (db=KEGG) (RBH)	ctr:CT269 murE UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase (EC:6.3.2.13) K01928 UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase [EC:6.3.2.13] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414628	NP_220275	eco:b0086 murF, ECK0087, JW0084, mra; UDP-N-acetylmuramoyl-tripeptide:D-alanyl-D-alanine ligase (EC:6.3.2.10); K01929 UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl-D-alanine ligase [EC:6.3.2.10] (db=KEGG) (RBH)	ctr:CT756 murF UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase K01929 UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl-D-alanine ligase [EC:6.3.2.10] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414629	NP_220276	ecm:EcSMS35_0092 mraY; phospho-N-acetylmuramoyl-pentapeptide-transferase (EC:2.7.8.13); K01000 phospho-N-acetylmuramoyl-pentapeptide-transferase [EC:2.7.8.13] (db=KEGG) (RBH)	ctz:CTB_7621 mraY phospho-N-acetylmuramoyl-pentapeptide-transferase K01000 phospho-N-acetylmuramoyl-pentapeptide-transferase [EC:2.7.8.13] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414630	NP_220277	ecd:ECDH10B_0070 murD; UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase; K01925 UDP-N-acetylmuramoylalanine--D-glutamate ligase [EC:6.3.2.9] (db=KEGG) (RBH)	ctz:CTB_7631 murD UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase K01925 UDP-N-acetylmuramoylalanine--D-glutamate ligase [EC:6.3.2.9] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414632	NP_220280	ebw:BWG_0085 murG; undecaprenyldiphospho-muramoylpentapeptide beta-N-acetylglucosaminyltransferase; K02563 UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase [EC:2.4.1.227] (db=KEGG) (RBH)	cta:CTA_0831 murG undecaprenyldiphospho-muramoylpentapeptide beta-N-acetylglucosaminyltransferase (EC:2.4.1.227) K02563 UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase [EC:2.4.1.227] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414633	NP_220281	ebw:BWG_0086 murC; UDP-N-acetylmuramate--L-alanine ligase; K01924 UDP-N-acetylmuramate--alanine ligase [EC:6.3.2.8] (db=KEGG) (RBH)	ctr:CT762 murC, ddlA bifunctional D-alanyl-alanine synthetase A/UDP-N-acetylmuramate--L-alanine ligase K01921 D-alanine-D-alanine ligase [EC:6.3.2.4] K01924 UDP-N-acetylmuramate--alanine ligase [EC:6.3.2.8] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414711	NP_220199	efe:EFER_0191 rpsB; 30S ribosomal protein S2; K02967 small subunit ribosomal protein S2 (db=KEGG) (RBH)	ctr:CT680 rpsB 30S ribosomal protein S2 K02967 small subunit ribosomal protein S2 (db=KEGG) (RBH)

Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414712	NP_220198	ssn:SSON_0182 tsf; elongation factor Ts; K02357 elongation factor EF-Ts (db=KEGG) (RBH)	ctr:CT679 tsf elongation factor Ts K02357 elongation factor EF-Ts (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414713	NP_220197	sbc:SbBS512_E0164 pyrH; uridylate kinase (EC:2.7.4.22); K09903 uridylate kinase [EC:2.7.4.22] (db=KEGG) (RBH)	ctr:CT678 pyrH uridylate kinase K09903 uridylate kinase [EC:2.7.4.22] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414714	NP_220196	ssn:SSON_0184 frr; ribosome recycling factor; K02838 ribosome recycling factor (db=KEGG) (RBH)	ctr:CT677 frr ribosome recycling factor K02838 ribosome recycling factor (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414719	NP_219746	ssn:SSON_0189 yaeT; outer membrane protein assembly factor YaeT; K07277 outer membrane protein (db=KEGG) (RBH)	ctr:CT241 yaeT OMP85 family membrane protein K07277 outer membrane protein (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414721	NP_219748	sdy:SDY_0195 lpxD; UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase; K02536 UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase [EC:2.3.1.-] (db=KEGG) (RBH)	cta:CTA_0265 lpxD UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC:2.3.1.-) K02536 UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase [EC:2.3.1.-] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414722	NP_220047	ssn:SSON_0192 fabZ; (3R)-hydroxymyristoyl-ACP dehydratase; K02372 3R-hydroxymyristoyl ACP dehydrase [EC:4.2.1.-] (db=KEGG) (RBH)	cta:CTA_0581 fabZ (3R)-hydroxymyristoyl-ACP dehydratase (EC:4.2.1.-) K02372 3R-hydroxymyristoyl ACP dehydrase [EC:4.2.1.-] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414723	NP_220046	ssn:SSON_0193 lpxA; UDP-N-acetylglucosamine acyltransferase (EC:2.3.1.129); K00677 UDP-N-acetylglucosamine acyltransferase [EC:2.3.1.129] (db=KEGG) (RBH)	cta:CTA_0580 lpxA UDP-N-acetylglucosamine acyltransferase (EC:2.3.1.129) K00677 UDP-N-acetylglucosamine acyltransferase [EC:2.3.1.129] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414948	NP_220249	eco:b0414 ribD, ECK0408, JW0404, ribG, ybaE; fused diaminohydroxyphosphoribosylaminopyrimidine deaminase and 5-amino-6-(5-phosphoribosylamino) uracil reductase (EC:3.5.4.26 1.1.1.193); K11752 diaminohydroxyphosphoribosylaminopyrimidine deaminase / 5-amino-6-(5-phosphoribosylamino)uracil reductase [EC:3.5.4.26 1.1.1.193] (db=KEGG) (RBH)	ctr:CT730 ribD riboflavin deaminase K11752 diaminohydroxyphosphoribosylaminopyrimidine deaminase / 5-amino-6-(5-phosphoribosylamino)uracil reductase [EC:3.5.4.26 1.1.1.193] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414949	NP_220251	sfv:SFV_0380 ribH; 6,7-dimethyl-8-ribityllumazine synthase; K00794 riboflavin synthase beta chain [EC:2.5.1.-] (db=KEGG) (RBH)	ctj:JALI_7371 ribH 6,7-dimethyl-8-ribityllumazine synthase K00794 riboflavin synthase beta chain [EC:2.5.1.-] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis	NP_414971	NP_220225	ssn:SSON_0420 clpP; ATP-dependent Clp protease proteolytic subunit (EC:3.4.21.92);	cta:CTA_0767 clpP ATP-dependent Clp protease proteolytic subunit (EC:3.4.21.92) K01358 ATP-

	D/UW-3/CX			K01358 ATP-dependent Clp protease, protease subunit [EC:3.4.21.92] (db=KEGG) (RBH)	dependent Clp protease, protease subunit [EC:3.4.21.92] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_414972	NP_220224	ssn:SSON_0421 clpX; ATP-dependent protease ATP-binding subunit ClpX; K03544 ATP-dependent Clp protease ATP-binding subunit ClpX (db=KEGG) (RBH)	ctr:CT705 clpX ATP-dependent protease ATP-binding subunit ClpX K03544 ATP-dependent Clp protease ATP-binding subunit ClpX (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_415254	NP_219557	ecy:ECSE_0785 sucA; 2-oxoglutarate dehydrogenase E1 component; K00164 2-oxoglutarate dehydrogenase E1 component [EC:1.2.4.2] (db=KEGG) (RBH)	ctr:CT054 sucA 2-oxoglutarate dehydrogenase E1 component (EC:1.2.4.2) K00164 2-oxoglutarate dehydrogenase E1 component [EC:1.2.4.2] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_415255	NP_219558	eck:EC55989_0710 sucB; dihydrolipoamide succinyltransferase (EC:2.3.1.61); K00658 2-oxoglutarate dehydrogenase E2 component (dihydrolipoamide succinyltransferase) [EC:2.3.1.61] (db=KEGG) (RBH)	ctl:CTLon_0306 sucB dihydrolipoamide succinyltransferase K00658 2-oxoglutarate dehydrogenase E2 component (dihydrolipoamide succinyltransferase) [EC:2.3.1.61] (db=KEGG) (RBH)
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_415256**	**NP_220342**	**efe:EFER_2385 sucC; succinyl-CoA synthetase subunit beta (EC:6.2.1.5); K01903 succinyl-CoA synthetase beta subunit [EC:6.2.1.5] (db=KEGG) (RBH); B**	**ctr:CT821 sucC succinyl-CoA synthetase subunit beta (EC:6.2.1.5) K01903 succinyl-CoA synthetase beta subunit [EC:6.2.1.5] (db=KEGG) (RBH)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_415257**	**NP_220343**	**eci:UTI89_C0724 sucD; succinyl-CoA synthetase subunit alpha (EC:6.2.1.5); K01902 succinyl-CoA synthetase alpha subunit [EC:6.2.1.5] (db=KEGG) (RBH)**	**cta:CTA_0896 sucD succinyl-CoA synthetase subunit alpha (EC:6.2.1.5) K01902 succinyl-CoA synthetase alpha subunit [EC:6.2.1.5] (db=KEGG) (RBH)**
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_415609	NP_219744	ssn:SSON_1111 fabH; 3-oxoacyl-(acyl carrier protein) synthase III (EC:2.3.1.41); K00648 3-oxoacyl-[acyl-carrier-protein] synthase III [EC:2.3.1.180] (db=KEGG) (RBH)	cta:CTA_0261 fabH 3-oxoacyl-(acyl carrier protein) synthase III (EC:2.3.1.41) K00648 3-oxoacyl-[acyl-carrier-protein] synthase III [EC:2.3.1.180] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_415610	NP_219743	ecw:EcE24377A_1213 fabD; malonyl CoA-acyl carrier protein transacylase (EC:2.3.1.39); K00645 [acyl-carrier-protein] S-malonyltransferase [EC:2.3.1.39] (db=KEGG) (RBH)	ctj:JALI_2331 fabD malonyl-CoA-[acyl-carrier-protein] transacylase K00645 [acyl-carrier-protein] S-malonyltransferase [EC:2.3.1.39] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_415611	NP_219742	sbc:SbBS512_E2231 fabG; 3-ketoacyl-(acyl-carrier-protein) reductase (EC:1.1.1.100); K00059 3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100] (db=KEGG) (RBH)	ctr:CT237 fabG 3-ketoacyl-(acyl-carrier-protein) reductase (EC:1.1.1.100) K00059 3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_415759	NP_219702	eco:b1243 oppA, ECK1237, JW1235; oligopeptide transporter subunit; K02035 peptide/nickel transport system substrate-binding protein (db=KEGG) (RBH)	ctr:CT198 oppA_3 oligopeptide binding protein K02035 peptide/nickel transport system substrate-binding protein (db=KEGG) (RBH)

Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_415760	NP_219703	sbo:SBO_1823 oppB; oligopeptide transporter permease; K02033 peptide/nickel transport system permease protein (db=KEGG) (RBH)	ctr:CT199 oppB_1 oligopeptide permease K02033 peptide/nickel transport system permease protein (db=KEGG) (RBH)
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_415761**	**NP_219704**	**eoj:ECO26_1756 oppC; oligopeptide transporter subunit OppC; K02034 peptide/nickel transport system permease protein (db=KEGG) (RBH)**	**ctz:CTB_1941 oppC oligopeptide transport system membrane permease K02034 peptide/nickel transport system permease protein (db=KEGG) (RBH)**
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_415776	NP_219674	sfl:SF1263 trpA; tryptophan synthase subunit alpha (EC:4.2.1.20); K01695 tryptophan synthase alpha chain [EC:4.2.1.20] (db=KEGG) (RBH)	ctr:CT171 trpA tryptophan synthase subunit alpha K01695 tryptophan synthase alpha chain [EC:4.2.1.20] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_415777	NP_219673	sfl:SF1264 trpB; tryptophan synthase subunit beta (EC:4.2.1.20); K01696 tryptophan synthase beta chain [EC:4.2.1.20] (db=KEGG) (RBH)	ctr:CT170 trpB tryptophan synthase subunit beta (EC:4.2.1.20) K01696 tryptophan synthase beta chain [EC:4.2.1.20] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_416195	NP_220206	eco:b1680 sufS, csdB, ECK1676, JW1670, ynhB; cysteine desulfurase, stimulated by SufE; selenocysteine lyase, PLP-dependent (EC:4.4.1.16); K11717 cysteine desulfurase / selenocysteine lyase [EC:2.8.1.7 4.4.1.16] (db=KEGG) (RBH)	ctr:CT687 yfhO_1 cysteine desulfurase K11717 cysteine desulfurase / selenocysteine lyase [EC:2.8.1.7 4.4.1.16] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_416197	NP_220204	eco:b1682 sufC, ECK1678, JW1672, ynhD; component of SufBCD complex, ATP-binding component of ABC superfamily; K09013 Fe-S cluster assembly ATP-binding protein (db=KEGG) (RBH)	cta:CTA_0746 ABC transporter ATP-binding protein K09013 Fe-S cluster assembly ATP-binding protein (db=KEGG) (RBH)
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_416229**	**NP_220357**	**ecr:ECIAI1_1770 pheS; phenylalanyl-tRNA synthetase subunit alpha (EC:6.1.1.20); K01889 phenylalanyl-tRNA synthetase alpha chain [EC:6.1.1.20] (db=KEGG) (RBH)**	**ctr:CT836 pheS phenylalanyl-tRNA synthetase subunit alpha (EC:6.1.1.20) K01889 phenylalanyl-tRNA synthetase alpha chain [EC:6.1.1.20] (db=KEGG) (RBH)**
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_416231	NP_220356	ribosomal protein L20; K02887 large subunit ribosomal protein L20 (db=KEGG evalue=8.0e-50 bit_score=198.0 identity=99.15 coverage=99.1525423728814) (BLAST)	rplT 50S ribosomal protein L20 K02887 large subunit ribosomal protein L20 (db=KEGG evalue=5.0e-65 bit_score=249.0 identity=100.0 coverage=99.1869918699187) (BLAST)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_416233	NP_220354	eoi:ECO111_2227 infC; translation initiation factor IF-3; K02520 translation initiation factor IF-3 (db=KEGG) (RBH)	ctr:CT833 infC translation initiation factor IF-3 K02520 translation initiation factor IF-3 (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_416375	NP_220016	ecx:EcHS_A1954 ruvA; Holliday junction DNA helicase RuvA; K03550 holliday junction DNA helicase RuvA (db=KEGG)	ctr:CT501 ruvA Holliday junction DNA helicase RuvA K03550 holliday junction DNA helicase RuvA (db=KEGG) (RBH)

				(RBH)	
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_416377	NP_220017	sbc:SbBS512_E1109 ruvC; Holliday junction resolvase (EC:3.1.22.4); K01159 crossover junction endodeoxyribonuclease RuvC [EC:3.1.22.4] (db=KEGG) (RBH)	ctb:CTL0764 ruvC Holliday junction resolvase K01159 crossover junction endodeoxyribonuclease RuvC [EC:3.1.22.4] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_416737	NP_220348	ecj:JW2228 nrdA; ribonucleoside diphosphate reductase 1, alpha subunit; K00525 ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1] (db=KEGG) (RBH)	ctr:CT827 nrdA ribonucleotide-diphosphate reductase subunit alpha (EC:1.17.4.1) K00525 ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_416738	NP_220349	sfv:SFV_2308 nrdB; ribonucleotide-diphosphate reductase subunit beta (EC:1.17.4.1); K00526 ribonucleoside-diphosphate reductase beta chain [EC:1.17.4.1] (db=KEGG) (RBH)	ctr:CT828 nrdB ribonucleotide-diphosphate reductase subunit beta (EC:1.17.4.1) K00526 ribonucleoside-diphosphate reductase beta chain [EC:1.17.4.1] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417097	NP_219530	rplS; 50S ribosomal protein L19; K02884 large subunit ribosomal protein L19 (db=KEGG evalue=4.0e-60 bit_score=232.0 identity=100.0 coverage=99.1304347826087) (BLAST)	rplS 50S ribosomal protein L19 K02884 large subunit ribosomal protein L19 (db=KEGG evalue=4.0e-63 bit_score=243.0 identity=100.0 coverage=99.1735537190083) (BLAST)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417098	NP_219529	sfl:SF2667 trmD; tRNA (guanine-N(1)-)-methyltransferase (EC:2.1.1.31); K00554 tRNA (guanine-N1-)-methyltransferase [EC:2.1.1.31] (db=KEGG) (RBH)	ctr:CT027 trmD tRNA (guanine-N(1)-)-methyltransferase/unknown domain fusion protein (EC:2.1.1.31) K00554 tRNA (guanine-N1-)-methyltransferase [EC:2.1.1.31] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417101	NP_219527	sbo:SBO_2746 ffh; signal recognition particle protein; K03106 signal recognition particle subunit SRP54 (db=KEGG) (RBH)	ctr:CT025 ffh signal recognition particle, subunit FFH/SRP54 K03106 signal recognition particle subunit SRP54 (db=KEGG) (RBH)
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417635**	**NP_219597**	**eum:ECUMN_3648 truB; tRNA pseudouridine synthase B (EC:4.2.1.70); K03177 tRNA pseudouridine synthase B [EC:5.4.99.12] (db=KEGG) (RBH)**	**ctr:CT094 truB tRNA pseudouridine synthase B K03177 tRNA pseudouridine synthase B [EC:5.4.99.12] (db=KEGG) (RBH)**
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417637	NP_219599	ecz:ECS88_3552 infB; translation initiation factor IF-2; K02519 translation initiation factor IF-2 (db=KEGG) (RBH)	ctr:CT096 infB translation initiation factor IF-2 K02519 translation initiation factor IF-2 (db=KEGG) (RBH)
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417638**	**NP_219600**	**ecz:ECS88_3553 nusA; transcription elongation factor NusA; K02600 N utilization substance protein A (db=KEGG) (RBH)**	**cta:CTA_0103 nusA transcription elongation factor NusA K02600 N utilization substance protein A (db=KEGG) (RBH)**
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417650	NP_219928	ecd:ECDH10B_3357 obgE; GTPase ObgE; K03979 GTP-binding protein (db=KEGG) (RBH)	ctr:CT418 obgE, cgtA, obg, yhbZ GTPase ObgE K03979 GTP-binding protein (db=KEGG) (RBH)
Escherichia	Chlamydia	NP_417652	NP_219929	rpmA; 50S ribosomal protein L27; K02899	rpmA 50S ribosomal protein L27 K02899 large

coli K12	trachomatis D/UW-3/CX			large subunit ribosomal protein L27 (db=KEGG evalue=6.0e-41 bit_score=169.0 identity=100.0 coverage=98.8235294117647) (BLAST)	subunit ribosomal protein L27 (db=KEGG evalue=2.0e-40 bit_score=167.0 identity=100.0 coverage=98.7951807228916) (BLAST)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417653	NP_219930	rplU; 50S ribosomal protein L21; K02888 large subunit ribosomal protein L21 (db=KEGG evalue=4.0e-52 bit_score=206.0 identity=99.03 coverage=99.0291262135922) (BLAST)	rplU 50S ribosomal protein L21 K02888 large subunit ribosomal protein L21 (db=KEGG evalue=2.0e-46 bit_score=187.0 identity=100.0 coverage=99.0654205607477) (BLAST)
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417697**	**NP_219629**	**rpsI; 30S ribosomal protein S9; K02996 small subunit ribosomal protein S9 (db=KEGG evalue=9.0e-69 bit_score=261.0 identity=100.0 coverage=99.2307692307692) (BLAST)**	**rpsI 30S ribosomal protein S9 K02996 small subunit ribosomal protein S9 (db=KEGG evalue=2.0e-57 bit_score=224.0 identity=100.0 coverage=99.2248062015504) (BLAST)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417698**	**NP_219628**	**rplM; 50S ribosomal protein L13; K02871 large subunit ribosomal protein L13 (db=KEGG evalue=5.0e-78 bit_score=292.0 identity=100.0 coverage=99.2957746478873) (BLAST)**	**ctz:CTB_1241 rplM 50S ribosomal protein L13 K02871 large subunit ribosomal protein L13 (db=KEGG) (RBH)**
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417721	NP_219626	accB; acetyl-CoA carboxylase biotin carboxyl carrier protein subunit (EC:6.4.1.2); K02160 acetyl-CoA carboxylase biotin carboxyl carrier protein (db=KEGG evalue=1.0e-53 bit_score=211.0 identity=100.0 coverage=99.3589743589744) (BLAST)	ctb:CTL0378 accB acetyl-CoA carboxylase biotin carboxyl carrier protein subunit K02160 acetyl-CoA carboxylase biotin carboxyl carrier protein (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417722	NP_219627	ecp:ECP 3341 biotin carboxylase C-terminal domain; region: Biotin_carb_C; cl08365 (EC:6.4.1.2); K01961 acetyl-CoA carboxylase, biotin carboxylase subunit [EC:6.4.1.2 6.3.4.14] (db=KEGG) (RBH)	ctr:CT124 accC acetyl-CoA carboxylase biotin carboxylase subunit (EC:6.4.1.2) K01961 acetyl-CoA carboxylase, biotin carboxylase subunit [EC:6.4.1.2 6.3.4.14] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417753	NP_220021	rplQ; 50S ribosomal protein L17; K02879 large subunit ribosomal protein L17 (db=KEGG evalue=6.0e-68 bit_score=258.0 identity=100.0 coverage=99.2125984251969) (BLAST)	rplQ 50S ribosomal protein L17 K02879 large subunit ribosomal protein L17 (db=KEGG evalue=1.0e-76 bit_score=287.0 identity=100.0 coverage=99.290780141844) (BLAST)
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417754**	**NP_220022**	**stt:t4090 rpoA; DNA-directed RNA polymerase subunit alpha (EC:2.7.7.6); K03040 DNA-directed RNA polymerase subunit alpha [EC:2.7.7.6] (db=KEGG) (RBH)**	**ctr:CT507 rpoA DNA-directed RNA polymerase subunit alpha (EC:2.7.7.6) K03040 DNA-directed RNA polymerase subunit alpha [EC:2.7.7.6] (db=KEGG) (RBH)**
Escherichia coli K12	**Chlamydia trachomatis**	**NP_417756**	**NP_220023**	**rpsK; 30S ribosomal protein S11; K02948 small subunit ribosomal protein S11**	**rpsK 30S ribosomal protein S11 K02948 small subunit ribosomal protein S11 (db=KEGG**

	D/UW-3/CX			**(db=KEGG evalue=3.0e-69 bit_score=263.0 identity=100.0 coverage=99.2248062015504) (BLAST)**	**evalue=3.0e-55 bit_score=216.0 identity=100.0 coverage=99.2424242424242) (BLAST)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417757**	**NP_220024**	**rpsM; 30S ribosomal protein S13; K02952 small subunit ribosomal protein S13 (db=KEGG evalue=3.0e-61 bit_score=236.0 identity=100.0 coverage=99.1525423728814) (BLAST)**	**rpsM 30S ribosomal protein S13 K02952 small subunit ribosomal protein S13 (db=KEGG evalue=1.0e-54 bit_score=214.0 identity=100.0 coverage=86.8852459016393) (BLAST)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417759**	**NP_220025**	**sbc:SbBS512_E3685 secY; preprotein translocase subunit SecY; K03076 preprotein translocase subunit SecY (db=KEGG) (RBH)**	**ctr:CT510 secY preprotein translocase subunit SecY K03076 preprotein translocase subunit SecY (db=KEGG) (RBH)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417762**	**NP_220027**	**sfv:SFV_3323 rpsE; 30S ribosomal protein S5; K02988 small subunit ribosomal protein S5 (db=KEGG) (RBH)**	**ctb:CTL0774 rpsE 30S ribosomal protein S5 K02988 small subunit ribosomal protein S5 (db=KEGG) (RBH)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417764**	**NP_220029**	**sfl:SF3337 rplF; 50S ribosomal protein L6; K02933 large subunit ribosomal protein L6 (db=KEGG) (RBH)**	**ctz:CTB_5171 rplF 50S ribosomal protein L6 K02933 large subunit ribosomal protein L6 (db=KEGG) (RBH)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417765**	**NP_220030**	**rpsH; 30S ribosomal protein S8; K02994 small subunit ribosomal protein S8 (db=KEGG evalue=2.0e-69 bit_score=263.0 identity=100.0 coverage=99.2307692307692) (BLAST)**	**rpsH 30S ribosomal protein S8 K02994 small subunit ribosomal protein S8 (db=KEGG evalue=5.0e-72 bit_score=272.0 identity=100.0 coverage=99.2481203007519) (BLAST)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417767**	**NP_220031**	**enc:ECL_04685 50S ribosomal protein L5; K02931 large subunit ribosomal protein L5 (db=KEGG) (RBH)**	**ctj:JALI_5191 rplE 50S ribosomal protein L5 K02931 large subunit ribosomal protein L5 (db=KEGG) (RBH)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417769**	**NP_220033**	**rplN; 50S ribosomal protein L14; K02874 large subunit ribosomal protein L14 (db=KEGG evalue=2.0e-64 bit_score=247.0 identity=100.0 coverage=99.1869918699187) (BLAST)**	**rplN 50S ribosomal protein L14 K02874 large subunit ribosomal protein L14 (db=KEGG evalue=6.0e-63 bit_score=242.0 identity=100.0 coverage=99.1803278688525) (BLAST)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417770**	**NP_220034**	**rpsQ; 30S ribosomal protein S17; K02961 small subunit ribosomal protein S17 (db=KEGG evalue=2.0e-41 bit_score=171.0 identity=100.0 coverage=98.8095238095238) (BLAST)**	**rpsQ 30S ribosomal protein S17 K02961 small subunit ribosomal protein S17 (db=KEGG evalue=1.0e-39 bit_score=164.0 identity=100.0 coverage=98.7951807228916) (BLAST)**
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417772	NP_220036	rplP; 50S ribosomal protein L16; K02878 large subunit ribosomal protein L16 (db=KEGG evalue=2.0e-72 bit_score=274.0 identity=100.0 coverage=99.2647058823529) (BLAST)	rplP 50S ribosomal protein L16 K02878 large subunit ribosomal protein L16 (db=KEGG evalue=7.0e-74 bit_score=278.0 identity=100.0 coverage=99.2753623188406) (BLAST)
Escherichia coli K12	**Chlamydia trachomatis**	**NP_417773**	**NP_220037**	**sfv:SFV_3334 rpsC; 30S ribosomal protein S3; K02982 small subunit**	**ctr:CT522 rpsC 30S ribosomal protein S3 K02982 small subunit ribosomal protein S3**

	D/UW-3/CX			**ribosomal protein S3 (db=KEGG) (RBH)**	**(db=KEGG) (RBH)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417774**	**NP_220038**	**rplV; 50S ribosomal protein L22; K02890 large subunit ribosomal protein L22 (db=KEGG evalue=4.0e-49 bit_score=196.0 identity=99.09 coverage=99.0909090909091) (BLAST)**	**rplV 50S ribosomal protein L22 K02890 large subunit ribosomal protein L22 (db=KEGG evalue=7.0e-58 bit_score=225.0 identity=100.0 coverage=99.0990990990991) (BLAST)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417775**	**NP_220039**	**rpsS; 30S ribosomal subunit protein S19; K02965 small subunit ribosomal protein S19 (db=KEGG evalue=4.0e-47 bit_score=189.0 identity=100.0 coverage=98.9130434782609) (BLAST)**	**rpsS 30S ribosomal protein S19 K02965 small subunit ribosomal protein S19 (db=KEGG evalue=6.0e-44 bit_score=179.0 identity=100.0 coverage=98.8636363636364) (BLAST)**
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417776**	**NP_220040**	**sfl:SF3349 rplB; 50S ribosomal protein L2; K02886 large subunit ribosomal protein L2 (db=KEGG) (RBH)**	**ctb:CTL0787 rplB 50S ribosomal protein L2 K02886 large subunit ribosomal protein L2 (db=KEGG) (RBH)**
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417778	NP_220042	sty:STY4359 rplD; 50S ribosomal protein L4; K02926 large subunit ribosomal protein L4 (db=KEGG) (RBH)	ctj:JALI_5301 rplD 50S ribosomal protein L4 K02926 large subunit ribosomal protein L4 (db=KEGG) (RBH)
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417779**	**NP_220043**	**sfx:S4410 rplC; 50S ribosomal protein L3; K02906 large subunit ribosomal protein L3 (db=KEGG) (RBH)**	**ctj:JALI_5311 rplC 50S ribosomal protein L3 K02906 large subunit ribosomal protein L3 (db=KEGG) (RBH)**
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417785	NP_220086	ecx:EcHS_A3520 gspE1; general secretory pathway protein E; K02454 general secretion pathway protein E (db=KEGG) (RBH)	ctj:JALI_5741 gspE general secretion pathway protein E K02454 general secretion pathway protein E (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417786	NP_220085	ecj:JW3289 gspF; general secretory pathway component, cryptic; K02455 general secretion pathway protein F (db=KEGG) (RBH)	ctr:CT570 gspF general secretion pathway protein F K02455 general secretion pathway protein F (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_417799	NP_219949	sbc:SbBS512_E3714 fusA; elongation factor G; K02355 elongation factor EF-G [EC:3.6.5.3] (db=KEGG) (RBH)	ctr:CT437 fusA elongation factor G K02355 elongation factor EF-G [EC:3.6.5.3] (db=KEGG) (RBH)
Escherichia coli K12	**Chlamydia trachomatis D/UW-3/CX**	**NP_417800**	**NP_219950**	**ebw:BWG_3032 rpsG; 30S ribosomal protein S7; K02992 small subunit ribosomal protein S7 (db=KEGG) (RBH)**	**ctz:CTB_4391 rpsG 30S ribosomal protein S7 K02992 small subunit ribosomal protein S7 (db=KEGG) (RBH)**
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418155	NP_219577	ssn:SSON_3650 recF; recombination protein F; K03629 DNA replication and repair protein RecF (db=KEGG) (RBH)	ctr:CT074 recF recombination protein F K03629 DNA replication and repair protein RecF (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418156	NP_219578	sbc:SbBS512_E4224 dnaN; DNA polymerase III subunit beta (EC:2.7.7.7); K02338 DNA polymerase III subunit beta [EC:2.7.7.7] (db=KEGG) (RBH)	ctl:CTLon_0326 dnaN DNA polymerase III subunit beta K02338 DNA polymerase III subunit beta [EC:2.7.7.7] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418407	NP_219827	sbc:SbBS512_E4468 tuf1; elongation factor Tu; K02358 elongation factor EF-Tu [EC:3.6.5.3] (db=KEGG) (RBH)	ctz:CTB_3171 tufA elongation factor Tu K02358 elongation factor EF-Tu [EC:3.6.5.3] (db=KEGG) (RBH)

Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418409	NP_219825	sfx:S3680 nusG; transcription antitermination protein NusG; K02601 transcriptional antiterminator NusG (db=KEGG) (RBH)	ctj:JALI_3151 nusG transcription antitermination protein NusG K02601 transcriptional antiterminator NusG (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418410	NP_219824	rplK; 50S ribosomal protein L11; K02867 large subunit ribosomal protein L11 (db=KEGG evalue=3.0e-68 bit_score=259.0 identity=100.0 coverage=99.2957746478873) (BLAST)	rplK 50S ribosomal protein L11 K02867 large subunit ribosomal protein L11 (db=KEGG evalue=8.0e-61 bit_score=235.0 identity=100.0 coverage=99.290780141844) (BLAST)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418411	NP_219823	ssn:SSON_4157 rplA; 50S ribosomal protein L1; K02863 large subunit ribosomal protein L1 (db=KEGG) (RBH)	ctl:CTLon_0566 rplA 50S ribosomal protein L1 K02863 large subunit ribosomal protein L1 (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418414	NP_219820	sfx:S3675 rpoB; DNA-directed RNA polymerase subunit beta (EC:2.7.7.6); K03043 DNA-directed RNA polymerase subunit beta [EC:2.7.7.6] (db=KEGG) (RBH)	ctr:CT315 rpoB DNA-directed RNA polymerase subunit beta (EC:2.7.7.6) K03043 DNA-directed RNA polymerase subunit beta [EC:2.7.7.6] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418415	NP_219819	sfl:SF4061 rpoC; DNA-directed RNA polymerase subunit beta' (EC:2.7.7.6); K03046 DNA-directed RNA polymerase subunit beta' [EC:2.7.7.6] (db=KEGG) (RBH)	cta:CTA_0336 rpoC DNA-directed RNA polymerase subunit beta' (EC:2.7.7.6) K03046 DNA-directed RNA polymerase subunit beta' [EC:2.7.7.6] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418566	NP_219614	groES; co-chaperonin GroES; K04078 chaperonin GroES (db=KEGG evalue=9.0e-47 bit_score=188.0 identity=100.0 coverage=98.9690721649485) (BLAST)	groES co-chaperonin GroES K04078 chaperonin GroES (db=KEGG evalue=2.0e-50 bit_score=201.0 identity=100.0 coverage=99.0196078431373) (BLAST)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418567	NP_219613	sfx:S4564 groEL; chaperonin GroEL; K04077 chaperonin GroEL (db=KEGG) (RBH)	ctr:CT110 groEL chaperonin GroEL K04077 chaperonin GroEL (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418577	NP_220106	sfx:S4576 frdB; fumarate reductase iron-sulfur subunit (EC:1.3.99.1); K00245 fumarate reductase iron-sulfur protein [EC:1.3.99.1] (db=KEGG) (RBH)	ctr:CT591 sdhB succinate dehydrogenase iron-sulfur subunit (EC:1.3.99.1) K00240 succinate dehydrogenase iron-sulfur protein [EC:1.3.99.1] (db=KEGG) (RBH)
Escherichia coli K12	Chlamydia trachomatis D/UW-3/CX	NP_418578	NP_220107	sdy:SDY_4398 frdA; fumarate reductase flavoprotein subunit (EC:1.3.99.1); K00244 fumarate reductase flavoprotein subunit [EC:1.3.99.1] (db=KEGG) (RBH)	ctr:CT592 sdhA succinate dehydrogenase flavoprotein subunit (EC:1.3.5.1) K00239 succinate dehydrogenase flavoprotein subunit [EC:1.3.99.1] (db=KEGG) (RBH)

Table S7: List of organisms used in pairwise comparisons of GOC and sequence divergence. Genomes were downloaded from the NCBI database or were from the in-house acid mine drainage dataset.

Organism name
Aplasma
Acaryochloris_marina_MBIC11017
Acholeplasma_laidlawii_PG_8A
Acidithiobacillus_ferrooxidans_ATCC_23270
Acidithiobacillus_ferrooxidans_ATCC_53993
Acidothermus_cellulolyticus_11B
Acidovorax_citrulli_AAC00_1
Aciduliprofundum_boonei_T469
Acinetobacter_baumannii_AB307_0294
Acinetobacter_sp_ADP1
Actinobacillus_pleuropneumoniae_serovar_3_JL03
Actinobacillus_pleuropneumoniae_serovar_5b_L20
Actinobacillus_pleuropneumoniae_serovar_7_AP76
Actinobacillus_succinogenes_130Z
Aeromonas_hydrophila_ATCC_7966
Aeropyrum_pernix_K1
Agrobacterium_tumefaciens_C58
Akkermansia_muciniphila_ATCC_BAA_835
Alcanivorax_borkumensis_SK2
Alkalilimnicola_ehrlichii_MLHE_1
Alkaliphilus_metalliredigens_QYMF
Alkaliphilus_oremlandii_OhILAs
Alteromonas_macleodii__Deep_ecotype_
Anaeromyxobacter_dehalogenans_2CP-C
Anaeromyxobacter_Fw109-5
Anaeromyxobacter_K
Anaplasma_marginale_Maries
Anaplasma_phagocytophilum_HZ
Anoxybacillus_flavithermus_WK1

Aquifex_aeolicus_VF5
Archaeoglobus_fulgidus
Arcobacter_butzleri_RM4018
Aromatoleum_aromaticum_EbN1
Arthrobacter_aurescens_TC1
Aster_yellows_witches_broom_phytoplasma_AYWB
Azoarcus_BH72
Azorhizobium_caulinodans_ORS_571
Bacillus_amyloliquefaciens_FZB42
Bacillus_anthracis_Ames
Bacillus_anthracis_str_Sterne
Bacillus_cereus_ATCC_10987
Bacillus_cereus_B4264
Bacillus_cereus_E33L
Bacillus_cereus_G9842
Bacillus_clausii_KSM-K16
Bacillus_halodurans_C_125
Bacillus_licheniformis_ATCC_14580
Bacillus_licheniformis_DSM_13
Bacillus_pumilus_SAFR_032
Bacillus_subtilis
Bacillus_thuringiensis_konkukian
Bacteroides_fragilis_NCTC_9343
Bacteroides_thetaiotaomicron_VPI-5482
Bacteroides_vulgatus_ATCC_8482
Bartonella_bacilliformis_KC583
Bartonella_henselae_Houston_1
Bartonella_quintana_Toulouse
Baumannia_cicadellinicola_Hc__Homalodisca_coagulata_
Bdellovibrio_bacteriovorus

Bifidobacterium_adolescentis_ATCC_15703
Bifidobacterium_animalis_lactis_AD011
Bifidobacterium_longum_infantis_ATCC_15697
Bifidobacterium_longum_NCC2705
Bordetella_avium_197N
Bordetella_bronchiseptica_RB50
Bordetella_parapertussis
Bordetella_pertussis_Tohama_I
Bordetella_petrii
Borrelia_hermsii_DAH
Borrelia_recurrentis_A1
Borrelia_turicatae_91E135
Bradyrhizobium_japonicum
Bradyrhizobium_ORS278
Brucella_abortus_bv__1_9_941
Brucella_abortus_S19
Brucella_canis_ATCC_23365
Brucella_melitensis_biovar_Abortus
Brucella_melitensis_bv__1_16M
Brucella_ovis
Brucella_suis_1330
Buchnera_aphidicola_5A__Acyrthosiphon_pisum_
Buchnera_aphidicola_Bp__Baizongia_pistaciae_
Buchnera_aphidicola_Cc_Cinara_cedri
Buchnera_aphidicola_Sg
Buchnera_aphidicola_Tuc7__Acyrthosiphon_pisum_
Burkholderia_383
Burkholderia_ambifaria_MC40_6
Burkholderia_cenocepacia_AU_1054
Burkholderia_cenocepacia_HI2424
Burkholderia_cenocepacia_J2315
Burkholderia_cenocepacia_MC0_3
Burkholderia_mallei_ATCC_23344

Burkholderia_mallei_NCTC_10229
Burkholderia_mallei_NCTC_10247
Burkholderia_multivorans_ATCC_17616_Tohoku
Burkholderia_phymatum_STM815
Burkholderia_phytofirmans_PsJN
Burkholderia_pseudomallei_1106a
Burkholderia_pseudomallei_1710b
Burkholderia_pseudomallei_668
Burkholderia_pseudomallei_K96243
Burkholderia_thailandensis_E264
Burkholderia_xenovorans_LB400
Caldicellulosiruptor_saccharolyticus_DSM_8903
Campylobacter_fetus_82-40
Campylobacter_jejuni_81116
Campylobacter_jejuni_doylei_269_97
Campylobacter_jejuni_NCTC_11168
Campylobacter_jejuni_RM1221
Candidatus_Amoebophilus_asiaticus_5a2
Candidatus_Blochmannia_floridanus
Candidatus_Blochmannia_pennsylvanicus_BPEN
Candidatus_Desulforudis_audaxviator_MP104C
Candidatus_Korarchaeum_cryptofilum_OPF8
Candidatus_Koribacter_versatilis_Ellin345
Candidatus_Methanoregula_boonei_6A8
Candidatus_Methanosphaerula_palustris_E1_9c
Candidatus_Pelagibacter_ubique_HTCC1062
Candidatus_Phytoplasma_australiense
Candidatus_Phytoplasma_mali
Candidatus_Protochlamydia_amoebophila_UWE25
Candidatus_Ruthia_magnifica_Cm__Calyptogena_magnifica_
Candidatus_Sulcia_muelleri_GWSS
Candidatus_Vesicomyosocius_okutanii_HA
Carboxydothermus_hydrogenoformans_Z_2901

Caulobacter_crescentus
Cellvibrio_japonicus_Ueda107
Chlamydia_trachomatis_434_Bu
Chlamydia_trachomatis_A_HAR-13
Chlamydia_trachomatis_D_UW_3_CX
Chlamydia_trachomatis_L2b_UCH_1_proctitis
Chlamydophila_abortus_S26_3
Chlamydophila_caviae
Chlamydophila_felis_Fe_C-56
Chlamydophila_pneumoniae_AR39
Chlamydophila_pneumoniae_CWL029
Chlamydophila_pneumoniae_J138
Chlamydophila_pneumoniae_TW_183
Chlorobaculum_parvum_NCIB_8327
Chlorobium_chlorochromatii_CaD3
Chlorobium_limicola_DSM_245
Chlorobium_luteolum_DSM_273
Chlorobium_phaeobacteroides_BS1
Chlorobium_phaeobacteroides_DSM_266
Chlorobium_tepidum_TLS
Chloroflexus_aggregans_DSM_9485
Chloroflexus_aurantiacus_J_10_fl
Chloroherpeton_thalassium_ATCC_35110
Chromobacterium_violaceum_ATCC_12472
Chromohalobacter_salexigens_DSM_3043
Citrobacter_koseri_ATCC_BAA_895
Clostridium_beijerinckii_NCIMB_8052
Clostridium_botulinum_A
Clostridium_botulinum_A_ATCC_19397
Clostridium_botulinum_A_Hall
Clostridium_botulinum_B_Eklund_17B
Clostridium_botulinum_E3_Alaska_E43
Clostridium_botulinum_F_Langeland

Clostridium_novyi_NT
Clostridium_perfringens_ATCC_13124
Clostridium_perfringens_SM101
Clostridium_phytofermentans_ISDg
Clostridium_tetani_E88
Clostridium_thermocellum_ATCC_27405
Colwellia_psychrerythraea_34H
Coprothermobacter_proteolyticus_DSM_5265
Corynebacterium_diphtheriae_NCTC_13129
Corynebacterium_efficiens_YS-314
Corynebacterium_glutamicum_ATCC_13032_Bielefeld
Corynebacterium_glutamicum_ATCC_13032_Kitasato
Corynebacterium_glutamicum_R
Corynebacterium_urealyticum_DSM_7109
Coxiella_burnetii_CbuG_Q212
Coxiella_burnetii_RSA_493
Cupriavidus_taiwanensis
Cyanobacteria_bacterium_Yellowstone_B-Prime
Cyanothece_PCC_7424
Cytophaga_hutchinsonii_ATCC_33406
Dechloromonas_aromatica_RCB
Dehalococcoides_BAV1
Dehalococcoides_CBDB1
Dehalococcoides_ethenogenes_195
Deinococcus_geothermalis_DSM_11300
Delftia_acidovorans_SPH-1
Desulfatibacillum_alkenivorans_AK_01
Desulfitobacterium_hafniense_DCB_2
Desulfitobacterium_hafniense_Y51
Desulfococcus_oleovorans_Hxd3
Desulfotalea_psychrophila_LSv54
Desulfotomaculum_reducens_MI_1
Desulfovibrio_desulfuricans_G20

Desulfovibrio_vulgaris__Miyazaki_F_
Desulfovibrio_vulgaris_Hildenborough
Desulfurococcus_kamchatkensis_1221n
Dictyoglomus_thermophilum_H_6_12
Dictyoglomus_turgidum_DSM_6724
Dinoroseobacter_shibae_DFL_12
Eplasma
Ehrlichia_canis_Jake
Ehrlichia_chaffeensis_Arkansas
Ehrlichia_ruminantium_Gardel
Ehrlichia_ruminantium_str._Welgevonden_CIRAD
Ehrlichia_ruminantium_Welgevonden_UPSA
Elusimicrobium_minutum_Pei191
Enterobacter_sakazakii_ATCC_BAA-894
Enterococcus_faecalis_V583
Erwinia_carotovora_atroseptica_SCRI1043
Erythrobacter_litoralis_HTCC2594
Escherichia_coli_536
Escherichia_coli_55989
Escherichia_coli_APEC_O1
Escherichia_coli_C_ATCC_8739
Escherichia_coli_CFT073
Escherichia_coli_ED1a
Escherichia_coli_HS
Escherichia_coli_IAI1
Escherichia_coli_IAI39
Escherichia_coli_K_12_substr__DH10B
Escherichia_coli_K_12_substr__MG1655
Escherichia_coli_O127_H6_E2348_69
Escherichia_coli_O157_H7_EDL933
Escherichia_coli_S88
Escherichia_coli_UMN026
Escherichia_fergusonii_ATCC_35469

Ferroplasma acidarmanus I
Ferroplasma acidarmanus II
Fervidobacterium_nodosum_Rt17-B1
Flavobacterium_johnsoniae_UW101
Flavobacterium_psychrophilum_JIP02_86
Francisella_tularensis_FSC198
Francisella_tularensis_holarctica_FTNF002_00
Francisella_tularensis_holarctica_LVS
Francisella_tularensis_holarctica_OSU18
Francisella_tularensis_mediasiatica_FSC147
Francisella_tularensis_novicida_U112
Francisella_tularensis_tularensis
Francisella_tularensis_WY96-3418
Frankia_alni_ACN14a
Frankia_CcI3
Frankia_EAN1pec
Fusobacterium_nucleatum_ATCC_25586
Gplasma
Geobacillus_thermodenitrificans_NG80-2
Geobacter_bemidjiensis_Bem
Geobacter_lovleyi_SZ
Geobacter_sulfurreducens_PCA
Geobacter_uraniumreducens_Rf4
Gloeobacter_violaceus_PCC_7421
Gluconacetobacter_diazotrophicus_PAl_5
Gramella_forsetii_KT0803
Granulobacter_bethesdensis_CGDNIH1
Haemophilus_ducreyi_35000HP
Haemophilus_influenzae_86_028NP
Haemophilus_influenzae_PittEE
Haemophilus_influenzae_PittGG
Haemophilus_influenzae_Rd_KW20
Haemophilus_parasuis_SH0165

Haemophilus_somnus_2336
Hahella_chejuensis_KCTC_2396
Halobacterium_salinarum_R1
Haloquadratum_walsbyi_DSM_16790
Halorhodospira_halophila_SL1
Helicobacter_acinonychis_Sheeba
Helicobacter_hepaticus_ATCC_51449
Helicobacter_pylori_26695
Helicobacter_pylori_G27
Helicobacter_pylori_HPAG1
Helicobacter_pylori_J99
Helicobacter_pylori_P12
Helicobacter_pylori_Shi470
Heliobacterium_modesticaldum_Ice1
Herpetosiphon_aurantiacus_ATCC_23779
Hydrogenobaculum_Y04AAS1
Hyperthermus_butylicus
Hyphomonas_neptunium_ATCC_15444
Iplasma
Idiomarina_loihiensis_L2TR
Ignicoccus_hospitalis_KIN4_I
Janthinobacterium_Marseille
Kocuria_rhizophila_DC2201
Lactobacillus_acidophilus_NCFM
Lactobacillus_brevis_ATCC_367
Lactobacillus_casei
Lactobacillus_delbrueckii_bulgaricus_ATCC_11842
Lactobacillus_delbrueckii_bulgaricus_ATCC_BAA-365
Lactobacillus_fermentum_IFO_3956
Lactobacillus_gasseri_ATCC_33323
Lactobacillus_helveticus_DPC_4571
Lactobacillus_johnsonii_NCC_533
Lactobacillus_plantarum

Lactobacillus_reuteri_DSM_20016
Lactobacillus_reuteri_F275_Kitasato
Lactobacillus_sakei_23K
Lactococcus_lactis_cremoris_MG1363
Lactococcus_lactis_Il1403
Lawsonia_intracellularis_PHE_MN1-00
Legionella_pneumophila_Corby
Legionella_pneumophila_Philadelphia_1
Leifsonia_xyli_xyli_CTCB0
Leptospira_biflexa_serovar_Patoc__Patoc_1__Ames_
Leptospira_biflexa_serovar_Patoc__Patoc_1__Paris_
Leptospira_borgpetersenii_serovar_Hardjo-bovis_JB197
Leptospira_borgpetersenii_serovar_Hardjo-bovis_L550
Leptospira_interrogans_serovar_Lai
Leptospira_interrogans_serovar_Lai_56601
Leptothrix_cholodnii_SP_6
Listeria_innocua
Listeria_monocytogenes
Listeria_monocytogenes_4b_F2365
Listeria_monocytogenes_HCC23
Listeria_welshimeri_serovar_6b_SLCC5334
Magnetococcus_MC_1
Magnetospirillum_magneticum_AMB_1
Mannheimia_succiniciproducens_MBEL55E
Maricaulis_maris_MCS10
Marinomonas_MWYL1
Mesoplasma_florum_L1
Mesorhizobium_loti
Metallosphaera_sedula_DSM_5348
Methanobrevibacter_smithii_ATCC_35061
Methanococcoides_burtonii_DSM_6242
Methanococcus_aeolicus_Nankai_3
Methanococcus_jannaschii

Methanococcus_maripaludis_C5
Methanococcus_maripaludis_C6
Methanococcus_maripaludis_C7
Methanococcus_maripaludis_S2
Methanococcus_vannielii_SB
Methanocorpusculum_labreanum_Z
Methanoculleus_marisnigri_JR1
Methanopyrus_kandleri
Methanosaeta_thermophila_PT
Methanosarcina_acetivorans_C2A
Methanosarcina_mazei
Methanosphaera_stadtmanae_DSM_3091
Methanospirillum_hungatei_JF_1
Methanothermobacter_thermautotrophicus_Delta_H
Methylacidiphilum_infernorum_V4
Methylibium_petroleiphilum_PM1
Methylobacillus_flagellatus_KT
Methylobacterium_chloromethanicum_CM4
Methylobacterium_extorquens_PA1
Methylocella_silvestris_BL2
Methylococcus_capsulatus_Bath
Microcystis_aeruginosa_NIES_843
Moorella_thermoacetica_ATCC_39073
Mycobacterium_avium_104
Mycobacterium_avium_paratuberculosis
Mycobacterium_bovis_AF2122_97
Mycobacterium_bovis_BCG_Pasteur_1173P2
Mycobacterium_gilvum_PYR_GCK
Mycobacterium_JLS
Mycobacterium_leprae_TN
Mycobacterium_MCS
Mycobacterium_smegmatis_MC2_155
Mycobacterium_tuberculosis_F11

Mycobacterium_tuberculosis_H37Ra
Mycobacterium_ulcerans_Agy99
Mycobacterium_vanbaalenii_PYR-1
Mycoplasma_agalactiae_PG2
Mycoplasma_arthritidis_158L3_1
Mycoplasma_capricolum_ATCC_27343
Mycoplasma_gallisepticum
Mycoplasma_genitalium_G37
Mycoplasma_hyopneumoniae_232
Mycoplasma_hyopneumoniae_7448
Mycoplasma_hyopneumoniae_J
Mycoplasma_mobile_163K
Mycoplasma_mycoides
Mycoplasma_penetrans_HF_2
Mycoplasma_pneumoniae
Mycoplasma_pulmonis
Mycoplasma_synoviae_53
Myxococcus_xanthus_DK_1622
Nanoarchaeum_equitans_Kin4_M
Natronomonas_pharaonis
Neisseria_gonorrhoeae_FA_1090
Neisseria_meningitidis_FAM18
Neisseria_meningitidis_MC58
Neisseria_meningitidis_Z2491
Neorickettsia_sennetsu_Miyayama
Nitratiruptor_SB155_2
Nitrobacter_winogradskyi_Nb_255
Nitrosomonas_europaea_ATCC_19718
Nitrosopumilus_maritimus_SCM1
Nitrosospira_multiformis_ATCC_25196
Nocardia_farcinica_IFM_10152
Nostoc_punctiforme_PCC_73102
Novosphingobium_aromaticivorans_DSM_12444

Oceanobacillus_iheyensis_HTE831
Ochrobactrum_anthropi_ATCC_49188
Oenococcus_oeni_PSU-1
Oligotropha_carboxidovorans_OM5
Onion_yellows_phytoplasma
Opitutus_terrae_PB90_1
Orientia_tsutsugamushi_Boryong
Orientia_tsutsugamushi_Ikeda
Parabacteroides_distasonis_ATCC_8503
Paracoccus_denitrificans_PD1222
Parvibaculum_lavamentivorans_DS_1
Pasteurella_multocida_Pm70
Pediococcus_pentosaceus_ATCC_25745
Pelobacter_carbinolicus
Pelodictyon_phaeoclathratiforme_BU_1
Pelotomaculum_thermopropionicum_SI
Petrotoga_mobilis_SJ95
Photorhabdus_luminescens
Picrophilus_torridus_DSM_9790
Pirellula_sp
Polaromonas_JS666
Polynucleobacter_necessarius_asymbioticus_QLW_P1DMWA_1
Polynucleobacter_necessarius_STIR1
Porphyromonas_gingivalis_ATCC_33277
Porphyromonas_gingivalis_W83
Prochlorococcus_marinus_AS9601
Prochlorococcus_marinus_CCMP1375
Prochlorococcus_marinus_MIT_9211
Prochlorococcus_marinus_MIT_9215
Prochlorococcus_marinus_MIT_9301
Prochlorococcus_marinus_MIT_9303
Prochlorococcus_marinus_MIT_9312
Prochlorococcus_marinus_MIT_9313

Prochlorococcus_marinus_MIT_9515
Prochlorococcus_marinus_NATL1A
Prochlorococcus_marinus_NATL2A
Prochlorococcus_marinus_pastoris_CCMP1986
Propionibacterium_acnes_KPA171202
Prosthecochloris_aestuarii_DSM_271
Prosthecochloris_vibrioformis_DSM_265
Proteus_mirabilis
Pseudoalteromonas_atlantica_T6c
Pseudoalteromonas_haloplanktis_TAC125
Pseudomonas_aeruginosa_LESB58
Pseudomonas_aeruginosa_PA7
Pseudomonas_aeruginosa_PAO1
Pseudomonas_aeruginosa_UCBPP_PA14
Pseudomonas_fluorescens_Pf_5
Pseudomonas_fluorescens_Pf0_1
Pseudomonas_mendocina_ymp
Pseudomonas_putida_F1
Pseudomonas_putida_GB_1
Pseudomonas_putida_KT2440
Pseudomonas_putida_W619
Pseudomonas_stutzeri_A1501
Pseudomonas_syringae_phaseolicola_1448A
Pseudomonas_syringae_pv_B728a
Pseudomonas_syringae_tomato_DC3000
Psychrobacter_arcticus_273_4
Psychromonas_ingrahamii_37
Pyrobaculum_aerophilum
Pyrobaculum_arsenaticum_DSM_13514
Pyrobaculum_calidifontis_JCM_11548
Pyrobaculum_islandicum_DSM_4184
Pyrococcus_furiosus_DSM_3638
Pyrococcus_horikoshii

Ralstonia_pickettii_12J
Ralstonia_solanacearum_GMI1000
Renibacterium_salmoninarum_ATCC_33209
Rhizobium_etli_CIAT_652
Rhodobacter_sphaeroides_ATCC_17025
Rhodococcus_jostii_RHA1
Rhodopseudomonas_palustris_BisA53
Rhodopseudomonas_palustris_BisB18
Rhodopseudomonas_palustris_BisB5
Rhodopseudomonas_palustris_CGA009
Rhodopseudomonas_palustris_HaA2
Rhodopseudomonas_palustris_TIE_1
Rhodospirillum_centenum_SW
Rickettsia_akari_Hartford
Rickettsia_bellii_OSU_85_389
Rickettsia_bellii_RML369-C
Rickettsia_canadensis_McKiel
Rickettsia_conorii_Malish_7
Rickettsia_felis_URRWXCal2
Rickettsia_prowazekii
Rickettsia_rickettsii_Iowa
Rickettsia_rickettsii_Sheila_Smith
Rickettsia_typhi_Wilmington
Roseiflexus_castenholzii_DSM_13941
Roseiflexus_RS_1
Roseobacter_denitrificans_OCh_114
Rubrobacter_xylanophilus_DSM_9941
Ruegeria_pomeroyi_DSS_3
Ruegeria_TM1040
Saccharophagus_degradans_2-40
Saccharopolyspora_erythraea_NRRL_2338
Salinibacter_ruber_DSM_13855
Salinispora_arenicola_CNS_205

Salinispora_tropica_CNB-440
Salmonella_enterica_arizonae_serovar_62_z4_z23__
Salmonella_enterica_serovar_Enteritidis_P125109
Salmonella_enterica_serovar_Gallinarum_287_91
Salmonella_enterica_serovar_Paratyphi_A_AKU_12601
Salmonella_enterica_serovar_Paratyphi_A_ATCC_9150
Salmonella_enterica_serovar_Paratyphi_B_SPB7
Salmonella_enterica_serovar_Typhi_Ty2
Salmonella_typhi
Salmonella_typhimurium_LT2
Shewanella_amazonensis_SB2B
Shewanella_baltica_OS195
Shewanella_baltica_OS223
Shewanella_denitrificans_OS217
Shewanella_frigidimarina_NCIMB_400
Shewanella_halifaxensis_HAW_EB4
Shewanella_loihica_PV_4
Shewanella_MR_4
Shewanella_oneidensis_MR_1
Shewanella_pealeana_ATCC_700345
Shewanella_piezotolerans_WP3
Shewanella_putrefaciens_CN-32
Shewanella_sediminis_HAW_EB3
Shewanella_W3-18-1
Shewanella_woodyi_ATCC_51908
Shigella_dysenteriae_Sd197
Shigella_flexneri_2a
Shigella_flexneri_2a_2457T
Shigella_flexneri_5_8401
Shigella_sonnei_Ss046
Sodalis_glossinidius__morsitans_
Solibacter_usitatus_Ellin6076
Sorangium_cellulosum__So_ce_56_

Staphylococcus_aureus_COL
Staphylococcus_aureus_Mu3
Staphylococcus_aureus_Mu50
Staphylococcus_aureus_MW2
Staphylococcus_aureus_N315
Staphylococcus_aureus_NCTC_8325
Staphylococcus_aureus_Newman
Staphylococcus_aureus_RF122
Staphylococcus_epidermidis_ATCC_12228
Staphylococcus_epidermidis_RP62A
Staphylococcus_haemolyticus
Staphylococcus_saprophyticus_ATCC_15305
Stenotrophomonas_maltophilia_K279a
Stenotrophomonas_maltophilia_R551_3
Streptococcus_agalactiae_2603V_R
Streptococcus_agalactiae_A909
Streptococcus_agalactiae_NEM316
Streptococcus_equi_zooepidemicus_MGCS10565
Streptococcus_gordonii_Challis_substr_CH1
Streptococcus_mutans_UA159
Streptococcus_pneumoniae_CGSP14
Streptococcus_pneumoniae_D39
Streptococcus_pneumoniae_G54
Streptococcus_pneumoniae_Hungary19A_6
Streptococcus_pneumoniae_R6
Streptococcus_pneumoniae_TIGR4
Streptococcus_pyogenes_M1_GAS
Streptococcus_pyogenes_MGAS10270
Streptococcus_pyogenes_MGAS10394
Streptococcus_pyogenes_MGAS10750
Streptococcus_pyogenes_MGAS315
Streptococcus_pyogenes_MGAS5005
Streptococcus_pyogenes_MGAS8232

Streptococcus_pyogenes_NZ131
Streptococcus_sanguinis_SK36
Streptococcus_suis_05ZYH33
Streptococcus_suis_98HAH33
Streptococcus_thermophilus_CNRZ1066
Streptococcus_thermophilus_LMG_18311
Streptomyces_avermitilis_MA_4680
Streptomyces_coelicolor_A3_2_
Streptomyces_griseus_NBRC_13350
Sulfolobus_acidocaldarius_DSM_639
Sulfolobus_solfataricus_P2
Sulfolobus_tokodaii
Sulfurihydrogenibium_YO3AOP1
Sulfurovum_NBC37_1
Symbiobacterium_thermophilum_IAM14863
Synechococcus_CC9311
Synechococcus_CC9605
Synechococcus_CC9902
Synechococcus_elongatus_PCC_6301
Synechococcus_JA_3_3Ab
Synechococcus_RCC307
Synechococcus_sp_WH8102
Synechococcus_WH_7803
Synechocystis_PCC6803
Syntrophobacter_fumaroxidans_MPOB
Syntrophomonas_wolfei_Goettingen
Syntrophus_aciditrophicus_SB
Thermoanaerobacter_pseudethanolicus_ATCC_33223
Thermoanaerobacter_tengcongensis
Thermoanaerobacter_X514
Thermobifida_fusca_YX
Thermococcus_kodakaraensis_KOD1
Thermococcus_onnurineus_NA1

Thermodesulfovibrio_yellowstonii_DSM_11347
Thermoplasma_acidophilum
Thermoplasma_volcanium
Thermoproteus_neutrophilus_V24Sta
Thermosipho_africanus_TCF52B
Thermosipho_melanesiensis_BI429
Thermosynechococcus_elongatus_BP_1
Thermotoga_lettingae_TMO
Thermotoga_maritima
Thermotoga_petrophila_RKU-1
Thermotoga_RQ2
Thermus_thermophilus_HB27
Thermus_thermophilus_HB8
Thiobacillus_denitrificans_ATCC_25259
Thiomicrospira_crunogena_XCL_2
Thiomicrospira_denitrificans_ATCC_33889
Treponema_denticola_ATCC_35405
Treponema_pallidum_Nichols
Treponema_pallidum_SS14
Trichodesmium_erythraeum_IMS101
Tropheryma_whipplei_TW08_27
Tropheryma_whipplei_Twist
uncultured_methanogenic_archaeon_RC-I
uncultured_Termite_group_1_bacterium_phylotype_Rs_D17
Ureaplasma_parvum_serovar_3_ATCC_27815
Ureaplasma_parvum_serovar_3_ATCC_700970
Ureaplasma_urealyticum_serovar_10_ATCC_33699
Vibrio_cholerae
Vibrio_fischeri_ES114
Vibrio_parahaemolyticus_RIMD_2210633
Vibrio_splendidus_LGP32
Vibrio_vulnificus_CMCP6
Wolbachia_endosymbiont_of_Culex_quinquefasciatus_Pel

Wolbachia_endosymbiont_of_Drosophila_melanogaster
Wolbachia_endosymbiont_TRS_of_Brugia_malayi
Wolinella_succinogenes
Xanthomonas_axonopodis_citri_306
Xanthomonas_campestris_8004
Xanthomonas_campestris_ATCC_33913
Xanthomonas_campestris_B100
Xanthomonas_oryzae_KACC10331
Xanthomonas_oryzae_MAFF_311018
Xanthomonas_oryzae_PXO99A
Xylella_fastidiosa
Xylella_fastidiosa_M12
Xylella_fastidiosa_M23
Yersinia_pestis_biovar_Microtus_91001
Yersinia_pestis_KIM_10
Yersinia_pseudotuberculosis_PB1_
Yersinia_pseudotuberculosis_YPIII
Zymomonas_mobilis_ZM4

Table S8: List of organisms used in pairwise comparisons of percentage of syntenous genes that are functionally related and GOC. Genomes were downloaded from the STRING database.

STRING organism name
Methanothermobacter_thermautotrophicus_str._Delta_H
Fusobacterium_nucleatum_subsp._nucleatum_ATCC_25586
Leptospira_interrogans_serovar_Lai_str._56601
Methanopyrus_kandleri_AV19
Tropheryma_whipplei_str._Twist
Desulfotalea_psychrophila_LSv54
Leuconostoc_mesenteroides_subsp._mesenteroides_ATCC_8293
Acinetobacter_sp._ADP1
Caulobacter_crescentus_CB15
Saccharophagus_degradans_2-40
Wigglesworthia_glossinidia_endosymbiont_of_Glossina_brevipalpis
Ehrlichia_chaffeensis_str._Arkansas
Rhodococcus_sp._RHA1
Bacillus_clausii_KSM-K16
Dechloromonas_aromatica_RCB
Desulfovibrio_desulfuricans_G20
Shewanella_sp._MR-7
Bradyrhizobium_sp._ORS278
Nostoc_sp._PCC_7120
Synechococcus_elongatus_PCC_7942
Haemophilus_influenzae_Rd_KW20
Bacillus_anthracis_str._Ames
Brucella_suis_1330
Corynebacterium_efficiens_YS-314
Streptococcus_agalactiae_A909
Salmonella_typhimurium_LT2
Pyrococcus_furiosus_DSM_3638
Pseudomonas_syringae_pv._syringae_B728a
Prochlorococcus_marinus_subsp._marinus_str._CCMP1375
Methanosarcina_acetivorans_C2A

Campylobacter_jejuni_RM1221
Synechococcus_sp._WH_7803
Synechococcus_sp._CC9605
Helicobacter_pylori_26695
Shewanella_sp._ANA-3
Xylella_fastidiosa_9a5c
Prochlorococcus_marinus_str._NATL1A
Xanthomonas_axonopodis_pv._citri_str._306
Prochlorococcus_marinus_str._MIT_9312
Shigella_flexneri_2a_str._2457T
Mycobacterium_sp._JLS
Mycobacterium_tuberculosis_CDC1551
Prochlorococcus_marinus_str._MIT_9515
Clostridium_perfringens_ATCC_13124
Staphylococcus_aureus_subsp._aureus_NCTC_8325
Francisella_tularensis_subsp._holarctica
Staphylococcus_epidermidis_ATCC_12228
Streptococcus_pneumoniae_R6
Neisseria_meningitidis_MC58
Chlamydophila_pneumoniae_J138
Escherichia_coli_K12
Streptococcus_pyogenes_MGAS315
Prochlorococcus_marinus_str._MIT_9215
Chlamydophila_pneumoniae_AR39
Staphylococcus_aureus_subsp._aureus_COL
Streptococcus_pyogenes_M1_GAS
Thermococcus_kodakarensis_KOD1
Desulfitobacterium_hafniense_Y51
Alkalilimnicola_ehrlichei_MLHE-1
Halobacterium_sp._NRC-1
Clostridium_thermocellum_ATCC_27405

Acidobacteria_bacterium_Ellin345
Pyrobaculum_aerophilum_str._IM2
Bifidobacterium_longum_NCC2705
Syntrophus_aciditrophicus_SB
Oenococcus_oeni_PSU-1
Pseudomonas_putida_KT2440
Mycobacterium_sp._MCS
Colwellia_psychrerythraea_34H
Yersinia_pestis_KIM
Wolbachia_endosymbiont_of_Drosophila_melanogaster
Nocardioides_sp._JS614
Listeria_monocytogenes_EGD-e
Herminiimonas_arsenicoxydans
Desulfovibrio_vulgaris_subsp._vulgaris_str._Hildenborough
Vibrio_vulnificus_YJ016
Xanthobacter_autotrophicus_Py2
Trichodesmium_erythraeum_IMS101
Prochlorococcus_marinus_str._MIT_9303
Haemophilus_somnus_129PT
Staphylococcus_aureus_subsp._aureus_N315
Agrobacterium_tumefaciens_str._C58
Corynebacterium_glutamicum_ATCC_13032
Streptococcus_pyogenes_str._Manfredo
Escherichia_coli_CFT073
Pyrococcus_horikoshii_OT3
Pseudomonas_fluorescens_PfO-1
Prochlorococcus_marinus_str._MIT_9313
Methanosarcina_mazei_Go1
Campylobacter_jejuni_subsp._jejuni_NCTC_11168
Synechococcus_sp._CC9311
Synechococcus_sp._WH_8102
Helicobacter_pylori_J99
Shewanella_sp._MR-4

Xylella_fastidiosa_Temecula1
Prochlorococcus_marinus_str._NATL2A
Xanthomonas_campestris_pv._campestris_str._ATCC_33913
Prochlorococcus_marinus_str._AS9601
Shigella_flexneri_2a_str._301
Mycobacterium_sp._KMS
Mycobacterium_tuberculosis_H37Rv
Prochlorococcus_marinus_subsp._pastoris_str._CCMP1986
Clostridium_perfringens_str._13
Staphylococcus_aureus_subsp._aureus_Mu50
Francisella_tularensis_subsp._tularensis_SCHU_S4
Staphylococcus_epidermidis_RP62A
Streptococcus_pneumoniae_TIGR4
Neisseria_meningitidis_Z2491
Chlamydophila_pneumoniae_CWL029
Escherichia_coli_O157:H7_EDL933
Streptococcus_pyogenes_MGAS8232
Prochlorococcus_marinus_str._MIT_9301
Chlamydophila_pneumoniae_TW-183
Staphylococcus_aureus_subsp._aureus_MW2
Streptococcus_pyogenes_SSI-1
Thermosynechococcus_elongatus_BP-1
Frankia_sp._CcI3
Chlorobium_tepidum_TLS

Text S1: Detailed explanation of methods.

Sampling and genome reconstruction

Samples for metagenomic sequencing were taken from biofilms collected from the Richmond Mine, Iron Mountain, California (40° 40' 38.42" N and 122" 31' 19.90" W, Elevation ~ 3,100') in March, 2002 (5-way site) and the Ultraback A drift (UBA site) in June, 2005. A third biofilm was collected from a few meters away from the UBA site (UBA BS) in November, 2005. The 5-way site biofilm was floating on a pH 0.83 solution at 42 °C. The UBA sample was a pink biofilm partially submerged in pH 1.1, 38 °C solution, whereas the UBA BS sample was a thick, pink, gelatinous, biofilm floating on pH 1.5, 39 °C solution.

After DNA extraction, ~700 base pair-long mate-paired reads were generated by random shotgun sequencing of small insert (~3 kb) plasmid libraries (Joint Genome Institute). Each dataset, comprising ~130 Mb of sequence from the 5-way biofilm, ~110 Mb from the UBA biofilm, and ~100 Mb from the UBA BS biofilm, was assembled independently using Phred/Phrap and manually curated using Consed to resolve assembly errors. The resulting composite genome fragments (contigs) were tentatively binned based on read depth and GC content to generate well-defined genomic bins for *Leptospirillum* Group II and *Ferroplasma* Type II. Manually curated genomic datasets were generated for *Leptospirillum* Group II from the 5-way site (Simmons et al, 2008) and *Leptospirillum* Group II and III from the UBA site (Lo I, 2007; Goltsman DSA, 2009). Contigs from *Ferroplasma* Type I and Type II populations assembled largely independently, but were binned by comparison to the *Ferroplasma acidarmanus* isolate genome (Allen EE, 2007; Eppley JM, 2007). Remaining contigs were determined to derive from novel lineages of ARMAN archaea (Baker BJ, 2006) and multiple populations of lower abundance *Thermoplasmatales* archaea and bacteria based on sampling of their 16S rRNA genes. Genome fragments of ARMAN2, a relatively abundant organism in the UBA BS dataset, were separated from other archaea based on GC content and read depth (Comolli L, 2009). Contigs from low abundance archaea were separated from bacteria based on their GC content (bacteria 52-68% vs. archaea 34-47%) and read depth, manually curated to resolve assembly errors, and scaffolded into fragments of up to ~ 200 kb in length based on paired end sequence placement in Consed (Gordon, 2004).

Comparative method for correlation analysis

According to phylogenetic comparative method, genome pairs were chosen to maximize the number of pairs (Maddison WP, 2009). In order to prevent pseudo-replication, pairs were chosen from non-overlapping branches. These non-overlapping evolutionary histories preserve independence in genome rearrangements because rearrangements are not shared across separate evolutionary trajectories. It should be noted that though this method preserves phylogenetic independence, it does not necessarily find statistically independent pairs.

Genome shearing

The Fer1 isolate genome was sheared into pieces and included in the genome comparisons in order to test the applicability of gene and genome evolution trends to incomplete genomes. Gene calls were made on the newly sheared fragments and the synteny analysis was carried out between this organism and organisms from the NCBI genome database. In order to better represent the distribution of contig lengths found in our dataset, the number and length of

fragments was chosen based on the lengths of the fragments in genomes reconstructed from all of the AMD metagenomic data. The range of fragmentation applied was from seven to 132 fragments.

16S rRNA distance divergence measurement

16S rRNA gene phylogeny was generated using the ARB software (Ludwig W, 2004) that calculated distance trees based on a neighbor joining method. Alignments were prepared with fast aligner in ARB and were then manually refined. A pairwise identity matrix was generated to calculate percent similarities between sequences, which are referred to as 16S rRNA distance/divergence in this study. Bootstrap values were determined using a distance-based method with default parameters.

Gene Annotations

Our annotation pipeline begins with gene prediction by Prodigal (ref). Protein sequences are then blasted again the KEGG database and the UniRef90 database. Reciprocal best BLAST hits with a bit score of at least 300 to a known gene in KEGG or UniRef 90. Hits to a gene containing any of the following terms were flagged as "unknowns": hypothetical, unknown, unassigned, unclassified, undetermined, uncharacterized, putative, probable, or predicted. One-way BLAST hits with a minimum bit score of 60 are ranked next and these matches to unknown proteins were noted. Finally, we used InterproScan to identify protein family and domain hits. Rankings are assigned using the following scheme: Rank A – reserved for any gene that is manually annotated; Rank B – a gene with a reciprocal best BLAST hit to a known gene in KEGG/ UniRef90; Rank C – a gene with a BLAST hit to KEGG/UniRef90; Rank D - a gene with only InterProScan annotation; Rank E – a predicted gene with no supporting annotation (pure hypothetical).

Operon prediction

We assigned genes to operons based on a very conservative approach, where adjacent genes have both the same direction of transcription and are separated by no more than thirty bases of intergenic space. Synteny is commonly used in operon prediction, but we chose not to make use of this information to avoid circularity in our analysis.

CHAPTER 2.

Comparative genomics in acid mine drainage biofilm communities reveals metabolic and structural differentiation of co-occurring archaea

Authors: Alexis P. Yelton[5], Nicholas Justice[6], Luis Comolli[7], Cindy Castelle[3], Brian C. Thomas[1], Jillian F. Banfield[1,8]

[5]Department of Environmental Science, Policy, and Management, University of California, Berkeley, California 94720, USA

[6] Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA

[7] Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA

[8] Department of Earth and Planetary Sciences, University of California, Berkeley, California 94720, USA

Abstract

Acid mine drainage (AMD) creates an environment that is inhospitable to most life. Despite dominance by a small number of bacteria, AMD microbial biofilm communities contain a surprising variety of coexisting and closely related *Euryarchaea*. We analyzed variation evident in genomes reconstructed from AMD metagenomic data that may contribute to niche differentiation in *Thermoplasmatales* archaea that we call A-, E-, and Gplasma as well as *Ferroplasma* type I and II and a novel organism, Iplasma. Our analyses reveal that all are facultative aerobic heterotrophs with the ability to use many of the same carbon substrates, including methanol. They also all have genes for toxic metal resistance and surface-layer production. Only Aplasma and Eplasma have a full suite of flagellar genes whereas all but the *Ferroplasma* spp. have genes for pili production. Cryogenic-electron microscopy (cryo-EM) and tomography (cryo-ET) strengthen these metagenomics-based ultrastructural predictions. Notably, only Aplasma, Gplasma and the *Ferroplasma* spp. have predicted iron oxidation genes and Eplasma and Iplasma lack most genes for cobalamin, valine, (iso)leucine and histidine synthesis. The closely related, co-occurring AMD archaea we examined here (A-, E-, and Gplasma as well as the *Ferroplasma* spp.) share a large number of metabolic capabilities and the uncultivated organisms (A-, E-, G-, and Iplasma) are very metabolically similar to characterized *Ferroplasma* spp., differentiating themselves mainly in their potential genetic capabilities for biosynthesis, iron oxidation, and motility. These results indicate that subtle, but important genomic differences, coupled with unknown differences in gene expression distinguish these organisms enough to allow for co-existence.

Introduction

Recent advances in sequencing technologies have led to a near exponential increase in the number of sequenced genomes of bacteria and archaea, but to date, most sequenced archaeal isolates come from disparate environments and therefore tell us little about niche differentiation within environments. Notable exceptions include isolates and draft genomes from metagenomic sequencing projects in hot springs, hypersaline environments [54-58] and genomes of different strains of one gut methanogen [59]. Metagenomics allows us to examine the genomes of closely related archaea in the same community and make inferences about physiological differences that allow them to coexist. This can aid in the understanding of population dynamics, community ecology, and biogeography. Spatial and temporal distributions of populations may be related to differences in geochemical conditions, in nutrients, or in other resources that different strains and species can utilize. Finally, if the intention is to isolate organisms with particular metabolic capacities, metagenomic insights can aid in the determination of the vitamins, nutrients, cofactors, and environmental conditions necessary for the growth of potential isolates.

A number of archaea of the *euryarchaeal* order *Thermoplasmatales* have been described. This order currently comprises five genera: *Ferroplasma*, *Thermoplasma*, *Picrophilus*, *Thermogymnomonas*, and *Acidiplasma*. All of the isolates from this order are obligate or facultative aerobes and extreme acidophiles that were isolated from acidic, high sulfur environments. However, there is some phenotypic variation within this clade. The *Picrophilus* spp. are characterized by a single cell membrane surrounded by a surface layer, whereas the species in the other *Thermoplasmatales* genera have no cell walls. The *Thermoplasma* spp. and *Picrophilus* spp. are moderate thermophiles with temperature optima around 60º C, whereas the *Ferroplasma* spp. are mesophiles with temperature optima around 40º C [9, 12, 60-66]. All of the isolates from the *Thermoplasmatales* order except for *Ferroplasma acidiphilum* are heterotrophs. All of the *Ferroplasma* spp. are Fe-oxidizers and grow anaerobically via Fe respiration, whereas the *Thermoplasma* spp. are capable of S^0 respiration.

In this study, we compare the near-complete genomes of the two *Ferroplasma acidarmanus* types (the isolate Fer1 sequence and the environmental Fer2 sequence) with newly reported genomes of A-, E-, G-, and Iplasma (APL, EPL, GPL, and IPL). These organisms coexist in biofilm communities sampled from within the Richmond Mine at Iron Mountain in Redding, California. Of these organisms, only Fer1 has been isolated [12], though the other genomes have been studied in previous metagenomic analyses [47, 67, 68]. The comparative genomic analysis presented here provides new insights into acid mine drainage (AMD) community function and genomic differentiation among these organisms.

Results and discussion

(i) Phylogeny: We previously published a phylogenetic tree of the 16S rRNA gene of the AMD plasmas (Chapter 1) [47, 68]. Here we improve upon that tree with the addition of a number of new taxa. This tree illustrates that the Richmond Mine AMD plasmas form the following clades: A-, B-, and Cplasma, E- with G-plasma, Dplasma with a number of environmental clones, I-plasma with a number of environmental clones, and the *Ferroplasma* spp. with *Acidiplasma aeolicum*. All of the 16S rRNA gene sequences, other than those of Fer1 and Fer2 (which have identical sequences), share less than 97% nucleotide identity. The Iplasma gene is the most divergent, and it is almost certainly not a member of the order *Thermoplasmatales* or the class *Thermoplasmata* (Figure 1, Table S1, Table S2). We found evidence for this classification in the phylogenetic analysis for both 16S rRNA and ribosomal protein S15 genes, where Iplasma groups outside of the *Thermoplasmata* clade (Figure 1 and Figure S1) as observed previously [5, 47, 68, 69]. In the case of the 16S tree, Iplasma forms a monophyletic group with a number of environmental clones from acidic solfataric mud and acidic springs (Genbank) [70]. Because archaeal phylogeny is still unresolved, it is impossible to exactly determine the phylogeny of new taxa [71]. However, the branch length separating Iplasma and the *Thermoplasmata* organisms is greater than 0.25, supporting the separation of Iplasma into a new class of *Euryarchaea* as we previously suggested in Justice *et al*, 2012 [5].

We examined a number of whole-genome measures of relatedness to further investigate evolutionary relationships. First, we identified the fraction of predicted orthologs in pairwise comparisons, and then determined their average amino acid identity. The normalization step involved dividing the number of orthologs by the average number of genes in the pair of genomes considered. Iplasma shares a lower percentage of orthologs, and a lower average amino acid identity with each of the other AMD plasma genomes than the other AMD plasma genomes share with each other (Table S3), consistent with a divergent phylogenetic placement. Fer1 vs. Fer2 has the highest amino acid identity, as expected for closely related species. Note that it was previously suggested that the genomes of Fer1 and Fer2 are different enough to merit classification as separate species based on analysis of recombination rates [72]. Eplasma and Gplasma are relatively closely related, as are Aplasma and Gplasma (Table S4). Second, we looked at conserved gene order as a measure of evolutionary distance [68]. For each genome pair, we determined the number of syntenous orthologs and divided this by the number of shared orthologs. The Iplasma genome has the lowest synteny with the other AMD plasma genomes, Fer1 vs. Fer2 displays the highest synteny, followed by Eplasma vs. Gplasma (Table S5). The same trend holds true for another measure of synteny, the average length of syntenous blocks of genes in pairwise comparisons (Table S6). This whole-genome data supports the tree topology and evolutionary distances assigned to the 16S rRNA genes in our phylogenetic analysis.

(ii) General genome features: Genome features, including the number of tRNA synthetases and ribosomal genes, are summarized in Chapter 1 [68]. All of the genomes contain the full suite of tRNAs and most or all orthologous marker genes [38, 68], consistent with a high degree of genome completeness (Table S7). Important metabolic and structural features of each genome are listed and illustrated in Figure 2 and Figure 3.

(iii) Genomic island (unique) in G-plasma: A potential genomic island was identified in the Gplasma genome. It consists of a block of nine genes that have virtually no orthologs in any of the other *Thermoplasmatales* genomes and is made up primarily of proteins of unknown function (Figure 4, Table S8). All nine of the proteins are represented in a whole community proteomic dataset reported previously [73], and three are among the most highly detected proteins of this organism in that dataset. The motifs and domains identified suggest that a number of these proteins are membrane associated, including a protein containing an AAA+ FtsH ATPase domain (gene number 13327_0053) (found in a membrane-integrated metalloprotease [74]), a protein containing six transmembrane motifs and a signal peptide (13327_0056), and another with fourteen transmembrane motifs and a signal peptide (13327_0059). Additionally, three of these proteins include a rhodanese-like domain possibly involved in phosphatase or sulfurtransferase activity and another contains an armadillo repeat region, often used to bind large substrates such as peptides or nucleic acids (13327_0058).

The absence of orthologs to this block of hypothetical proteins in other *Thermoplasmatales* genomes is a strong indication that it may have been acquired by horizontal gene transfer. Many flanking genes have syntenous orthologs in other closely-related genomes. However, the lack of GC skew in the nucleotide signature of these genes suggests that the transfer event was not recent or that the donor had a similar GC content to Gplasma.

(iv) Cell wall biosynthesis and imaging: *Thermoplasmatales* cells are generally bounded by a single membrane. However, there are two *Picrophilus* species that have a single membrane surrounded by a surface-layer (S-layer) [64]. We characterized archaeal-rich biofilm communities via cryo-electron microscopy and identified surface layers on many single membrane bound cells (Figure 5, Movie S1). Thus, we looked for the genes needed for surface layer structural proteins and their post-translation modifications (i.e., N-glycosylation). We found putative S-layer genes in all of the AMD plasma genomes (except Fer1) that are homologous with the predicted *P. torridus* S-layer genes (Table S9) [75], but found no homology to the predicted S-layer genes in their next closest relative, *Acidiloprofundum boonei* [76]. We also found genes potentially involved in archaeal S-layer protein N-glycosylation. Of particular interest were homologs to the AglD and AglB genes of *Haloferax volcanii*, which have been shown to be essential to S-layer protein N-glycosylation in that organism [77]. Many of the Iplasma S-layer-related genes occur in a cluster, and several have conserved gene order in distant relatives, including several enzymes that attach sugars to a dolichol that might serve as a membrane anchor for the formation of an oligosaccharide during N-glycosylation. The Iplasma genome contains a gene cluster syntenous with distant relatives that encodes all of the proteins in the ADP-L-glycero-β-D-manno-heptose (AGMH) biosynthesis pathway (Table S9). AGMH is attached to S-layer proteins in gram-positive bacteria [78-80], suggesting that this may be involved in S-layer glycosylation in Iplasma as well. Finally, in the same genomic region genes are found for the biosynthesis of GDP-L-fucose, a glycoprotein component, and dTDP-L-rhamnose, a lipopolysaccharide component, indicating that these may make up part of the AMD plasma S-layer polysaccharides.

(v) Energy metabolism *(a) Iron oxidation*: Ferric iron produced by biotic iron oxidation drives metal sulfide mineral dissolution, and thus iron oxidation is one of the most important biochemical processes that occurs in acid mine drainage systems [81-83]. In order to assess

which of the AMD plasmas were involved in this process, we looked for potential iron oxidation genes. Allen *et al.* [37] inferred that a sulfocyanin blue-copper protein is involved in iron oxidation in *Ferroplasma* spp. (e.g. Fer1), and Dopson *et al.* provided proteomic and spectrophotometric evidence that support this inference [8]. Fer2's genome contains a sulfocyanin homolog, whereas Aplasma and Gplasma contain homologs to the rusticyanin gene, which encodes a blue-copper protein implicated in iron oxidation in *Acidithiobacillus ferrooxidans* (Table S9) [84]. E- and Iplasma do not appear to have a rusticyanin or a sulfocyanin gene, suggesting that they are not iron oxidizers. It is important to note that the rusticyanin gene found in the A- and Gplasma genomes is most closely related to rusB, which does not have a proven iron oxidizing function.

All of the AMD plasma blue-copper proteins (BCPs) contain a type I copper-binding site, consisting of two histidines, one cysteine, one methionine and a cupredoxin fold, identified by a 7 or 8-stranded β-barrel fold [85-87] (Figure S2). Previous research has indicated that phylogenetic trees of BCPs are consistent with structural relatedness of the proteins [88]. Our phylogenetic analysis grouped Aplasma's gene with the rusticyanins, whereas the Fer1 and Fer2 genes grouped with the sulfocyanins (Figure S3). Interestingly, the Gplasma gene is so divergent that it does not consistently group with the other iron-oxidation blue-copper proteins. Its divergence seems to stem from two more β-strands than most of the other rusticyanin-like proteins (Figure S2). The tree also provides evidence for the horizontal transfer of the BCP genes. For example, related rusticyanin-like genes are found in the *Gammaproteobacteria* and in a variety of *Euryarchaea*. Similarly, closely related sulfocyanin-like genes are found in *Euryarchaea* and *Crenarchaea*. This suggests that both sulfocyanin and rusticyanin genes have been horizontally transferred. Additional evidence for the function of these genes was found in their inferred protein structure and copper-binding residues (Figure S2). Vivekanandan Giri *et al.* identified amino acid signatures for both sulfocyanin and rusticyanin based on conserved sequence in copper-binding regions of the proteins [89]. The Fer1 and Fer2 BCPs include one of the sulfocyanin motifs, FNFNGTS, as well as imperfect conservation of the motifs identified for both sulfocyanin and rusticyanin (Table S10). Conversely, the Aplasma and Gplasma blue-copper proteins do not contain any of the conserved sulfocyanin-specific motifs. They contain imperfect matches to the rusticyanin-specific motif. This is consistent with the inferences we have made based on the phylogenetic tree of these genes.

In addition to iron oxidases, all of the AMD plasma genomes also contain an analog to the complex III/cytochrome bc complex used during iron oxidation (and aerobic respiration) in *A. ferrooxidans* [90]. Both the cytochrome b and rieske Fe-S protein subunits of the cytochrome bc analog found in archaea were identified. However, like other archaea, they do not have genes for the cytochrome c subunit of the bacterial-type complex [91, 92]. None of the genomes contain homologs to any of the other genes in the *A. ferrooxidans* rus operon [93].

In general, the absence of blue-copper protein genes suggests that E- and Iplasma lack the Fe-oxidation capability entirely, whereas the other AMD plasmas utilize two different pathways to carry out this metabolism. However, the BCP genes may have novel functions in these oganisms and it is possible that E- and Iplasma do have blue-copper proteins in their genomes because gaps remain in their assemblies. We took steps to rule out this possibility (see Methods section). Because Fe(II) is an abundant electron donor in the AMD environment, this observed genetic variation in Fe oxidation potential may be important in niche differentiation.

(v) Energy metabolism *(b) Carbon monoxide dehydrogenase*: The Iplasma, Fer1 and Fer2 genomes encode genes for a possible carbon monoxide dehydrogenase, (CODH) (Table S9), including genes for all three subunits of the CoxMLS complex. Recent research suggests that aerobic CO oxidation may be a widespread metabolism among bacteria [94]. Thus, it is not surprising to find organisms with this capability in AMD systems. In fact, up to 50 ppm of CO has been measured in air within the Richmond Mine (M. Jones, *pers. comm.* 2011).

A phylogenetic tree of the catalytic subunits of CODH indicates that all but one of the AMD plasma complexes is more closely related to the aerobic type than the anaerobic type (Figure S4). The active site encoded by these genes also suggests that they are aerobic CODH proteins closely related to the form II CODH, which has the motif: AYRGAGR (Figure S5) [94, 95]. This enzyme can be used to make CO_2 either for C fixation or to make reducing equivalents. The AMD plasma genomes do not contain any of the genes for the known archaeal C fixation pathways. Based on these observations, we hypothesize that these CODH proteins are used solely to make electrons available for aerobic respiration. However, it is possible that they use a novel C fixation pathway that incorporates this CODH such as that suggested by Cardenas *et al*. [96].

Interestingly, our CODH phylogenetic tree suggests that there is another AMD plasma gene that encodes a Ni-CODH, Fer2 scaffold 31 gene 47. Ni-CODHs are anaerobic and reduce CO_2 to CO. This enzyme is generally involved in C fixation via the Wood-Ljungdahl pathway, the genes for which are not found in the AMD plasma genomes. However, additional evidence for the annotation of this gene as a Ni-CODH is provided in its structural alignment with known Ni-CODH proteins (Figure S6). As a whole, the genomic evidence suggests CO oxidation capacity among Fer1, Fer2, and Iplasma and a potential for CO_2 reduction in Fer2.

(v) Energy metabolism *(c) Aerobic respiration*: Fer1 and *T. acidophilum* are known to be facultative anaerobes [7, 12, 97, 98], whereas *T. volcanium* and *P. torridus* are aerobes. Therefore, it is not surprising that all of the Richmond Mine AMD plasmas have the genetic capacity for aerobic respiration and catabolism of organic compounds via two glycolytic pathways, pyruvate dehydrogenase, the TCA cycle and an aerobic electron transport chain (Table S9). Some AMD plasma genes in the aerobic electron transport chain have been observed in proteomic analyses [5].

The AMD plasmas' electron transport chain genes are similar to that of other archaea in that they do not contain all of the subunits of the NADH ubiquinone-oxidoreductase complex [99]. All of the AMD plasmas except Aplasma are missing the NuoEFG subunit genes found in the bacterial type complex I and instead have the subunits found in the archaeal-type complex I, NuoABCDHIJKLMN. Fer2 is missing NuoIJKLM possibly because the genes for this complex are found at the end of an incomplete contig. Eplasma, Gplasma and Fer1 maintain the Nuo gene order found in a number of other archaea including, *Halobacterium sp.*, *Sulfolobus solfataricus*, and *T. acidophilum* [100]. All contain succinate dehydrogenase complex genes (Table S9). In the case of A-, E-, and Gplasma, the complex is missing SdhD, and many of the SdhC genes have annotations with low confidence. This finding is congruent with previous research that shows that the genes for the membrane anchor subunits of the complex are poorly conserved, possibly due to low selective pressure [101]. As mentioned previously in section (v)(a), the AMD plasmas have genes homologous to several predicted archaeal complex III/cytochrome bc complex genes (Table S9).

Archaeal-type aerobic terminal oxidases include cytochrome c oxidases (CCOs) and cytochrome bd oxidases. Genes for the cytochrome bd complex are found in *P. torridus*, *T. acidophilum* and *T. volcanium* [102]. All of the AMD plasma genomes contain the two genes for this complex. They also all contain the two essential genes for the archaeal heme-copper oxidase/CCO complex (subunit I and II) [102], and we confirmed that subunit II contains the Cu-binding motif generally found in CCOs [103] (Figure S7). Like the other CCO genes in *B. subtilis* and *E. coli,* the two cytochrome c genes in the AMD plasmas occur in a gene cluster with a protoheme IX farnesyltransferase, required for synthesis of the heme type used in aa(3) type CCOs [104]. The subunit II gene shares a high amino acid identity with several oxidases of this type, further indicating an aa(3) type CCO (Table S11).

Archaea use A-type ATP synthases to generate ATP from an electrochemical gradient. All of the AMD archaeal genomes contain the AhaABCDEFIK genes that comprise this complex in *Methanosarcina mazei*, although they are missing an ortholog to AhaG. All but Eplasma and Iplasma contain a putative AhaH gene. AhaG is also absent in *T. acidophilum*, indicating that it may not be necessary for ATP synthesis in these organisms.

(v) Energy metabolism ***(d) Alternative electron acceptors:*** In addition to aerobic respiratory capabilities, some *Thermoplasmatales* organisms are able to respire anaerobically [98]. Anaerobic reduction of S^0 or sulfur ions could allow archaea in AMD systems to survive under anoxic conditions deep inside floating biofilms or in sunken biofilms and sediment, where many sulfur compounds are present [105]. Iplasma's genome contains several genes that are homologous to AsrA and B, known sulfite reduction protein genes (13606_0515 and 13606_0514). These proteins comprise two of the three subunits of the AsrABC dissimilatory sulfite reductase complex found in *Salmonella typhimurium* [106]. However, the Iplasma genome does not contain the AsrC subunit, which contains the siroheme-binding motif and thus is thought to contain the active site for sulfite reduction. As the Asr proteins are not well characterized in many organisms, it is possible that these genes are misannotated. Synteny-based annotation ties these two genes to an adjacent FdhF formate dehydrogenase alpha subunit gene, indicating a possible involvement of these genes in formate dehydrogenase activity. In fact, in *Methanosaeta thermophila* the orthologs to these genes are annotated as 4Fe-4S ferredoxin, iron-sulfur binding instead of sulfite reductases (NCBI), and one is structurally related to the HycB hydrogenase 3 Fe-S protein formate dehydrogenase subunit based on CBLAST against the NCBI protein structure database. Additional protein modeling suggests that one of the proteins in Iplasma could be a subunit of the formate dehydrogenase complex (unpublished). Thus, we suggest that these two proteins are functionally related to formate dehydrogenase in Iplasma.

Interestingly, Iplasma's genome contains homologs to all of the genes overexpressed under anaerobic conditions for *T. volcanium* as well as all of the genes overexpressed or over-transcribed under anaerobic conditions for *T. acidophilum* (except for their predicted sulfur respiration gene Ta1129) in two previous studies [107, 108] (Table S12). The other AMD archaea also share most, but not all, of these genes. Although there is no direct genomic evidence for anaerobic respiration, novel anaerobic respiratory pathways are possible. In fact, there is evidence that Fer1 can grow via anaerobic Fe reduction [7], and enrichment cultures of Fer1 and Aplasma reduce iron [5].

(v) Energy metabolism *(e) Heterotrophy*: Chemolithoautotrophy is a common lifestyle in AMD communities (e.g., of *Leptospirillum* spp.) [109]. However, the *Thermoplasmatales* archaea are mostly heterotrophs (only *F. acidiphilum* has been shown to have any autotrophic capability [62]). The AMD plasma genomes encode genes for a wide variety of heterotrophic metabolisms, both aerobic and anaerobic. These genomes have the genes necessary for the catabolism of organic compounds for energy generation, including fatty acids, sugars, starch, and glycogen, but not refractory organic matter such as cellulose (Table S9).

All of the AMD plasmas have genes for sugar and polysaccharide catabolism, including glucoamylase genes required to break down starch and alpha-amylase genes for glycogen catabolism into glucose and dextrin. They have the conventional Embden-Meyerhoff (EM) glycolytic pathway (Table S9). Moreover, they have the genes for the non-phosphorylative Entner-Doudoroff (NPED) pathway for glucose degradation also found in a number of (hyper)thermophilic archaea, including *T. acidophilum*, *P. torridus*, *S. solfataricus*, *Sulfolobus acidocaldarius*, *Sulfolobus tokodai* and *Thermoproteus tenax* [110-113]. The AMD plasma genomes contain homologs to all of the genes in this pathway, including a homolog to the proven *P. torridus* KDG aldolase [114]. Thus, the AMD plasmas are similar to their *Thermoplasmatales* relatives, all of which have genes homologous to those of both pathways. Previously published proteomic data indicates that all of the AMD plasma organisms express some of these genes in these two pathways [5].

Another potential carbon source for the AMD plasmas is lipids from lysed cells. All of the AMD plasma genomes contain a full set of orthologs to the genes for the aerobic fatty acid oxidation pathway from *E. coli* (Table S9), and these genes are often detected in proteomic data [5]. Because many of these proteins are acyl-CoA dehydrogenases, which are known to have undergone frequent gene duplication and horizontal transfer events [115], it is difficult to discern which genes are actually involved in the pathway. However the number of β-oxidation-related annotations suggests that the AMD plasmas are capable of fatty acid breakdown, and many of the proteins from this pathway have been identified by proteomics. Lipids may be important carbon sources for the AMD plasmas, as more readily metabolized sugars tend to be consumed rapidly, and hence are likely in low concentrations in the environment.

Interestingly, the AMD plasmas have the capacity to catabolize one-carbon compounds such as methanol. All except for Gplasma have several genes for subunits of a formate dehydrogenase. A number of these genes are found in gene clusters with biosynthesis genes for their specific molybdopterin cofactor [68]. The AMD plasma genomes lack genes for the formate hydrogen lyase complex and the reductive acetyl-CoA pathway of carbon fixation. Thus, we surmise that the formate dehydrogenase is involved in an oxidative pathway for methanol, methylotrophy (i.e., methanol degradation to formaldehyde, formaldehyde to formate, and formate oxidation to CO_2). The AMD plasmas have homologs to all of the enzymes in this pathway, including the enzyme used by all thermotolerant methanol-oxidizing bacteria, a NAD-linked methanol dehydrogenase [116] (Table S9). Among the AMD plasmas, only Iplasma appears to have the genes necessary for the ribulose monophosphate cycle, which is commonly used for carbon assimilation from formaldehyde [116]. None of the genomes contain the genes necessary for the other known formaldehyde assimilation pathway, the serine cycle. The ubiquity of methylotrophic genes among the AMD plasmas suggests a substantial concentration of methanol in the AMD solution. Fer1 has been shown to produce methanethiol during cysteine degradation [117]. Methanol in the AMD biofilm may be a product of methanethiol catabolism.

(v) Energy metabolism *(f) Fermentation and the use of fermentation products*: Archaea are typically more abundant in thick, mature AMD biofilms [118] where they may encounter anoxic microenvironments [105]. Thus, we looked for potential fermentation genes in their genomes. They all have the genes for fermentation of pyruvate to acetate found in *Pyrococcus furiosus* and a number of other anaerobic fermentative and aerobic archaea [119-122] (Table S9). This pathway is unique in that it converts acetyl-CoA to acetate in only one step, with an ADP-forming acetyl-CoA synthetase. It is the only phosphorylating step of pyruvate fermentation via the NPED pathway. Previously this enzyme had been detected in hyperthermophilic and mesophilic archaea as well as some eukaryotes [122]. In anaerobic archaea this enzyme is involved in fermentation, whereas in aerobic archaea it makes acetate that is then catabolized via aerobic respiration [123]. The AMD plasmas have the genes necessary for fermentation to acetate under anaerobic conditions and for acetate respiration under aerobic conditions via an acetate-CoA ligase or the reversal of the direction of the acetate-CoA synthetase.

(vi) Putative hydrogenase 4 genes: Several AMD plasma genomes contain a number of putative group 4 hydrogenase genes (Figure S8). A group 4 hydrogenase complex and formate dehydrogenase comprise the formate hydrogen lyase that catalyzes non-syntrophic growth on formate and production of H_2 in hyperthermophilic archaea (*T. onnurineus*) [124, 125]. The putative group 4 hydrogenases, though closely related to the group 4 type, lack the two conserved hydrogen and Ni-binding motifs that are thought to be necessary for H_2 formation [125, 126], possibly indicating some other function.

(vii) Toxic metal resistance: The Richmond Mine solutions contain extremely high (mM) concentrations of arsenic, cadmium, copper, and zinc [127]. Genomic evidence indicates that the AMD plasmas utilize multiple strategies to protect themselves from these elements, such as oxidation/reduction to less toxic forms and efflux (Table S9) [9, 11]. All of the AMD plasmas have at least two genes from the arsenic resistance (ArsRABC) operon. Only Gplasma has all of the genes in the operon, but Fer1 has previously been shown to have resistance to both arsenate and arsenite, despite lacking the arsenate reductase gene [11]. All of the AMD plasmas except for Fer2 have two of the genes in the mercury resistance operon (MerTPCAD), MerA and MerP (mercuric reductase and the mercuric ion-binding protein, respectively). All of the genomes also contain some putative copper resistance genes in the CopABCD operon or the CopYBZ loci that was identified previously in Fer1 [128]. Specifically they all have homologs to CopB. This gene has been shown to be involved in copper sequestration as a copper resistance strategy in *Pseudomonas syringae* [129]. The heavy metal transporter genes found in the AMD plasma genomes group into two different clades in a phylogenetic tree of metal resistance P-type ATPases. All of the genomes except for that of Iplasma contain two types of metal resistance transporters according to this phylogenetic analysis, a Cu/Ag transporter related to CopA or CopBZ and a Zn/Cd transporter related to CadA.

(viii) Biosynthesis: Because the AMD plasmas live in dense biofilms, they can rely on other organisms for certain biomolecules such as cofactors, amino acids, etc. We previously demonstrated a lack of known genes for *de novo* cobalamin biosynthesis in A-, E-, G-, and

Iplasma in Chapter 1 [68]. Here we examined the AMD plasma genomes for other biosynthetic pathways.

(viii) Biosynthesis *(a) Glyoxylate shunt*: Only Eplasma has the genes for the glyoxylate shunt, a pathway associated with the TCA cycle that allows the use of organic compounds that are degraded to acetyl-CoA (i.e. fatty acids) for biosynthesis (Table S9). One of the proteins encoded in this pathway, the malate synthase, has been detected in proteomic analyses [5].

(viii) Biosynthesis *(b) Amino acid synthesis*: The *Thermoplasmatales* archaea exhibit differential abilities to synthesize amino acids, suggesting that some of them rely more heavily on organic compound uptake than others. The genomes of E-, G- and Iplasma do not contain most of the histidine synthesis pathway genes. Eplasma and Iplasma also lack many of the genes necessary for the valine and (iso)leucine synthesis pathway (Table S9). They are also among the subset of organisms that do not make their own cobalamin [68]. This group of organisms may rely on amino acid and cobalamin scavenging to avoid the energetic costs of *de novo* synthesis.

(viii) Biosynthesis *(c) Trehalose biosynthesis*: Compatible solutes allow organisms to maintain osmotic balance under high salt conditions or to protect against heat shock and cold shock [130]. A number of archaea make organic solutes for this purpose. *T. acidophilum* and a number of *Sulfolobales* archaea have been shown to produce trehalose as a compatible solute. In these organisms it has also been suggested that it is used to thermostabilize macromolecules and as a carbon storage molecule [130]. All of the AMD plasmas except for Iplasma have the genes necessary for trehalose biosynthesis from maltose (Table S9). The monophyletic group of A-, E-, and Gplasma also has the genetic potential for trehalose synthesis from glycogen.

(ix) Motility: Motility can provide a competitive advantage for archaea in aquatic environments by allowing them to colonize new sites and move across environmental gradients. To determine potential for motility, we looked for flagellar, chemotaxis and pili genes in the AMD plasma genomes.

Both the A- and Gplasma genomes contain the full flagella FlaBCDEFGHIJ operon found in *Methanococcus voltae* [131-133] and *Halobacterium salinarum* [134] (Table S9). Thus, these organisms are predicted to be motile, yet they lack identifiable chemotaxis genes.

No flagellar genes are found in the other AMD plasma genomes, suggesting differences in motility. We used cryo-EM to confirm the existence of flagella on cells identified as AMD plasma archaea based on the presence of a single cell membrane, cell shape and size, and sample location (Figure 6, Movie S2). The flagella-like structures found in these cells have diameters of about 10-14 nm, far thicker than the pili observed in similar AMD plasma organisms. In Figure 6 we see a high-electron density area inside the cytoplasm immediately adjacent to the flagella that may be part of the associated protein motor complex.

In addition to flagellar assembly genes, a number of the AMD plasma genomes contain genes for Type II secretion or Type IV pili that are used in twitching motility, or possibly conjugation or attachment to the biofilm or other surfaces. All of the genomes except for Fer1 and Fer2 contain some of these genes, and in Eplasma, Gplasma, and Iplasma they are in a cluster with conserved gene order among the AMD plasmas (Table S13). Cryo-EM confirms the

existence of pili, and shows attachment of the pili from the original cell to other cells (Figure 7, Movie S2).

(x) Vesicle-like cavities: Cryo-EM imaging demonstrates that a number of the AMD plasma cells harbor low electron-density inclusions within what appears to be a lipid membrane (Figure 7). These are similar in appearance to the gas vesicles that some extreme halophiles use for buoyancy [135], though those gas vesicles are enclosed in a proteinaceous casing. We did not find genomic evidence of gas vesicle formation in the AMD plasmas by performing BLASTP searches of their genomes against the gas vesicle protein (gvp) genes of *Haloarchaea* [136]. Novel vesicle formation genes are expected.

Conclusion: Metagenomic and phylogenetic analyses reveal evolutionary, metabolic and cell structural differences among the AMD plasmas. We recognize Iplasma as a representative of a phylogenetically distinct class. All AMD plasmas have the capacity to grow both aerobically and anaerobically. Although we do not see differences in the gene content required for catabolism of organic carbon, differences in regulation of the genes in these pathways may differentiate these organisms. Important differences in their potential abilities to oxidize iron and for biosynthesis rather than reliance on scavenging of cofactors and amino acid precursors may allow the coexisting AMD plasmas to take advantage of microniches that occur in complex biofilms. Similarly, differences in motility may allow some AMD plasmas to colonize new sites or move along physicochemical gradients. Comparative genomic analyses also provide new information indicating the importance of methylotrophy, fermentation potential, and other heterotrophic metabolisms to the AMD plasmas.

Materials and methods

(i) DNA sequencing and assembly: The new genomes presented here are composite assemblies of DNA extracted from a number of biofilm samples from the Richmond Mine, Iron Mountain, CA. Sample collection, DNA extraction, sequencing, genome assembly, and automated annotation were described previously [36, 68, 137, 138]. All of the genomes were automatically assembled using velvet [139] and then manually curated, using the Consed software [140] to correct misassemblies and join contigs across gaps. Assembly statistics were published in Yelton, *et al*. 2011 [68].

(ii) Gene annotation: In addition to the automated annotation pipeline for the genomes described [68], we used a synteny-based method to improve the annotations of poorly annotated genes. This method was described previously [68], and provides either specific or general functional annotations based on gene context in distantly related genomes.

We manually curated all annotations that are specifically cited in this paper in the following manner. Genes were aligned against the interpro and nr databases with a BLASTP algorithm. Genes were then annotated if they had a TIGR or Pfam domain hit that predicted a specific function with an e-value of at least 1×10^{-10} and coverage of more than 70% of the protein. Genes were given a "putative" annotation if they met the previous criteria except they had an e-value between 1×10^{-4} and 1×10^{-10}, and matched 50-70% of the protein, or if their domain-based hits provided only general functional information. In these cases, additional evidence from hits from the nr database was used if possible to provide a specific functional annotation. Genes were given a "probable" annotation if they had annotated hits in the nr database with greater than 30% amino acid identity over 70% of the length of the gene. For incomplete metabolic and structural pathways, BLASTP searches were carried out against the entire Richmond Mine metagenomic database for missing genes based on the amino acid sequence of their closest relative. In the case where significant hits were uncovered, maximum-likelihood amino acid trees were used to place these genes within the AMD plasma group of archaea and this placement was used to associate the genes with a specific AMD plasma genome or outside the group altogether.

(iii) Phylogenetic analyses: Phylogenetic analyses of certain genes were used to help place them in evolutionary context (e.g. 16S rRNA, blue-copper proteins). In these cases, the genes were aligned using the MAFFT alignment tool and default parameters [141, 142]. The alignment was then manually corrected if needed. For protein trees, the completed alignment was used to make a phylogenetic tree with the FastTree [143, 144] maximum likelihood-based tree software. In the case of the 16S rRNA gene, the phylogenetic tree was made using RaxML for improved accuracy based on the taxonomy of isolate organisms [145]. Support values were calculated for each branch split via the Shimodaira-Hasegawa test provided by the –boot option set to 1000 bootstraps for FastTree trees and using the rapid bootstrap for the RaxML tree.

(iv) Cryo-EM Specimen Preparation: For cryo-EM, aliquots of 5 µl were taken directly from the fresh biofilm samples and placed onto lacey carbon grids (Ted Pella 01881) that were pre-treated by glow-discharge. For cryo-ET, samples were deposited onto support grids pre-loaded with 10 nm colloidal gold particles. The Formvar support was not removed from the lacey

carbon. The grids were manually blotted and plunged into liquid ethane by a compressed air piston, then stored in liquid nitrogen.

(v) Electron Tomography Imaging: Images were acquired on a JEOL–3100 electron microscope equipped with a FEG electron source operating at 300 kV, an Omega energy filter, a Gatan 795 2Kx2K CCD camera, and cryo-transfer stage. The stage was cooled to 80 K with liquid nitrogen. For more information on imaging and analysis see Text S1.

Acknowledgements

We thank Mr. Ted Arman (President, Iron Mountain Mines), Mr. Rudy Carver, and Mr. Richard Sugarek for site access and other assistance. This work was supported by DOE Genomics:GTL project Grant No. DE-FG02-05ER64134 (Office of Science). APY acknowledges NSF Graduate Research Fellowship Program support. LRC also acknowledges support by the Director, Office of Science, Office of Biological and Environmental Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Figures

Figure 1: 16S rRNA tree indicating the possibility of a candidate class that includes Iplasma. Bootstrap values are shown at branch splits. Genbank accession numbers are listed after organism names.

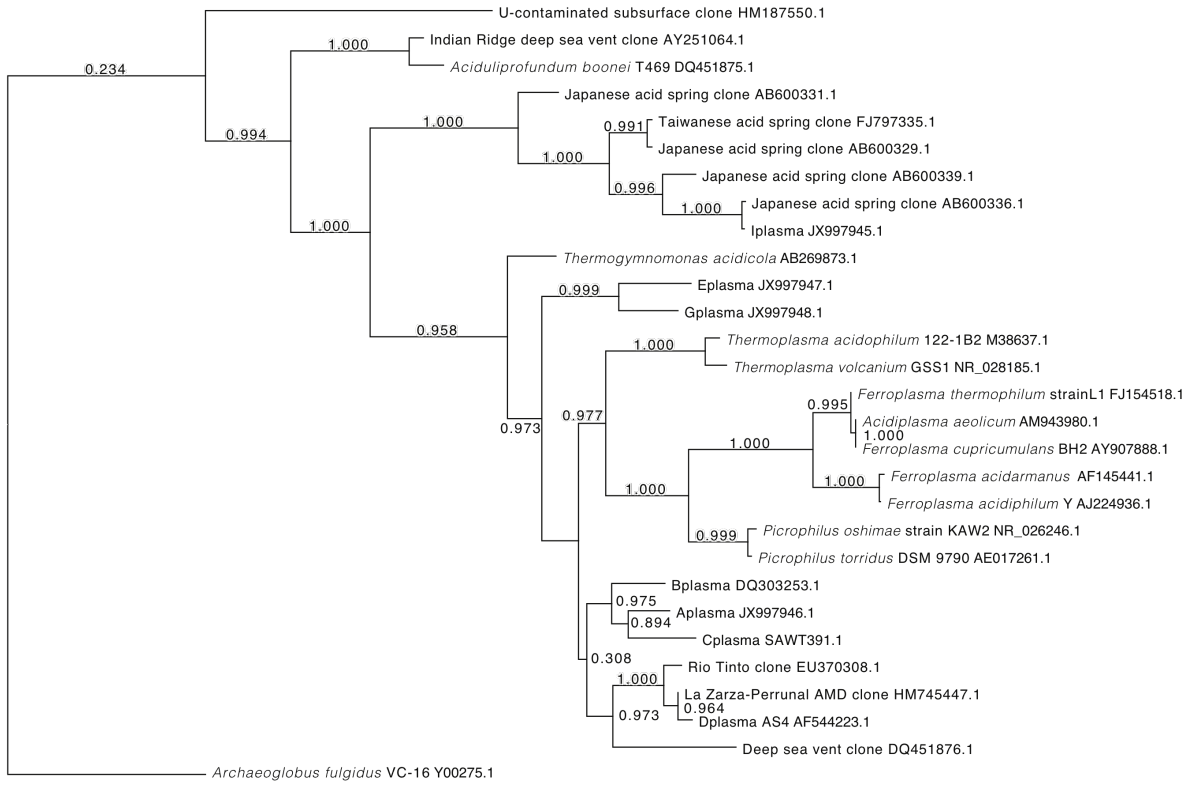


Figure 2: General overview of metabolic differences within the AMD plasmas: Y indicates that the pathway is found in the genome, whereas N indicates that it is not.

Function	APL	EPL	GPL	FER1	FER2	IPL
Aerobic metabolisms						
Aerobic respiration	Y	Y	Y	Y	Y	Y
Fe oxidation (blue-copper protein)	Y	N	Y	Y	Y	N
Aerobic CODH	N	N	N	Y	Y	Y
Anaerobic CODH	N	N	N	N	Y	N
Anaerobic metabolisms						
Formate dehydrogenase	Y	Y	N	Y	Y	Y
Putative hydrogenase complex	Y	Y	Y	Y	Y	N
Fermentation to acetate	Y	Y	Y	Y	Y	Y
Carbon catabolism						
Glycolysis	Y	Y	Y	Y	Y	Y
Entner-Doudoroff	Y	Y	Y	Y	Y	Y
Beta oxidation	Y	Y	Y	Y	Y	Y
Methylotrophy	Y	Y	Y	Y	Y	Y
Biosynthesis						
Cobalamin biosynthesis	N	N	N	Y	Y	N
Molybdopterin biosynthesis	Y	N	N	Y	Y	Y
Histidine synthesis	Y	N	Y	Y	Y	N
Leucine/Isoleucine synthesis	Y	N	Y	Y	Y	N
Glyoxylate shunt	N	Y	N	N	N	N
Motility						
Flagella	Y	Y	N	N	N	N
Chemotaxis	N	N	N	N	N	N
Toxic metal resistance						
Arsenic resistance	Y	Y	Y	Y	Y	Y
Copper resistance	Y	Y	Y	Y	Y	Y
Mercury resistance	Y	Y	Y	Y	N	Y
Structure						
S-layer	Y	Y	Y	Y	Y	Y
Ether-linked lipids	Y	Y	Y	Y	Y	Y
Cellulose/cell wall polysaccharides	N	N	N	N	N	N
Pili	N	Y	Y	N	N	Y

Figure 3: Metabolic and structural features of the AMD plasma organisms. The surface layer proteins are pink. Pili are blue. Flagella are brown. The electron transport chain is yellow. The metal resistance proteins are blue. The archaeal type ATP synthase is yellow and rusticyanin is blue.

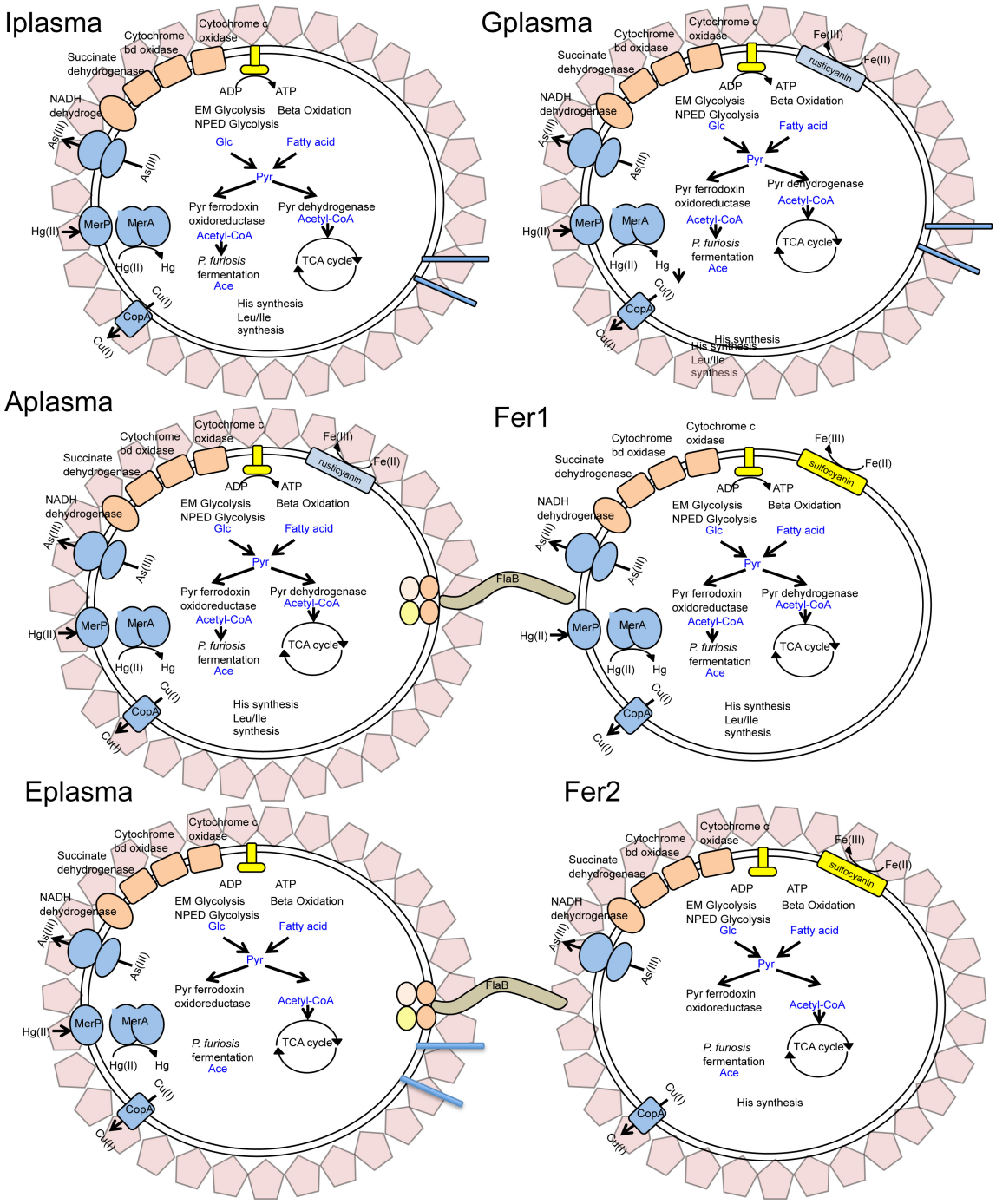


Figure 4: Cluster of unique genes in Gplasma. Arrows are proportional to the length of each gene and indicate its direction of transcription. The gene numbers are shown inside the arrows. All genes are from contig number 13327. Motif and domain-based annotations are shown above the arrows. Rhod indicates a rhodanese-like domain.

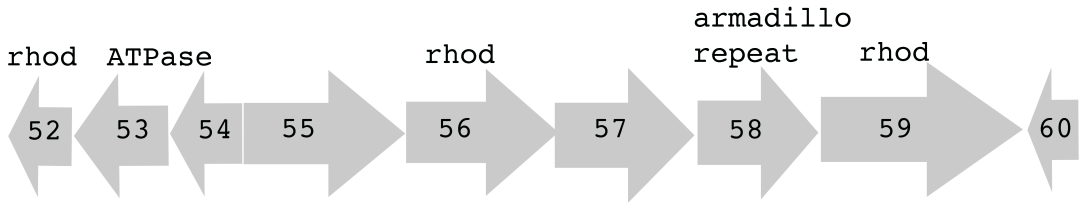


Figure 5: Cryo-EM of surface-layer on an AMD plasma cell from the Richmond Mine. Insets show a higher magnification. Arrows point to putative surface layer proteins.

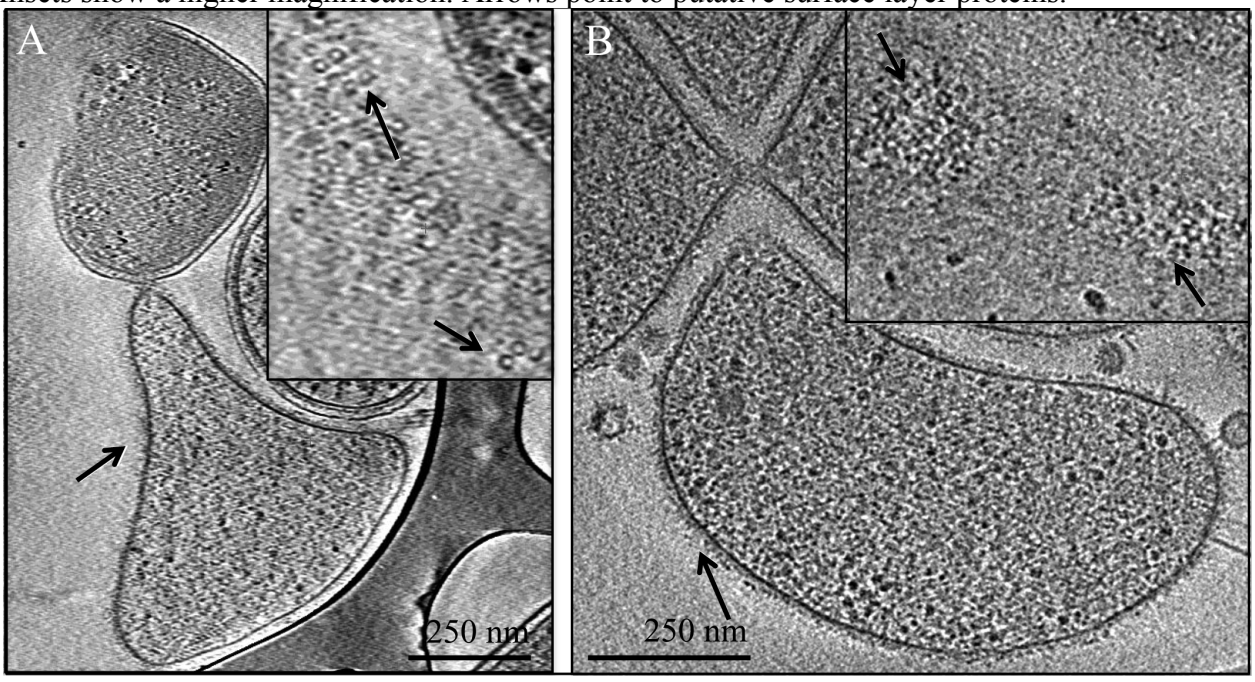


Figure 6: Cryo-electron microscopy of AMD plasma cells. Arrows point to flagella. The blue box is around a potential motor protein complex.

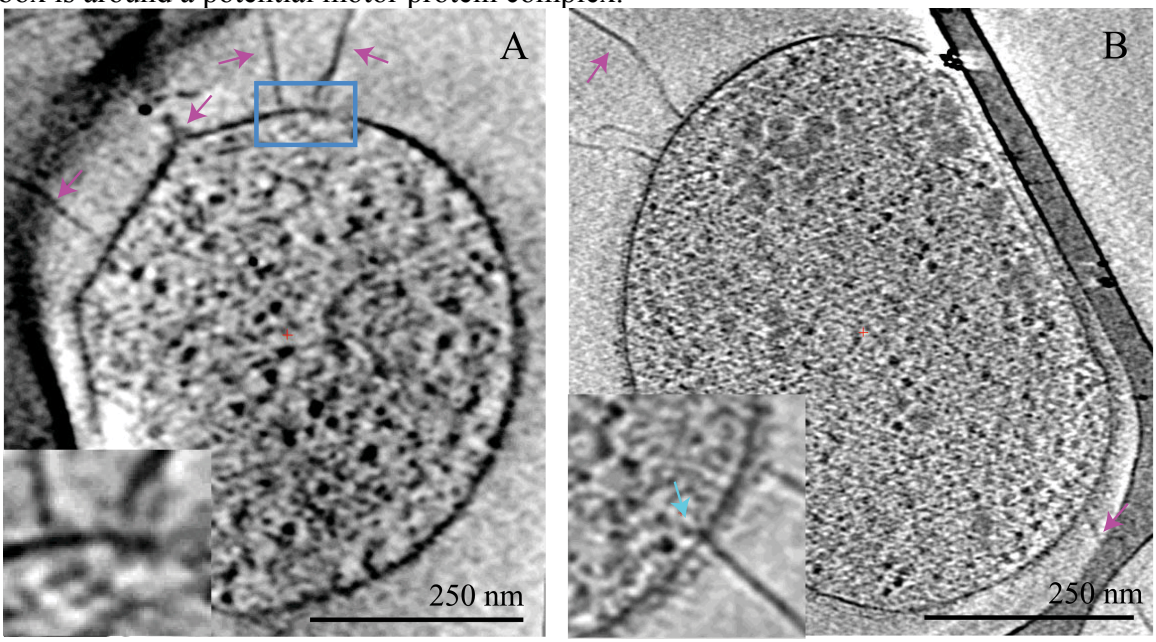
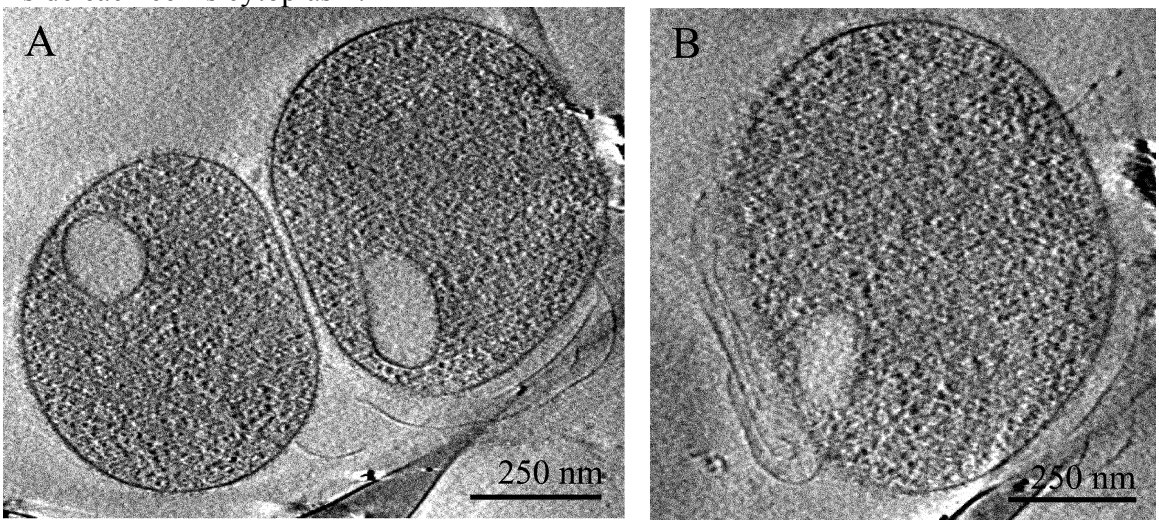


Figure 7: Cryo-electron microscopy of AMD plasma cells with putative pili. Arrows point to pili. Vesicle-like structures are delineated by a singe membrane layer around an ovoid shape inside each cell's cytoplasm.



Supplementary Materials

Table S1: Percent nucleotide identity of 16S rRNA genes in all of the AMD plasmas relative to one another.

16S rRNA Nucleotide Identity	Aplasma	Eplasma	Gplasma	Fer1/Fer2	Iplasma
Aplasma		91.67%	91.64%	88.06%	84.75%
Eplasma			92.95%	89.47%	83.99%
Gplasma				88.06%	86.65%
Fer1/Fer2					84.88%
Iplasma					

Table S2: 16S nucleotide identity for AMD Thermoplasmatales organisms and close relatives. Note that all of the organisms in the first column except for *Aciduliprofundum boonei* are classified as *Thermoplasmatales*.

	APL	EPL	GPL	Fer1/Fer2	IPL
***Ferroplasma acidiphilum* Y**	88	86.5	86.3	99.3	83.4
***Ferroplasma cupricumulans* BH2**	88.5	87.6	87.4	96.4	84.5
***Ferroplasma thermophilum strain*L1**	90.3	86.5	87.1	95.8	84.1
Acidiplasma aeolicum	90.7	88.9	87.4	96.4	85.4
***Picrophilus oshimae strain* KAW2**	90.9	88.4	90	92.3	84.3
***Picrophilus torridus* DSM 9790**	90.9	88.6	90	92.4	84.4
***Thermoplasma acidophilum* DSM 1728**	92.3	88.9	89.6	89.1	84.1
***Thermoplasma volcanium* GSS1**	91.7	88.7	89.4	89.2	85
Thermogymnomonas acidicola	92.6	90.9	90.7	88	84.7
***Aciduliprofundum boonei* T469**	84.1	85.8	85.5	86.2	87.3

Table S3: Percentage of shared orthologs between the AMD plasma genomes.

Percentage of genes that are shared orthologs	Aplasma	Eplasma	Fer1	Fer2	Gplasma	Iplasma
Aplasma	100%					
Eplasma	50%	100%				
Fer1	50%	49%	100%			
Fer2	48%	45%	68%	100%		
Gplasma	51%	56%	49%	45%	100%	
Iplasma	44%	47%	43%	39%	42%	100%

Table S4: Average amino acid identity of shared orthologs between the AMD plasma genomes.

% AA ID	Aplasma	Eplasma	Fer1	Fer2	Gplasma	Iplasma
Aplasma	100					
Eplasma	54	100				

Fer1	51	49	100			
Fer2	51	51	82	100		
Gplasma	56	56	49	50	100	
Iplasma	46	45	44	46	45	100

Table S5: Gene order conservation between the AMD plasma genomes. Synt/Orth indicates the number of syntenous orthologs divided by the total number of orthologs.

Synt/Orth	Aplasma	Eplasma	Fer1	Fer2	Gplasma	Iplasma
Aplasma	1					
Eplasma	0.74	1				
Fer1	0.58	0.57	1			
Fer2	0.58	0.56	0.93	1		
Gplasma	0.71	0.79	0.56	0.55	1	
Iplasma	0.44	0.44	0.38	0.39	0.42	1

Table S6: Average syntenous blocks of genes between the AMD plasma genomes. Synt Block indicates the average number of genes of syntenous blocks of genes in each pairwise comparison.

Synt Block	Aplasma	Eplasma	Fer1	Fer2	Gplasma	Iplasma
Aplasma	NA					
Eplasma	4.1	NA				
Fer1	3.1	3.2	NA			
Fer2	3.1	3	8.2	NA		
Gplasma	4.2	4.5	3.1	3.1	NA	
Iplasma	2.9	2.9	2.8	2.8	2.9	NA

Table S7: Estimate of genome completeness based on orthologous marker gene homologs. Note that genome estimates of 100% are not exact. These genomes still contain gaps between contigs.

35 Orthologous group marker genes	KEGG orthology	APL	GPL	EPL	IPL	Fer1	Fer2
COG0012 Predicted GTPase, probable translation factor	K06942	00004_0007	13459_0273	15243_190	13624_0052	1277	56_0001
COG0016 Phenylalanyl-tRNA synthetase alpha subunit	K01889	17445_0016	12302_0027	15243_601	15911_0403	1493	1_0019
COG0048 Ribosomal protein S12	K02950	5769_0003	13459_0179	15243_193	13624_0346	1276	163_0005
COG0049 Ribosomal protein S7	K02992	5769_0004	13459_0178	15243_194	13624_0347	1275	40_0045
COG0052 Ribosomal protein S2	K02967	17087_0034	13287_0131	17965_528	13606_0443	888	11_0047
COG0080 Ribosomal protein L11	K02867	17068_0016, 17068_0017 (split)	13459_0231	15243_111	13606_0475	722	31_0009
COG0081 Ribosomal protein L1	K02863	17068_0015	13459_0230	15243_112	13606_0474	721	31_0008

COG0085 DNA-directed RNA polymerase, beta subunit/140 kD subunit	K13798	17357_0065	13296_0047	15243_499	15911_0239	1179	15_0042
COG0087 Ribosomal protein L3	K02906	17357_0005	13296_0096	15243_450	15911_0332	1134	60_009
COG0088 Ribosomal protein L4	K02930	17357_0006	13296_0095	15243_451	15911_0331	1135	60_0008
COG0090 Ribosomal protein L2	K02886	17357_0008	13296_0093	15243_453	15911_0329	1137	60_0006
COG0091 Ribosomal protein L22	K02890	17357_0010	13296_0091	15243_455	15911_0327	1139	60_0004
COG0092 Ribosomal protein S3	K02982	17357_0011	13296_0090	15243_456	15911_0326	1140	15_0001
							60_0003
COG0093 Ribosomal protein L14	K02874	17357_0015	13296_0085	15243_461	15911_0321	1145	15_0006
COG0094 Ribosomal protein L5	K02931	17357_0018	13296_0082	15243_464	15911_0317 (split)	1148	15_0009
COG0096 Ribosomal protein S8	K02994	17357_0020	13296_0080	15243_465	15911_0316	1149	15_0010
COG0097 Ribosomal protein L6P/L9E	K02933	17357_0021	13296_0079	15243_466	15911_0315	1150	15_0011
COG0098 Ribosomal protein S5	K02988	17357_0025	13296_0075	15243_470	15911_0311	1154	15_0015
COG0099 Ribosomal protein S13	K02952	17357_0081	13296_0034	15243_517	13606_0273	1249	29_0045
COG0100 Ribosomal protein S11	K02948	17357_0083	13296_0032	15243_519	13606_0271	1251	29_0043
COG0102 Ribosomal protein L13	K02871	17498_0005	13296_0005	15243_533	15911_0225	701	12_0017
COG0103 Ribosomal protein S9	K02996	17498_0004	13296_0004	15243_534	15911_0226	700	12_0018
COG0124 Histidyl-tRNA synthetase	K01892	17087_0042	13287_0136	15243_316	15911_0559	883	11_0051
COG0184 Ribosomal protein S15P/S13E	K02956	17298_0054	13374_0148	17965_271	13624_0100	1231	29_0027
COG0185 Ribosomal protein S19	K02965	17357_0009	13296_0092	15243_454	15911_0328	1138	60_0005
COG0186 Ribosomal protein S17	K02961	17357_0014	13296_0086	15243_460	15911_0322	1144	15_0005
COG0197 Ribosomal protein L16/L10E	K02866	13214_0046	13374_0030	17965_211	13624_0037	1203	15_0066
COG0200 Ribosomal protein L15	K02883	17498_0006	13296_0006	15243_532	15911_0224	702	12_0016
COG0201 Preprotein translocase subunit SecY	K03076	17357_0028	13296_0072	15243_473	15911_0308	1157	15_0018
COG0256 Ribosomal protein L18	K02881	17357_0024	13296_0076	15243_469	15911_0312	1153	15_0014
COG0495 Leucyl-tRNA synthetase	K01869	17442_0015	13374_0135	17965_267	13606_0177	1297	9_0012
COG0522 Ribosomal protein S4 and related proteins	K02986	17357_0082	13296_0033	15243_518	13606_0272	1250	29_0044
COG0525 Valyl-tRNA synthetase	K01873	14887_0061	13334_0038	15243_007	13624_0071	961	24_0049

COG0533 Metal-dependent proteases with possible chaperone activity (TIGR gcp: metalloendopeptidase) (PFAM Peptidase M22, glycoprotease)	K07174	17306_0016	13455_0078	15243_429	13249_0144	1107	93_0006
COG0541 Signal recognition particle GTPase (Ffh in bacteria, SRP54 in archaea)	K03106	17481_0001	13459_0112	17965_500	13606_0269	168	18_0019
Estimated percentage of genome assembled		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table S8: Cluster of unique genes in Gplasma. PUF indicates a protein of unknown function. Bold font indicates gene numbers for proteins detected in proteomic data.

GPL gene	Annotation
13327_52	Rhodanese-like (probable phosphatase)*#
13327_53	AAA family ATPase, 1 CDC48 subfamily domain*#
13327_54	PUF
13327_55	PUF
13327_56	PUF, signal peptide, 6 transmembrane domains, P-loop containing NTP hydrolase and a Rhodanese-like domain/probable phosphatase*
13327_57	PUF
13327_58	PUF with armadillo repeat domain*
13327_59	Rhodanese-like probable phosphatase, signal peptide, 14 transmembrane helices*
13327_60	PUF

Table S9: Genes of metabolic and structural importance in the AMD plasma genomes. * indicates a putative annotation. ** indicates a probable annotation. *** indicates a possible annotation. Gray indicates additional evidence of function via synteny analysis. Bold font indicates gene numbers for proteins detected in proteomic data. "split" indicates a split gene. "Fusion" indicates a fused gene.

Enzyme or pathway name	EC number (if available)	KEGG orthology (if available)	Aplasma	Eplasma	Gplasma	Fer1	Fer2	Iplasma
S-layer protein P. torridus			17346_0011**	15243_0842**	13290_0155**		**29_0023*****	15911_0456
S-layer protein P. torridus					13296_0127**			**13606_0447**
AglB oligosaccharyltransferase			17298_0060***	17965_0276** split	**13374_0141****	1211***	85_0013** split	**13606_0191****
AglB oligosaccharyltransferase				17965_275** split			85_0014** split	
AglD oligosaccharyltransferase			17500_0046	15243_0859**	13477_0002**	307**	38_0043**	13606_0253*
AglD oligosaccharyltransferase				15243_0850**				13606_0252
ADP-L-glycero-β-D-manno-heptose biosynthesis								
D-sedoheptulose 7-phosphate isomerase	5.3.1.28	K03271	**17462_0013****	**15243_0772**		293		13624_0097**
Fused heptose 7-phosphate kinase/heptose 1-phosphate adenyltransferase	2.7.1.167	K03272	17336_0028**			0294**		13606_0256**
D,D-heptose 1,7-bisphosphate phosphatase	3.1.3.82	K03273						13606_0255**
Fused heptose 7-phosphate kinase/heptose 1-phosphate adenyltransferase	2.7.7.70	K03272	17336_0028**			0294**		13606_0256**
ADP-L-glycero-D-mannoheptose-6-epimerase	5.1.3.20	K03274		**15243_0763**		881	11_0053	13606_0259
ADP-L-glycero-D-mannoheptose-6-epimerase	5.1.3.20	K03274					38_0034	
GDP-L-fucose biosynthesis								

Mannose-1-phosphate guanyltransferase	2.7.7.13	K00966		10685_7				13606_0258
GDP-mannose 4,6-dehydratase	4.2.1.47	K01711		15243_0869				13606_0251
GDP-fucose synthase	1.1.1.271	K02377						**13606_0254****
Dolichol-sugar synthesis from dolichol and galactose								
UDP-glucose-4-epimerase	5.1.3.2	K01784	17336_0029	**15243_0765***	**13334_0006**	316	166_0012	**13606_0257**
UDP-glucose-4-epimerase	5.1.3.2	K01784	**17300_0022**					
UDP-glucose:dolichyl-phosphate glucosyltransferase	2.4.1.117	K00729						13606_0253* split
UDP-glucose:dolichyl-phosphate glucosyltransferase	2.4.1.117	K00729	13083_0006***					13606_0252 split
Dolichol-phosphate-mannosyltransferase	2.4.1.83	K00721		15243_0773	13477_0015	1949	38_0043**	13624_0181
dTDP-L-rhamnose biosynthesis								
dTDP-glucose pyrophosphorylase	2.7.7.24	K00973	17487_0015	10685_0003	**13459_0056**	1942		**13624_0077**
dTDP-glucose pyrophosphorylase	2.7.7.24	K00973			13459_0020			
dTDP-glucose 4,6-dehydratase	4.2.1.46	K01710		15243_0770		1944		
dTDP-4-dehydrorhamnose 3,5-epimerase	5.1.3.13	K01790		10685_0005		1945		13606_0243
dTDP-4-dehydrorhamnose 3,5-epimerase	5.1.3.13	K01790		10685_0009				
dTDP-4-dehydrorhamnose reductase	1.1.1.133	K00067		10685_0004	**13459_0055**	1943		
Electron Transport Chain								
NADH dehydrogenase								
NuoA	1.6.5.3	K00330	17428_0014	15243_400	**13455_0039**	1639	1_0005	13606_0102
NuoB	1.6.5.3	K00331	17428_0015	15243_401	**13455_0040**	1640	1_0004	**13606_0100**
NuoC	1.6.5.3	K00332	17428_0016	15243_402	**13455_0041**	1641	1_0003	13606_0101

NuoD	1.6.5.3	K00333	**17428_0017**	15243_403	**13455_0042**	1642	1_0001	**13606_0099**
NuoE	1.6.5.3	K00334						
NuoF	1.6.5.3	K00335	17306_0015					
NuoG	1.6.5.3	K00336						
NuoH	1.6.5.3	K00337	17428_0018	15243_404	13455_0043	1643	1_0002	13606_0098
NuoI	1.6.5.3	K00338	**17428_0019**	15243_405	**13455_0044**	1644		13606_0097
NuoJ	1.6.5.3	K00339	17428_0020	15243_406	13455_0045	1645		13606_0095
NuoJ	1.6.5.3	K00339	17428_0021	15243_407	13455_0046	1646		13606_0096
NuoK	1.6.5.3	K00340	17428_0022	15243_408	13455_0047	1647		13606_0094
NuoL	1.6.5.3	K00341	17428_0023	15243_409	13455_0048, 13455_0050	1648		**13606_0093**
NuoM	1.6.5.3	K00342	**17428_0024**	15243_410	13455_0051	1649		13606_0092
NuoN	1.6.5.3	K00343	17428_0025	15243_411	13455_0052	1650	19_0001	13606_0091
Succinate dehydrogenase								
Succinate dehydrogenase	1.3.99.1	K00239	17500_0036	17965_344	**13287_0168**	**1743**	**19_0043**	13606_0106, 13606_0110
SdhA	1.3.99.1	K00240	17500_0037	17965_345	**13287_0167**	**1742**	19_0042	13606_0103, 13606_0109
SdhC	1.3.99.1	K00241	17500_0038	17965_346**	**13287_0166****	**1741****	19_0041*	13606_0105
SdhD	1.3.99.1	K00242	17500_0039	**15243_0016*****	**13334_0105*****	1740**	**19_0040****	13606_0104
Cytochrome bc complex								
CytB/PetB	1.10.2.2	K00412	17325_0015	17965_146*	**13374_0019**	**1373***	104_0011*	13606_0323

CytB/PetB	1.10.2.2	K00412	17442_0026					**13606_0418**
CytB/PetB	1.10.2.2	K00412	17298_0104**					
Rieske/PetA	1.10.2.2	K00411	17442_0024**	17965_147**	**13374_0018**	1374*	**104_0012****	13606_0324
Rieske/PetA	1.10.2.2	K00411						13606_0419
Iron oxidation proteins								
Rusticyanin	1.16.9.1		00084_0003		13459_0280*			
SoxE sulfocyanin						1129	60_0014	
Cytochrome C oxidase								
CoxAC (catalytic subunit)	1.9.3.1	K02274	13214_0024	15243_561	**13374_0156**	1127	**60_0016**	13606_0425
CoxB	1.9.3.1	K02275	13214_0025	15243_560	13374_0155	1130	60_0013	13606_0426
Cytochrome C oxidase/CoxC	1.9.3.1	K02276						
Cytochrome C oxidase/CoxD	1.9.3.1	K02277						
Cupredoxins				15243_562	13374_0157			13606_0424
Cytochrome bd complex								
CydA	1.10.3.-	K00425	12042_0002		**13287_0064***	671	20_0021	13624_0232
CydA	1.10.3.-	K00425		311313_0002	13287_0111*	1031	12_0062	**13624_0233**
CydB	1.10.3.-	K00426	12042_0001***	311313_0003	13287_0063***	1032***	20_0020***	731028_0012
CydB	1.10.3.-	K00426			13287_0110***	672***	12_0061***	
ATPase A-type								
NtpD/AhaD	3.6.3.14	K02120	17300_0014	15833_14	**13334_0012**	934	97_0003	**13606_0199**
NtpB/AhaB	3.6.3.14	K02118	**17300_0012**	15833_15	**13334_0013**	935	**97_0004**	**13606_0200**
NtpB/AhaB	3.6.3.14	K02118	**17300_0013**					
NtpA/AtpA	3.6.3.14	K02117	**17300_0011**	15833_16	**13334_0014**	**936**	**97_0005****	**13606_0201**

NtpG/AhaF	3.6.3.14	K02122	**17300_0010**	15833_17	**13334_0015**	937	97_0006	13606_0202
AhaC/NtpC	3.6.3.14	K02119	17300_0009	15833_18	13334_0016, 13334_0018	938	**97_0007**	13606_0203
NtpE/AhaE	3.6.3.14	K02121	**17300_0008**	15833_19	**13334_0017**	**932****, 939*	**97_0008**	13606_0204
NtpI/AhaI	3.6.3.14	K02123	**17300_0017**	15833_12	**13334_0010**	931	**97_0001**	**13606_0196**
AtpE/H/AhaK	3.6.3.14	K02110	17300_0007***	15833_20	13334_0018**	940	97_0009	13606_0205
AhaG	3.6.3.14							
AhaH	3.6.3.14		**17300_0016****	15833_13***		932	97_0002	
Aerobic CODH complex								
Large/catalytic subunit/CoxL	1.2.99.2	K03520				**1676**	**10_0056**	**15911_0062**
Large/catalytic subunit/CoxL	1.2.99.2	K03520					**64_0005**	
Small subunit/CoxS	1.2.99.2	K03518				**1675**	10_0055	15911_0063
Medium subunit/CoxM	1.2.99.2	K03519	00387_0001			**1674**	**10_0054**	**15911_0064**
Medium subunit/CoxM	1.2.99.2	K03519				1769		
Xanthine and CO dehydrogenase maturation factor/CoxF		K07402	00387_0002			1752	19_0052	15911_0059
Accessory protein/CoxG		K09386	13214_0018			467	**138_0007**	15911_0060
Accessory protein/CoxG		K09386	17518_0012			**288**		
Anaerobic CODH complex								
Beta/catalytic subunit/CooS	1.2.99.2C	K00198					**31_0047**	
Maturation factor/CooC		K07321					**31_0046*****	
Entner Doudoroff pathway								
Glucose dehydrogenase	1.1.1.47	K00034	**17462_0066****	15243_607**	**12303_0024***	**109****	107_0012**	13624_0024*

Glucose dehydrogenase	1.1.1.47	K00034		15243_608**		298**		
Gluconate dehydratase	4.2.1.39	K05308	**14887_0067****	15251_40*	**13334_0053****	**930****	**11_0004**	**13624_0026***
KDG aldolase	4.1.2.-	K11395	00419_0007**	17965_249***	**13374_0117****	**1727****	19_0027	13624_0195**
GADH/aldehyde dehydrogenase	1.2.1.3		**12840_0002****	15243_377***	**13374_0164****	**706**	**12_0012*****	**13624_0267**
Glycerate kinase	2.7.1.165	K11529	17336_0043	17965_54	**13477_0025**	1867**	**10_0060****	15911_0118*
Glycerate kinase	2.7.1.165	K11529	17336_0044***+					
Enolase	4.2.1.11	K01689	17500_0031	17965_47	**13374_0178**	**1336**	152_0007	**13606_0442**
Enolase	4.2.1.11	K01689					**50_0011**	
Pyruvate kinase	2.7.1.40	K00873	17498_0009	15243_541	**13296_0009**	707	**12_0011**	459686_0003
Glycolysis								
Glucokinase/glk	2.7.1.2	K00845	17498_0018	17965_155*	**13287_0076****	1266**	**40_0033**	**13624_0027***
Glucose-6-phosphate isomerase (archaeal bifunctional enzyme)	5.3.1.9	K06859	**13214_0016**	15243_568	**13287_0025**	**1583**	**1_0124**	13624_0322
Fructose 1,6-bisphosphate aldolase/phosphatase	4.2.1.13	K01622					322_0003	
Fructose 1,6-bisphosphate aldolase/phosphatase (archaeal type)/fba	4.2.1.13	K01624	13214_0029	15243_586	**13374_0081**	**1320**	50_0014**	**15911_0401**
Fructose 1,6-bisphosphate aldolase/phosphatase (archaeal type)/fba	4.2.1.13		13214_0030				322_0003	
Triosephosphate isomerase (TIM)/tpiA	5.3.1.1	K01803	AP2_3650_0002	15243_442	**13296_0103**	**1082**	47_0036	13606_0341
Glyceraldehyde-3-phosphate dehydrogenase (NAD(P))/gap2	1.2.1.59	K00150	17087_0075	17965_400	**13459_0136**	1571	1_0111	**15911_0284**
Phosphoglycerate kinase /pgk	2.7.2.3	K00927	17487_0016	15243_230	13459_0014	**1387**	104_0026	**15911_0422**
Phosphoglycerate mutase/gpm	5.4.2.1	K01834	**17298_0091**	17965_440	13290_0025	**1458**	**5_0046**	13606_0398

Phosphoglycerate mutase/gpm	5.4.2.1	K01834	17336_0054**		**13459_0073**			
Enolase /eno	4.2.1.11	K01689	17500_0031	17965_47	**13374_0178**	**1336**	152_0007	**13606_0442**
Pyruvate kinase /pyk	2.7.1.40	K00873	17498_0009	15243_541	**13296_0009**	707	**12_0011**	459686_0003
Pyruvate dehydrogenase complex								
Pyruvate dehydrogenase E1 component subunit alpha/pdhA	1.2.4.1	K00161	**17068_0073**	15833_38	**13459_0211**	**866**	11_0062	**13606_0032**
Pyruvate dehydrogenase E1 component subunit beta/pdhB	1.2.4.1	K00162	**17068_0074**	15833_37	**13459_0212**	**867**	**11_0061***	**13606_0031**
Pyruvate dehydrogenase E1 component subunit beta/pdhB	1.2.4.1	K00162				**371**		
Pyruvate dehydrogenase E2 component/pdhC	2.3.1.12	K00627	**17068_0075***	15833_36*	**13459_0213**	868**	**11_0060**	13606_0030
Pyruvate dehydrogenase E3/dihydrolipoamide dehydrogenase	1.8.1.4	K00382	**17068_0076**	15833_35	**13459_0214**	**869**	11_0059	**13606_0029**
TCA cycle								
Citrate synthase	2.3.3.1	K01647	17433_0013	15243_18	**13334_0108**	**185**	**89_0012**	**13624_0118**
Citrate synthase	2.3.3.1	K01647		15243_526	**13459_0255**	**249**		**15911_0546**
Aconitase	4.2.1.3	K01681	**14887_0066**	15251_41	**13459_0162**	**1728**	**19_0023**	**13606_0363**
Isocitrate dehydrogenase	1.1.1.42	K00031	17087_0014	15833_57	**13459_0219**	**368**	31_0020	**15911_0558**
2-oxoacid:ferredoxin oxidoreductase a/g-subunit	1.2.7.3	K00174	17112_0056	15243_130	**13374_0086**	**1588**	**1_0130**	**13606_0265**
2-oxoacid:ferredoxin oxidoreductase a/g-subunit	1.2.7.3	K00174		17965_194	**13459_0183***	**1738**		**15911_0268**
2-oxoacid:ferredoxin oxidoreductase b-subunit	1.2.7.3	K00175	17112_0057	15243_129	**13374_0087**	**1589**	**1_0131**	**13606_0264**
2-oxoacid:ferredoxin oxidoreductase b-subunit	1.2.7.3	K00175		17965_193	**13459_0182**	**1739**		**15911_0267**
Succinyl-CoA synthetase alpha subunit /sucD	6.2.1.5	K01902	17433_0052	17965_291	**13374_0106**	**1736**	19_0036	**13249_0151**

Succinyl-CoA synthetase alpha subunit /sucD	6.2.1.5	K01902						**13249_0152**
Succinyl-CoA synthetase beta subunit	6.2.1.5	K01903	17433_0051	17965_292	**13374_0105**	**1735**	**19_0035**	**13249_0153**
Succinate dehydrogenase flavoprotein subunit	1.3.99.1	K00239	17500_0036	17965_344	**13287_0168**	**1743**	**19_0043**	**13606_0106**
Succinate dehydrogenase flavoprotein subunit	1.3.99.1	K00239						**13606_0110**
Succinate dehydrogenase iron-sulfur protein	1.3.99.1	K00240	17500_0037	17965_345	**13287_0167**	**1742**	19_0042	**13606_0103**
Succinate dehydrogenase iron-sulfur protein	1.3.99.1	K00240						**13606_0109**
Succinate dehydrogenase (ubiquinone) cytochrome b subunit	1.3.5.1	K00241	**17500_0038**	17965_346**	**13287_0166****	**1741****	19_0041*	13606_0105**
Succinate dehydrogenase cytochrome b small subunit	1.3.5.1	K00242	17500_0039	**15243_0016*****	**13334_0105*****	1740**	**19_0040****	13606_0104**
Fumarate hydratase	4.2.1.2	K01679	17112_0058	15243_131	**13334_0165**	**734**	**31_0019**	13249_0077*
Malate dehydrogenase	1.1.1.37	K00024	17428_0038	17965_423	**13459_0040**	**1744**	**19_0044**	**13624_0119**
Glyoxylate shunt enzymes not found in the TCA cycle								
Malate synthase/aceB	2.3.3.9	K01638		15243_643				
Isocitrate lyase/aceA	4.1.3.1	K01637		15243_642				
Carbohydrate and sugar catabolism								
Alpha-amylase	3.2.1.1	K07405	17068_0053			1351	50_0027	15911_0567
Glucoamylase	3.2.1.3	K01178	17481_0012	17965_50	**13327_0014**	103	1_0037	13624_0127
Glucoamylase	3.2.1.3	K01178					12_0053	
Glucan 1,4-alpha-glucosidase	3.2.1.3	K01178	13083_0001	17965_532	**13287_0155**			
Glucan 1,4-alpha-glucosidase	3.2.1.3	K01178	17481_0011					
Glycoside hydrolase 15-related	3.2.1.3		13083_0002**					
Alpha-glucosidase	3.2.1.20	K01187		15243_200	**13455_0099**			

Alpha-glucosidase	3.2.1.20	K01187			**13374_0167**			
Beta-galactosidase	3.2.1.23	K01190			13290_0163* split	136	20_0056	
Beta-galactosidase	3.2.1.23	K01190			13290_0164 split			
Beta-galactosidase	3.2.1.23	K01190			**13455_0091**	1310		
Glucan 1,4-α-maltohydrolase	3.2.1.133	K05992			**13455_0098*****			
GlgX: glycogen debranching enzyme isoform 1	2.4.1.25, 3.2.1.33	K02438	17068_0051	17965_317	13459_0253	1349	split 50_0023	**15911_0569**
GlgX: glycogen debranching enzyme isoform 1	2.4.1.25, 3.2.1.33	K02438					split 50_0024	
GlgX: glycogen debranching enzyme isoform 1	2.4.1.25, 3.2.1.33	K02438					split 50_0025	
GlgA: glycogen synthase	2.4.1.21	K00703/K16148	**17068_0052**		13374_0215***	1350	**50_0026**	15911_0568
Oligosaccharide amylase						1348	**50_0022**	15911_0570
Beta oxidation								
Fatty acyl-CoA synthetase/FadD	6.2.1.3	K01897	**17298_0042**	**15243_0438**	**13477_0036**	**250**	19_0031	**13606_0347**
Fatty acyl-CoA synthetase/FadD	6.2.1.3	K01897	00419_0016	**15243_0203**	13374_0237	75	**1_0063**	**15911_0156**
Fatty acyl-CoA synthetase/FadD	6.2.1.3	K01897	17082_0014	**15243_0207**	**13459_0328**	376	**10_0030**	
Fatty acyl-CoA synthetase/FadD	6.2.1.3	K01897			**13334_0078**	1724	17_0005	
Fatty acyl-CoA synthetase/FadD	6.2.1.3	K01897			**13290_0075**			
Fatty acyl-CoA synthetase/FadD	6.2.1.3	K01897			**13334_0033**			
Acyl-CoA dehydrogenase/FadE	1.3.99.-		**17393_0003**	**15243_0396**	**13455_0035**	1636	1_0009	**13624_0325**
Acyl-CoA dehydrogenase/FadE	1.3.99.-		**17428_0009**			661	**103_0022**	
Acyl-CoA dehydrogenase/FadE	1.3.99.-		**17068_0062**					

Delta-3-cis-delta-2-trans-enoyl-CoA isomerase/FadB	4.2.1.17		13186_0039	**15243_0060**	**13459_0259**	**1530**	**1_0062**	**13606_0355**
Delta-3-cis-delta-2-trans-enoyl-CoA isomerase/FadB	4.2.1.17		17500_0016	17965_0036	**13287_0080**	**1784**	**64_0015**	**13249_0160**
Delta-3-cis-delta-2-trans-enoyl-CoA isomerase/FadB	4.2.1.17		**17068_0085**		**13290_0086**	1758		
Delta-3-cis-delta-2-trans-enoyl-CoA isomerase/FadB	4.2.1.17				**13290_0087**			
Enoyl-CoA hydratase/PaaZ	4.2.1.17		17487_0025	**15243_0377**	**13374_0164**	**706**	**247_0005**	**13624_0267**
Enoyl-CoA hydratase/PaaZ	4.2.1.17		**12840_0002**	**15243_0206**	**13287_0036**	1560	**12_0012**	
Enoyl-CoA hydratase/PaaZ	4.2.1.17		**17462_0060**					
3-ketoacyl-CoA thiolase/FadA	2.3.1.16	K00632	**17068_0064**	**15833_0048**	**13334_0182**	326	38_0026	**13606_0350**
3-ketoacyl-CoA thiolase/FadA	2.3.1.16	K00632	00123_0001	**17965_0179**	**13290_0085**	**1532**	1_0064	**13624_0269**
3-ketoacyl-CoA thiolase/FadA	2.3.1.16	K00632	**14887_0034**	17965_0037	**13334_0127**	**323**	**38_0027**	**13249_0161**
3-ketoacyl-CoA thiolase/FadA	2.3.1.16	K00632	13186_0040					
Methylotrophy								
Methanol:N,N-dimethyl-4-nitrosoaniline oxidoreductase	1.1.99.37		17518_0009**	**17965_0490****	**13290_0028***	158**	**18_0013****	**15911_0146****
Methanol:N,N-dimethyl-4-nitrosoaniline oxidoreductase	1.1.99.37					200**	39_0027**	
NADP-dependent methylene tetrahydromethanopterin dehydrogenase/Methenyl tetrahydrofolate cyclohydrolase	1.5.1.5/3.5.4.9	K01491	17298_0025	**15243_0609**	**12303_0025****	297	107_0011	15911_0223
Formate-tetrahydrofolate ligase	6.3.4.3	K01938	**17068_0040**	**17965_0511**	**13477_0075**	**653**	103_0010	**15911_0504**
Formate-tetrahydrofolate ligase	6.3.4.3	K01938			13477_0076 split			
Pyruvate fermentation								
Pyruvate ferredoxin oxidoreductase/PorA	1.2.7.1	K00169	17112_0057	15243_130	**13459_0183**	**1588**	**1_0130**	**13606_0171**
Pyruvate ferredoxin oxidoreductase/PorA	1.2.7.1	K00169		17965_194				

Pyruvate ferredoxin oxidoreductase/PorB	1.2.7.1	K00170	17112_0056	15243_129	**13459_0182**	**1589**	**1_0131**	**13606_0170**
Pyruvate ferredoxin oxidoreductase/PorB	1.2.7.1	K00170		17965_193				
Pyruvate ferredoxin oxidoreductase/PorD	1.2.7.1	K00171						13606_0172
Pyruvate ferredoxin oxidoreductase/PorG	1.2.7.1	K00172						13606_0173
Acetyl-CoA synthetase/AcdA	6.2.1.13	K01905	**17300_0003**	15243_443	**13459_0208**	**837**	90_0003	**15911_0417**
Acetyl-CoA synthetase/AcdB	6.2.1.13	K01905	17300_0002**	15243_444**	**13459_0208**** fusion	**837**** fusion	90_0003** fusion	15911_0416**
Copper resistance								
Cu2+-exporting ATPase/CopA	3.6.3.4	K01533			**13459_0335**	**1516**	104_0018	
Cu2+-exporting ATPase/CopA	3.6.3.4	K01533					1_0042	
MarR				15243_166				
Copper resistance protein/CopD				15243_201	13459_0176	1274	40_0042	**15911_0138**
Copper resistance protein/CopC								
Heavy metal transport associated protein/CopZ			17466_0019			1379	fer2_104_0017	13606_0461
CopY			17466_0020		**13290_0168**	1378	fer2_104_0016	13606_0462
CopB			00084_0009	15243_167	13290_0169	**1380**	fer2_104_0018	**13606_0460**
CopB			17466_0021	15243_168				
CopB			17466_0022					
Mercury resistance								
Hg reductase/MerA	1.16.1.1	K00520	10741_0002	15243_51	13287_0073	1595, 1909		15911_0564
Hg reductase/MerA	1.16.1.1	K00520	10741_0003					
Hg2+ binding protein/MerP		K08364	10741_0001**	15243_52	**13287_0074****	1596		15911_0565*

Mercury resistance regulatory protein/MerR								
Arsenate resistance								
Arsenite transporter ATPase/ArsA	3.6.3.16	K01551	17112_0009		**13287_0034**	624	29_0052	13624_0235
Arsenite transporter ATPase/ArsA	3.6.3.16	K01551						**13606_0505**
Transcriptional repressor/ArsR		K03892/K07721			**13334_0172**	0909**	**11_0023**	**15911_0300**
Arsenite transporter/ArsB	3.6.3.16	K03893	17466_0012	17965_494	13334_0174	0910*	11_0022	15911_0301
Arsenate reductase/ArsC	1.20.4.1	K00537/K03741		17965_495	13334_0173**			
Histidine synthesis								
HisG	2.4.2.17	K00765	17082_0045		13475_0072	**531**	9_0035	
HisI	3.6.1.31/3.5.4.19	K11755	17082_0039			525	9_0042	
HisA	5.3.1.16	K01814	17082_0041			527	9_0040*	
HisF	2.4.2.-	K02500	17082_0040			**526**	9_0041	
HisH	2.4.2.-	K02501	17082_0042**			528	9_0038** split	
HisH	2.4.2.-	K02501					9_0039** split	
HisB	4.2.1.19	K01693	17082_0043			529	9_0037	
HisC	2.6.1.9	K00817	17082_0038	15833_10		**469**	**138_0005**	
	3.1.3.15	K01089						
HisD	1.1.1.23	K00013	17082_0044	17965_171		530	9_0036	
Valine/(Iso)Leucine synthesis								
3-isopropylmalate/(R)-2-methylmalate dehydratase large subunit	4.2.1.35/4.2.1.33	K01703	17082_0054		**13455_0074**	**950**	11_0026	
3-isopropylmalate/(R)-2-methylmalate dehydratase large subunit	4.2.1.35/4.2.1.33	K01703	17445_0006		13475_0088	**903**	**11_0027**	

3-isopropylmalate/(R)-2-methylmalate dehydratase large subunit	4.2.1.35/4.2.1.33	K01703			13475_0089		**205_0009**	
3-isopropylmalate/(R)-2-methylmalate dehydratase small subunit	4.2.1.35/4.2.1.33	K01704	17082_0053		13475_0087	904	**205_0008**	
3-isopropylmalate/(R)-2-methylmalate dehydratase small subunit	4.2.1.35/4.2.1.33	K01704	17445_0005		**13455_0073**			
3-isopropylmalate dehydrogenase	1.1.1.85	K00052	17082_0052		13447_0027	**952**	56_0013	
3-isopropylmalate dehydrogenase small subunit	1.1.1.85	K00052			13475_0086	951	**205_0007**	
Threonine dehydratase	4.3.1.19	K01754	17433_0010		**13459_0223**	668	18_0005	**15911_0095**
Threonine dehydratase	4.3.1.19	K01754						**15911_0503**
Acetolactate synthase I/II/III large subunit	2.2.1.6	K01652	17087_0069*	**15243_0744**	**13459_0095**	1562	**38_0021**	**13624_0240**
Acetolactate synthase I/II/III large subunit	2.2.1.6	K01652				1781	38_0022	
Acetolactate synthase I/II/III large subunit	2.2.1.6	K01652				**331**		
Ketol-acid reductoisomerase	1.1.1.86	K00053				333	**38_0019**	
Dihydroy-acid dehydratase	4.2.1.9	K01687	00457_0002			**330**	**38_0023**	
Dihydroy-acid dehydratase	4.2.1.9	K01687					**38_0024**	
Pyruvate dehydrogenase	see above							
2-isopropylmalate synthase	2.3.3.13	K01649	17082_0051			911	205_0010**	
2-isopropylmalate synthase	2.3.3.13	K01649	17445_0004		**13455_0071**	949		
Branched-chain amino acid aminotransferase	2.6.1.42	K00826	17082_0015	15833_40	**13334_0132**			
Branched-chain amino acid aminotransferase	2.6.1.42	K00826			13475_0034			
Trehalose synthesis								
Isoamylase	3.2.1.68	K01214	**17462_0041*****	17965_317***	13459_0253			

Maltooligosyl-trehalose synthase	5.4.99.15	K06044	17462_0042*	12303_035				
Maltooligosyl-trehalose synthase	5.4.99.15	K06044	17462_0043		12303_0035			
Maltooligosyl-trehalose trehalohydrolase	3.2.1.141	K01236	**17462_0044**	17965_321***	**12303_0036**			
Maltooligosyl-trehalose trehalohydrolase	3.2.1.141	K01236	**17462_0035**					
Trehalose synthase	5.4.99.16	K05343	17462_0038 split	17965_318 split	13334_0119 split	56**	1_0074**	
Trehalose synthase	5.4.99.16	K05343	17462_0039 split	17965_319 split	13334_0120 split			
Flagellar proteins								
Archaeal flagellin/FlaB		K07325	17298_0067**	17965_280				
Archaeal flagellar protein/FlaC		K07822	**17298_0068**	17965_281				
Archaeal flagellar protein/FlaDE		K07327/K07328	**17298_0069**	17965_282**				
Archaeal flagellar protein/FlaF		K07329	17298_0070**	17965_283**				
Archaeal flagellar protein/FlaG		K07330	17298_0071**	17965_284**				
Archaeal flagellar protein/FlaH		K07331	17298_0072**	17965_285**				
Archaeal flagellar protein/FlaI		K07332	17298_0074**	17965_286				
Archaeal flagellar protein/FlaJ		K07333	17298_0075**	17965_287**				
Archaeal flagellin/FlaB		K07325		17965_24				

Table S10: Blue-copper protein motifs found in AMD plasma genes.

Protein	Motif	Aplasma	Gplasma	Fer1	Fer2
Rusticyanin	[G][M][YF][G][K][I][V][V]	GMYGRIVV	GMYGKISV	GMYAFVIV	GMYVFVIV
Sulfocyanin	[C][G][I][AL][G][H][A][VAQ][SA][G][M][W]			CGLTTHAEAGMY	CGLTTHAEAGMY
Sulfocyanin	FNFNGTS			FNFNGTS	FNFNGTS

Table S11: Amino acid identity of AMD plasma cytochrome c oxidase subunit II genes with closely related genes.

Plasma gene	Related gene	Amino acid identity (%)
fer1_1130	aa3_cyt_C_oxidase_subunit_II_Aeropyrum_pernix	38.89
fer2_60_0013	aa3_cyt_C_oxidase_subunit_II_Aeropyrum_pernix	39.29
EPL_15243_0560	aa3_cyt_C_oxidase_subunit_II_Sulfolobus_acidocaldarius	34.31
GPL_13374_0155	aa3_cyt_C_oxidase_subunit_II_Sulfolobus_acidocaldarius	35.35
APL_13214_0025	aa3_cyt_C_oxidase_subunit_II_Aeropyrum_pernix	36.54
IPL_13606_0436	aa3_cyt_C_oxidase_subunit_II_Sulfolobus_acidocaldarius	48.7

Table S12: AMD plasma gene homologs to genes overexpressed or overtranscribed under anaerobic conditions in *T. volcanium* and *T. acidophilum*. Bold font indicates gene numbers for proteins detected in proteomic data.

T. volcanium gene	*T. acidophilum* gene	APL	EPL	GPL	Fer1	Fer2	IPL
TVG1406112	Ta0773	17112_0057	**15243_0130**	**13459_0183**	**1588**	**1_0130**	**13606_0265**
TVG1405258	Ta0260	17112_0056	**15243_0129**	**13459_0182**	**1589**	**1_0131**	**15911_0267**
TVG0773051	Ta0635	00325_0005	**17965_0324**	**13327_0008**	**1453**	**5_0059**	**13249_0140**
TVG0774569	Ta0635		**17965_0324**	**13327_0008**	**1453**	**5_0059**	**13249_0140**
TVG1348694	Ta0832	17518_0009	**17965_0490**	**13290_0028**	158	464_0002	**15911_0146**
TVG0004359	Ta0046						13606_0515
TVG0005131	Ta0414		**17965_0233**	**12302_0038**	336	38_0016	13606_0514
TVG1211894	Ta0414	**17298_0094**	**17965_0437**	**13459_0089**			**15911_0460**
TVG1225518	Ta0403	17365_0054	**17965_0233**	**12302_0038**	336	38_0016	13249_0164
TVG1343229	Ta0327	17306_0008	**15243_0434**	13455_0086	244	89_0017	13606_0021
TVG0881751	Ta0773	13214_0036	**17965_0194**	**13374_0086**	**1738**	**19_0038**	**15911_0268**
TVG0882664	Ta0772	13214_0037	**17965_0193**	**13374_0087**	**1739**	19_0039	**15911_0267**
TVG0870807							13606_0172
TVG0870514	Ta0626						13606_0173
TVG0105141		**17068_0074**	**15833_0037**	**13459_0212**	**867**	**11_0061**	**13606_0031**
TVG0103933		**17068_0075**	15833_0036	**13459_0213**	868	**11_0060**	13606_0030
TVG1458424	Ta1435	**17068_0076**	15833_0035	**13459_0343**	**1839**	11_0059	**13606_0029**
TVG1508405		13214_0006	15243_0056	13334_0083	**1329**	97_0019	13624_0133
TVG0102621	Ta1435	**17068_0076**	15833_0035	**13459_0214**	**869**	11_0059	**13606_0029**
TVG0808604	Ta0810	17112_0007					**15911_0480**

TVG1304136	Ta1142	17112_0002					**13606_0513**
TVG0563624	Ta1012	00011_0003	**15243_0626**	**13459_0331**	1328	15_0060	**13606_0169**
TVG0563052	Ta1013	00011_0004	**15243_0625**	**13459_0332**	1327	15_0061	13606_0168
TVG0003442	Ta0045m	17306_0014			1916	17_0055	13606_0516
TVG1345079	Ta0329	17306_0006	**15243_0436**	**13455_0088**	**246**	**89_0015**	**13606_0023**
TVG1344033	Ta0328	17306_0007	**15243_0435**	**13455_0087**	**246**	**89_0015**	**13606_0022**
	Ta0626						13606_0173
	Ta0629	17428_0003					**13606_0170**
	Ta1438	**17068_0073**	**15833_0038**	**13459_0211**	**866**	756_0002	**13606_0032**
	Ta0326	17306_0010	**15243_0433**	13455_0085	**766**	**31_0042**	13606_0020
	Ta0776	00325_0005	**17965_0324**	**13327_0008**	**1453**	**5_0059**	**13249_0140**
	Ta0047		**17965_0233**	**12302_0038**	336	38_0016	13606_0514
	Ta0837	17433_0004	**15243_0537**	**13327_0064**	306	38_0044	15911_0564

Table S13: Pili genes in the AMD plasmas. * indicates a putative annotation. ** indicates a probable annotation. *** indicates a possible annotation. Gray indicates additional evidence of function via synteny analysis. "split" indicates a split gene. Bold font indicates gene numbers for proteins detected in proteomic data.

Pili/Type II secretion	Annotation	APL	EPL	GPL	IPL
ATPase	Type II secretion system protein E/ heli_sec_ATPase[TIGR03819], helicase/secretion neighborhood ATPase	17298_0019*	**15243_0612**	**12303_0028**	15911_0076
Inner membrane scaffold-like	GSPII_F[pfam00482], Bacterial type II secretion system protein F domain	17298_0020** split	**15243_0613**	**12303_0029****	15911_0077*
	GSPII_F[pfam00482], Bacterial type II secretion system protein F domain	17298_0021** split	**15243_0614****	**12303_0030****	15911_0078*
ATPase	Type II secretion system protein E/ heli_sec_ATPase[TIGR03819], helicase/secretion neighborhood ATPase	17298_0022*	**15243_0615****	**12303_0031****	**15911_0079***
DEAH-box helicase-like		**17298_0087****	15243_639**		**15911_0042*****

Figure S1: Ribosomal protein S15 tree of the AMD plasma archaea and their close relatives.

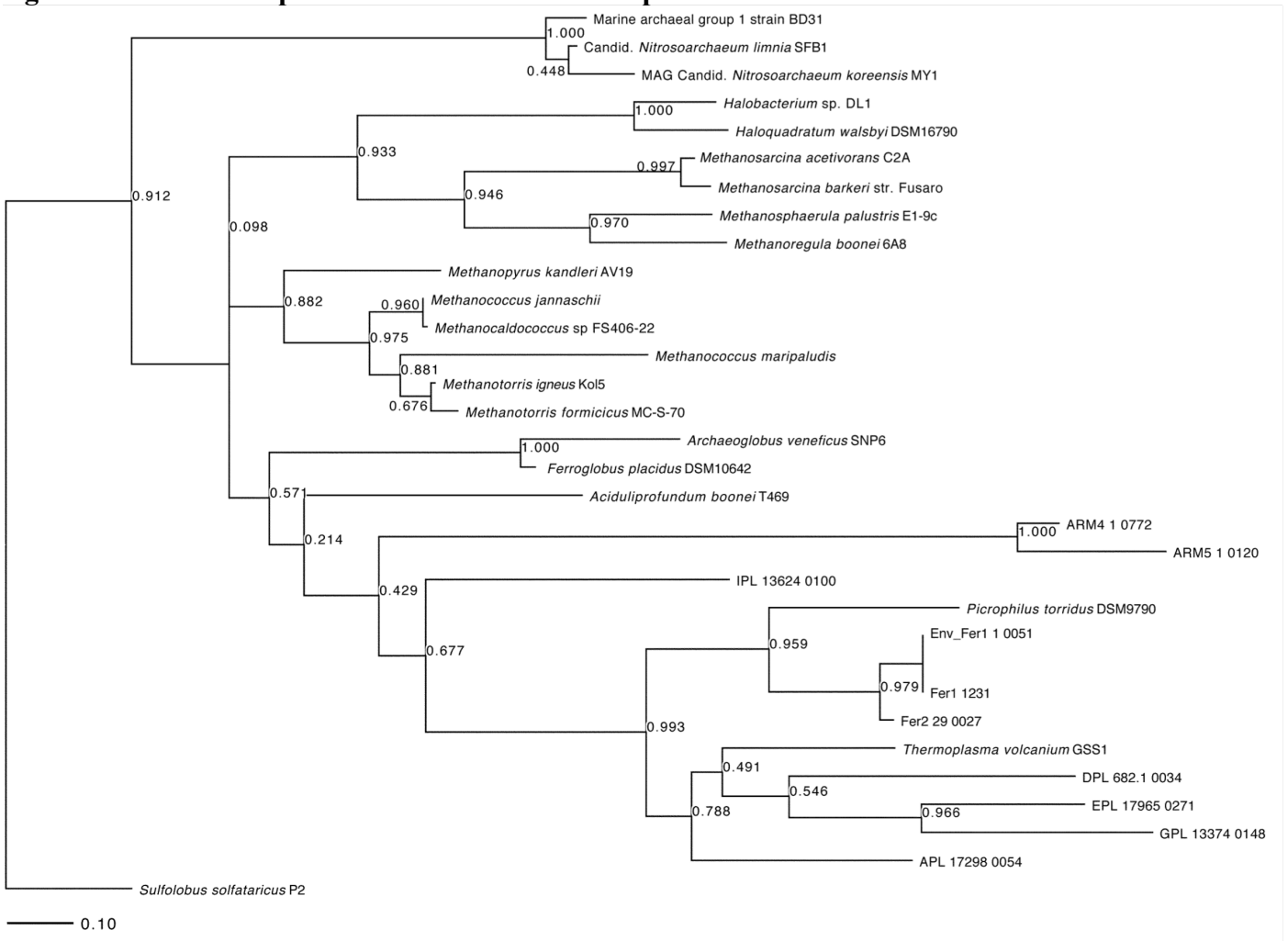


Figure S2: Structural alignment of blue copper proteins. β-Strands (cupredoxin fold) predicted by YASPIN [146] are highlighted (cyan for β-strand 1, yellow and light green for β-strand 2, pink for β-strand 3, dark blue for β-strand 4, dark green for β-strand 5, purple for β-strand 6 and red for β-strand 7). Amicyanin from *Paracoccus denitrificans* (PDB:1AAC) and Plastocyanin from *Synechococcus elongatus* (PCC 7942) serve as references. Red circles indicate copper-binding ligands. Residues highlighted by light grey correspond to additional β-strands and those in bold orange correspond to α-helices. Sulfocyanin-specific motifs are boxed in red. Black arrows indicate copper-binding ligands. Additional loops are indicated at the bottom of the alignment by a light orange line.

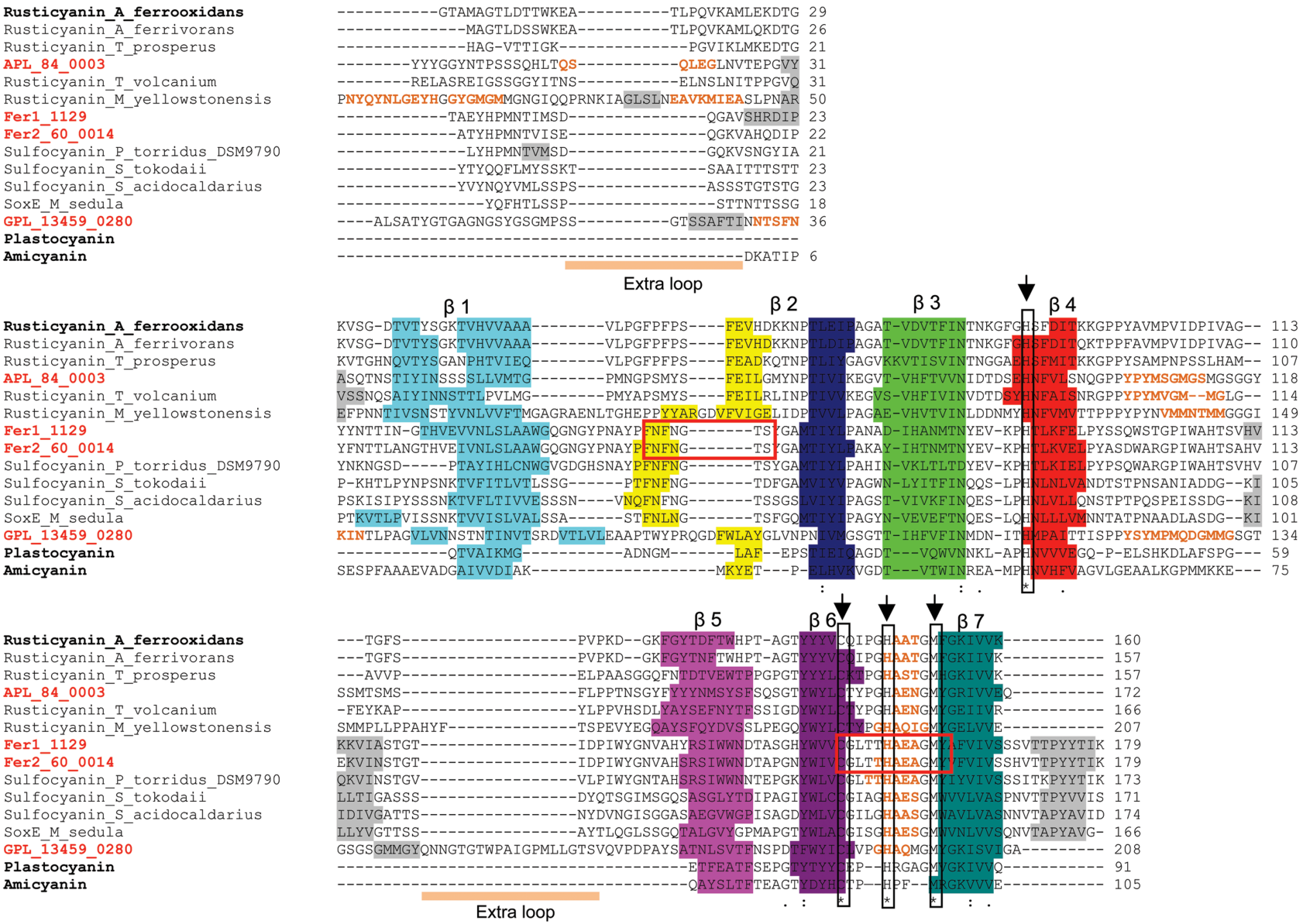

β 1
β 2
β 3
β 4
β 5
β 6
β 7
Extra loop
Extra loop

Figure S3: AMD plasma blue-copper protein tree. bcp indicates a blue-copper protein of unknown function.

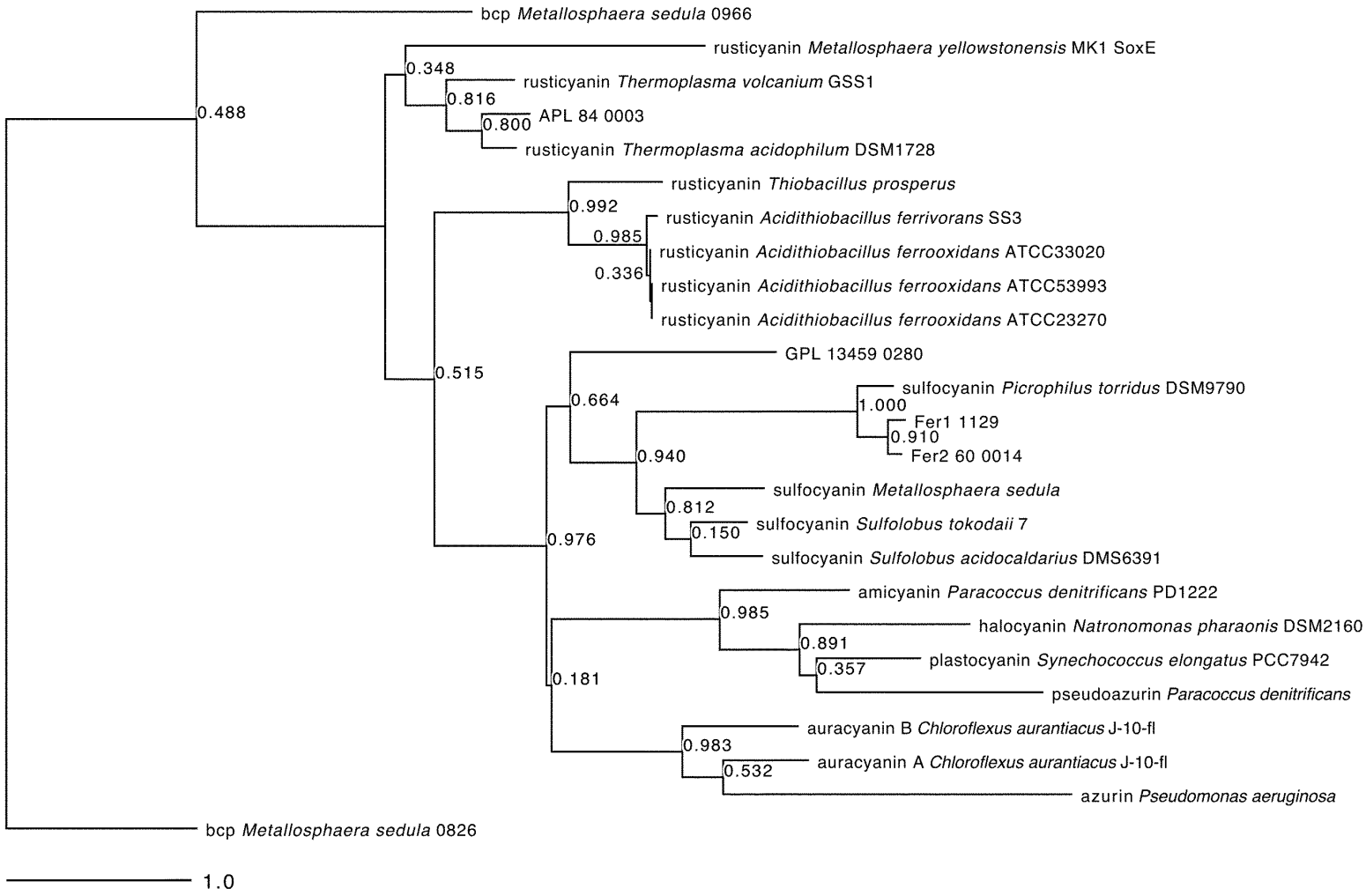


Figure S4: AMD plasma CODH gene tree.

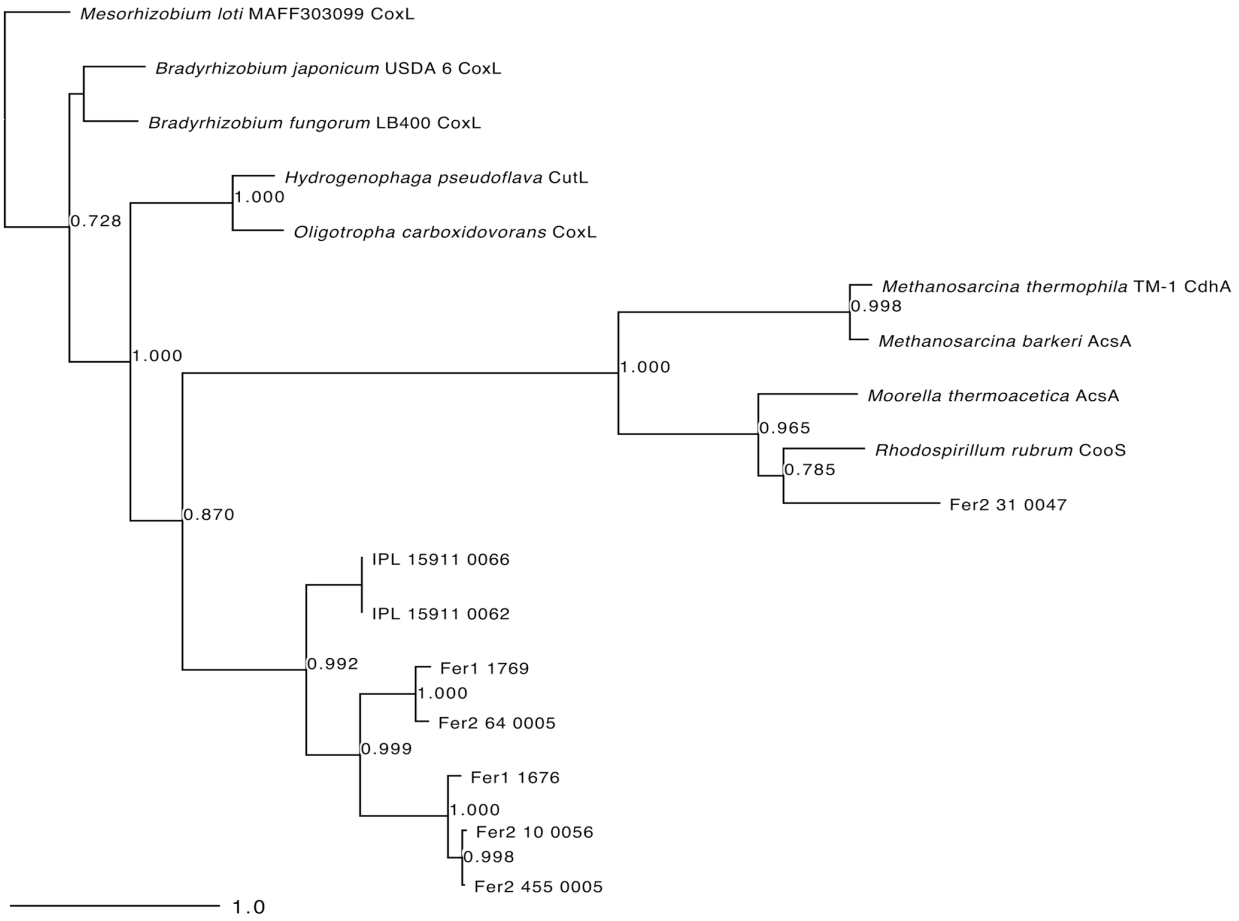


Figure S5: Active site alignment of aerobic CODH catalytic subunit genes. The red box indicates the active site residues. *H. pseudoflava* is *Hydrogenophaga pseudoflava*, *O. carboxidovorans* is *Oligotropha carboxidovorans*, *M. loti* is *Mesorhizobium loti*, *B. japonicum* is *Bradyrhizobium japonicum*, and *B. fungorum* is *Burkholderia fungorum*.

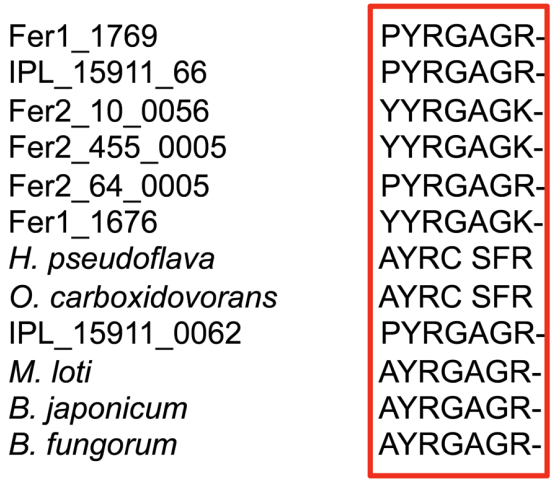


Figure S6: Ni-CODH catalytic subunit alignment. Genes in this alignment are the Ni-CODH catalytic subunits from *R. rubrum* (CooS, PDB:1JQK), *M. thermoacetica* (AcsA, PDB:1MJG) and Fer2 (fer2_31_0047). fer2_31_0047's secondary structure was predicted by YASPIN [146]. β-strands are shown in green and α-helices are highlighted in cyan. Residues belonging to the D-cluster are boxed in yellow (Cys41 and Cys49). Ligands of the B-cluster are boxed in black (Cys50, Cys53, Cys58 and Cys72). Catalytic residues binding the Ni-Fe-S cluster from C-cluster are boxed in purple (His265, Cys300, Cys338, Cys451, Cys481, and Cys531) and catalyze the oxidation of carbon. His95 and Lys568 (boxed in dark red) are non-coordinating residues conserved in Ni-CODHs and have been suggested to be involved in facilitating the reaction [147]. Residue numbering is from the *R. rubrum* Ni-CODH.

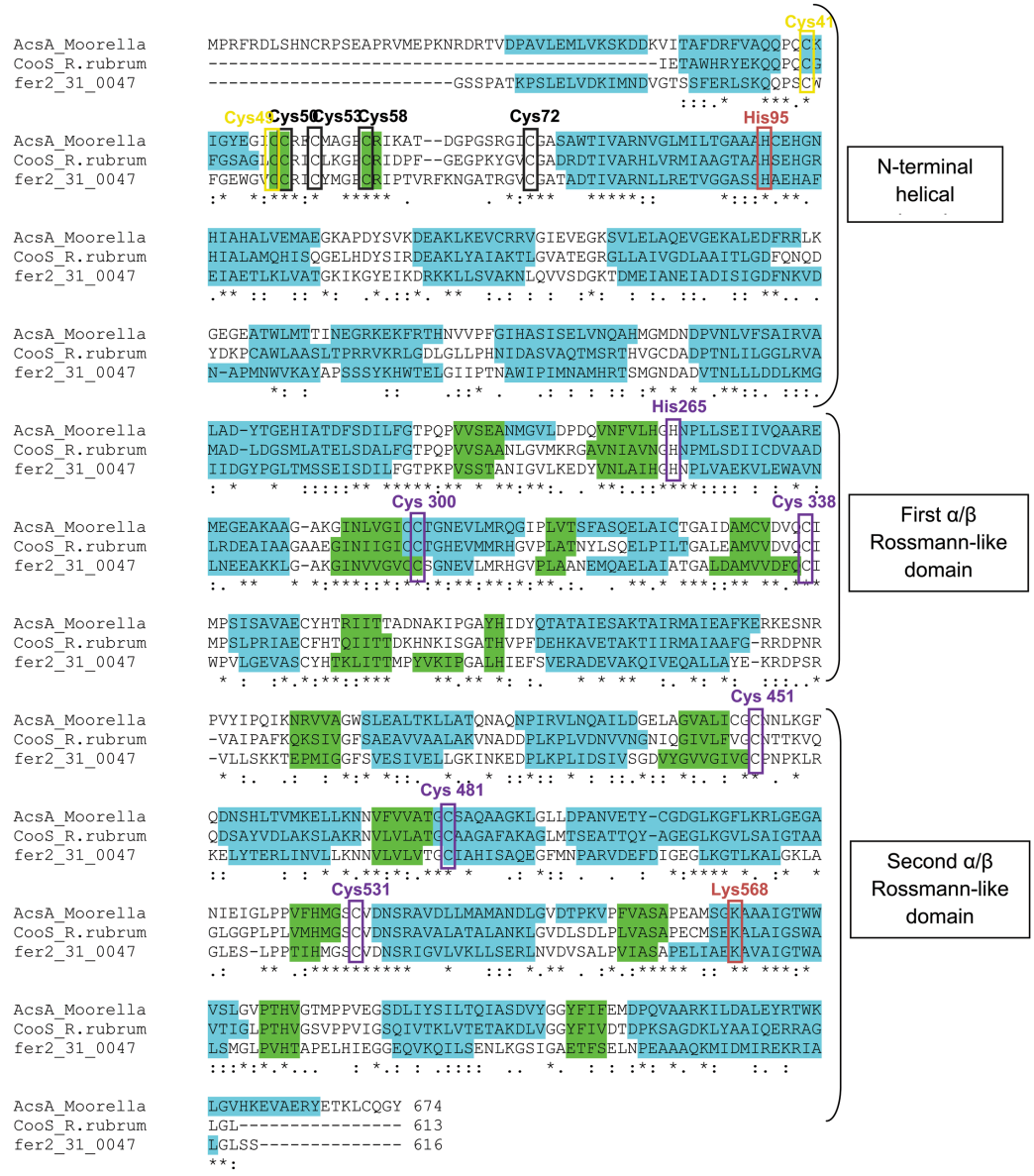
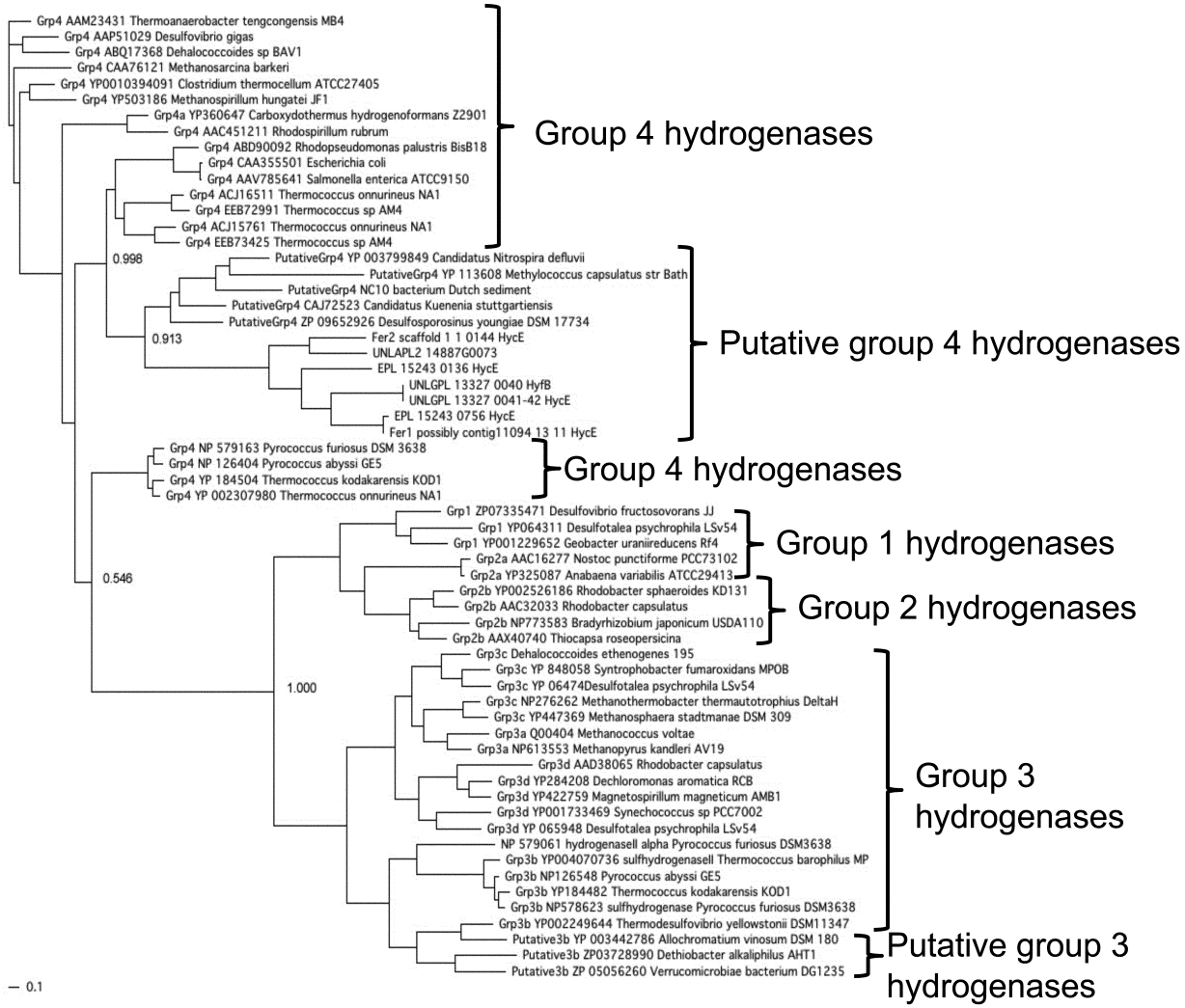


Figure S7: Cytochrome c oxidase subunit II alignment. * indicates the copper-binding motif found in other cytochrome c oxidase proteins. *S. acidocaldarius* is *Sulfolobus acidocaldarius*, *A. pernix* is *Aeropyrum pernix*, *P. oguniense* is *Pyrobaculum oguniense*, *T. thermophilus* is *Thermus thermophilus*, *P. denitrificans* is *Paracoccus denitrificans*.

```
                                                  * * *   *  *
aa3 cytochrome c oxidase S. acidocaldarius CAEFCGPGHYTM
ba3 cytochrome c oxidase A. pernix         CNEYCGVGHQFM
aa3 cytochrome c oxidase A. pernix         CYELCGIGHSLM
aa3 cytochrome c oxidase P. oguniense TE7  CAEFCGAGHYLM
aa3 cytochrome c oxidase H. volcanii DS2   CTEYCGVAHSQM
ba3 cytochrome c oxidase T. thermophilus   CNQYCGLGHQNM
ba3 cytochrome c oxidase P. denitrificans  CSELCGINHAYM
                         IPL_136060_0436   CVEYCGEYHYEM
                         Fer2_60_0013      CVELCGLGHAHM
                         EPL_15243_0560    CVEYCGYDHYKM
                         GPL_13374_0155    CVEYCGYDHYLM
                         Fer1_1130         CVELCGLGHAHM
                         APL_13214_0025    CVEYCGEDHYLM
```

Figure S8: AMD plasma putative hydrogenase 4 gene tree. Accession numbers are to the left of the species names.



Movie S1: Cryo-EM movie of AMD plasma cell with S-layer proteins. See attached.

Movie S2: Cryo-EM movie of AMD plasma cells with flagella, pili, and viruses. See attached.

CHAPTER 3.

Isolation and genome sequencing of a vanadium-reducing *Simplicispira* sp. from a vanadium and uranium-contaminated aquifer

Authors: Alexis P. Yelton[9], Kelly C. Wrighton[10], Kim M. Handley[2,11], Birgit Luef[2], Robert G. Butler[1], Jillian F. Banfield[1,2]

[9] Department of Environmental Science, Policy, and Management, University of California, Berkeley, California 94720, USA
[10] Earth and Planetary Science, University of California, Berkeley, California 94720, USA
[11] Computation Institute, Argonne National Laboratory, Argonne, Illinois 60439, USA

Abstract

Vanadium is a ubiquitous trace metal mobilized by mining and fossil fuel burning activities. Because vanadium is redox-sensitive it is susceptible to biotransformation by bacteria. Bacterial vanadium reducers have been isolated from the environment. However, none have been expressly associated with vanadium bioremediation. In order to assess the potential for vanadium bioreduction in shallow subsurface environments, we enriched for vanadium reducers from groundwater from a uranium and vanadium-contaminated site, the Old Rifle Mill in Rifle, Colorado, USA. From these enrichments we isolated a strain of *Comamonadacaea* in the genus *Simplicispira* (str. BDI) on vanadium as the sole terminal electron acceptor. BDI is a facultative anaerobic respirer and nitrate reducer. The organism's genome was sequenced via the Illumina HiSeq method. The genome contains a high number of potential chemotaxis, toxic metal and metalloid resistance, and conjugal transfer-related genes. The occurrence of these genes in the genome is consistent with a lifestyle based on resistance to high levels of contaminants and adaptability established via plasmid transfer of genes such as those that confer resistance to toxic metals. Motility was confirmed via cryogenic electron microscopy. Laboratory experiments established str. BDI's ability to reduce vanadium (V^{5+}), but also indicated that vanadium is toxic to str. BDI at concentrations in the 1-20 mM range. Vanadium addition at less toxic concentrations did not appear to competitively inhibit growth on nitrate, suggesting that vanadium does not compete with nitrate to bind to the nitrate reductase in this organism.

Introduction

Vanadium (V) is the 22nd most abundant metal in the earth's crust, more abundant than other common pollutants such as copper and zinc [148]. It is found at particularly high concentrations in seawater ~ 30 nM, making it the second most abundant transition metal in marine environments (surpassed by molybdenum) [148]. It is known to act as a cofactor in certain prokaryotic enzymes including some nitrogenases [149]. V is moderately toxic, possibly because of its ability to substitute for phosphate in proteins [148, 150-153]. Because it is a widespread metal and contaminant, it is notable that so little research has addressed the understanding of vanadium transformations in soils and sediments. Of particular interest is the biotransformation of vanadium at contaminated sites.

Vanadium can be reduced or oxidized in contaminated soils and sediments. Researchers have previously observed the removal of vanadium from groundwater concurrent with the addition of carbon compounds to stimulate the microbial community [18]. Because biostimulation results in reducing conditions, this removal is thought to be due to vanadium bioreduction and subsequent precipitation. Vanadium has three oxidation states at circumneutral pH, V^{5+}, V^{4+}, and V^{3+} [148]. Under oxic conditions some form of the V^{5+} vanadate (VO_2^+) ion dominates, whereas under reducing conditions V^{5+} can be reduced to vanadyl (VO^{2+}) [148]. The published redox potential versus the normal hydrogen electrode for the V^{5+}/V^{4+} couple at pH 7 is VO_2+/VO^{2+} 0.16 V [154]. This sets the V^{5+} to V^{4+} reduction at a slightly lower voltage than the potential of the Mn(IV)/Mn(II) couple [155].

Biotic V^{5+} reduction has previously been demonstrated [18-21, 148, 156, 157], but no vanadium respirers have been isolated from contaminated systems on vanadate, and very little is known about the importance of vanadium reducers in their natural environments. Isolates from contaminated environments may prove to be important if bioremediation or bioextraction of vanadium is the goal because these isolates will be capable of surviving in the environments in question and reducing high concentrations of vanadium. The known vanadium respirers include *Pseudomonas vanadium-reductans,* and *Pseudomonas isachenkovii*, *Shewanella oneidensis* MR-1, *Vibrio parahaemolyticus, Geobacter metallireducens*, and various bacterial strains from deep-sea hydrothermal vents [18-20, 148, 157]. Of these, the *Pseudomonas* spp. were isolated from the tissue of the ascidian, a sea-dwelling filter-feeding animal known to use vanadium as a cofactor for the oxygen-binding proteins in its blood. Neither the subsurface species that respire V, *S. oneidensis* and *G. metallireducens*, nor *V. parahaemolyticus*, were isolated on vanadium. Several other bacteria have demonstrated vanadium reduction but not V respiration (*Enterobacter cloacae* EV-SA01, and *Micrococcus lactilyticus*, *Acidothiobacillus* spp., *Fervidicella metallireducens*) [21, 22, 156, 158], suggesting that this is a widespread metabolism potentially used a means of vanadium resistance in addition to energy conservation.

Here we report the isolation of a bacteria from the Old Rifle Mill in Rifle, Colorado, USA, a former vanadium and uranium mill. This site was contaminated by leachate from mill tailings piles, which were removed between 1992 and 1996. The site is located on an alluvial flood plain of the Colorado River, and comprises an unconfined shallow aquifer that is contaminated by residual vanadium, uranium, arsenic, and selenium. This aquifer has been the site of bioremediation experiments since 1999, involving injection of a carbon-source, acetate, into the

aquifer, using 5.1 cm diameter wells. We isolated a vanadium-reducing bacteria from groundwater samples taken from one of these wells.

The bacteria isolated from the Rifle site is of the genus *Simplicispira*. This genus is comprised of three isolate species, *Simplicispira psychrophila, Simplicispira metamorpha*, and *Simplicispira limi* str. EMB325 [159-162]. Thirty-one additional *Simplicispira* spp. 16S rRNA sequences have been submitted to the NCBI database (http://www.ncbi.nlm.nih.gov/). No genomes for this genus have been reported. All of the characterized species are aerobes or facultative anaerobic denitrifiers. They are chemoorganoheterotrophs and are motile, with bipolar flagella [160]. No *Simplicispira* spp. has previously been shown to have vanadium reduction capability. Here we report the isolation, physiological characterization, and genome of a new vanadium-reducing *Simplicispira* strain and demonstrate its cosmopolitan distribution in both contaminated and pristine sites.

Materials and methods

Culturing and isolation: Enrichment cultures were grown in 10 ml of the carbonate-buffered freshwater medium with N_2:CO_2 (80:20) headspace described by Lovley *et al*. [163, 164] with 100 uM or 1 mM sodium metavanadate (Sigma Aldrich) and 5 mM sodium acetate (Sigma). Cultures were inoculated with 1 ml (0.7%) Rifle groundwater from a well bore that has never received added carbon, U01 (See Chapter 4, supplementary figure 2) [165]. Controls for no vanadium, no acetate, and heat-killed cells were run concurrently with enrichment cultures. All treatments had two replicates. Cultures were then serially diluted and grown for three weeks in agar shake tubes prepared according to Coates *et al*. and containing the same media as the enrichment cultures [164]. After three weeks, colonies were picked and grown up in 2 ml of freshwater medium on acetate and vanadate. Cultures were grown in vanadium concentrations greater than or equal to 1 mM for two reasons: firstly, these concentrations aid in the isolation of a bacteria capable of vanadium reduction in highly-contaminated environments such as that found after fossil fuel spills, and secondly because colorimetric detection was not precise for concentrations below 10-20 uM, making measurement of reduction of environmental concentrations of vanadium difficult.

Growth of the isolate was characterized on different electron donor and acceptors by measurement of changes in optical density at 600 nm of the original culture and after one transfer. All terminal electron acceptors were tested with acetate as the electron donor, whereas all electron donors were tested with nitrate as the terminal electron acceptor.

Vanadium(V) colorimetry: A colorimetric V^{5+} assay based on the interaction of V^{5+} with diphenylcarbazide was used to assess aqueous V^{5+} concentration and reduced V concentrations [20, 166]. The diphenylcarbazide reagent was newly prepared for each sampling time point to prevent degradation. Reduced V species were estimated based on initial V^{5+} concentrations minus current V^{5+} concentrations.

Cell growth: Cell counts were performed with the Baclight LIVE/DEAD cell viability kit (Life Technologies). Cells were counted per field of view for ten fields of view per sample at 100X. Relative cell density was measured via optical density at a wavelength of 600 nm. Because vanadium reduction results in a color change of the media, we used the optical density of the culture minus the optical density of the filtered media for each sample to determine cell growth.

Cell suspensions: We monitored vanadium reduction in high-density cell suspensions and a heat-killed control. In order to obtain a high density of cells, a one liter culture of str. BDI was grown on sodium nitrate and sodium acetate and cells were pelleted, washed in nitrate-free, vitamin-free media, and resuspended in 10 ml of 5 mM vanadate, 5 mM acetate, and freshwater media with or without vitamins and minerals to allow for growth and non-growth conditions. All suspensions were carried out in triplicate.

Genomics: DNA was extracted with the PowerSoil extraction kit (Mo Bio) from liquid freshwater medium cultures. Extracted DNA was then made into metagenomic libraries

according to the University of Illinois protocol. The DNA was sequenced with Illumina HiSeq technology as paired-end reads with a target length of 100 bases and an insert size of approximately 300 bases. The resulting reads were trimmed of bases with a quality score of three or less and automatically assembled with velvet [139], using a kmer size of 49-mers. The assembly was manually curated to correct for mis-assemblies and close gaps with the consed software [140].

16S rRNA amplification: DNA was extracted with the PowerMax extraction kit (Mo Bio) from 10 ml of liquid isolate cultures for PCR amplification. DNA was then precipitated and eluted in 50 µl of elution buffer. PCR was performed to amplify the full 16S rRNA genes of the isolate with 27F and 1492R universal primers for 25 cycles.

Results

Isolation: Colonies grown in acetate-vanadate medium were white with a diameter of approximately 1 mm. Colonies were picked and then grown in a liquid vanadate/acetate freshwater medium.

Phylogenetic analysis and cryogenic electron microscopy: PCR and Sanger sequencing was used to resolve strain BDI's 16S ribosomal RNA sequence. This sequence was 99% identical with that of an isolate obtained from river water in Germany and isolated on pivalate and nitrate, *Comamonadaceae* str. PIV-8-2 [167]. The sequence was 97% identical to *Simplicispira psychrophila* LMG 5408, *Simplicispira metamorpha* strain D-416, *Simplicispira metamorpha* strain DSMZ 1837, and *Simplicispira limi* strain ST3. It was 96% identical to the 16S sequences of *Variovorax paradoxus*, *Variovorax ginsengisoli* Gsoil 3165, *Variovorax boronicumulans* BAM48, and *Variovorax koreensis* GH9-3. This is a good indication that strain BDI belongs in the *Betaproteobacteria* subphylum, the *Comamonadaceae* family, and the *Simplicispira* genus. Maximum likelihood trees placed BDI in the *Simplicispira* genus, with its closest relative being *Simplicispira psychrophila* (Figure 1).

Cryogenic electron microscopy (cryo-EM) revealed that cells were motile (flagellate) gram-negative, rods (length ~2-2.5 um width ~500-600 nm). The cryo-EM micrographs consistently showed cells containing 1-2 electron-dense inclusions located at the opposite poles of the cells (Figure 2).

Vanadium reduction: In order to demonstrate vanadium reduction by str. BDI, cells were condensed in high-density cell suspensions. These suspensions were able to reduce vanadate with or without vitamin and mineral mix (growth and non-growth conditions) (Figure 3). A heat-killed control did not reduce vanadium. After six days 66% of V^{5+} was reduced in the non-growth cultures, while 78% of V^{5+} was reduced in growth cultures. Ion chromatography indicated that acetate removal also occurred in both growth and non-growth conditions (Figure 4). However, peak interference made the errors in the acetate analysis large.

The isolate proved to be a facultative anaerobic respirer capable of growth on oxygen and nitrate (with acetate as carbon source). It could not use ferric citrate, ferric pyrophosphate, hydrous ferric oxide, ferric-NTA, nitrite, sulfate, sulfite, perchlorate, chlorate, arsenate, thiosulfate, manganese, or selenate as terminal electron acceptors. It was capable of growth on the following carbon sources: acetate, lactate, and proprionate, but it did not grow on citrate, phenol, butyrate, glucose, fumarate, trehalose, formate, sucrose, or pyruvate. The isolate was not capable of growth via fermentation on 1 g/L yeast extract, 1g/L yeast extract with 1 g/L casamino acids, and 10 mM glucose, 1 g/L casamino acids alone, 10 mM glucose alone, or 10 mM fumarate alone.

Growth on vanadium and toxicity: Cultures of BDI in medium with acetate and vanadium did not show growth over the course of seven days, as determined via cell counts (data not shown). We hypothesize that this is due to vanadium toxicity at high concentrations because vanadium is known to have toxic effects on *Pseudomonas aeruginosa*, a related bacterium [168]. To test this

hypothesis, we examined str. BDI's growth on nitrate with varying concentrations of vanadate in the media. This test demonstrated increasing doubling times with increasing vanadate concentrations from 0 mM to 5 mM V (Figure 5). At 20 mM vanadate, BDI did not grow after 200 h. Among the cultures amended with vanadium, vanadium reduction ranged from 8% to 80% of the vanadium added (Figure 6).

Growth with nitrate: We tested the vanadium interference with growth on nitrate by adding 1 mM vanadate to cultures in freshwater medium with 0-5 mM nitrate. The amount of vanadium reduced was not affected by the nitrate concentration in the growth medium (Figure 7), despite differences in growth rates (Figure 8).

Prevalence: Previously published microarray data [169] indicate that a bacteria with 99% 16S rRNA nucleotide identity to str. BDI is found in both groundwater and sediment samples *in situ* as well as in *ex situ* flow-through sediment columns and sulfate-reducing enrichment cultures. In Chapter 4 it is shown that strain BDI increases in abundance with the addition of acetate and vanadate to in-well flow-through sediment columns. Furthermore, an essentially identical genotype responds strongly to acetate-only addition to in-well Rifle sediment columns, especially during the iron reduction phase that precedes accumulation of sulfide in groundwater (unpublished data). 16S ribosomal RNA sequences with more than 97% nucleotide identity to the str. BDI sequence were found in the NCBI nr database (http://www.ncbi.nlm.nih.gov/) from studies conducted in the US, Germany, Spain, Japan, Taiwan, and China from subsurface and aquatic habitats as well as activated sludge. These sequences include several from another uranium-contaminated aquifer in Richland, WA, USA (AY532568.1, AY532569.1, AY532540.1) and a trichloroethene-contaminated site in Livermore, CA, USA (AF422662.1, AF422643.1) [170]. The site in Richland, WA is contaminated with vanadium [171], whereas vanadium concentrations at the Livermore, CA site are not published [172].

Genome: Basic genome assembly information is reported in Table 1. The genome was approximately 4.2 megabases in length on 234 contigs. We estimate the completeness of the genome assembly using orthologous marker genes [38] and find that the assembly contains all 35 of these genes in one copy in addition to one copy of the recA and gyrA genes, indicating that the genome is near complete (Supplementary table 1). Annotation of the genome indicates that it contains 21 heavy metal-associated genes, 9 copper resistance genes, 38 efflux pump-associated genes, the full ars arsenic resistance operon, 121 chemotaxis-related genes, 37 flagellar assembly genes, 24 conjugal transfer-related genes, 6 plasmid replication-related genes, 42 type II secretion-related genes, 7 polyhydroxyalconoate and cyanophycin-related genes, 20 denitrification genes, and 3 sulfite oxidase genes (Supplementary file 1).

Discussion

These results demonstrate that *Simplicispira* strain BDI can reduce high concentrations of V(V) when grown on acetate. Strain BDI is the first vanadium-reducer associated with bioremediation of vanadium at the site of its isolation. Laboratory results suggest that strain BDI can reduce up to 3 mM vanadate at high cell densities, the highest concentration of vanadium bioreduction by bacteria yet described.

Very little research on vanadium reductases has been published. To date, a nitrate reductase from the V-reducer *Pseudomonas isachenkovii* has been shown to have vanadium reduction activity [173]. CymA and OmcB, multiheme cytochromes involved in iron reduction from *Shewanella oneidensis* have been shown to be necessary for vanadium reduction in that organism [19, 174]. Because strain BDI is a nitrate reducer and does not respire iron, we hypothesized that its nitrate reductase was involved in vanadium reduction. However, growth of BDI on varying concentrations of nitrate was not affected by the presence of vanadium (Figure 8). These results suggest that vanadium does not competitively inhibit nitrate reduction.

We speculate that strain BDI uses vanadium reduction as a detoxification mechanism and potentially as a terminal electron-accepting process. Because it was isolated on vanadate as the sole terminal electron acceptor, it is likely capable of vanadium respiration. However, as our results demonstrate, vanadium toxicity to strain BDI increases with increasing concentrations. This hindered growth experiments and makes it unclear whether strain BDI can grow on vanadate.

Strain BDI's ability to grow aerobically as well via nitrate respiration helps to explain the intercontinental distribution of 16S rRNA sequences > 97% similar to that of strain BDI.. It is also interesting to note that some of these sequences were found at other contaminated sites, including a site contaminated by uranium and another by trichloroethene. These findings, along with genomic and microscopic evidence of motility, potential for chemotaxis, multiple metal resistance systems, and genetic malleability support the conclusion that BDI is well adapted to living in highly contaminated environments.

Acknowledgements

Funding was provided by Environmental and Remediation Sciences Program, Office of Science, Biological and Environmental Research, US Department of Energy. The Rifle, Colorado, Integrated Field Research Center Project is managed by Lawrence Berkeley National Laboratory for the U.S. DOE (contract no. DE-AC02-05CH11231). APY acknowledges NSF Graduate Research Fellowship Program support. The authors thank Susan Spaulding for laboratory support.

Tables and Figures

Table 1: General genome information.

N50	32,693 b
Average sequence length	18,073.13 b
Total number of contig	234
Total number of base pairs	4,229,113

Figure 1: Phylogenetic tree of 16S ribosomal RNA of strain BDI and close relatives. Bootstrapping values are centered on the branch junctions. Start and stop positions of the gene are listed after each species name followed by the Genbank accession number. *S.* str. BDI is shown in bold.

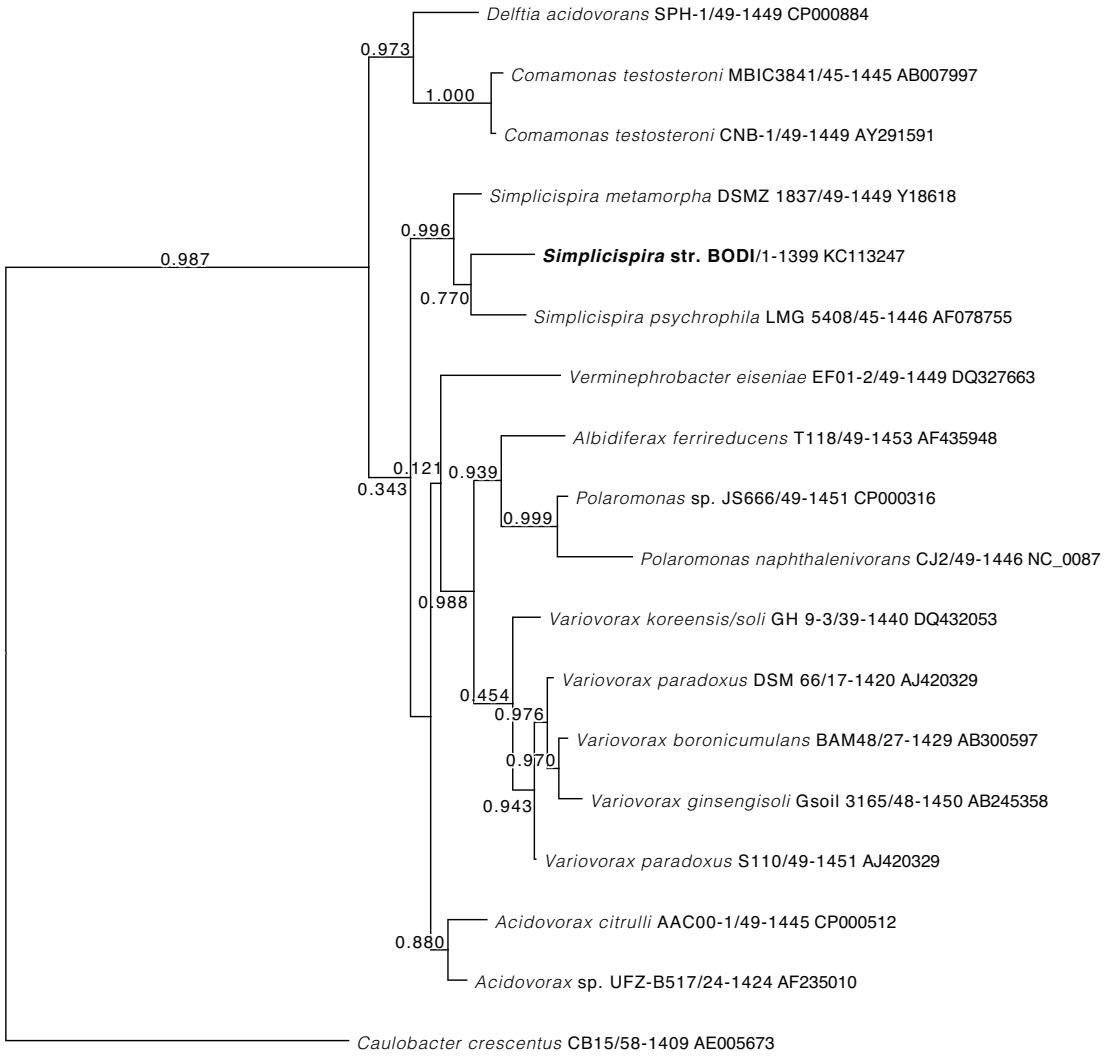


Figure 2: CryoEM of a BDI cell grown on nitrate. The arrows point to a single polar flagellum and two electron-dense inclusions on either pole of the cell.

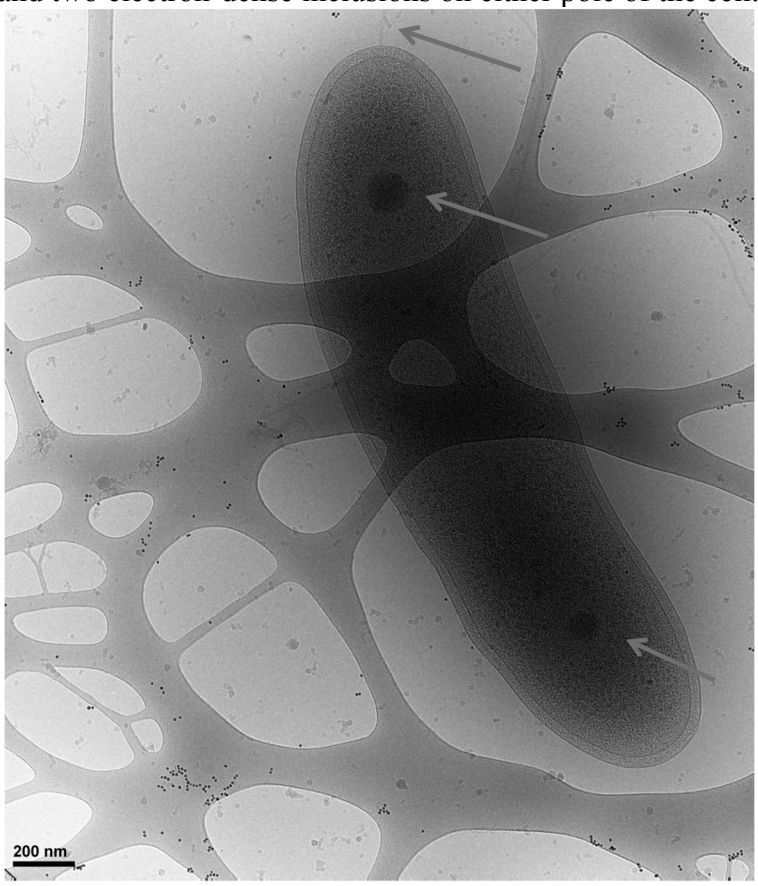


Figure 3: Vanadate concentration in high-density cell suspension of BDI cells.

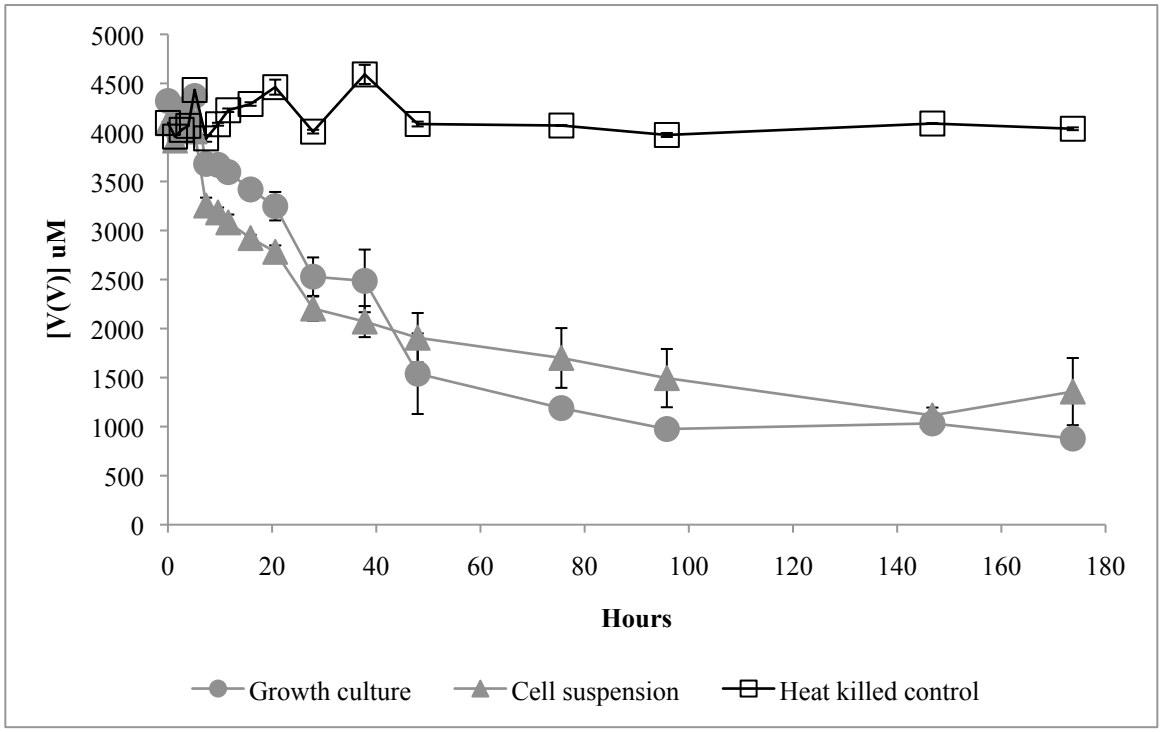


Figure 4: Acetate concentration in high-density cell suspension of BDI cells.

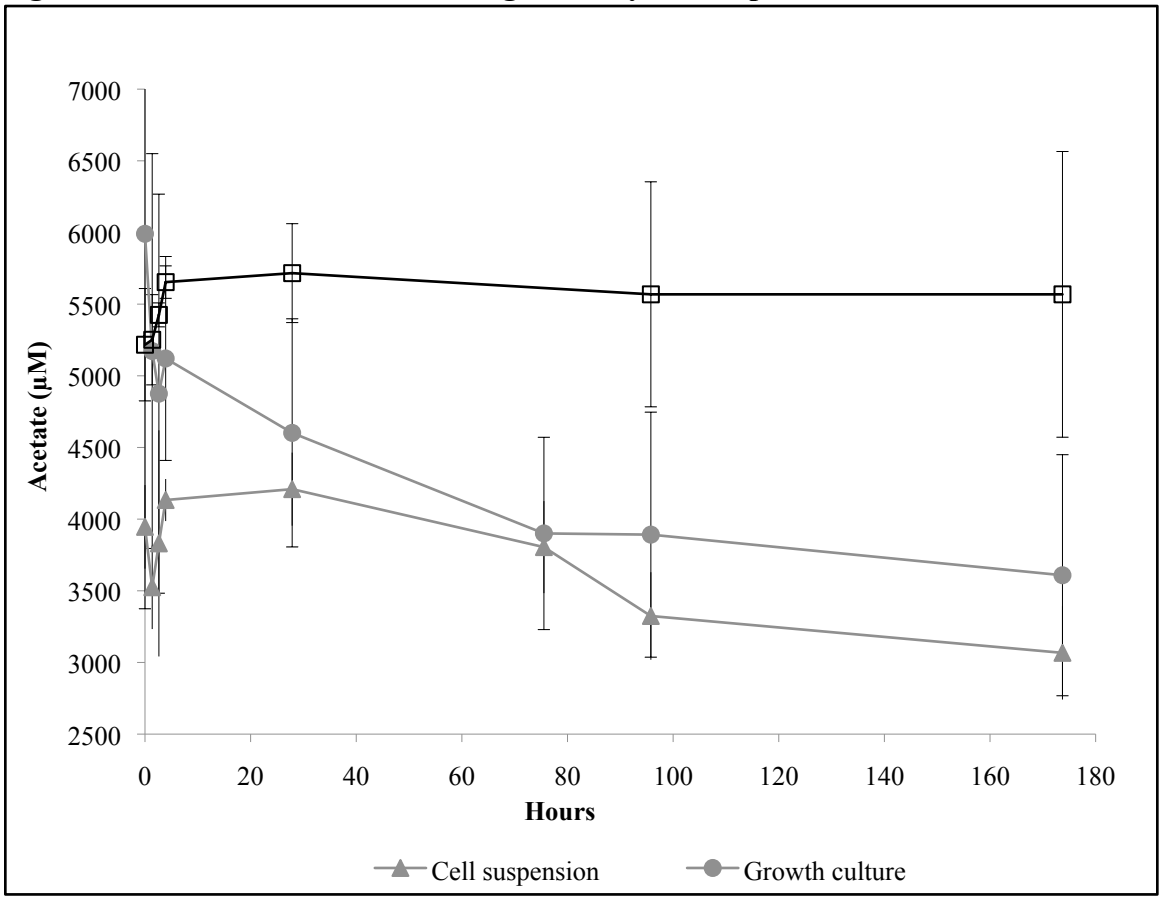


Figure 5: Optical density of BDI cultures in nitrate media amended with varying vanadium concentrations.

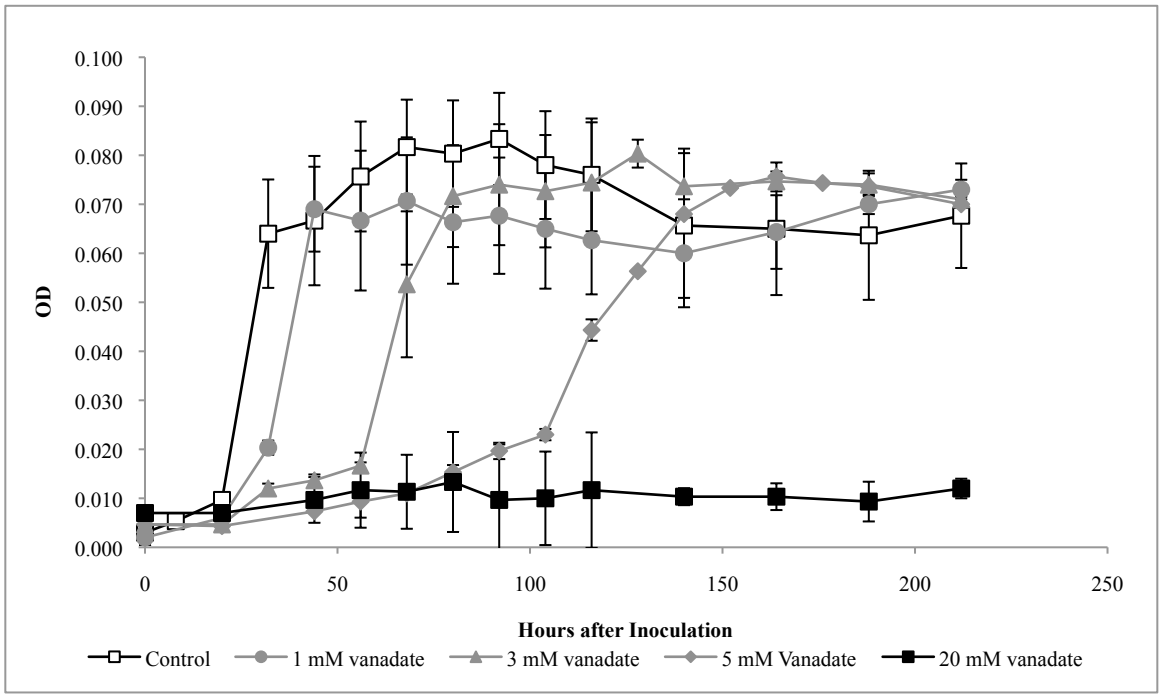


Figure 6: Vanadate concentration of cultures in nitrate media amended with varying vanadium concentrations.

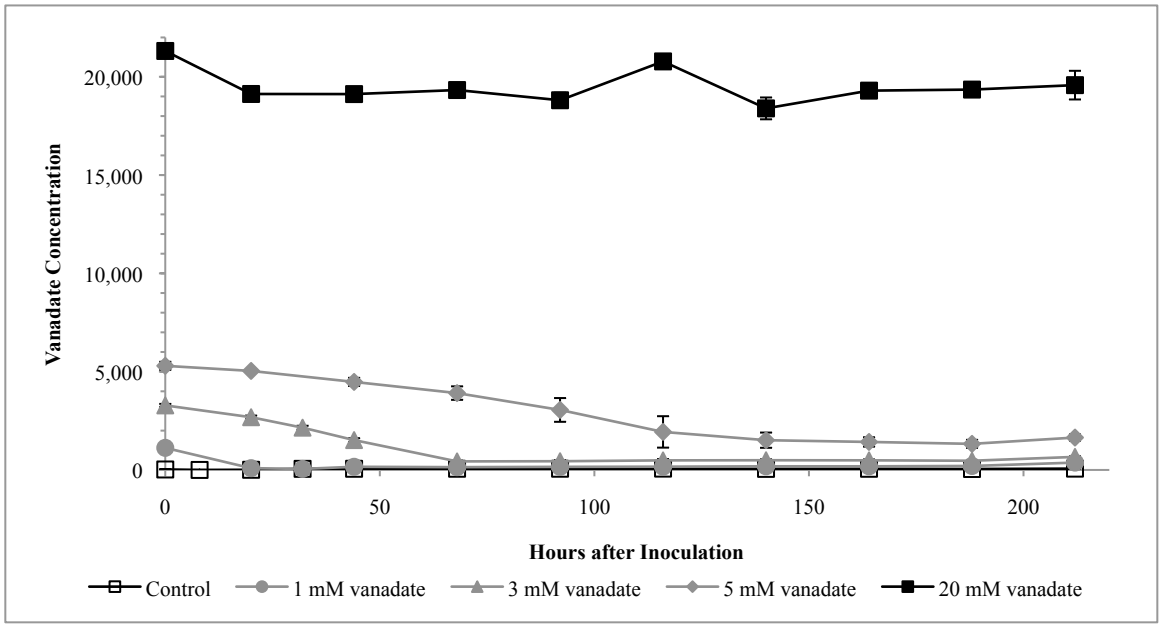


Figure 7: Vanadate concentration of str. BDI cultures in media with varying concentrations of nitrate amended with 1 mM vanadate.

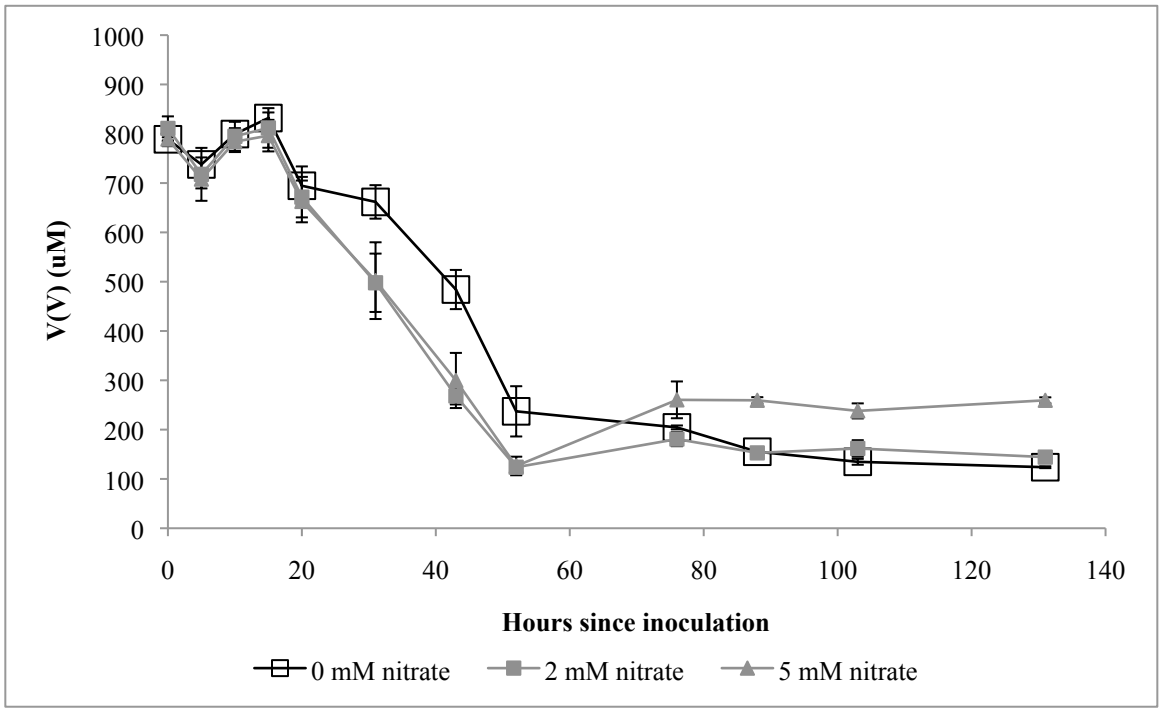
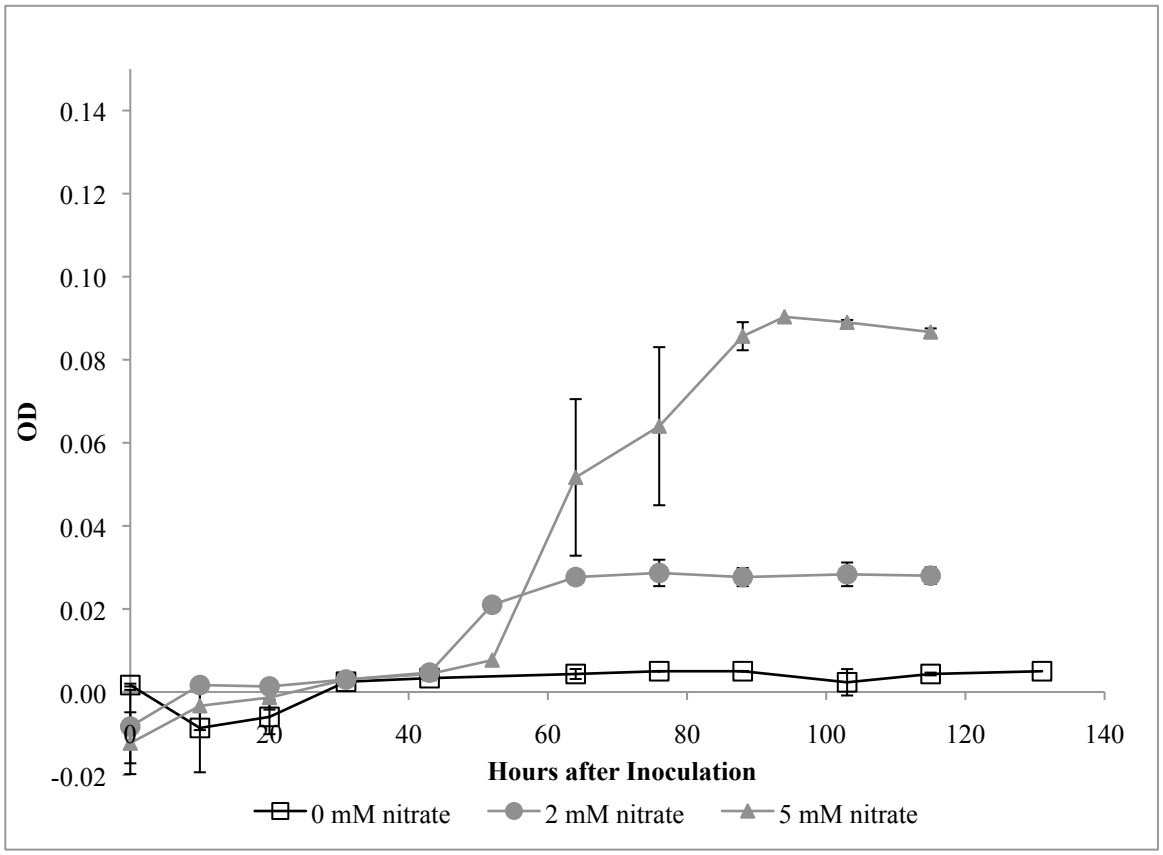


Figure 8: Optical density of str. BDI cultures in media with varying concentrations of nitrate amended with 1 mM vanadate.



Supplementary materials

Table S1: Estimated percentage of genome assembled based on occurrence of orthologous marker genes.

35 Orthologous group markers	**Gene number**
COG0012 Predicted GTPase, probable translation factor	UCBBDI_158_6
COG0016 Phenylalanyl-tRNA synthetase alpha subunit	UCBBDI_2_16
COG0048 Ribosomal protein S12	UCBBDI_116_4
COG0049 Ribosomal protein S7	UCBBDI_116_3
COG0052 Ribosomal protein S2	UCBBDI_192_16
COG0080 Ribosomal protein L11	UCBBDI_184_27
COG0081 Ribosomal protein L1	UCBBDI_184_26
COG0085 DNA-directed RNA polymerase, beta subunit/140 kD subunit, RpoB	UCBBDI_184_23
COG0087 Ribosomal protein L3	UCBBDI_131_30
COG0088 Ribosomal protein L4	UCBBDI_131_29
COG0090 Ribosomal protein L2	UCBBDI_131_27
COG0091 Ribosomal protein L22	UCBBDI_131_25
COG0092 Ribosomal protein S3	UCBBDI_131_24
COG0093 Ribosomal protein L14	UCBBDI_384_11
COG0094 Ribosomal protein L5	UCBBDI_384_13
COG0096 Ribosomal protein S8	UCBBDI_384_15
COG0097 Ribosomal protein L6P/L9E	UCBBDI_384_16
COG0098 Ribosomal protein S5	UCBBDI_384_18
COG0099 Ribosomal protein S13	UCBBDI_384_23
COG0100 Ribosomal protein S11	UCBBDI_384_24
COG0102 Ribosomal protein L13	UCBBDI_200_10
COG0103 Ribosomal protein S9	UCBBDI_200_11
COG0124 Histidyl-tRNA synthetase	UCBBDI_11_51
COG0184 Ribosomal protein S15P/S13E	UCBBDI_59_20
COG0185 Ribosomal protein S19	UCBBDI_131_26
COG0186 Ribosomal protein S17	UCBBDI_131_21
COG0197 Ribosomal protein L16/L10E	UCBBDI_131_23
COG0200 Ribosomal protein L15	UCBBDI_384_20
COG0201 Preprotein translocase subunit SecY	UCBBDI_384_21
COG0256 Ribosomal protein L18	UCBBDI_384_17
COG0495 Leucyl-tRNA synthetase	UCBBDI_53_14
COG0522 Ribosomal protein S4 and related proteins	UCBBDI_384_25
COG0525 Valyl-tRNA synthetase	UCBBDI_36_5
COG0533 Metal-dependent proteases with possible chaperone activity (TIGR gcp: metalloendopeptidase) (PFAM Peptidase M22, glycoprotease)	UCBBDI_151_36
COG0541 Signal recognition particle GTPase (Ffh in bacteria, SRP54 in archaea)	UCBBDI_158_15
RecA	UCBBDI_136_4
GyrA, DNA gyrase subunit A	UCBBDI_4_28
Estimated percentage of genome assembled	**100**

File S1: Amino acid sequences and annotations of genes in *S.* str. BDI's genome. See attached.

CHAPTER 4.

Vanadate and acetate biostimulation of contaminated sediments decreases diversity, selects for specific taxa and decreases aqueous V^{5+} concentration

Authors: Alexis P. Yelton[12], Kenneth H. Williams[13], Kelly C. Wrighton[14], Kim M. Handley[3,15], Jillian F. Banfield[1,3]

[12] Department of Environmental Science, Policy, and Management, University of California, Berkeley, California 94720, USA
[13] Division of Earth Sciences, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
[14] Earth and Planetary Science, University of California, Berkeley, California 94720, USA
[15] Computation Institute, Argonne National Laboratory, Argonne, Illinois 60439, USA

Abstract

Vanadium (V) is a commercially important metal that is released into the environment by fossil fuel combustion and mining. Despite its prevalence as a contaminant, the potential for bioremediation and biorecovery of vanadium has not been widely studied. Injection of organic carbon (acetate) directly into an aquifer to biostimulate contaminated sediments in Rifle, CO, USA, resulted in a two-year removal of contaminant vanadium from groundwater. To further investigate this process, we simultaneously added acetate and vanadate (V^{5+}) to columns that were packed with aquifer sediment and inserted into wells installed on the Colorado River floodplain. This allowed evaluation of the microbial response to amendments in columns that received influx of natural groundwater. Our results demonstrate *in situ* removal of up to 99% of the added V^{5+}(aq), suggesting microbially-mediated removal of vanadate. Most probable number measurements demonstrate up to a 50-fold increase in numbers of cells able to reduce V^{5+} in vanadium-amended columns compared to controls. 16S rRNA gene sequencing indicates a decrease in diversity in columns that received vanadate compared to those that did not. These findings also indicate increased representation of groups of closely related taxa in vanadate and acetate-amended columns as compared to acetate-amended columns and un-amended background sediment. Relative abundances computed from amplicon data reveal a relative increase in known vanadium reducers and related taxa, suggesting selection for V-tolerant and V-reducing species. Overall, our results demonstrate that limited organic carbon amendment (for a period of two months) can be an effective strategy for prolonged V removal, indicating that V bioremediation and may be a viable future technology.

Introduction

Vanadium (V) is both a widespread environmental contaminant from mining and fossil fuel combustion and a commercially important metal. Due to its value to the steel industry and its status as an impurity in fossil fuels, vanadium contamination of the environment continues to grow with increasing industrialization. Substantial vanadium contamination in the US has been recorded in 283 superfund sites (EPA, http://www.epa.gov/superfund/), but its remediation has generally not been addressed, and little is known about its biogeochemistry. Vanadium is moderately toxic to animals [150-153]. At concentrations greater than 1-10 nm, vanadium becomes toxic to animal cells [175-179]. Despite these findings, it is not currently regulated under the U.S. Environmental Protection Agency's Safe Drinking Water Act.

Vanadium naturally occurs in three oxidation states: V^{3+}, V^{4+}, and V^{5+} [180]. Under oxic conditions V^{5+} is usually found in aqueous solution as vanadate (VO_4^{3-}) [181]. In shallow aquifer sediments, where conditions fluctuate between oxic and anoxic, vanadium may be reduced (either from V^{5+} to V^{4+} or V^{3+}). Reduction to V^{4+} or V^{3+} can result in precipitation of V-bearing minerals [18, 20], making V less bioavailable and reducing its toxicity.

Previous studies have demonstrated microbial reduction of soluble V^{5+} [18-22, 148, 156, 157]. Several of these studies also report precipitation of an insoluble V^{4+} phase [18, 20]. However, little is known about biological reduction in terms of importance to environmental systems and remediation of V-contaminated sites. All research to date has made use of laboratory microbial isolates. In the current study we investigate the impact of vanadium addition on a complex microbial community. We document V removal during stimulation, demonstrate a decrease in bacterial community diversity, and identify native V-reducing microorganisms implicated in *in situ* remediation.

Previous research on uranium-contaminated aquifers and soils has shown decreases in diversity associated with carbon amendment and with high U concentrations [182, 183]. Both carbon amendment and U contamination are associated with increased relative abundances of organisms associated with U-reduction (i.e. *Geobacteraceae*, *Anaeromyxobacter*) [165, 169, 184-188]. Other studies have come to similar conclusions with laboratory enrichments of soils, aquifer sediments or slurries [183, 189, 190]. Islam *et al.* also compared uranium (U) ore communities to surrounding soil communities with 16S rRNA clones libraries (< 100 clones/sample) and found decreased diversity in the U-rich ores [182].

If we use uranium-contaminated environments as a model system for V-contamination, we would expect reduced diversity and increased relative abundance of select taxa with increasing V-contamination and biostimulation (e.g. organic carbon amendment). In fact, several studies have demonstrated this to a limited extent. A study performed on samples from a vanadium- and uranium-contaminated aquifer indicated a reduction in richness and evenness in microbial communities grown on hematite in the contaminated area as compared to nearby pristine sediments [191]. However, this study utilized clone libraries with only ~100 clones, resulting in undersampling of the community diversity. Decreases in diversity have also been documented at the field site used in this study, the Old Rifle Mill site, subsequent to acetate amendment [169, 184]. Additional research is needed to determine how vanadium contamination affects diversity during biostimulation, which taxa are enriched with increased vanadium and

carbon availability, and to determine the full extent of changes in diversity with increased sampling of taxa.

In order to study the effect of vanadium and acetate addition on a contaminated sediment community, we carried out research at the Old Rifle Mill site in Colorado, USA. Vanadium-bearing ore was mined and milled at this site from 1924 to 1958 [192]. Leachate derived from the mill tailings contaminated alluvial sediments and groundwater in an aquifer discharging to the Colorado River, which is adjacent to the southern boundary of the mill site. The aquifer is shallow and comprised of alluvium (sands, silts, and gravels) deposited by the Colorado River. The saturated thickness of the aquifer varies seasonally due to water level excursions tied to river stage, with an average thickness of ca. 3 m. Pore water velocities are typically between 0.3-0.8 m/day, with aquifer recharge resulting from precipitation infiltrating highlands to the north of the site [184]. Conditions in shallow un-stimulated sediments are usually oxic or suboxic, with dissolved oxygen levels rarely exceeding 30 μM in the saturated zone. Background V concentrations vary across the width of the floodplain, ranging from 30-100 μM, with excursions of 10-20 μM accompanying seasonal water level fluctuations. Groundwater geochemical conditions suggest that aqueous vanadate species will predominate. The aquifer maintains a pH of about 7.0-7.2 over the course of the year with groundwater conditions at or near saturation with respect to calcite. The ionic strength at Rifle is high for groundwater (~ 0.04 M), but is much lower than that of seawater. Combined with the circumneutral pH, this should result in the dominance of diprotonated vanadate species [193].

A number of biostimulation experiments have been carried out *in situ* at the Rifle site in order to assess bioremediation potential for uranium contamination [165, 184, 188]. To stimulate uranium reduction, a carbon source (acetate) is pumped into the subsurface. In response to acetate stimulation, the bacterial community changes, accompanied by reduction of iron and uranium [169, 184-188]. Short-term vanadium removal during biostimulation at the Rifle site was previously reported in association with acetate injection targeting uranium removal [18]. Here we carried out a 23 month assessment of vanadium removal after biostimulation and subsequently studied the effects of acetate and vanadium addition on vanadium geochemistry and the native bacterial community *in situ* over a period of 22 days.

Results

V removal and V^{5+} concentration: Inductively coupled plasma mass spectroscopy (ICP-MS) results indicate that two successive acetate additions lead to the removal of total vanadium from solution for a period of 23 months or longer (Figure 1); ongoing groundwater monitoring continues to indicate prolonged removal of V from groundwater in the absence of additional acetate loadings (data not shown).

To further to assess the mechanism behind vanadium removal and to evaluate the effect of V on the bacterial community, we conducted an *in situ* experiment in which flow-through sediment-packed columns were placed in well MNA-01 at a depth of 5 m. Vanadate and acetate were added directly to the column in the well (Figure S1), natural groundwater was allowed to flow through the columns, and effluent was pumped out (see Materials and methods section for details). We compared the geochemistry and microbiology of the vanadium and acetate addition treatment to the acetate-only treatment and to control columns that only received ambient groundwater. The microbial communities in these experiments were compared to those in background, un-stimulated sediment (Table 1).

Vanadium and acetate addition to *in situ* sediment columns resulted in a decrease in V^{5+} in the column groundwater (Figure 2). The groundwater turned blue, the color diagnostic of aqueous V^{4+} [18], after one day of injection (Figure 3), and the color persisted for the duration of the experiment.

Although inflow rates differed between the columns due to different degrees of sediment accumulation in the column tubing, the rates of V^{5+} removal after stabilization of the signal were similar (Table 2). After breakthrough (first detection of bromide tracer in the column), the vanadium influent was on average ~ 6.6 mM in V1 (± 2.3 mM) and 4.5 mM in V2 (± 1.7 mM). Thus, these samples allow us to explore the effects of different vanadium (and acetate) amendment levels on bacterial community structure.

Up to 99% of V^{5+} in the column influent was not subsequently found in the column effluent. 14.7 and 17.2 μM V^{5+}/h/g sediment were removed on average from V1 and V2, respectively (Table 2). Vanadium removal was concurrent with acetate oxidation as measured via ion chromatography (data not shown), with oxidation rates of 12.2 and 12.8 μM/h/g. The maximum amount of V^{5+} removal was 7.6 mM for V1 and 6.6 mM for V2.

MPN counts: To estimate enrichment of V-reducers in column sediment we carried out most probable number counts (MPN) of these organisms by inoculating acetate/vanadate media with sediment samples and serially diluting the media seven times. The results indicate that the V-amended sediments contain between 2.4×10^5 and 1.1×10^6 cells/g sediment. These MPN counts indicate enrichment for V reducers in V-stimulated columns as compared to acetate-stimulated columns (Figure 4). This difference was significant for V2 vs. A1 and C1, with $p < 0.05$.

Community composition and diversity: After 22 days the columns were sacrificed and DNA was extracted from the sediment. The 16S ribosomal RNA gene DNA was PCR-amplified and the amplicons were sequenced with Hiseq Illumina technology (see Materials and methods

section for details).

The 16S rRNA genes were assembled from sequencing reads, using the EMIRGE algorithm [194]. Full EMIRGE 16S ribosomal RNA sequences are available at http://genegrabber.berkeley.edu/SOM/yelton2012/. The reconstructed genes were used to characterize the bacterial communities in each sample (see Materials and methods section for details). Rarefaction curves were generated using the relative abundance data from the normalized priors of each EMIRGE operational taxonomic unit (OTU) (Figure 5). These relative abundances were divided by a minimal unit of abundance, 0.00003 relative abundance (where average read coverage is 10x), in order to normalize for detection limit.

The 16S rRNA rank abundance curves indicate differences in community structure between the treatments (Figure 6 and 7). The background community (B) has a much longer tail of rare taxa than the other treatments. Fewer rare taxa and a higher level of dominance were detected in the acetate samples (A1, A2), whereas even fewer rare taxa and even more dominant OTUs were detected in the vanadium samples (V1, V2)..

The horizontal line in Figure 6 delineates the abundance cut-off of 0.00003, used to filter out genes with low sequence coverage. Differences between the treatments above this cut-off still exist, but are less pronounced than differences in unfiltered data. A Kolmogorov-Smirnov test to determine if the filtered data's rank abundance curves came from different distributions indicates that all of the samples' curves are from different distributions except for those of the treatment replicates and A2 vs. B (Table 3). The Bonferroni multiple testing correction was applied to these p-values, but did not change the number of distributions that are significantly different at a p-value of 0.05.

Presence of known V-reducers: Previously, we isolated a strain of *Simplicispira* (str. BDI) (See Chapter 3) from un-amended groundwater that is capable or reducing high concentrations of vanadium. Community analyses of the stimulated sediments indicate the presence of strain BDI. Strain BDI was found in sediments receiving vanadium-addition at a higher abundance than in acetate addition or background sediments (on average 2.01%, 0.89%, and 0.11% respectively) (Table 4).

Biostimulation of column sediments with acetate and vanadium resulted in a relative abundance increase in a number of bacterial families (Figure 8), most of which fall into the *Proteobacteria* phylum. These include the *Geobacteraceae*, *Comamonadaceae*, and *Pseudomonadaceae* families of known vanadium reducers (Chapter 3) [18, 157]. The dominant *Comamonadaceae* in the vanadium-amended samples had > 97% nucleotide identity to an uncultured *Albidiferax* sp., whereas the sequence of the dominant *Geobacteraceae* in all samples was most closely related to an environmental clone sequence (HM141865.1) 97.1% identical to the 16S rRNA sequence of *Geobacter bemidjiensis* Bem (NR042769.1). Of the families of known vanadium respirers, *Vibrioaceae* was not detected and *Shewanellaceae* was detected at very low abundances (< 0.04%). Among the other families that increased with biostimulation were *Neisseraceaea* and *Erysipelotricaceae*, which are families that contain many pathogens. However, the majority of the sequences in these families were more than 98% identical to environmental species or strains (EU431736.1 and HQ012841.1) [195-199].

Alpha and beta diversity: Estimates of overall richness, Pielou's evenness, Shannon's diversity and Simpson's diversity all suggest decreasing diversity from the background to the vanadium treatment (Figure 9). Estimates of true/standard species diversity [200] indicate that the

magnitude of the decrease in diversity from the background to the acetate and to the vanadium samples is large (Table S1 and S2). For example, the average standard detectable richness in the acetate samples is 46% of that of the background sample, whereas the standard richness in the vanadium samples is only 23% of that in the background sample.

We used Fast Unifrac to create hierarchical clusters of the communities in each sample, taking phylogeny into account. Each treatment sample clustered with its replicate (Figure 10). For filtered data (no abundance weighting), the Unifrac significance test indicates that all samples are significantly different except the samples with the same treatments A1/A2, V1/V2. For filtered data with normalized abundance weightings the Unifrac significance test indicates that only B/A1 and B/A2 are significantly different.

We detected immense species richness in the background sediment. Interestingly, even given the high level of sampling in this study, the organisms identified in the amended samples are not a subset of those identified in the background community, based on overlap of OTUs at 97% nucleotide identity (Figure 11A). This is probably due to lack of detection of very low abundance OTUs. However, overlap of the communities between samples at lower phylogenetic resolution is substantial. At a 90% nucleotide identity OTU cut-off, more than 50% of taxa in each sample is shared with the other samples of the same treatment (vanadium and acetate or acetate-only) Figure 11B, and more than 60% of each treatment sample is shared with the background sample (data not shown). At this OTU cut-off, more than 50% of the background sediment OTUs are represented in the treatment samples. The Bray-Curtis beta diversity index indicates that there is more OTU overlap between the treatment samples than between the treatments and the background sample (Table S3).

Evidence for clustering: Faith's phylogenetic diversity (PD), a diversity index that takes into account phylogenetic distance, shows decreasing diversity detection from the background to acetate treatment to the vanadium treatment. We used the nearest taxon index (NTI) and the net relatedness index (NRI) [201] to differentiate between the case in which the community as a whole is more clustered than expected by chance (NRI) and the case in which there is more clustering of taxa among close relatives (NTI) than expected. Our data indicates an overall increase in clustering among close relatives in the vanadium treatments as compared to any other treatments in terms of Faith's phylogenetic diversity, NRI, and NTI without abundance weighting (Figure 12A, 12B, 12C). Non-abundance-weighted results for NRI were consistent with abundance-weighted results (data not shown). The abundance-weighted NTI results indicate that all samples' communities are more clustered than expected in a randomly assembled community (Figure 12D). However, the overall community is less clustered in the vanadium samples than in the other treatments. This differs from the Faith's PD results and NTI without abundance weighting, which suggest more clustering in the V samples.

Discussion

V removal and V^{5+} concentration: Both V^{4+} and V^{5+} can sorb to mineral surfaces. Thus, V removal from solution could be due to either sorption or V-mineral precipitation. In fact, prior work has shown vanadium adsorption onto iron oxides and clay minerals has a significant effect on V flux in groundwater systems [181, 202-206]. V desorption likely accounts for much of the persistent V plume at the Rifle site long after removal of tailings and contaminated surficial sediments. Sorption needs to be considered in this system for both oxidized and reduced forms of V. Here we detected an initial increase in V following acetate amendment (Figure 1), likely due to release from Fe-oxide sorbents as these minerals are reductively dissolved via dissimilatory Fe-reducing bacteria. Ion exchange accompanying increases in alkalinity associated with biostimulation [165] may also contribute to the initial V release into solution. The two-year removal detected after acetate amendment suggests that biostimulation may be effective in management of V contamination. Interestingly, acetate simulation may be more effective for vanadium than it is for uranium remediation, as uranium often returns to aqueous solution within weeks after the end of carbon addition [207]. As such, prolonged vanadium removal may not require sustained organic carbon loadings. This may be due to the formation of vanadium precipitates with very slow dissolution rates, even under the slightly oxidizing conditions that develop after acetate amendment ceases or to the continuous reduction of vanadium by bacteria for detoxification. The prolonged stability of the immobilized V towards re-oxidation over very long times scales (e.g. decades) remains to be assessed.

We infer that the decrease in V^{5+} (Figure 2) was partly the result of vanadium reduction, based on the concurrent color change of the effluent solution (Figure 3). We suggest that this reduction may result in the removal of vanadium from the groundwater during *in situ* biostimulation (Figure 1), presumably by increasing precipitation or adsorption of vanadium.

MPN counts: The MPN counts of vanadium reducers indicate enrichment for vanadium-reducing cells in the vanadium amendment columns (Figure 4). The differences seen in the number of vanadium reducing cells in V1 and V2 are likely due to the increased V concentration in V1, which may have a toxicity effect (see Chapter 3). We recognize that reduction rates of different V-reducing taxa may vary substantially, biasing the MPN estimate of cell numbers. Keeping this in mind, we can use the MPN cell abundance numbers to provide a preliminary estimate of V-reduction rates per cell in the column experiment. Dividing the average reduction rates of each column by the number of cells/g sediment in the MPN counts gives us 61.3 and 15.6 pM V reduced/h/cell for columns V1 and V2 respectively.

Community composition and diversity: The normalized rarefaction curves from the 16S rRNA sequences begin to level off, suggesting a high (though not complete) level of sampling of the community diversity (Figure 5). The differences observed between the rank abundance curves are small, indicating that samples from the same treatment are good replicates in terms of community structure. The long tail of rare taxa in the background community is typical of sediment samples and indicates a large resident diversity. The increased dominance and decreased richness in the treatment samples points to selection for acetate oxidizers (A1, A2) and acetate oxidizers that are vanadium tolerant (V1, V2). The calculated alpha diversity indices suggest that vanadium addition (with acetate) and acetate addition alone lead to decreases in

standard diversity of up to 54% and 77% respectively, though further research is necessary to determine the significance of these decreases (Figures 9A, 9B, 9C, and 9D). OTU abundances differentiate the V1 and V2 samples from one another. These samples were also subjected to differing levels of acetate and vanadium and thus are not perfect replicates. However, the diversity indices in the vanadium treatment are consistently lower than in other treatments.

One of the primary findings of this study is an increase in relative abundance of three families of known vanadium reducers in one or both vanadium-amended samples (Figure 7 and 8). The study adds to the evidence indicating an increase in *Geobacteraceae* following acetate amendment [165, 169, 184-188], but also indicates that *Geobacteraceae* dominates in V2, the 4.5 mM vanadium addition sample. Interestingly, *Geobacteraceae* decreased to a very low relative abundance (~0.5%) in V1, the high vanadium concentration sample. This may be because of vanadium toxicity to *Geobacteraceae* or greater competition with the dominant family in V2, *Comamonadaceae*, at high V concentrations. Though *Comamonadaceae* have been observed at the Rifle site before, this family has not previously been shown to respond to biostimulation there [184]. The vanadium-reducing strain that we isolated in the laboratory is also a member of this family. We observed an increase in relative abundance of one other known vanadium reducer family in V1, *Pseudomonadaceae* [157]. In contrast to *Geobacteraceae*, *Comamonadaceae* and *Pseudomonadaceae* increased in abundance in the high vanadium concentration sample (V1) as compared to the moderate vanadium concentration sample (V2), indicating that they may be better adapted to high vanadium concentrations. Overall the dominance of vanadium-amendment samples by families containing known vanadium reducers suggests that this metabolism, or at least vanadium tolerance, is widespread within these families, and that selection for these specific taxa is possible with vanadium and acetate amendment.

Evidence for clustering: The Faith's PD, NRI, and NTI without abundance weighting demonstrate more clustering in vanadium-amended communities than in the other communities (Figures 12A, 12B, 12C). This indicates a possible selection for related taxa. We interpret the divergent abundance-weighted NTI result (Figure 12D) to mean that the vanadium treatments lead to proliferation of groups of closely related organisms, but highly abundant taxa are less clustered than average. This may occur because several distantly related types of taxa are vanadium tolerant or can reduce vanadium (such as *Comamonadaceae* and *Pseudomonadaceae* spp.), and thus are more abundant in the vanadium-amended samples. Discrete taxa from these groups dominate in the vanadium samples, and do not tend to co-occur with close relatives, possibly because of competition. These patterns are evident in 16S rRNA phylogenetic trees constructed with sequences recovered in this study (data not shown).

Conclusions: The results of this study indicate that two acetate additions lasting a month each to a vanadium-contaminated aquifer can remove vanadium for two years or more, indicating its utility in V bioremediation. We establish that vanadium reducers can be detected in contaminated sediments and respond to increased concentrations of vanadate and acetate. Stimulation of the aquifer with vanadium and acetate results in a decrease in bacterial community richness and evenness, increased clustering of taxa overall, with dominance by discrete taxa, an increase in the abundance of *Comamonadaceae*, *Geobacteraceae*, and *Pseudomonadaceae*, close relatives of isolate vanadium reducers, as well as an increase in the abundance of *Simplicispira* str. BDI, our isolated V reducer.

Materials and methods

Biostimulation: An acetate, bromide, and deuterium oxide solution was injected into the subsurface via nine pump wells upgradient to a gallery of sampling wells (Figure S2) placed sequentially along the flow path of the aquifer as described previously [165]. The acetate was added twice for a period of one month each time, in July, 2010 and in August, 2011, with a target acetate concentration of 5 mM. Groundwater samples were taken from 5 m below the surface from these monitoring wells, filtered with 0.2 μm PTFE (Teflon) syringe filters (Alltech Associates Inc.), acidified with nitric acid (Baker Chemical Co), and then analyzed for total vanadium concentrations via ICP-MS.

In-well Columns: For community analysis experiments, flow-through columns were dispatched into a monitoring well (MNA-01) on the flood plain of the Colorado River in Rifle, CO, USA that had not previously been acetate amended. Sediments used for the field column study were recovered from the Rifle aquifer at a depth of ca. 4 m below ground surface using a backhoe excavator. Upon recovery, sediments were sieved (< 4.5 mm), loaded into a permeable mesh housing, and placed into well MNA-01 for a period of ca. 1-yr prior to beginning the experiments. As the sediments were located below the static water level (ca. 4 m) in MNA-01, they were fully saturated and presumed to be well-equilibrated with groundwater conditions prior to their use. Sediments used for the experiments described here were removed from the mesh inserts within MNA-01 and loaded into the columns.

The columns were made of PVC piping 16 cm in length (with a 3.175 cm inner diameter). Columns received influent solution treatments as well as inflow from natural groundwater surrounding the well. This was accomplished by pumping influent into the columns at a constant rate ~20% of the volume of outflow pumping (36 μl/min). See Figure S1 for details. Column treatments included addition of approximately 5 mM sodium metavanadate (Aldrich) and sodium acetate (Sigma-Aldrich) with a potassium bromide tracer (Sigma-Aldrich), a control with a sodium acetate and potassium bromide-only influent, and control sediment columns with no flow. Vanadate was added at this relatively high concentration in order to ensure the stimulation of the vanadium tolerant community as opposed to acetate oxidizing, iron reducers and to allow for measurement of V^{5+} concentrations, using a colorimetric method, which lacks precision at ambient environmental concentrations.

MPN Counts: One gram of sediment from in-well columns was added to 9 ml anoxic (N_2 headspace) minimal freshwater media [163], containing 5 mM sodium metavanadate as the sole terminal electron acceptor and 5 mM sodium acetate as a carbon source. 0.1% by volume sodium pyrophosphate was added and tubes were gently shaken for one hour at room temperature. The sediment slurry was then serially diluted (1/10 dilutions) and cultures were incubated at 30 °C for 8 weeks. V reduction was determined via the visible color change of the media to the blue color characteristic of vanadyl and compared to a no cell control.

16S rRNA sequencing: DNA was extracted from sediments using the Powermax soil DNA isolation kit (Mo Bio). The sediment samples were taken from *in situ* flow-through columns buried in sampling wells. Samples were from background sediment, sediment stimulated with

carbon and vanadium addition, and sediment stimulated with carbon addition alone. To increase the number of bacterial taxa recovered, universal primers (27F, 1492R) and temperature gradient PCR with 11 different annealing temperatures (48-58° C) for 25 cycles were used to amplify the 16S rRNA gene from the organisms sampled.

HiSeq Illumina paired-end technology was used to sequence 2.7 megabases of PCR product at the University of California, Davis. The sequencing consisted of 26,954,412 100-base pair reads. Reads were mapped to reference sequences from the Silva database and the most probable sequences were inferred using the EMIRGE iterative algorithm [194]. Resulting EMIRGE OTUs were then filtered to include sequences with abundances of > 0.00003 to exclude low coverage OTUs (Figure 13) and those with a high proportion of Ns (Figure 14). The genes were aligned to each other, using the SSU-align software [208]. The alignment was automatically masked with the ssu-mask program. Bacterial OTUs were clustered at a 97% nucleotide identity cutoff, using Usearch [209]. A phylogenetic tree was constructed with the aligned sequences via the FastTree maximum likelihood method with options –gtr –nt and 1000 iterations of the FastTree bootstrap [143, 144].

Acknowledgements

Funding was provided by Environmental and Remediation Sciences Program, Office of Science, Biological and Environmental Research, US Department of Energy. The Rifle, Colorado, Integrated Field Research Center Project is managed by Lawrence Berkeley National Laboratory for the U.S. DOE (contract no. DE-AC02-05CH11231). APY acknowledges NSF Graduate Research Fellowship Program support, and would like to thank Chris Miller for support with the EMIRGE algorithm.

Tables and Figures

Table 1: Sample names and treatments for sediment columns and groundwater samples.

Sample	Treatment	Acetate	Vanadium
MNA-01	Upgradient well groundwater	No	No
V1	Flow-through column	Yes	~ 6.6 mM
V2	Flow-through column	Yes	~ 4.5 mM
A1	Flow-through column	Yes	No
A2	Flow-through column	Yes	No
C1	No flow column	No	No
C2	No flow column	No	No
B	Background sediment	No	No

Table 2: Removal rates of V5+ and acetate. Rates are in mM/h/g sediment and were generated based on the average vanadium concentration removal from 8/5/10 to 8/9/10 multiplied by the column flow rate and divided by the mass of dry sediment per .

Sample	Average V^{5+} removal rate (μM/h/g)	Peak V^{5+} removal rate (μM/h/g)	Average acetate removal rate (μM/h/g)	Peak acetate removal rate (μM/h/g)
V1	17.8	22.5	16.1	39.3
V2	20.7	23.7	16.5	33.7
A1	NA	NA	12.7	17.7
A2	NA	NA	7.7	47.6

Table 3: Kolmogorov-Smirnov significance tests of rank abundance. Numbers are p-values: ** indicates significantly different distributions with a p-value < 0.01. * indicates significantly different distributions with a p-value < 0.05.

Raw K-S	A1	A2	V1	V2	B
A1					
A2	6.44E-02				
V1	5.40E-06**	2.49E-08**			
V2	3.90E-03**	2.52E-05**	1.31E-01		
B	2.49E-08**	2.27E-01	2.18E-08**	1.60E-04**	
Bonferroni correction	A1	A2	V1	V2	B
A1					
A2	6.44E-01				
V1	5.40E-05**	2.49E-07**			
V2	3.90E-02*	2.52E-04**	1.00E+00		
B	2.49E-07**	1.00E+00	2.18E-07**	1.60E-03**	

Table 4: Relative abundance of Str. BDI in each sample.

	V1	V2	A1	A2	B
BDI	1.28%	2.74%	1.10%	0.67%	0.11%

Figure 1: Total vanadium, acetate, and bromide concentrations in Rifle groundwater after acetate addition. CU01 is an upgradient, un-amended control. CD01 is downgradient of the acetate-addition well (see Figure S2). Total vanadium was determined by ICP-MS. Total acetate was determined by IC. A is total vanadium for well CU01. B is total vanadium for well CD01. C is total acetate and bromide for well CU01. D is total acetate and bromide for well CD01.

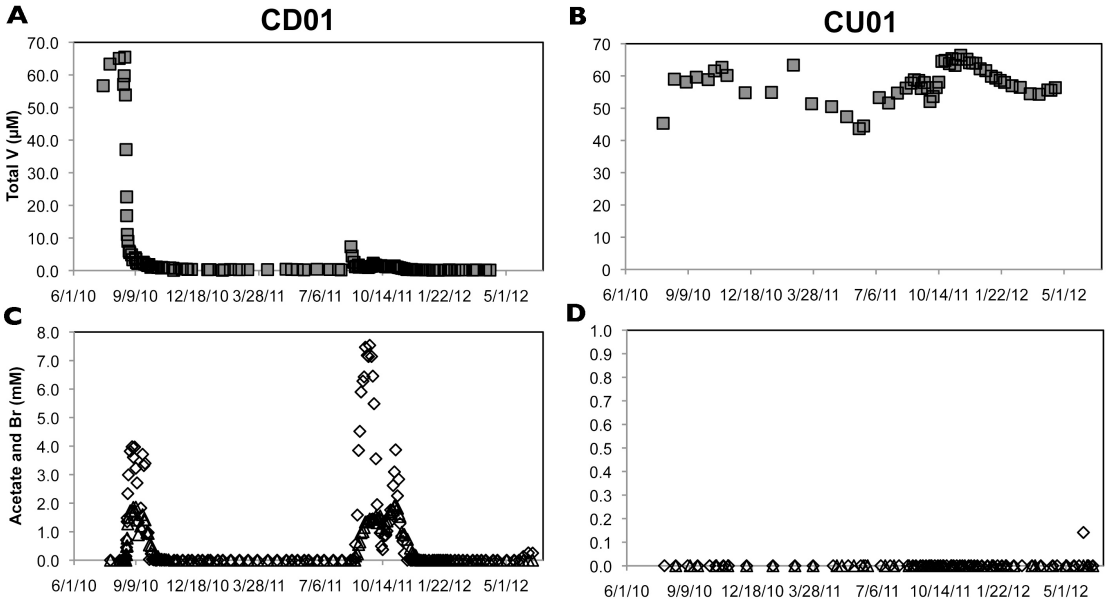


Figure 2: Aqueous V(V) concentrations in flow-through column effluent. Injection began on 7/23/10. MNA-01 is groundwater from the surrounding well. A1 and A2 are acetate addition columns. V1 and V2 are vanadium and acetate addition columns. C1 and C2 are no flow controls in the well.

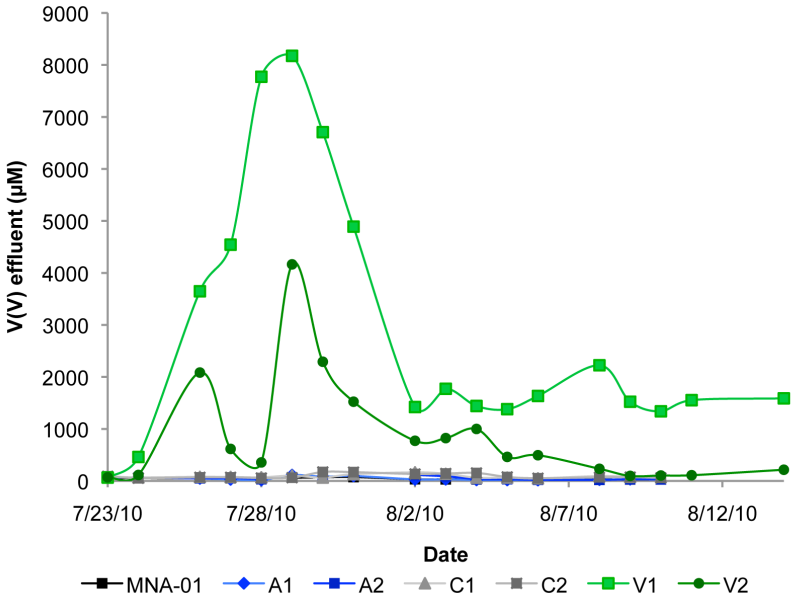


Figure 3: Visible evidence of vanadium reduction. Each centrifuge tube contains groundwater pumped from in-well columns. Ace1 and Ace2 refer to A1 and A2 column groundwater respectively. The blue color is indicative of aqueous vanadyl, V(IV).

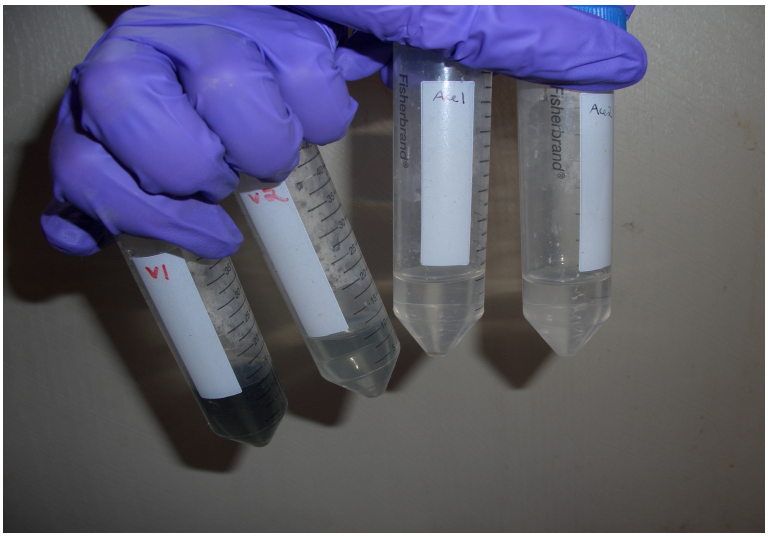


Figure 4: Most probable number estimates of cells of vanadium reducers per gram of sediment.

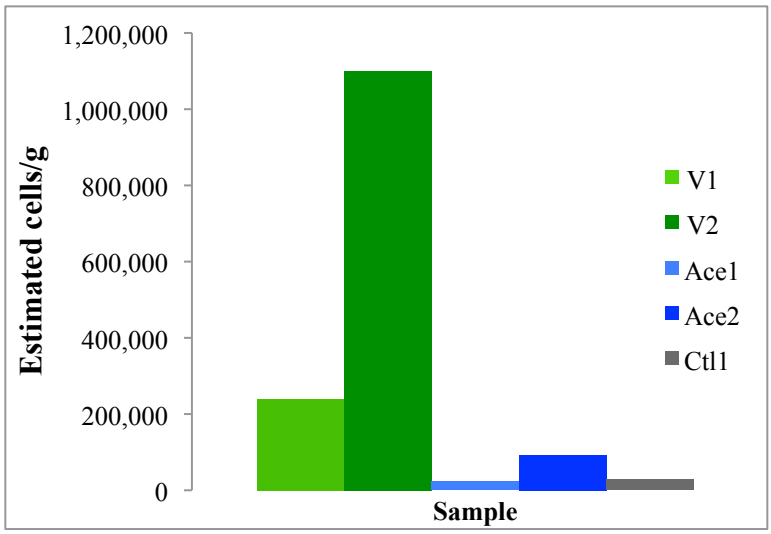


Figure 5: Rarefaction of 16S rRNA genes in each sample. Sample sizes are derived from normalized relative abundances calculated from normalized prior probabilities of each OTU [194] and are standardized by dividing by a minimum normalized prior value that corresponds to an average of 10x read coverage of the gene.

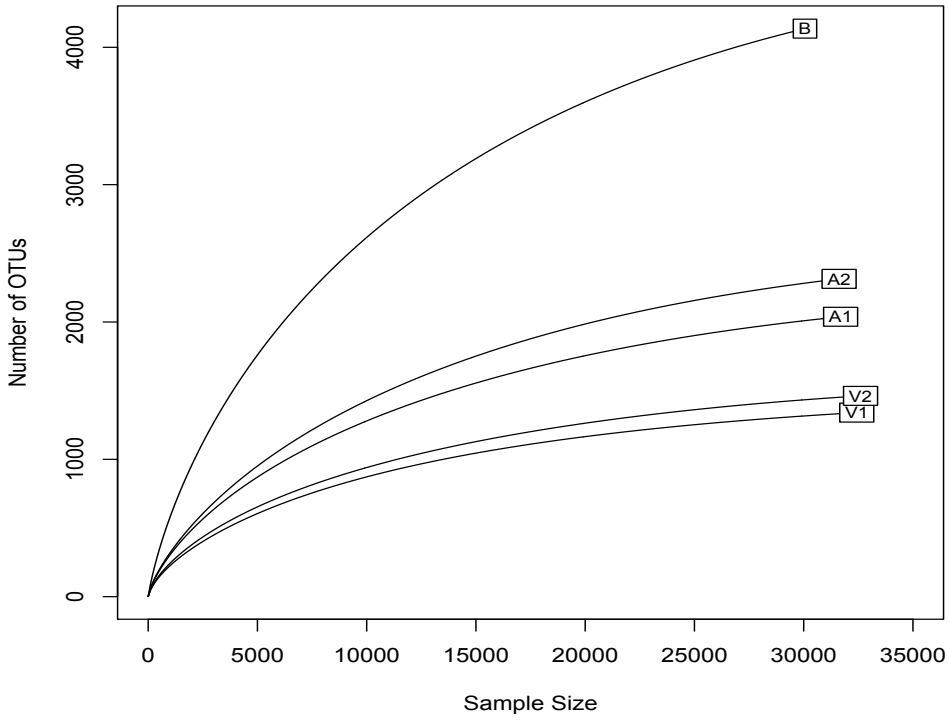


Figure 6: Community rank abundance curves. Black is background, dark blue is A2, light blue is A1, dark green is V2, and green is V1. The horizontal black line indicates the abundance threshold cutoff of a normalized prior of 0.00003 that was used to filter out poor quality sequences.

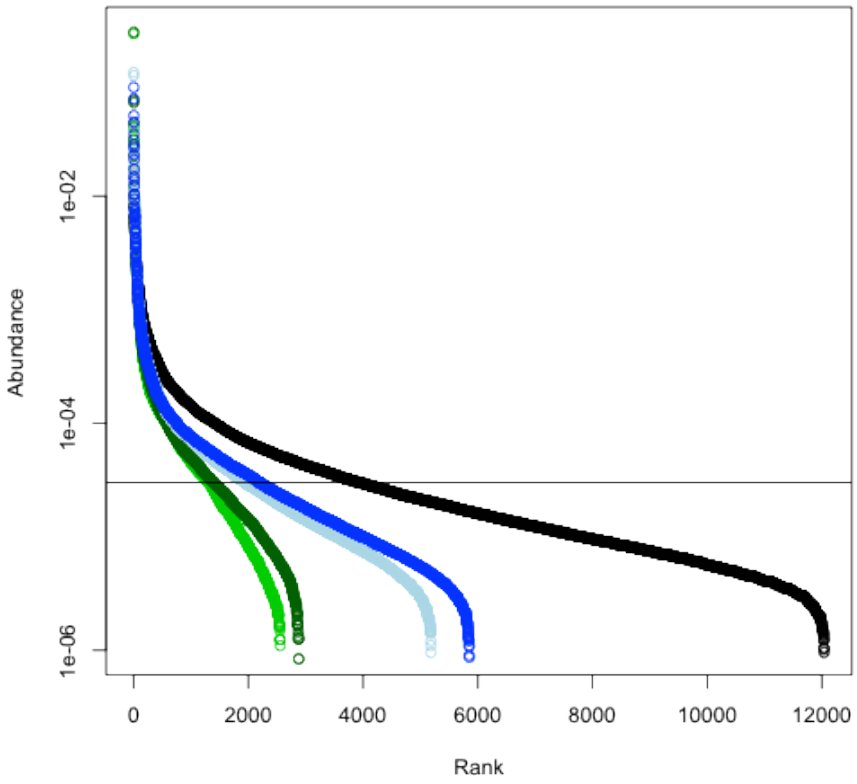


Figure 7: Rank abundance of ten most abundant OTUs in each column sample. 1 – *Geobacter* sp., 2 – *Albidiferax* sp., 3 – *Vogesella* sp., 4 – *Clostridiales* sp., 5 – *Dechloromonas* sp., 6 – *Lutibacter* sp., 7 – *Pseudomonas* sp., 8 – *Desulfurivibrio* sp. Black is background, dark blue is A2, light blue is A1, dark green is V2, and green is V1.

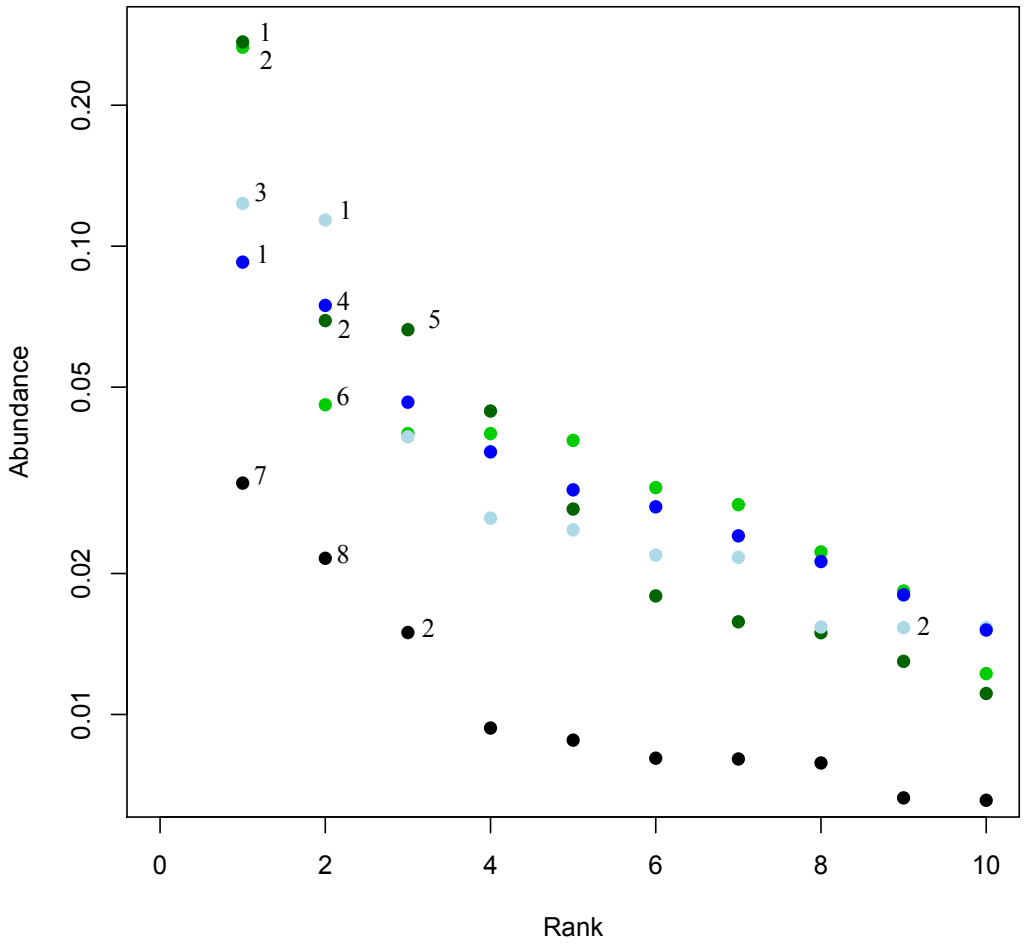


Figure 8: Relative abundance of bacterial families in column sediment. 1 – *Comamonadaceae*, 2 – *Geobacteraceae*, 3 – *Oxalobacteraceae*, 4 – *Pseudomondaceae*, 5 – *Rhodocyclaceae*, 6 – *Flavobacteriaceae*, 7 – *Nitrosomondaceae*, 8 – *Hydrogenophilaceae*, 9 - *Gallionelaceae*, 10 - *Anaerolinaceae*, 11 - *Desulfobulbaceae*, 12 – *Nitrospiraceae*, 13 – *Syntrophaceae*, 14 – *Desulfobacteraceae*, 15 – Unknown family, 16 – *Neisseraceae*, 17 – *Rhodobacteraceae*, 18 – *Rhizobiaceae*, 19 – *Erysipelotrichaceae*, 20 – *Caulobacteraceae*.

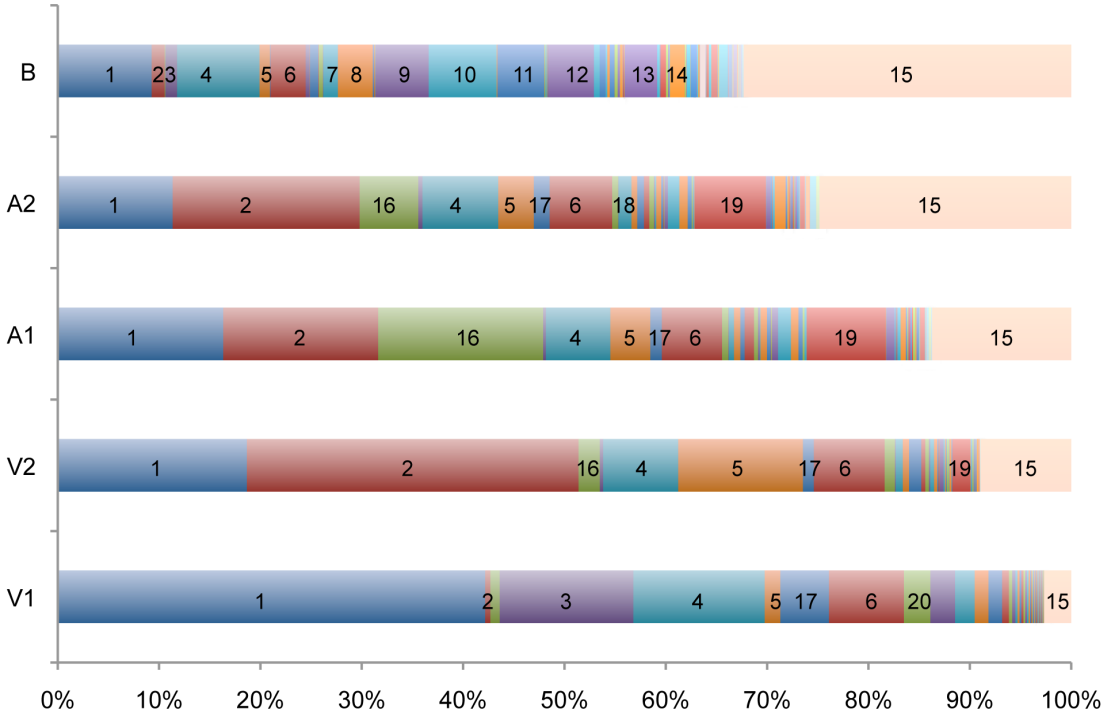


Figure 9: Diversity indices. Treatments are as follows: V is vanadium and acetate amendment. A is acetate amendment. B is background sediment. 9A – Richness as number of OTUs. 9B – Pielou's evenness. 9C – Shannon's diversity. 9D – 1- Simpson's diversity.

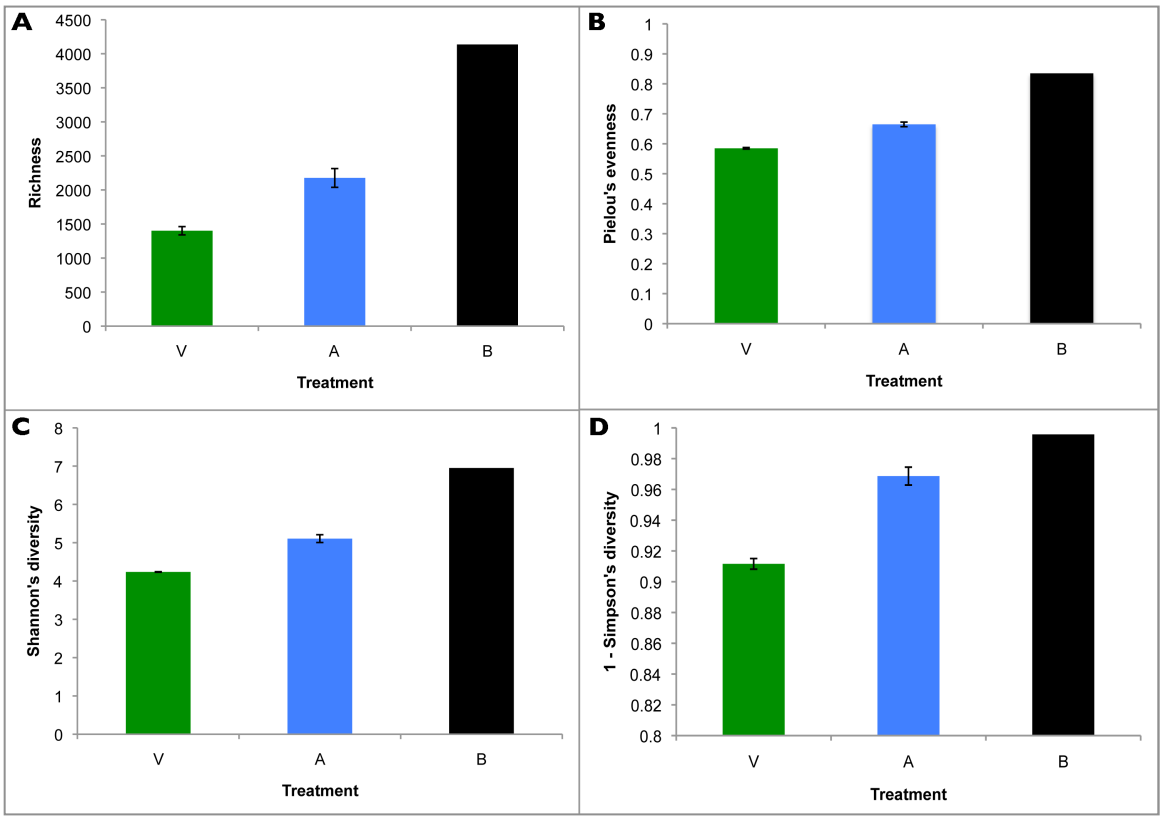


Figure 10: Unifrac hierarchical clustering of sample community data.

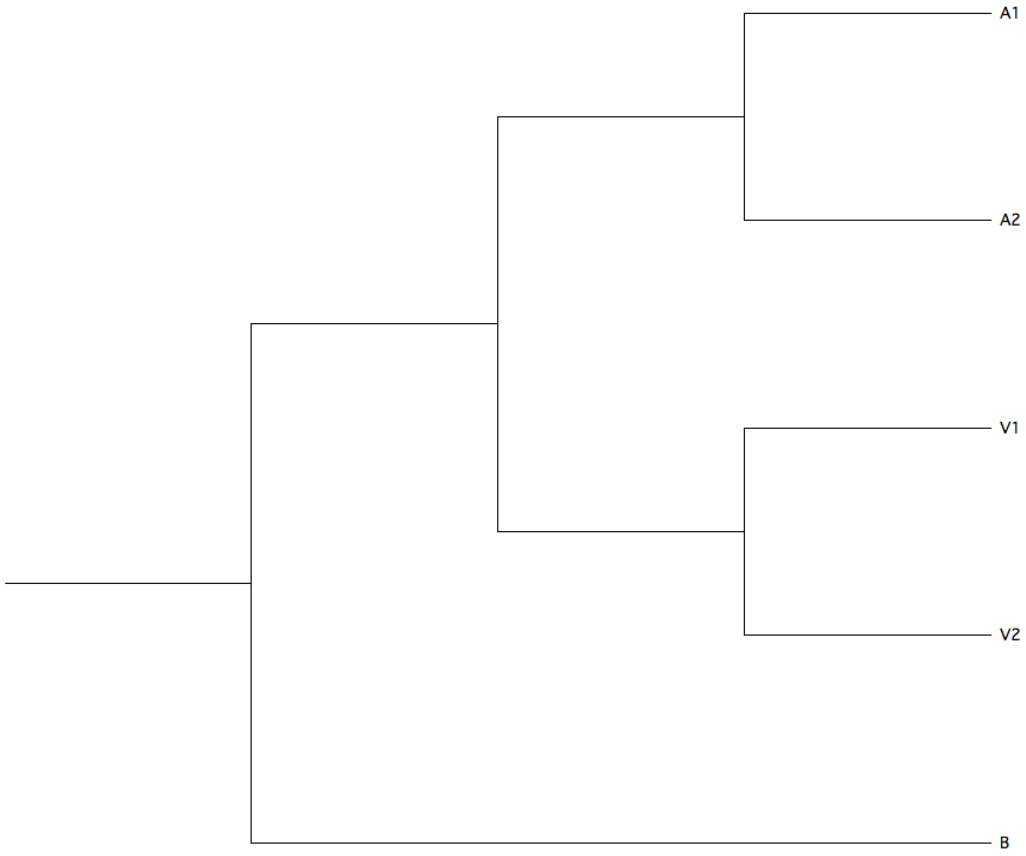


Figure 11: OTU overlap. 11A: OTUs unique to each community and shared between samples of the same treatment at a 97% nucleotide identity OTU designation. 11B: OTUs unique to each community and shared between samples of the same treatment at a 90% nucleotide identity OTU designation.

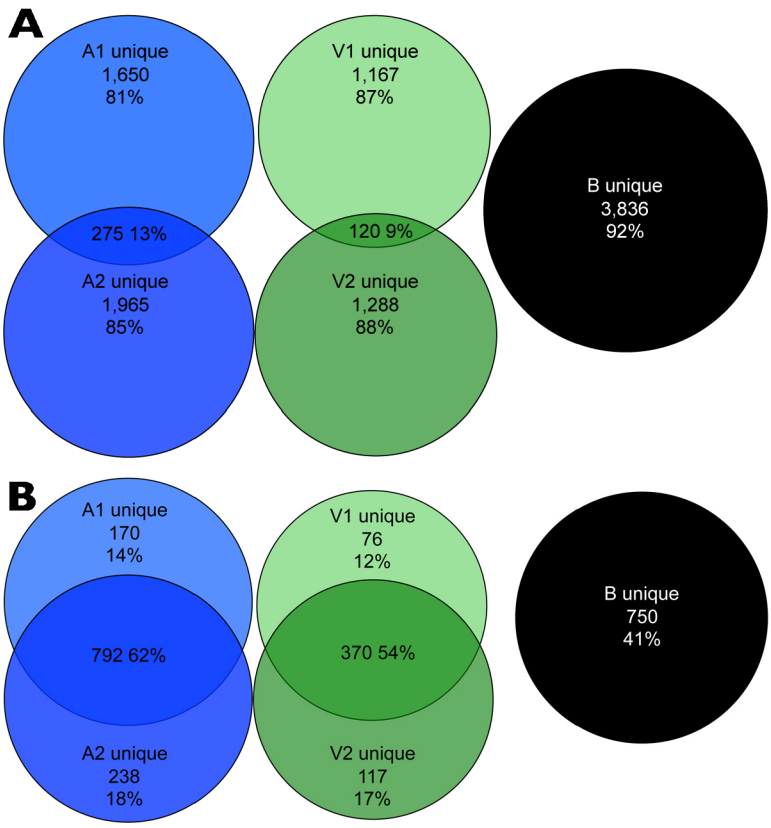


Figure 12: Evidence of clustering. 12A Faith's phylogenetic diversity. 12B Abundance weighted net relatedness index. 12C Nearest taxon index with no abundance weighting. 12D Abundance weighted nearest taxon index. Black is background, dark blue is A2, blue is A1, dark green is V2, and green is V1.

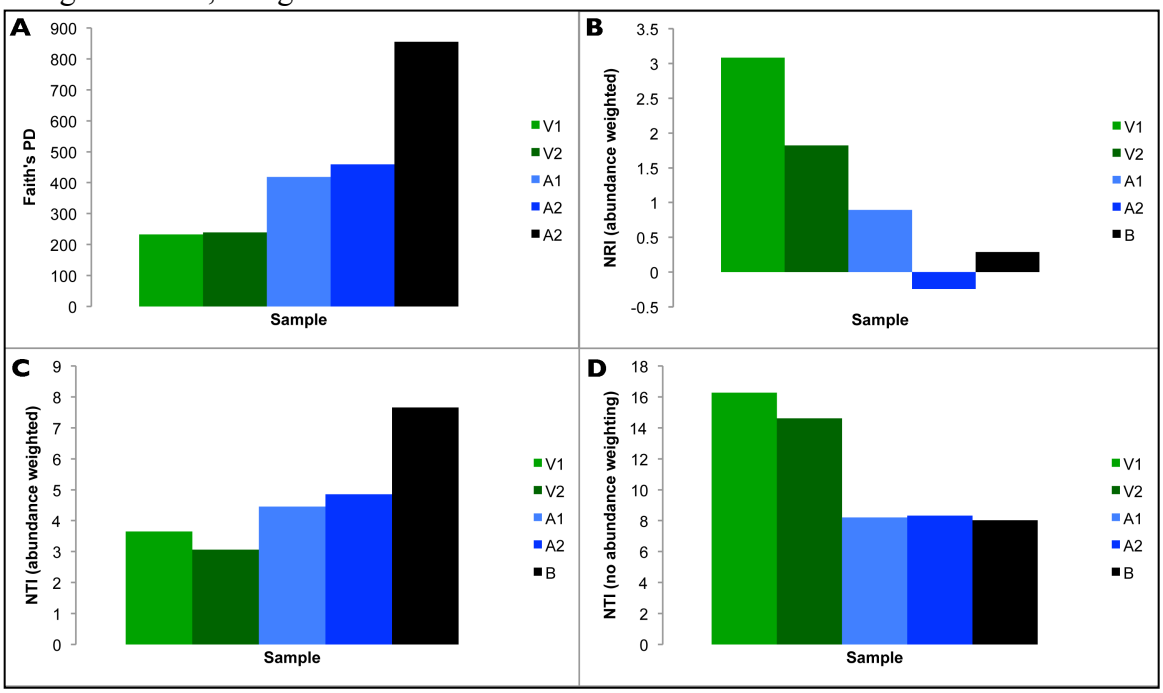


Figure 13: Average estimated read coverage of each OTU 16S rRNA gene compared to relative abundance in the sample. The lines indicate the relative abundance at approximately 10x read coverage, 0.00003.

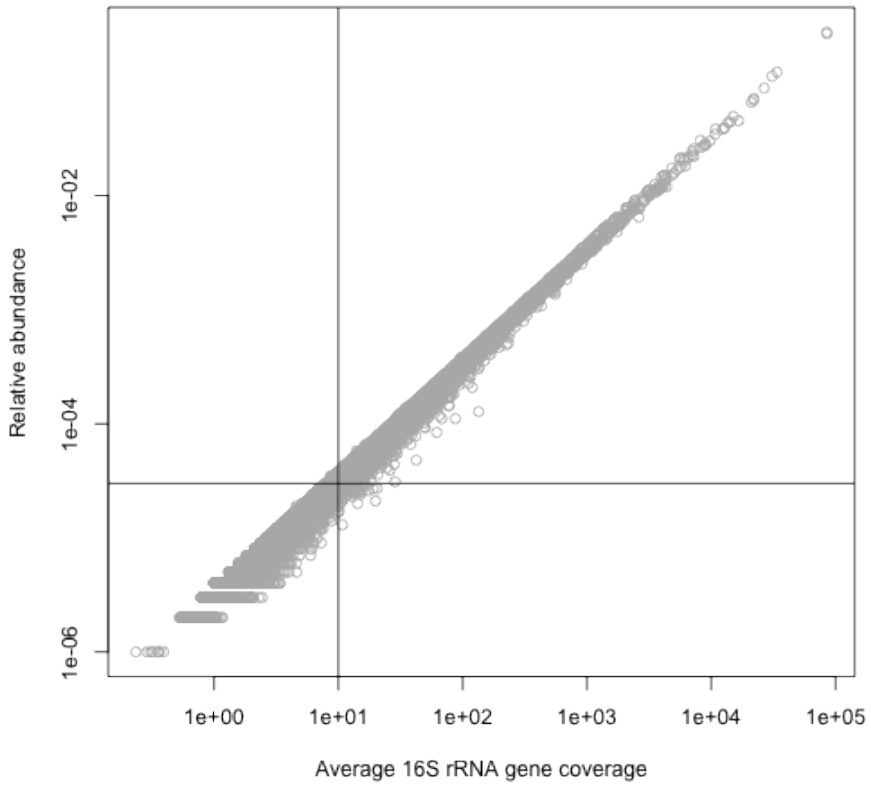
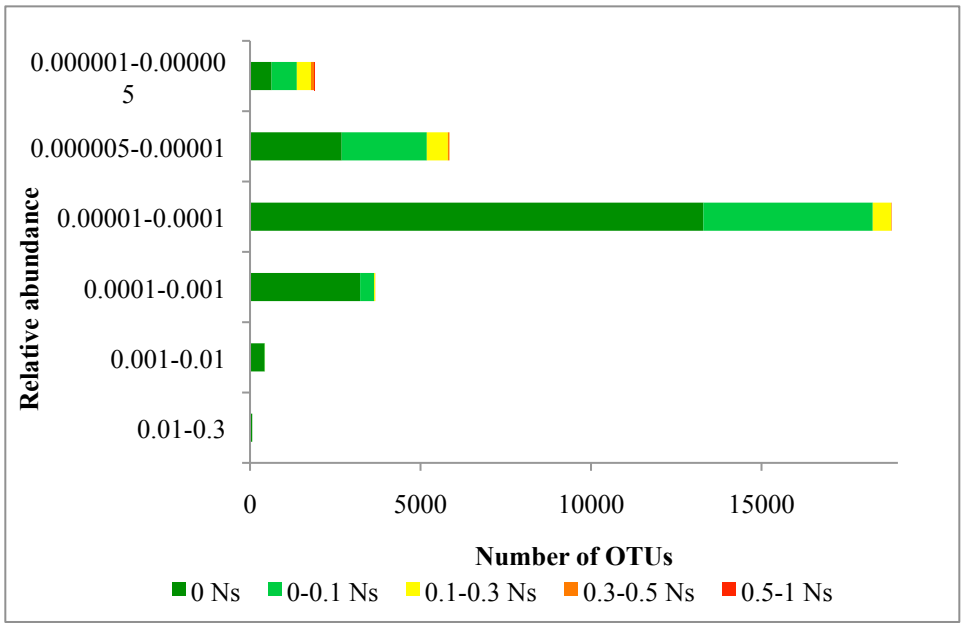


Figure 14: Distribution of proportional number of Ns per 16S rRNA gene compared to relative abundance in the sample.



Supplementary materials

Table S1: Standard diversity measures for each column community. Standard measures were calculated according to Jost, 2006 [200].

Sample	V1	V2	A1	A2	B
Richness	2,556	2,884	5,190	5,860	12,048
Shannon	4.32	4.34	5.24	5.46	7.38
1 - Simpson's	0.92	0.91	0.96	0.98	1.00

Table S2: Proportional standard diversity measures for each column community. ave indicates the average value for the treatment samples.

	V ave/ B	A ave/ B	V ave/ A ave
Richness	0.23	0.46	0.49
Shannon	0.59	0.73	0.81
1 - Simpson's	0.92	0.97	0.94

Table S3: Bray-Curtis dissimilarity between column communities.

	V1	V2	A1	A2
V2	0.78			
A1	0.88	0.71		
A2	0.87	0.71	0.60	
B	0.95	0.95	0.91	0.92

Figure S1: In-well flow-through injection column design. Arrows indicate the direction of flow. The clear rectangles represent tubing and the brown grains represent the sediment in the column.

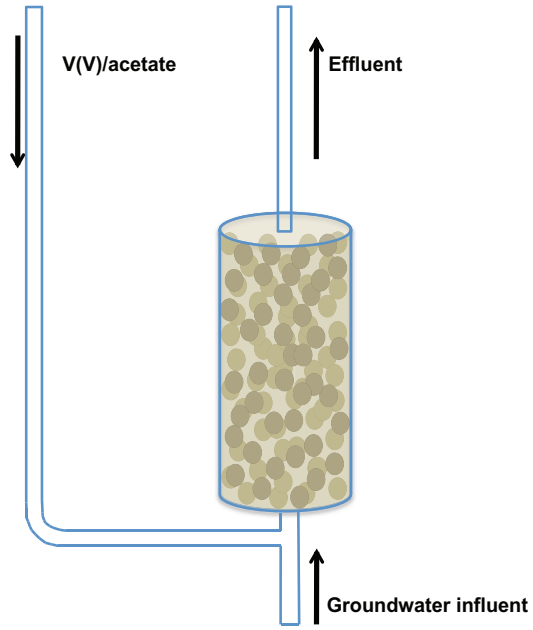
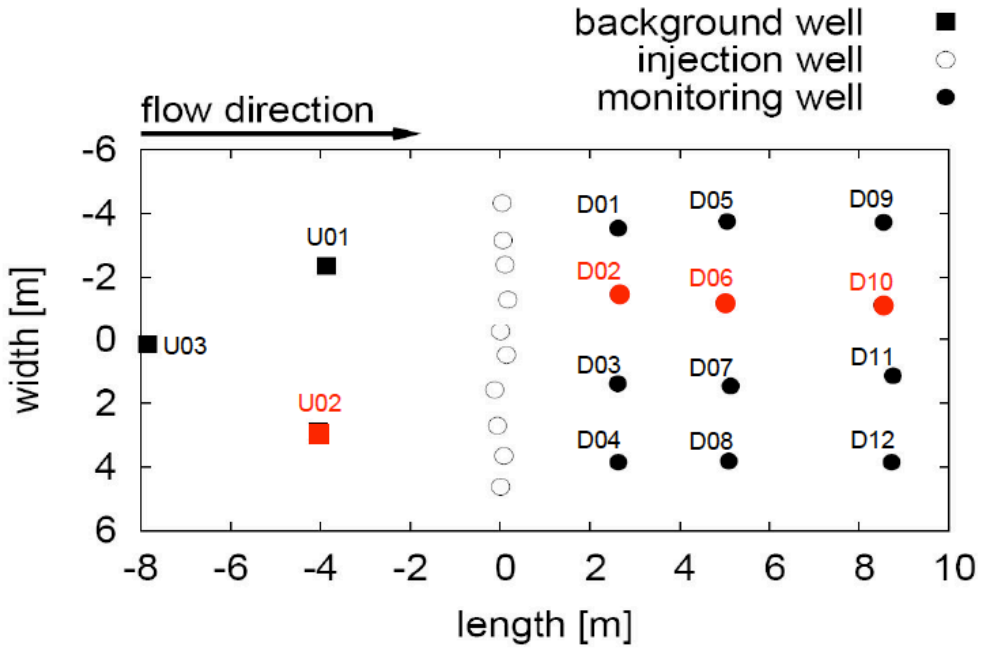


Figure S2: Orientation of bore wells at the Old Rifle Mill, CO, USA.



REFERENCES

1. **Silverman M, Ehrlich, HL: Microbial formation and degradation of minerals. In: *Adv Appl Microbiol.* Edited by Umbreit W, vol. 6. New York: Academic Press; 1964: 153-206.**
2. **Nordstrom D, Southam, G: Geomicrobiology of sulfide mineral oxidation. In: *Geomicrobiology: Interactions between Microbes and Minerals.* Edited by Nealson JBaK, vol. 35. Washington D.C.: Mineralogical Society of America; 1997: 361-390.**
3. **Baker BJ, Banfield JF: Microbial communities in acid mine drainage. *FEMS Microbiol Ecol* 2003, 44(2):139-152.**
4. **Denef VJ, Mueller RS, Banfield JF: AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME J* 2010, 4(5):599-610.**
5. **Justice N, Pan, C, Mueller, R, Spaulding, SE, Shah, V, Sun, CS, Yelton, AP, Miller, CS, Thomas, BC, Shah, M, VerBerkmoes, N, Hettich, R, Banfield, JF: Heterotrophic archaea contribute to carbon cycling in low-pH, suboxic biofilm communities. *Appl Environ Microbiol* 2012, 78:8321-8330.**
6. **Tyson GW, Lo I, Baker BJ, Allen EE, Hugenholtz P, Banfield JF: Genome-directed isolation of the key nitrogen fixer Leptospirillum ferrodiazotrophum sp nov from an acidophilic microbial community. *Appl Environ Microbiol* 2005, 71(10):6319-6324.**
7. **Dopson M, Baker-Austin C, Bond P: Towards determining details of anaerobic growth coupled to ferric iron reduction by the acidophilic archaeon '*Ferroplasma acidarmanus*' Fer1. *Extremophiles* 2007, 11(1):159-168.**
8. **Dopson M, Baker-Austin C, Bond PL: Analysis of differential protein expression during growth states of *Ferroplasma* strains and insights into electron transport for iron oxidation. *Microbiology* 2005, 151:4127-4137.**
9. **Dopson M, Baker-Austin C, Hind A, Bowman JP, Bond PL: Characterization of *Ferroplasma* isolates and *Ferroplasma acidarmanus* sp nov., extreme acidophiles from acid mine drainage and industrial bioleaching environments. *Appl Environ Microbiol* 2004, 70(4):2079-2088.**
10. **Bond PL, Dopson M, Baker-Austin C: Versatile metabolic capabilities and changing growth rates are important to development of bioleaching biofilms of Ferroplasma spp. *Abstracts of Papers of the American Chemical Society* 2003, 225:U916-U916.**
11. **Gihring TM, Bond PL, Peters SC, Banfield JF: Arsenic resistance in the archaeon "*Ferroplasma acidarmanus*": new insights into the structure and evolution of the ars genes. *Extremophiles* 2003, 7(2):123-130.**
12. **Edwards KJ, Bond PL, Gihring TM, Banfield JF: An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. *Science* 2000, 287(5459):1796-1799.**
13. **Okibe N, Gericke, M, Halberg, KB, Johnson, DB: Enumeration and characterization of acidophilic microorganisms isolated from a pilot plant stirred-tank bioleaching operation. *Appl Environ Microbiol* 2003, 69:1936-1943.**

14. **Seyrig G: Uranium bioremediation: current knowledge and trends. *Basic Biotechnology* 2010, 6:19-24.**
15. **Wall JD, Krumholz, L.: Uranium reduction. *Annu Rev Microbiol* 2006, 60:149-166.**
16. **Cardenas E, Wu, W., Leigh, M., Carley, J., Carroll, S., Gentry, T., Luo, J., Watson, D., Gu, B., Ginder-Vogel, M.: Microbial communities in contaminated sediments, associated with bioremediation of uranium to submicromolar levels. *Appl Environ Microbiol* 2008, 74:3718-3729.**
17. **Wu WM, Carley, J., Gentry, T., Ginder-Vogel, M., Fienen, M., Mehlhorn, T., Yan, H., Caroll, S., Pace, M., Nyman, J.: Pilot-scale in situ bioremedation of uranium in a highly contaminated aquifer. 2. Reduction of U(VI) and geochemical control of U(VI) bioavailability. *Environ Sci Technol* 2006, 40:3986-3995.**
18. **Ortiz-Bernad I, Anderson RT, Vrionis HA, Lovley DR: Vanadium respiration by *Geobacter metallireducens*: Novel strategy for *in situ* removal of vanadium from groundwater. *Appl Environ Microbiol* 2004, 70(5):3091-3095.**
19. **Carpentier W, De Smet L, Van Beeumen J, Brige A: Respiration and growth of *Shewanella oneidensis* MR-1 using vanadate as the sole electron acceptor. *J Bacteriol* 2005, 187(10):3293-3301.**
20. **Carpentier W, Sandra K, De Smet I, Brige A, De Smet L, Van Beeumen J: Microbial reduction and precipitation of vanadium by *Shewanella oneidensis*. *Appl Environ Microbiol* 2003, 69(6):3636-3639.**
21. **Van Marwijk J, Opperman DJ, Piater LA, Van Heerden E: Reduction of vanadium(V) by *Enterobacter cloacae* EV-SA01 isolated from a South African deep gold mine. *Biotechnol Lett* 2009, 31(6):845-849.**
22. **Ogg CD, Patel BKC: *Fervidicella metallireducens* gen. nov., sp nov., a thermophilic, anaerobic bacterium from geothermal waters. *Int J Syst Evol Microbiol* 2010, 60:1394-1400.**
23. **Friedberg I: Automated protein function prediction--the genomic challenge. *Brief Bioinform* 2006, 7(3):225-242.**
24. **Suyama M, Bork P: Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends in genetics : TIG* 2001, 17(1):10-13.**
25. **Dandekar T, Snel B, Huynen M, Bork P: Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 1998, 23(9):324-328.**
26. **von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B: STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 2003, 31(1):258-261.**
27. **Huynen M, Snel B, Lathe W, 3rd, Bork P: Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* 2000, 10(8):1204-1210.**
28. **Karimpour-Fard A, Leach SM, Gill RT, Hunter LE: Predicting protein linkages in bacteria: Which method is best depends on task. *BMC Bioinformatics* 2008, 9.**
29. **Rogozin IB, Makarova KS, Wolf YI, Koonin EV: Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. *Brief Bioinform* 2004, 5(2):131-149.**
30. **Mushegian AK, EV: Gene order is not conserved in bacterial evolution. *Trends Genet* 1996, 12:289-290.**

31. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* 1999, 96(6):2896-2901.
32. Overbeek R, Larsen N, Walunas T, D'Souza M, Pusch G, Selkov E, Jr., Liolios K, Joukov V, Kaznadzey D, Anderson I *et al*: The ERGO genome analysis and discovery system. *Nucleic Acids Res* 2003, 31(1):164-171.
33. Korbel JO, Jensen LJ, von Mering C, Bork P: Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol* 2004, 22(7):911-917.
34. Huynen MA, Snel B: Gene and context: integrative approaches to genome analysis. *Adv Protein Chem* 2000, 54:345-379.
35. Snel B, Lehmann G, Bork P, Huynen MA: STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* 2000, 28(18):3442-3444.
36. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 2004, 428(6978):37-43.
37. Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF: Genome dynamics in a natural archaeal population. *Proc Natl Acad Sci U S A* 2007, 104(6):1883-1888.
38. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P: Prediction of effective genome size in metagenomic samples. *Genome Biol* 2007, 8(1).
39. Rocha E: Inference and analysis of the relative stability of bacterial chromosomes. *Mol Biol Evol* 2006, 23:513-522.
40. Maddison W: Testing character correlation using pairwise comparisons on a phylogeny. *J Theor Biol* 2000, 202:195-204.
41. Lai D, Lluncor B, Schroder I, Gunsalus RP, Liao JC, Monbouquette HG: Reconstruction of the archaeal isoprenoid ether lipid biosynthesis pathway in Escherichia coli through digeranylgeranylglyceryl phosphate. *Metab Eng* 2009, 11(3):184-191.
42. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV: A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* 2006, 1:7.
43. Woodson JD, Escalante-Semerena JC: CbiZ, an amidohydrolase enzyme required for salvaging the coenzyme B12 precursor cobinamide in archaea. *Proc Natl Acad Sci U S A* 2004, 101(10):3591-3596.
44. Bevers LE, Hagedoorn PL, Santamaria-Araujo JA, Magalon A, Hagen WR, Schwarz G: Function of MoaB proteins in the biosynthesis of the molybdenum and tungsten cofactors. *Biochemistry (Mosc)* 2008, 47(3):949-956.
45. Proudfoot M, Kuznetsova E, Brown G, Rao NN, Kitagawa M, Mori H, Savchenko A, Yakunin AF: General enzymatic screens identify three new nucleotidases in Escherichia coli. Biochemical characterization of SurE, YfbR, and YjjG. *The Journal of biological chemistry* 2004, 279(52):54687-54694.

46. Shah SA, Garrett RA: CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res Microbiol* 2011, 162(1):27-38.
47. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Yelton AP, Banfield JF: Community-wide analysis of microbial genome sequence signatures. *Genome Biol* 2009, 10(8):50.
48. Konstantinidis KT, Tiedje JM: Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* 2005, 187(18):6258-6264.
49. Brenner SE, Chothia C, Hubbard TJ: Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A* 1998, 95(11):6073-6078.
50. Felsenstein J: Phylogenies and the Comparative Method. *Am Nat* 1985, 125:1-15.
51. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ: Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010, 11:119.
52. Kari L, Goheen MM, Randall LB, Taylor LD, Carlson JH, Whitmire WM, Virok D, Rajaram K, Endresz V, McClarty G *et al*: Generation of targeted Chlamydia trachomatis null mutants. *Proc Natl Acad Sci U S A* 2011, 108(17):7189-7193.
53. Ogata H, Goto S, Fujibuchi W, Kanehisa M: Computation with the KEGG pathway database. *Bio Systems* 1998, 47(1-2):119-128.
54. Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE: *De novo* metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J* 2012, 6(1):81-93.
55. Guo L, Brugger K, Liu C, Shah SA, Zheng HJ, Zhu YQ, Wang SY, Lillestol RK, Chen LM, Frank J *et al*: Genome Analyses of Icelandic Strains of *Sulfolobus islandicus*, Model Organisms for Genetic and Virus-Host Interaction Studies. *J Bacteriol* 2011, 193(7):1672-1680.
56. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ: Biogeography of the *Sulfolobus islandicus* pan-genome. *Proc Natl Acad Sci U S A* 2009, 106(21):8605-8610.
57. Whitaker RJ, Grogan DW, Taylor JW: Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* 2003, 301(5635):976-978.
58. Inskeep WP, Rusch DB, Jay ZJ, Herrgard MJ, Kozubal MA, Richardson TH, Macur RE, Hamamura N, Jennings RD, Fouke BW *et al*: Metagenomes from High-Temperature Chemotrophic Systems Reveal Geochemical Controls on Microbial Community Structure and Function. *PLoS One* 2010, 5(3).
59. Hansen EE, Lozupone CA, Rey FE, Wu M, Guruge JL, Narra A, Goodfellow J, Zaneveld JR, McDonald DT, Goodrich JA *et al*: Pan-genome of the dominant human gut-associated archaeon, *Methanobrevibacter smithii*, studied in twins. *Proc Natl Acad Sci U S A* 2011, 108:4599-4606.
60. Huber H, Stetter KO: Thermoplasmatales: Springer, 233 Spring Street, New York, NY 10013, United States; 2006.
61. Zhou H, Zhang R, Hu P, Zeng W, Xie Y, Wu C, Qiu G: Isolation and characterization of *Ferroplasma thermophilum* sp. nov., a novel extremely acidophilic, moderately thermophilic archaeon and its role in bioleaching of chalcopyrite. *J Appl Microbiol* 2008, 105(2):591-601.

62. **Golyshina OV, Pivovarova TA, Karavaiko GI, Kondrat'eva TF, Moore ERB, Abraham WR, Lunsdorf H, Timmis KN, Yakimov MM, Golyshin PN: *Ferroplasma acidiphilum* gen. nov., sp nov., an acidophilic, autotrophic, ferrous-iron-oxidizing, cell-wall-lacking, mesophilic member of the *Ferroplasmaceae* fam. nov., comprising a distinct lineage of the Archaea. *Int J Syst Evol Microbiol* 2000, 50:997-1006.**
63. **Darland G, Brock TD, Samsonof.W, Conti SF: Thermophilic, acidophilic mycoplasma isolated from a coal refuse pile. *Science* 1970, 170(3965):1416-&.**
64. **Schleper C, Puhler G, Klenk HP, Zillig W: *Picrophilus oshimae* and *Picrophilus torridus* fam nov, gen nov, sp nov, two species of hyperacidophilic, thermophilic, heterotrophic, aerobic archaea. *Int J Syst Bacteriol* 1996, 46(3):814-816.**
65. **Hawkes RB, Franzmann PD, O'Hara G, Plumb JJ: *Ferroplasma cupricumulans* sp nov., a novel moderately thermophilic, acidophilic archaeon isolated from an industrial-scale chalcocite bioleach heap. *Extremophiles* 2006, 10(6):525-530.**
66. **Itoh T, Yoshikawa N, Takashina T: *Thermogymnomonas acidicola* gen. nov., sp nov., a novel thermoacidophilic, cell wall-less archaeon in the order *Thermoplasmatales*, isolated from a solfataric soil in Hakone, Japan. *Int J Syst Evol Microbiol* 2007, 57:2557-2561.**
67. **Jones DS, Albrecht HL, Dawson KS, Schaperdoth I, Freeman KH, Pi YD, Pearson A, Macalady JL: Community genomic analysis of an extremely acidophilic sulfur-oxidizing biofilm. *ISME J* 2012, 6(1):158-170.**
68. **Yelton AP, Thomas BC, Simmons SL, Wilmes P, Zemla A, Thelen MP, Justice N, Banfield JF: A Semi-Quantitative, Synteny-Based Method to Improve Functional Predictions for Hypothetical and Poorly Annotated Bacterial and Archaeal Genes. *PLoS Comp Biol* 2011, 7(10).**
69. **Durbin AM, Teske A: Archaea in organic-lean and organic-rich marine subsurface sediments: an environmental gradient reflected in distinct phylogenetic lineages. *Front Microbiol* 2012, 3:168-168.**
70. **Kato S, Itoh T, Yamagishi A: Archaeal diversity in a terrestrial acidic spring field revealed by a novel PCR primer targeting archaeal 16S rRNA genes. *FEMS Microbiol Lett* 2011, 319(1):34-43.**
71. **Pace NR: Mapping the Tree of Life: Progress and Prospects. *Microbiol Mol Biol Rev* 2009, 73(4):565-576.**
72. **Eppley JM, Tyson GW, Getz WM, Banfield JF: Genetic exchange across a species boundary in the archaeal genus *Ferroplasma*. *Genetics* 2007, 177(1):407-416.**
73. **Denef VJ, Kalnejais LH, Mueller RS, Wilmes P, Baker BJ, Thomas BC, VerBerkmoes NC, Hettich RL, Banfield JF: Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proc Natl Acad Sci U S A* 2010, 107(6):2383-2390.**
74. **Niwa H, Tsuchiya D, Makyio H, Yoshida M, Morikawa K: Hexameric ring structure of the ATPase domain of the membrane-integrated metalloprotease FtsH from *Thermus thermophilus* HB8. *Structure* 2002, 10(10):1415-1423.**
75. **Futterer O, Angelov A, Liesegang H, Gottschalk G, Schleper C, Schepers B, Dock C, Antranikian G, Liebl W: Genome sequence of *Picrophilus torridus* and its implications for life around pH 0. *Proc Natl Acad Sci U S A* 2004, 101(24):9091-9096.**

76. Reysenbach AL, Flores GE: Electron microscopy encounters with unusual thermophiles helps direct genomic analysis of *Aciduliprofundum boonei*. *Geobiology* 2008, 6(3):331-336.
77. Abu-Qarn M, Yurist-Doutsch S, Giordano A, Trauner A, Morris HR, Hitchen P, Medalia O, Dell A, Eichler J: *Haloferax volcanii* AglB and AglD are involved in N-glycosylation of the S-layer glycoprotein and proper assembly of the surface layer. *J Mol Biol* 2007, 374(5):1224-1236.
78. Kosma P, Wugeditsch T, Christian R, Zayni S, Messner P: Glycan structure of a heptose-containing S-layer glycoprotein of *Bacillus thermoaerophilus*. *Glycobiology* 1995, 5(8):791-796.
79. Wugeditsch T, Zachara NE, Puchberger M, Kosma P, Gooley AA, Messner P: Structural heterogeneity in the core oligosaccharide of the S-layer glycoprotein from *Aneurinibacillus thermoaerophilus* DSM 10155. *Glycobiology* 1999, 9(8):787-795.
80. Valvano MA, Messner P, Kosma P: Novel pathways for biosynthesis of nucleotide-activated glycero-manno-heptose precursors of bacterial glycoproteins and cell surface polysaccharides. *Microbiology-Sgm* 2002, 148:1979-1989.
81. Johnson DB, Hallberg KB: The microbiology of acidic mine waters. *Res Microbiol* 2003, 154(7):466-473.
82. Fowler TA, Holmes PR, Crundwell FK: Mechanism of pyrite dissolution in the presence of *Thiobacillus ferrooxidans*. *Appl Environ Microbiol* 1999, 65(7):2987-2993.
83. Rawlings DE: Characteristics and adaptability of iron- and sulfur-oxidizing microorganisms used for the recovery of metals from minerals and their concentrates. In: *Microb Cell Fact.* vol. 4; 2005.
84. Cobley JG, Haddock BA: Respiratory chain of *Thiobacillus ferrooxidans*: Reduction of cytochromes by Fe2+ and preliminary characterization of rusticyanin, a novel blue copper protein. *FEBS Lett* 1975, 60(1):29-33.
85. Ambler RP: Blue copper proteins as honorary cytochromes: The structure and evolution of blue copper proteins. *J Chem Soc Pak* 1999, 21(3):213-228.
86. Redinbo MR, Yeates TO, Merchant S: Plastocyanin: Structural and functional analysis. *J Bioenerg Biomembr* 1994, 26(1):49-66.
87. Gough J, Chothia C: The linked conservation of structure and function in a family of high diversity: The monomeric cupredoxins. *Structure* 2004, 12(6):917-925.
88. Murphy MEP, Lindley PF, Adman ET: Structural comparison of cupredoxin domains: Domain recycling to construct proteins with novel functions. *Protein Sci* 1997, 6(4):761-770.
89. Giri AV, Anishetty S, Gautam P: Functionally specified protein signatures distinctive for each of the different blue copper proteins. *BMC Bioinformatics* 2004, 5.
90. Elbehti A, Nitschke W, Tron P, Michel C, Lemesle-Meunier D: Redox components of cytochrome bc-type enzymes in acidophilic prokaryotes I. Characterization of the cytochrome bc(1)-type complex of the acidophilic ferrous ion-oxidizing bacterium *Thiobacillus ferrooxidans*. *J Biol Chem* 1999, 274(24):16760-16765.
91. Hiller A, Henninger T, Schafer G, Schmidt CL: New genes encoding subunits of a cytochrome bc1 - Analogous complex in the respiratory chain of the

hyperthermoacidophilic crenarchaeon *Sulfolobus acidocaldarius*. *J Bioenerg Biomembr* 2003, 35(2):121-131.

92. Dinarieva TY, Zhuravleva AE, Pavlenko OA, Tsaplina IA, Netrusov AI: Ferrous iron oxidation in moderately thermophilic acidophile *Sulfobacillus sibiricus* N1(T). *Can J Microbiol* 2010, 56(10):803-808.

93. Yarzabal A, Appia-Ayme C, Ratouchniak J, Bonnefoy V: Regulation of the expression of the *Acidithiobacillus ferrooxidans* rus operon encoding two cytochromes c, a cytochrome oxidase and rusticyanin. *Microbiology* 2004, 150:2113-2123.

94. King GA: Molecular and culture-based analyses of aerobic carbon monoxide oxidizer diversity. *Appl Environ Microbiol* 2003, 69(12):7257-7265.

95. Cunliffe M: Correlating carbon monoxide oxidation with cox genes in the abundant Marine *Roseobacter* Clade. *ISME J* 2011, 5(4):685-691.

96. Cardenas JP, Martinez V, Covarrubias P, Holmes DS, Quatrini R: Predicted CO/CO(2) fixation in Ferroplasma spp. via a novel chimaeric pathway. In: *Biohydrometallurgy: A Meeting Point between Microbial Ecology, Metal Recovery Processes and Environmental Remediation.* Edited by Donati ER, Viera MR, Tavani EL, Giaveno MA, Lavalle TL, Chiacchiarini PA, vol. 71-73; 2009: 219-222.

97. Ruepp A, Graml W, Santos-Martinez ML, Koretle KK, Volker C, Mewes HW, Frishman D, Stocker S, Lupas AN, Baumeister W: The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* 2000, 407(6803):508-513.

98. Segerer A, Langworthy TA, Stetter KO: *Thermoplasma acidophilum* and *Thermoplasma volcanium* sp nov from solfatara fields. *Syst Appl Microbiol* 1988, 10(2):161-171.

99. Kletzin A: General Characteristics and Important Model Organisms, vol. 1. Washington D.C.: ASM Press; 2007.

100. Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV: Evolution of mosaic operons by horizontal gene transfer and gene displacement *in situ*. *Genome Biol* 2003, 4(9).

101. Lemos RS, Fernandes AS, Pereira MM, Gomes CM, Teixeira M: Quinol : fumarate oxidoreductases and succinate : quinone oxidoreductases: phylogenetic relationships, metal centres and membrane attachment. *Biochim Biophys Acta* 2002, 1553(1-2):158-170.

102. Schroder I, and De Vries, S.: Archaea: new models for prokaryotic biology, vol. 1. Norfolk, UK: Caister Academic Press; 2008.

103. Paumann M, Lubura B, Regelsberger G, Feichtinger M, Kollensberger G, Jakopitsch C, Furtmuller PG, Peschek GA, Obinger C: Soluble Cu(A) domain of cyanobacterial cytochrome c oxidase. *J Biol Chem* 2004, 279(11):10293-10303.

104. Garciahorsman JA, Barquera B, Rumbley J, Ma JX, Gennis RB: The superfamily of heme-copper respiratory oxidases. *J Bacteriol* 1994, 176(18):5587-5600.

105. Ma S, Banfield JF: Micron-scale Fe(2+)/Fe(3+), intermediate sulfur species and O(2) gradients across the biofilm-solution-sediment interface control biofilm organization. *Geochim Cosmochim Acta* 2011, 75(12):3568-3580.

106. Huang CJ, Barrett EL: Sequence analysis and expression of the *Salmonella typhimurium* asr operon encoding production of hydrogen sulfide from sulfite. *J Bacteriol* 1991, 173(4):1544-1553.

107. Kawashima T, Yokoyama K, Higuchi S, Suzuki M: Identification of proteins present in the archaeon *Thermoplasma volcanium* cultured in aerobic or anaerobic conditions. *Proceedings of the Japan Academy Series B-Physical and Biological Sciences* 2005, 81(6):204-219.

108. Sun N, Pan CP, Nickell S, Mann M, Baumeister W, Nagy I: Quantitative Proteome and Transcriptome Analysis of the Archaeon Thermoplasma acidophilum Cultured under Aerobic and Anaerobic Conditions. *J Proteome Res* 2010, 9(9):4839-4850.

109. Golovacheva RS, Golyshina OV, Karavaiko GI, Dorofeev AG, Pivovarova TA, Chernykh NA: A new iron-oxidizing bacterium, *Leptospirillum thermoferrooxidans* sp. nov. *Microbiology* 1992, 61(6):744-750.

110. Reher M, Schonheit P: Glyceraldehyde dehydrogenases from the thermoacidophilic euryarchaeota *Picrophilus torridus* and *Thermoplasma acidophilu*m, key enzymes of the non-phosphorylative Entner-Doudoroff pathway, constitute a novel enzyme family within the aldehyde dehydrogenase superfamily. *FEBS Lett* 2006, 580(5):1198-1204.

111. Budgen N, Danson MJ: Metabolism of Glucose Via a Modified Entner-Doudoroff Pathway in the Thermoacidophilic Archaebacterium T*hermoplasma acidophilum*. *FEBS Lett* 1986, 196(2):207-210.

112. Derosa M, Gambacorta A, Nicolaus B, Giardina P, Poerio E, Buonocore V: Glucose-Metabolism in the Extreme Thermoacidophilic Archaebacterium *Sulfolobus solfataricus*. *Biochem J* 1984, 224(2):407-414.

113. Siebers B, Tjaden B, Michalke K, Dorr C, Ahmed H, Zaparty M, Gordon P, Sensen CW, Zibat A, Klenk HP *et al*: Reconstruction of the central carbohydrate metabolism of *Thermoproteus tenax* by use of genomic and biochemical data. *J Bacteriol* 2004, 186(7):2179-2194.

114. Reher M, Fuhrer T, Bott M, Schoenheit P: The nonphosphorylative Entner-Doudoroff pathway in the thermoacidophilic euryarchaeon *Picrophilus torridus* involves a novel 2-keto-3-deoxygluconate-specific aldolase. *J Bacteriol* 2010, 192(4):964-974.

115. Swigonova Z, Mohsen AW, Vockley J: Acyl-CoA Dehydrogenases: Dynamic History of Protein Family Evolution. *J Mol Evol* 2009, 69(2):176-193.

116. Lidstrom ME: Aerobic Methylotrophic Prokaryotes, vol. 2. New York, NY, USA: Springer; 2006.

117. Baumler DJ, Hung KF, Jeong KC, Kaspar CW: Production of methanethiol and volatile sulfur compounds by the archaeon "*Ferroplasma acidarmanus*". *Extremophiles* 2007, 11(6):841-851.

118. Wilmes P, Remis JP, Hwang M, Auer M, Thelen MP, Banfield JF: Natural acidophilic biofilm communities reflect distinct organismal and functional organization. *ISME J* 2009, 3(2):266-270.

119. Schafer T, Selig M, Schonheit P: Acetyl-Coa Synthetase (Adp Forming) in Archaea, a Novel Enzyme Involved in Acetate Formation and Atp Synthesis. *Arch Microbiol* 1993, 159(1):72-83.

120. Brasen C, Schonheit P: Unusual ADP-forming acetyl-coenzyme A synthetases from the mesophilic halophilic euryarchaeon *Haloarcula marismortui* and from the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. *Arch Microbiol* 2004, 182(4):277-287.
121. Sakuraba H, Ohshima T: Novel energy metabolism in anaerobic hyperthermophilic archaea: A modified Embden-Meyerhof pathway. *J Biosci Bioeng* 2002, 93(5):441-448.
122. Siebers B, Schonheit P: Unusual pathways and enzymes of central carbohydrate metabolism in Archaea. *Curr Opin Microbiol* 2005, 8(6):695-705.
123. Brasen C, Schonheit P: Regulation of acetate and acetyl-CoA converting enzymes during growth on acetate and/or glucose in the halophilic archaeon *Haloarcula marismortui*. *FEMS Microbiol Lett* 2004, 241(1):21-26.
124. Kim YJ, Lee HS, Kim ES, Bae SS, Lim JK, Matsumi R, Lebedinsky AV, Sokolova TG, Kozhevnikova DA, Cha SS *et al*: Formate-driven growth coupled with H(2) production. *Nature* 2010, 467(7313):352-U137.
125. Lim JK, Kang SG, Lebedinsky AV, Lee JH, Lee HS: Identification of a novel class of membrane-bound NiFe-hydrogenases in *Thermococcus onnurineus* NA1 by *in silico* analysis. *Appl Environ Microbiol* 2010, 76(18):6286-6289.
126. Wu LF, Mandrand MA: Microbial hydrogenases: Primary structure, classification, signatures and phylogeny. *FEMS Microbiol Rev* 1993, 104(3-4):243-270.
127. Edwards KJ, Gihring TM, Schrenk MO, Hamers RJ, Banfield JF: Microbial populations and distribution at an extreme acid mine drainage environment: A study using fluorescent in-situ hybridization. *Abstr Gen Meet Am Soc Microbiol* 1998, 98:382.
128. Baker-Austin C, Dopson M, Wexler M, Sawers RG, Bond PL: Molecular insight into extreme copper resistance in the extremophilic archaeon 'Ferroplasma acidarmanus' Fer1. *Microbiology* 2005, 151:2637-2646.
129. Cooksey DA: Molecular mechanisms of copper resistance and accumulation in bacteria. *FEMS Microbiol Rev* 1994, 14(4):381-386.
130. Martins LO, Huber R, Huber H, Stetter KO, Da Costa MS, Santos H: Organic solutes in hyperthermophilic archaea. *Appl Environ Microbiol* 1997, 63(3):896-902.
131. Jarrell KF, Bayley DP, Florian V, Klein A: Isolation and characterization of insertional mutations in flagellin genes in the archaeon *Methanococcus voltae*. *Mol Microbiol* 1996, 20(3):657-666.
132. Thomas NA, Mueller S, Klein A, Jarrell KF: Mutants in flaI and flaJ of the archaeon *Methanococcus voltae* are deficient in flagellum assembly. *Mol Microbiol* 2002, 46(3):879-887.
133. Thomas NA, Pawson CT, Jarrell KF: Insertional inactivation of the flaH gene in the archaeon *Methanococcus voltae* results in non-flagellated cells. *Mol Genet Genomics* 2001, 265(4):596-603.
134. Patenge N, Berendes A, Engelhardt H, Schuster SC, Oesterhelt D: The fla gene cluster is involved in the biogenesis of flagella in *Halobacterium salinarum*. *Mol Microbiol* 2001, 41(3):653-663.
135. Walsby AE: Gas vesicles. *Annu Rev Plant Physiol Plant Mol Biol* 1975, 26:427-439.

136. Chu LJ, Chen MC, Setter J, Tsai YS, Yang HY, Fang XF, Ting YS, Shaffer SA, Taylor GK, von Haller PD *et al*: New Structural Proteins of *Halobacterium salinarum* Gas Vesicle Revealed by Comparative Proteomics Analysis. *J Proteome Res* 2011, 10(3):1170-1178.

137. Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, Land ML, VerBerkmoes NC, Hettich RL, Banfield JF: Enigmatic, ultrasmall, uncultivated Archaea. *Proc Natl Acad Sci U S A* 2010, 107(19):8806-8811.

138. Goltsman DSA, Denef VJ, Singer SW, VerBerkmoes NC, Lefsrud M, Mueller RS, Dick GJ, Sun CL, Wheeler KE, Zemla A *et al*: Community Genomic and Proteomic Analyses of Chemoautotrophic Iron-Oxidizing "*Leptospirillum rubarum*" (Group II) and "*Leptospirillum ferrodiazotrophum*" (Group III) Bacteria in Acid Mine Drainage Biofilms. *Appl Environ Microbiol* 2009, 75(13):4599-4615.

139. Zerbino DR, Birney E: Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* 2008, 18(5):821-829.

140. Gordon D, Abajian C, Green P: Consed: A graphical tool for sequence finishing. *Genome Res* 1998, 8(3):195-202.

141. Katoh K, Toh H: Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 2008, 9(4):286-298.

142. Katoh K, Misawa K, Kuma K, Miyata T: MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002, 30(14):3059-3066.

143. Price MN, Dehal PS, Arkin AP: FastTree 2-Approximately maximum-likelihood trees for large alignments. *PLoS One* 2010, 5(3):e9490.

144. Price MN, Dehal PS, Arkin AP: FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 2009, 26(7):1641-1650.

145. Stamatakis A: RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006, 22(21):2688-2690.

146. Lin K, Simossis VA, Taylor WR, Heringa J: A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* 2005, 21(2):152-159.

147. Drennan CL, Heo JY, Sintchak MD, Schreiter E, Ludden PW: Life on carbon monoxide: X-ray structure of *Rhodospirillum rubrum* Ni-Fe-S carbon monoxide dehydrogenase. *Proc Natl Acad Sci U S A* 2001, 98(21):11973-11978.

148. Rehder D: Bioinorganic vanadium chemistry, 2 edn. West Sussex: John Wiley & Sons, Ltd; 2008.

149. Robson RL, Eady RR, Richardson TH, Miller RW, Hawkins M, Postgate JR: The alternative nitrogenase of *Azotobacter chroococcum* is a vanadium enzyme. *Nature* 1986, 322(6077):388-390.

150. Sabbioni E, Pozzi G, Pintar A, Casella L, Garattini S: Cellular retention, cytotoxicity and morphological transformation by vanadium (IV) and vanadium(V) in BALB/3T3 cell-lines. *Carcinogenesis* 1991, 12(1):47-52.

151. Shi XL, Jiang HG, Mao Y, Ye JP, Saffiotti U: Vanadium(IV)-mediated free radical generation and related 2'-deoxyguanosine hydroxylation and DNA damage. *Toxicology* 1996, 106(1-3):27-38.

152. Cortizo AM, Bruzzone L, Molinuevo S, Etcheverry SB: A possible role of oxidative stress in the vanadium-induced cytotoxicity in the MC3T3E1 osteoblast and UMR106 osteosarcoma cell lines. *Toxicology* 2000, 147(2):89-99.

153. Valko MM, H.; Cronin M.T.D.: Metals, toxicity and oxidative stress. *Current Medical Chemistry* 2005, 12(10):1161-1208.

154. Burgot J-L: Ionic Equilibria in Analytical Chemistry, vol. 24: Springer; 2012.

155. Weber KA, Achenbach LA, Coates JD: Microorganisms pumping iron: anaerobic microbial iron oxidation and reduction. *Nature Reviews Microbiology* 2006, 4(10):752-764.

156. Woolfolk CA, Whiteley HR: Reduction of inorganic compounds with molecular hydrogen by *Micrococcus Lactilyticus*. 1. Stoichiometry with compounds of arsenic, selenium, tellurium, transition and other elements. *J Bacteriol* 1962, 84(4):647-&.

157. Yurkova NA, Lyalikova NN: New facultative chemolithotrophic bacteria reducing vanadate. *Microbiology* 1990, 59(6):672-677.

158. Bredberg K, Karlsson HT, Holst O: Reduction of vanadium(V) with *Acidithiobacillus ferrooxidans* and *Acidithiobacillus thiooxidans*. *Bioresour Technol* 2004, 92(1):93-96.

159. Lu SP, Ryu SH, Chung BS, Chung YR, Park W, Jeon CO: *Simplicispira limi* sp. nov., isolated from activated sludge. *Int J Syst Evol Microbiol* 2007, 57:31-34.

160. Grabovich M, Gavrish E, Kuever J, Lysenko AM, Podkopaeva D, Dubinina G: Proposal of *Giesbergeria voronezhensis* gen. nov., sp nov and G. kuznetsovii sp nov and reclassification of *Aquaspirillum anulus*, *A. sinuosum* and *A. giesbergeri* as *Giesbergeria anulus* comb. nov., *G. sinuosa* comb. nov and *G. giesbergeri* comb. nov., and *Aquaspirillum metamorphum* and *A. psychrophilum* as *Simplicispira metamorpha* gen. nov., comb. nov and *S. psychrophila* comb. nov. *Int J Syst Evol Microbiol* 2006, 56:569-576.

161. Terasaki Y: On two new species of *Spirillum*. *Botanical Magazine-Tokyo* 1961, 74:220-227.

162. Terasaki Y: Studies on the genus *Spirillum Ehrenberg*. I. Morphological, physiological, and biochemical characteristics of water spirilla. *Bull Suzugamine Women's Coll Nat Sci* 1973, 16:1-71.

163. Lovley DR, Giovannoni SJ, White DC, Champine JE, Phillips EJP, Gorby YA, Goodwin S: *Geobacter metallireducens* gen. nov. sp. nov., a microorganism capable of coupling the complete oxidation of organic compounds to the reduction of iron and other metals. *Arch Microbiol* 1993, 159(4):336-344.

164. Coates JD, Phillips EJP, Lonergan DJ, Jenter H, Lovley DR: Isolation of *Geobacter* species from diverse sedimentary environments. *Appl Environ Microbiol* 1996, 62(5):1531-1536.

165. Williams KH, Long PE, Davis JA, Wilkins MJ, N'Guessan AL, Steefel CI, Yang L, Newcomer D, Spane FA, Kerkhof LJ *et al*: Acetate availability and its influence on sustainable bioremediation of uranium-contaminated groundwater. *Geomicrobiol J* 2011, 28(5-6):519-539.

166. Sandell EB: Determination of chromium, vanadium, and molybdenum in silicate rocks. *Industrial and Engineering Chemistry-Analytical Edition* 1936, 8:336-341.

167. Harder J, Probian C, Wulfing A: Anaerobic mineralization of quaternary carbon atoms: Isolation of denitrifying bacteria on pivalic acid (2,2-dimethylpropionic acid). *Appl Environ Microbiol* 2003, 69(3):1866-1870.

168. Baysse C, De Vos D, Naudet Y, Vandermonde A, Ochsner U, Meyer JM, Budzikiewicz H, Schafer M, Fuchs R, Cornelis P: Vanadium interferes with siderophore-mediated iron uptake in *Pseudomonas aeruginosa*. *Microbiology-Uk* 2000, 146:2425-2434.

169. Handley KM, Wrighton KC, Piceno YM, Andersen GL, DeSantis TZ, Williams KH, Wilkins MJ, N'Guessan AL, Peacock A, Bargar J *et al*: High-density PhyloChip profiling of stimulated aquifer microbial communities reveals a complex response to acetate amendment. *FEMS Microbiol Ecol* 2012, 81(1):188-204.

170. Lowe M, Madsen EL, Schindler K, Smith C, Emrich S, Robb F, Halden RU: Geochemistry and microbial diversity of a trichloroethene-contaminated Superfund site undergoing intrinsic in situ reductive dechlorination. *FEMS Microbiol Ecol* 2002, 40(2):123-134.

171. The U.S. Environmental Protection Agency R, Hanford Project Office: USDOE Hanford Site: First Five Year Review Report. In. Edited by USDOE. Richland, WA; 2001.

172. Dibley V, Valett, J, Gregory, S, Madrid, V: Five-Year Review Report for the Building 834 Operable Unit at Lawrence Livermore National Laboratory Site 300. In. Edited by USDOE. Livermore, CA; 2007.

173. Antipov AN, Lyalikova NN, Khijniak TV, L'Vov NP: Vanadate reduction by molybdenum-free dissimilatory nitrate reductases from vanadate-reducing bacteria. *Iubmb Life* 2000, 50(1):39-42.

174. Myers JM, Antholine WE, Myers CR: Vanadium(V) reduction by Shewanella oneidensis MR-1 requires menaquinone and cytochromes from the cytoplasmic and outer membranes 2. *Appl Environ Microbiol* 2004, 70(3):1405-1412.

175. Soares SS, Martins H, Gutierrez-Merino C, Aureliano M: Vanadium and cadmium in vivo effects in teleost cardiac muscle: Metal accumulation and oxidative stress markers. *Comparative Biochemistry and Physiology Part C Toxicology & Pharmacology* 2008, 147(2):168-178.

176. Zychlinski L, Byczkowski, JZ, Kulkarni, AP: Toxic effects of long- term intratracheal administration of vanadium pentoxide in rats. *Arch Environ Contam Toxicol* 1991, 20:295-298.

177. Zaporowska H, Wasilewski W: Haematological effects of vanadium on living organisms. *Comparative Biochemistry and Physiology C Comparative Pharmacology and Toxicology* 1992, 102(2):223-231.

178. Stohs S, Bagchi, D.,: Oxidative mechanisms in the toxicity of metal ions. *Free Radic Biol Med* 1995, 18:321-336.

179. Domingo J: Vanadium: a review of the reproductive and development toxicity. *Reprod Toxicol* 1996, 10:175-182.

180. Crans DCA, S.S.; Keramidas, A.D.: Chemistry of relevance to vanadium in the environment. In: *Vanadium in the environment Part I: Chemistry and biochemistry.* Edited by Nriagu JO. New York, NY: John Wiley & Sons, Inc.; 1998: 73-95.

181. Peacock CL, Sherman DM: Vanadium(V) adsorption onto goethite (alpha-FeOOH) at pH 1.5 to 12: A surface complexation model based on ab initio molecular geometries and EXAFS spectroscopy. *Geochim Cosmochim Acta* 2004, 68(8):1723-1733.

182. Islam E, Dhal PK, Kazy SK, Sar P: Molecular analysis of bacterial communities in uranium ores and surrounding soils from Banduhurang open cast uranium mine, India: A comparative study. *Journal of Environmental Science and Health Part a-Toxic/Hazardous Substances & Environmental Engineering* 2011, 46(3):271-280.

183. Islam E, Sar P: Molecular assessment on impact of uranium ore contamination in soil bacterial diversity. *Int Biodeterior Biodegrad* 2011, 65(7):1043-1051.

184. Anderson RT, Vrionis HA, Ortiz-Bernad I, Resch CT, Long PE, Dayvault R, Karp K, Marutzky S, Metzler DR, Peacock A *et al*: Stimulating the in situ activity of *Geobacter* species to remove uranium from the groundwater of a uranium-contaminated aquifer. *Appl Environ Microbiol* 2003, 69(10):5884-5891.

185. Holmes DE, O'Neil RA, Vrionis HA, N'Guessan LA, Ortiz-Bernad I, Larrahondo MJ, Adams LA, Ward JA, Nicoll JS, Nevin KP *et al*: Subsurface clade of Geobacteraceae that predominates in a diversity of Fe(III)-reducing subsurface environments. *ISME J* 2007, 1(8):663-677.

186. North NN, Dollhopf SL, Petrie L, Istok JD, Balkwill DL, Kostka JE: Change in bacterial community structure during in situ Biostimulation of subsurface sediment cocontaminated with uranium and nitrate. *Appl Environ Microbiol* 2004, 70(8):4911-4920.

187. Chang YJ, Long PE, Geyer R, Peacock AD, Resch CT, Sublette K, Pfiffner S, Smithgall A, Anderson RT, Vrionis HA *et al*: Microbial incorporation of C-13-labeled acetate at the field scale: Detection of microbes responsible for reduction of U(VI). *Environ Sci Technol* 2005, 39(23):9039-9048.

188. Vrionis HA, Anderson RT, Ortiz-Bernad I, O'Neill KR, Resch CT, Peacock AD, Dayvault R, White DC, Long PE, Lovley DR: Microbiological and geochemical heterogeneity in an in situ uranium bioremediation field site. *Appl Environ Microbiol* 2005, 71(10):6308-6318.

189. Mohanty SR, Kollah B, Brodie EL, Hazen TC, Roden EE: 16S rRNA Gene Microarray Analysis of Microbial Communities in Ethanol-Stimulated Subsurface Sediment. *Microbes Environ* 2011, 26(3):261-265.

190. Petrie L, North NN, Dollhopf SL, Balkwill DL, Kostka JE: Enumeration and characterization of iron(III)-reducing microbial communities from acidic subsurface sediments contaminated with uranium(VI). *Appl Environ Microbiol* 2003, 69(12):7467-7479.

191. Reardon CL, Cummings DE, Petzke LM, Kinsall BL, Watson DB, Peyton BM, Geesey GG: Composition and diversity of microbial communities recovered from surrogate minerals incubated in an acidic uranium-contaminated aquifer. *Appl Environ Microbiol* 2004, 70(10):6037-6046.

192. Final Site Observational Work Plan for the UMTRA Project Old Rifle Site. In. Edited by Energy UDo. Grand Junction, CO; 1999.

193. Rehder D: Bioinorganic Vanadium Chemistry. Hoboken, NJ: John Wiley & Sons, Inc.; 2008.

194. Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF: EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol* 2011, 12(5):R44.
195. Chou J-H, Chou Y-J, Arun AB, Young C-C, Chen CA, Wang J-T, Chen W-M: *Vogesella lacus* sp nov., isolated from a soft-shell turtle culture pond. *Int J Syst Evol Microbiol* 2009, 59:2629-2632.
196. Chou Y-J, Chou J-H, Lin M-C, Arun AB, Young C-C, Chen W-M: *Vogesella perlucida* sp nov., a non-pigmented bacterium isolated from spring water. *Int J Syst Evol Microbiol* 2008, 58:2677-2681.
197. Grimes DJ, Woese CR, MacDonell MT, Colwell RR: Systematic study of the genus *Vogesella* gen. nov. and its type species, *Vogesella indigofera* comb. nov. *Int J Syst Bacteriol* 1997, 47(1):19-27.
198. Jorgensen NOG, Brandt KK, Nybroe O, Hansen M: *Vogesella mureinivorans* sp. nov., a peptidoglycan-degrading bacterium from lake water. *Int J Syst Evol Microbiol* 2010, 60:2467-2472.
199. Ziv-El M, Delgado AG, Yao Y, Kang D-W, Nelson KG, Halden RU, Krajmalnik-Brown R: Development and characterization of DehaloR boolean AND 2, a novel anaerobic microbial consortium performing rapid dechlorination of TCE to ethene. *Appl Microbiol Biotechnol* 2011, 92(5):1063-1071.
200. Jost L: Entropy and diversity. *Oikos* 2006, 113(2):363-375.
201. Webb CO: Exploring the phylogenetic structure of ecological communities: An example for rain forest trees. *Am Nat* 2000, 156(2):145-155.
202. Northrop HR, Goldhaber MB, Landis GP, Unruh JW, Reynolds RL, Campbell JA, Wanty RB, Grauch RI, Whitney G, Rye RO: Genesis of the tabular-type vanadium-uranium deposits of the Henry Basin, Utah. 1. Geochemical and mineralogical evidence for the sources of the ore-forming fluids. 2. Mechanisms of ore and gague mineral formation at the interface between brine and meteoric water. 3. Evidence from the mineralogy and geochemistry of clay-minerals. *Economic Geology and the Bulletin of the Society of Economic Geologists* 1990, 85(2):215-269.
203. Wanty RB, Goldhaber MB: Thermodynamics and kinetics of reactions involving vanadium in natural systems - Accumulation of vanadium in sedimentary rocks. *Geochim Cosmochim Acta* 1992, 56(4):1471-1483.
204. Wanty RB, Goldhaber MB, Northrop HR: Geochemistry of vanadium in an epigenetic, sandstone-hosted vanadium-uranium deposit, Henry Basin, Utah. *Economic Geology and the Bulletin of the Society of Economic Geologists* 1990, 85(2):270-284.
205. Breit GN: Origin of clay-minerals associated with V-U deposits in the Entrada Sandstone, Placerville mining district, Southwestern Colorado. *Economic Geology and the Bulletin of the Society of Economic Geologists* 1995, 90(2):407-419.
206. Schwertmann U, Pfab G: Structural vanadium and chromium in lateritic iron oxides: Genetic implications. *Geochim Cosmochim Acta* 1996, 60(21):4279-4283.
207. N'Guessan AL, Vrionis HA, Resch CT, Long PE, Lovley DR: Sustained removal of uranium from contaminated groundwater following stimulation of dissimilatory metal reduction. *Environ Sci Technol* 2008, 42(8):2999-3004.

208. Nawrocki EP, Kolbe DL, Eddy SR: Infernal 1.0: inference of RNA alignments. *Bioinformatics* 2009, 25(10):1335-1337.
209. Edgar RC: Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010, 26(19):2460-2461.