Title
Community Genomic, Proteomic, and Transcriptomic Analyses of Acid Mine Drainage Biofilm Communities

Author
Goltsman, Daniela

Publication Date
2013

Peer reviewed|Thesis/dissertation

Community Genomic, Proteomic, and Transcriptomic Analyses of Acid Mine Drainage Biofilm Communities

by

Daniela Salome Goltsman

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 Banfield, Chair

Professor Mary Firestone

Professor John Taylor

Spring 2013

Abstract

Community Genomic, Proteomic, and Transcriptomic Analyses of Acid Mine Drainage Biofilm Communities

by

Daniela Salome Goltsman

Doctor of Philosophy in Environmental Science, Policy, and Management

University of California, Berkeley

Professor Jillian F. Banfield, Chair

Culturing isolated microorganisms can be challenging, not only because usually the detailed environmental conditions where organisms grow optimally are not known, but also because many of them need to grow in the presence of other organisms. High-throughput sequencing and other 'omics' technologies provide important approaches for the study of microorganisms in their natural environments. Specifically metagenomics methods enable culture-independent surveys of organisms and functions in microbial consortia, and can yield near-complete genomes of the most abundant community members, and partial genomes of lower abundance organisms. When coupled to community proteomic and/or transcriptomic analyses it is possible to predict what functions are being expressed within the community. Therefore, 'omics' technologies provide a means for the study of community physiology and ecology in natural systems.

Acid mine drainage (AMD) is a mining-related problem caused by sulfide mineral dissolution coupled to microbial iron oxidation, which leads to acidification and metal contamination of the environment. The Richmond Mine AMD community is currently the best-studied AMD system. Bacteria of the genus *Leptospirillum*, of the Nitrospira phylum, generally dominate Richmond Mine AMD microbial communities. Current studies show that *Leptospirillum rubarum* (group II) tends to dominate early-formed biofilms, and *Leptospirillum* group II 5way CG (a genotype related to *L. rubarum*) or *L. ferrodiazotrophum* (group III) increase in abundance as environmental conditions change.

In chapter 1, community genomics was used to reconstruct the near-complete genome of *Leptospirillum ferrodiazotrophum*, and report the genome annotation and metabolic reconstruction of *L. ferrodiazotrophum, Leptospirillum rubarum* and an extrachromosomal plasmid associated to these bacteria. In addition, proteomic analyses were used to evaluate protein expression patterns in three AMD biofilms. Results indicate that, despite sharing only 92% identity at the 16S rRNA level, *L. rubarum* and *L. ferrodiazotrophum* share more than half of their genes. Both bacteria are motile, acidophilic iron-oxidizers, as evidenced by the presence of cytochrome Cyt_{572} and an electron transport chain. They are chemoautotrophs, using a reverse tricarboxilic acid (TCA) cycle for carbon fixation. Their metabolic potential indicates that *L. rubarum* and *L. ferrodiazotrophum* are capable of amino acid and vitamin biosynthesis, fatty acid biosynthesis, flagella biosynthesis, synthesis of polymers such as cellulose, and the synthesis of compatible solutes for osmotic tolerance. Only *L.*

ferrodiazotrophum is capable of nitrogen fixation, although proteins were not detected by proteomics in the analyzed biofilms. Proteomic analyses indicate that core metabolic proteins are similarly expressed in both bacteria, however high expression of many hypothetical proteins unique to each *Leptospirillum* might contribute to their differentiation within the biofilms.

In chapter 2, the partial genome reconstruction of a new *Leptospirillum* bacterium, which is closely related to *L. ferrodiazotrophum*, is reported. The bacterium represents ~ 3% of the sequenced community, and comparison of its 16S rRNA gene sequence with those of other *Leptospirilli* identifies it as a new group within the *Leptospirillum* clade: *Leptospirillum* group IV UBA BS. The bacterium grows in unusually thick, Archaeal-dominated biofilms where other *Leptospirillum* spp. are found at very low abundance. It shares 98% 16S rRNA sequence identity and 70% amino acid identity between orthologs with *L. ferrodiazotrophum*. Its metabolic potential indicates that it too is a motile, iron oxidizing chemoautotroph capable of nitrogen fixation, although nitrogen fixation expression was not observed. *Leptospirillum* group IV UBA BS is distinguished from the other *Leptospirilli* in that it contains a unique multicopper oxidase likely involved in iron oxidation, and the presence of two clusters of hydrogenase genes. The cytoplasmic hydrogenase is likely used to take up H_2 during nitrogen fixation, while the membrane-bound hydrogenase might be involved in anaerobic H_2 oxidation for energy generation. Community transcriptomic and proteomic analyses confirm expression of the multicopper oxidase, as well as the expression of many hypothetical proteins and core metabolic genes. Transcription of hydrogenases in the only biofilm in which the nitrogen fixation operon in *L. ferrodiazotrophum* is transcribed points to potential cooperative interactions between the two bacteria.

AMD has long been considered a simple, low-diversity ecosystem. In chapter 3, a new view of the diversity of organisms in AMD was obtained by deep sequencing of the small subunit (SSU) rRNA from 13 biofilm communities. A total of 159 taxa, including Archaea, Bacteria, and Eukaryotes, were identified. *Leptospirillum* spp. dominate the samples, and members of diverse phyla, such as Actinobacteria, Acidobacteria, Firmicutes, Alpha-, Beta-, Gamma-, and Delta-Proteobacteria, Chloroflexi, and Deferribacter were present at low abundance. Interestingly, members related to *Magnetobacterium* spp. of the Nitrospira phylum were detected. These bacteria have not been reported present in AMD environments, and they have not been identified in community genomics datasets. However, the presence of magnetosome-like structures observed by cryo-TEM in some AMD biofilms supports the transcriptomics results. The findings indicate that it is dominance by a few taxa, and not lack of complexity of the system that has made AMD environments model systems for the study of microbial physiology and ecology.

In Chapter 4, non-ribosomal transcriptomic reads were mapped to several genomes of AMD organisms, including Archaea, Bacteria, and viruses, in order to evaluate the expression profiles of genes and non-coding regions (ncRNAs) in biofilms at increasing stages of development. More than 95% of the genes in the most abundant *Leptospirillum* group II and group III bacteria were detected by at least one transcriptomic read, indicating that the whole genome is transcribed at some level. More than half of the genes in the Archaea G-plasma and *Ferroplasma* Type II, and a virus associated to *Leptospirillum* were also detected. Transposases, cytochromes, and ncRNAs were among the most highly expressed genes in all samples. Gene expression profiles indicate that *Leptospirillum* group II 5way CG and *L. ferrodiazotrophum* prefer growth at higher pH and lower temperature, conditions generally

present in bioreactors, while the opposite is true for *L. rubarum* and G-plasma, who prefer conditions found in early to mid-developmental stage environmental biofilms. High levels of expression were observed for a novel ectoine riboswitch predicted in the *Leptospirillum* group II genome, as well as for other non-coding RNAs. Results provide new insight into understanding functioning and adaptation of acidic ecosystems.

To my wonderful husband Женя,
for all the love and happiness you have given me.

To my mom and sister, Margarita and Paulina
for always being there for me.

To my father, Alberto (1951-1993)
I wish you could have seen this.

I love you, los amo.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to start by thanking my advisor and mentor, Jill Banfield, for her continued support and guidance since I joined her lab. I have learned so much from you, Jill, from assembling a genome, to writing publications and grants, to interacting with collaborators and journal editors. You are an incredibly understanding and caring PI, always showing interest in our projects. You are truly an inspiration to me, both as a successful researcher and as a wife and mother of three. I now feel prepared to continue my career wherever life takes me. By the way, going sampling to the mine was quite a fun experience!

I would also like to thank current and former Banfield lab members. Vincent Denef, Chris Miller, Brian Thomas, and Ryan Mueller, thank you for taking the time to meet with me whenever I needed help, and for your invaluable advice when writing grant applications, dissertation proposals and papers. Mauna Dasari, thank you for your hard work doing fluorescent microscopy. I need to thank our collaborators at Oak Ridge National Lab, at Lawrence Livermore National Lab, and at Lawrence Berkeley National Lab, Nathan VeBerkmoes, Manesh Shah, Loren Hauser, Bob Hettich, Steven Singer, Michael Thelen, and Luis Comolli. Thank you for your help in writing my papers.

I would like to thank Dr. David Holmes, my undergraduate advisor and mentor at Universidad de Santiago de Chile. Thank you for believing in me when applying for the Fundación Andes fellowship, and for all your support and help since I first came to your lab back in 2002. You encouraged me to get in the path of bioinformatics, to learn to program on my own, and you taught me to do accurate annotation of microbial genomes. I believe I would not be here if it wasn't for what I learned from you. I am truly grateful, Profe David. I'd like to also thank Dr. Andrei Osterman, for his guidance during my internship at Integrated Genomics (and for all those letters of recommendation that you keep writing for me, thank you Andrei!). I need to thank my dissertation committee, Dr. Mary Firestone and Dr. John Taylor, for their support and guidance during my qualifying examination and my dissertation work. Their input was extremely valuable to me. I would also like to thank Dr. Kimmen Sjölander; being a GSI for your class was a fun and rewarding experience.

I would like to thank the SACNAS at Berkeley student group, especially Patty, Galo, Joey, and Brandon. With you I realized the importance of increasing diversity in the sciences, and of giving back to the community. We had so much fun planning events and attending the SACNAS Conferences! I would like to thank Colette Patt and The Berkeley Edge for the summer fellowship I was awarded in 2007, for funding to attend the SACNAS National Conferences, and for allowing me to participate in the mentorship program in 2008. Jessica Osuna, thank you for guiding me through the initial steps of graduate school in ESPM.

Finally, I would like to thank those who have been with me at every step of my life and career. I want to thank my parents, Margarita and Alberto, for realizing that I wanted to become a scientist at an early age and encouraging me to follow that path (they gave me a microscope for my 10th birthday!). But most of all, thank you for teaching me to respect and value all forms of life and for teaching me to care for the environment. Thank you for making me who I am. I want to thank my sister and best friend, Paulina for always being there for me, listening and understanding even through the distance. And finally, Женя, you have seen me walk this long path of graduate school. Thank you for celebrating with me when things went well and smooth, and for cheering me up when things were challenging and stressful. Thank you for your unconditional support and love. I am excited about the new life we are about to start.

INTRODUCTION

Several decades ago some mining companies realized that by adding sulfuric acid to crushed mineral ore, desired precious and low grade metals were released in solution (1). It was later acknowledged that the process is mediated by microorganisms, and ten years later harnessing of bacteria for metal-recovery (biomining) was already widely used (reviewed in (2)). Research has shown that microorganisms involved in biomining, also involved in acid rock and mine drainage (ARD/AMD), are extremely acidophilic (grow at pH <3), moderate to extremely thermophilic organisms (2, 3). Among the organisms that have been consistently found in biomining systems, and have become the target for much of the research done in cultures, are bacteria of the genus *Acidithiobacillus* spp., and *Ferroplasma* archaea (reviewed in (2)).

It was in 1995 that Dr. Jill Banfield and her research lab started to study the microbiology of Richmond Mine AMD environments. Acid mine drainage is produced when sulfide mineral ores are exposed to air and water, promoting the dissolution of minerals (predominantly pyrite, FeS_2) and release of ferric iron, the byproduct of microbial metabolism (3, 4). Ferric iron enhances the dissolution of sulfide minerals and production of sulfuric acid, and the cycle continues rapidly until the mineral has been completely dissolved. Therefore, microorganisms are largely responsible for AMD generation. Fluorescent microscopy and 16S rRNA clone library surveys showed that, indeed, *Acidithiobacillus ferrooxidans* was present in AMD systems at low abundance, but *Leptospirillum* spp. were far more abundant (5). Over the years, reports indicate that microorganisms found in biomining are often the same bacteria and archaea found in AMD environments: *Acidithiobacillus* spp., *Leptospirillum* spp., *Acidiphilum* spp., *Sulfobacillus* spp., *Acidimicrobium* spp., and *Ferromicrobium* spp., and the *Ferroplasma, Thermoplasmatales*, ARMAN and *Sulfolobales* archaea (reviewed in (2, 3, 6)). Furthermore, some of these organisms were also found in low-temperature (8 °C) underground AMD waters in Northern Wales (reviewed in (6)). Still, much of the work done to identify these organisms involved cultivation-dependent techniques and 16S rRNA clone libraries.

Cultivation-independent community genomics (metagenomics) involves sequencing of total DNA of a community, and allows recovering information about what organisms are present and what their metabolic potential is (reviewed in (7)). However, genomic data alone do not reveal how organisms alter their activity within a community under changing conditions. Community proteomics, the extraction of total proteins and their identification by mass-spectrometry, informs which proteins are expressed by the organisms in their natural environment (reviewed in (7)). Using cultivation-independent metagenomics, in 2004 the Banfield lab published the near-complete and partial genomes of five organisms recovered directly from the Richmond Mine AMD environment (8). The study used ~ 76 million bp of Sanger sequencing. This was the first time cultivation-independent genome recovery had been accomplished for any microbial community (reviewed in (7)). A year later, mass-spectrometry based proteomics analyses allowed for the study of the total protein pool of AMD biofilms and demonstrated that physiology studies in the most abundant community member were possible without the need for cultivation (9, 10).

I joined the Banfield lab in 2006, working on the manual annotation and metabolic reconstruction of the most abundant bacteria in Richmond Mine AMD biofilms: *Leptospirillum* group II and group III. A year later I joined the PhD program and started working on a collaborative project with mining companies in Chile. Chile is one of the largest producers of

copper in the world, and much of the metal extraction is done using biomining (1). Studies of the ecology and physiology of AMD systems are relevant for the biomining industry because they provide information about the functioning of acidophilic microbial consortia. The project would look at the biogeography and evolutionary dynamics of acidophilic microorganism across a longitudinal transect, studying communities both in the USA and in Chile. I was awarded a National Science Foundation Graduate Research Fellowship for that project. Unfortunately, confidentiality agreement issues arose and, by the time I was taking my qualifying examination in 2009, the collaborations fell apart.

Meanwhile, research in AMD systems continued at a fast pace. Metagenomic sequencing was done on many more biofilms and, as of today, the genomes of many bacteria (8, 11-15, Justice *et al.*, in preparation, and unpublished data), archaea (8, 13, 16, Yelton *et al.*, in review), viruses (17), plasmids (12) and fungi (Miller *et al*., in preparation) are available. Noteworthy, many of these organisms belong to phylogenetic groups without cultivated representatives and some, although cultivated, lacked genome sampling. For example, the near-complete genomes of *Leptospirillum* group II (11) and group III (12) and their metabolic reconstruction (Chapter 1 of this dissertation) were the first detailed genomic reports for any member of the Nitrospira phylum. Moreover, genome reconstruction of a new low abundance (< 3% of the sequenced community) *Leptospirillum* species (Chapter 2 of this dissertation) was achieved from a combined metagenomic dataset of ~ 210 million bp. Analysis of this dataset was first used to demonstrate new methods for binning (assignment of assembled scaffolds to their organism of origin) (13). In addition, community proteomics have enabled studies of the physiology of the most abundant AMD community members, showing, for example, that adaptation of *Leptospirillum* groups II and III to different niches can be explained by protein expression profiles (12, 18, 19). Studies of stain variation at the DNA and protein levels were also achieved (10, 11, 14, 20).

Reports of deep sequencing of total RNA from the environment (community transcriptomics) provided new insights into the diversity of natural systems, and allowed studying important metabolic processes that take part in space and time. For example, the genes involved in key metabolic pathways, such as carbon fixation, nitrogen metabolism, and photosynthesis, were detected by transcriptomics in ocean water (21, 22), and the flow of carbon between organisms could be tracked by looking at the expression of genes in microbial mats (23). While proteomics provide valuable information about the expression of protein functions, the dynamic range of proteomics is rather limited (24). Deep-sequencing transcriptomics enables detection of expression levels of a larger number of genes, including those from low abundance community members. In addition, community transcriptomics can provide information about known and novel non-coding RNAs, as observed in ocean waters where expression of novel small RNAs varies with depth (25). Small non-coding RNAs (ncRNA) are important gene and protein regulators in microorganisms: some small RNAs function by base-pairing with their target RNA and regulate their expression; some others bind to proteins complexes to up- or down-regulate transcription (26). Other ncRNAs act as sensors of changes in the environment (e.g., riboswitches, (27)), and others are involved in microbial immunity against foreign DNA (CRISPR, (28)). Therefore, studying the total transcriptome of a community provides information about the functioning and regulation of genes and pathways in microorganisms in their natural environment.

My dissertation focused on using community transcriptomics to study the gene expression profiles of organisms in AMD biofilms at different maturation stages and under

different growth conditions. Developing the transcriptomics protocol was challenging, at times frustrating, however obtaining and analyzing the data was extremely exciting. Indeed, analyses of total RNA show that the AMD communities are much more diverse than previously though (Chapter 3 of this dissertation). It was considered by many that research in AMD systems have been possible due to its simplicity (low richness, (29)). The number of genomes recovered from metagenomics shows that these communities have a larger richness than the initial five organisms reported in 2004, and transcriptomics analyses now indicate that AMD consortia are much more complex, having detected up to 159 different taxa (Chapter 3). In addition, transcriptomics enabled identification of ncRNAs, some of which are among the most highly expressed RNAs in the organism. For example, different expression patterns of a novel ectoine riboswitch predicted in the genomes of two *Leptospirillum* group II genotypes suggest that expression of non-coding regulatory RNAs might contribute to their ecological differentiation (Chapter 4 of this dissertation).

Overall, this dissertation improved our understanding of the ecology of AMD environments, and provided clues into how other environments might operate. Specifically, it informs of the diversity and function of dominant and low abundance community members growing under different environmental conditions, and addresses ecological adaptation of organisms as explained by gene and protein expression profiles, and expression of non-coding RNAs.

CHAPTER 1

Community genomic and proteomic analysis of chemoautotrophic, iron-oxidizing "*Leptospirillum rubarum*" (Group II) and *Leptospirillum ferrodiazotrophum* (Group III) in acid mine drainage biofilms.

ABSTRACT

We analyzed near-complete population (composite) genomic sequences for coexisting acidophilic iron-oxidizing *Leptospirillum* Groups II and III bacteria (phylum Nitrospirae) and an extrachromosomal plasmid from a Richmond Mine, CA acid mine drainage (AMD) biofilm. Community proteomic analysis of the genomically characterized sample and two other biofilms identified 64.6% and 44.9% of the predicted proteins of *Leptospirillum* Groups II and III, respectively and 20% of the predicted plasmid proteins. The bacteria share 92% 16S rRNA gene sequence identity and > 60% of their genes, including integrated plasmid-like regions. The extrachromosomal plasmid encodes conjugation genes with detectable sequence similarity to genes in the integrated conjugative plasmid, but only those on the extrachromosomal element were identified by proteomics. Both bacteria have genes for community-essential functions, including carbon fixation, biosynthesis of vitamins, fatty acids and biopolymers (including cellulose); proteomic analyses reveal these activities. Both *Leptospirillum* types have multiple pathways for osmotic protection. Although both are motile, signal transduction and methyl-accepting chemotaxis proteins are more abundant in *Leptospirillum* Group III, consistent with its distribution in gradients within biofilms. Interestingly, *Leptospirillum* Group II uses a methyl-dependent and *Leptospirillum* Group III a methyl-independent response pathway. Although only *Leptospirillum* Group III can fix nitrogen, these proteins were not identified by proteomics. Abundances of core proteins are similar in all communities, but abundance levels of unique and shared proteins of unknown function vary. Some proteins unique to one organism were highly expressed and may be key to the functional and ecological differentiation of *Leptospirillum* Groups II and III.

INTRODUCTION

To understand how microorganisms contribute to biogeochemical cycling, it is necessary to determine the roles of uncultivated as well as cultivated groups and to establish how these roles vary during ecological succession and when environmental conditions change. Shotgun genomic sequencing (metagenomics) has opened new opportunities for culture-independent studies of microbial communities. Examples include investigations of acid mine drainage (AMD) biofilm communities (8, 11, 17), symbiosis in a marine worm involving sulfur-oxidizing and sulfate-reducing bacteria (30), and of enhanced biological phosphorous-removal by sludge communities (31). From these genomic datasets, it has been possible to reconstruct aspects of the metabolism of individual organisms (31) and coexisting community members (8, 32) and to identify which organisms contribute community-essential functions (8). An interesting question relates to how differences in metabolic potential of organisms from the same lineage allow them to occupy distinct niches. Identification of potentially adaptive traits in closely related organisms is also important from an evolutionary perspective.

Genomic data do not reveal how organisms alter their metabolisms in response to the presence of other organisms or environmental conditions. Proteomics methods for analysis of metabolic responses of isolates (33-37) have been extended to analyze the functioning of the dominant members of natural consortia (9) (38), with strain-level resolution (11, 39). In these studies, peptides are separated by liquid chromatography and identified by tandem mass spectrometry through reference to appropriate genomic databases. Proteomic analysis is

possible even if the genome sequences are not identical to those of the organisms present (40), however missing sequence information reduces the resolution of such proteogenomic studies.

Due to dominance by a few organism types, chemoautotrophic microbial acid mine drainage (AMD) biofilms from the Richmond Mine, California, are tractable model systems for development of cultivation-independent metagenomic and proteogenomic methods to analyze community structure, function, and ecology (41). Acidophilic *Leptospirillum* bacteria dominate this AMD system (42), other AMD systems (43), and bioleaching systems used for metals recovery (44-46). These bacteria play pivotal roles in sulfide mineral dissolution because they are iron oxidizers (8, 45), and ferric iron drives sulfide oxidation, leading to formation of metal-rich sulfuric acid solutions. Based on a recent microscopy-based study (47), *Leptospirillum* Group II are the first colonists in AMD biofilm communities whereas *Leptospirillum* Group III generally appear later, sometimes partitioned within biofilm interiors. Because only *Leptospirillum* Group III appear to be able to fix nitrogen, they may be keystone species in AMD ecosystems (8). This observation enabled the isolation of one representative, *Leptospirillum ferrodiazotrophum* (48). In prior work, we reported near-complete genome sequences of two *Leptospirillum* Group II types (11, 14), but detailed functional annotations and metabolic analyses have not been published. Genomic data have been used to explore the metabolism of *Leptospirillum* in one biofilm community (9), but proteomic and genomic analysis of the same biofilm community have not been studied.

Here we report the near complete genomic sequence of *Leptospirillum* Group III from a biofilm from the UBA site, Richmond mine, CA, the detailed functional annotation of the genomes of *Leptospirillum* Groups II and III, and the genomic and proteomic comparison of them. In addition, we report the sequence of an extrachromosomal plasmid associated with these organisms. This represents the first comprehensive genomics-based analysis of the metabolism of bacteria from the Nitrospirae phylum, and the first environmental community proteogenomic study where the genomic and proteomic data derived from the same sample. We compare the proteomic profiles of three different biofilm communities to evaluate the importance of shared and unique genes and pathways in environmental adaptation.

MATERIALS AND METHODS

Samples:

Biofilm samples were collected underground within the Richmond mine, Iron Mountain, CA (Figure 1). The UBA biofilm was collected from the surface of a slowly draining ~0.5 cm deep pool in a stream with a pH of 1.1 and temperature 41˚ C in the A drift in June 2005 (Figure S1). The thin (few 10s of μm thick estimated by microscopy, Figure S1) floating ABend biofilm was collected from the surface of a deeper pool in the AB drift in January 2004. Geochemical data and other information were reported by Ram (9). Briefly, the pH of the solution was 1.07, temperature 43˚ C. The ABfront biofilm was also collected in the AB drift, from about 2 meters from the ABend location, in June 2004 (Figure 1). The ABfront sample is inferred to be a much more mature biofilm than the ABend sample based on its thickness (~200 μm) (Figure S1). At the time of sampling, the pH at the ABfront location was 0.99 and temperature 39˚ C.

Assembly: Total DNA recovered from the UBA biofilm was cloned and sequenced (~3 kb library), as reported previously (11). Briefly, 100 Mb of sequence was obtained from the UBA site (Figure 1), sequences assembled (Phred/Phrap) and contigs manually curated to correct

misassemblies and remove errors such as co-assembly of fragments from different organism types (identified based on misplacement of mate paired sequences). Misassembles due to repetitive sequences were either resolved based on surrounding unique sequences using mate-pair information or allowed to terminate scaffolds. Contig fragmentation due to multiple genome paths for different individuals within a single population (strain variation) was identified so that larger scaffolds could be established. Contig editing was done using Consed (49).

The very near complete, deeply sampled (~25X) genome of *Leptospirillum* Group II recovered from the UBA genomic dataset (*Leptospirillum* Group II UBA) is in seven composite scaffolds (11). The population is genomically distinct from a *Leptospirillum ferriphilum*-like strain (*Leptospirillum* Group II 5-way CG), previously described from the 5-way CG site from the Richmond Mine (Figure 1) (8, 14). The two genomically characterized types (UBA and 5-way CG) share 99.7% 16S rRNA gene sequence identity and ~94% DNA sequence similarity for orthologs (50).

After manual curation of the assembly, contigs of *Leptospirillum* Group III from the UBA genomic dataset were separated from archaea based on GC content and from low abundance bacteria (average depth ~2X) based on sequence depth (average depth ~10X). Binning was verified using tetranucleotide sequence signatures (51) analyzed using emergent self-organizing maps (13).

We reconstructed what we believe to be a near-complete composite sequence (~250 kb) for a large extrachromosomal plasmid. Fragments were clustered by ESOM, with strong distance structure that definitively separated the sequences from any others in the community (13). The numerous small contigs were manually curated into 10 scaffolds.

Annotation: Protein gene predictions of *Leptospirillum* Groups II and III were made using a combination of FgenesB (Softberry Inc.), CRITICA (52) and Glimmer (53). Automated function predictions were generated by searching all predicted peptides against TIGRFAMs (HMMPfam trusted cutoffs), PRIAM (rpsblast 1e_30 cutoff), PFAM (HMMPfam trusted cutoffs), InterPro (interproscan default cutoffs), and COGs (rpsblast 1e_10 cutoff). Proteins that failed to return a definitive result with the aforementioned profile searches were annotated on the basis of BLASTP searches against the KEGG and Swissprot-TREMBL (1e_5 cutoff) peptide databases. The tRNAScanSE tool (54) was used to find tRNA genes, while ribosomal RNAs were found using BLASTn vs. the 16S and 23S ribosomal RNA databases. Other ‘‘standard’’ structural RNAs (e.g., 5S rRNA, rnpB, tmRNA, SRP RNA) were found using covariance models with the Infernal search tool (55).

Manual Curation of Gene Annotation: The automatic gene annotations were manually curated. Product descriptions were assigned when scores for matches to protein families in the PRIAM database were e-30 or more. Functions were also inferred based on the TIGRFAM or PFAM assignments as long as the protein had an ortholog in the public databases with >70% identity over >70% of the length of the protein alignment. “Putative” was added to product descriptions for proteins with PFAM assignments and an ortholog in the public databases with between 30% and 70% identity and alignments involved over 70% of the protein length. “Probable” was added to product descriptions for predicted proteins with >30% identity to proteins in the SwissProt database. For these cases, BLAST (56) matches in the non-redundant (NR) NCBI sequence database (http://blast.ncbi.nlm.nih.gov/Blast.cgi) were also considered.

The term "conserved protein of unknown function" was used when the predicted protein was a conserved hypothetical protein identified by proteomics in the current study or validated in previous studies. Similarly, "protein of unknown function" was used when the predicted protein was a hypothetical protein identified by proteomics in the current study, or in prior studies of these AMD biofilm communities. "Conserved hypothetical protein" was used when the predicted protein had an alignment of >70% to one or more hypothetical proteins and >30% identity with these. The term "Hypothetical protein" was used when there was no good alignment against predicted proteins in the NR database. In the specific case of a possible PEP carboxylase, the protein structure from *Leptospirillum* group II was modeled after Maize (PDB entry: 1jqo_A) and *E. coli* (PDB entry: 1jqn_A) crystal structure (57, 58).

Both *Leptospirillum* Group II and Group III genomes are composites in that they generally report a single genome path, although multiple paths exist in some regions. Due to the strain variation and gaps present in both genomes, a subset of sequencing reads, potentially carrying important genes, were not brought into the composite sequences. Consequently, analysis of gene content included consideration of the read databases as well as composite sequences and strain variant paths.

Proteomics: Proteomic data were obtained from the same UBA biofilm samples used for genomic library construction, as well as two other samples: the ABfront biofilm and the ABend biofilm. Complete descriptions of the ABend biofilm preparation and analyses were published previously (9, 11). The UBA and ABfront proteomes were prepared and analyzed via comparable methods. Briefly, proteins from the biofilms were released via sonication and fractionated based on cellular location (membrane, soluble, whole cell and extracellular). Proteome fractions (~3mg total protein per fraction) were denatured and reduced, digested using sequencing grade trypsin (Promega, Madison, WI), desalted, concentrated and frozen until analyses.

Two-dimensional nano LC-ES-MS/MS analyses of all samples were performed on a linear ion trap mass spectrometer (LTQ Thermo Fisher, San Jose, CA) as previously described (9, 36). 4 different fractions were analyzed using the same methodology on the same LC-MS system, with three technical replicates for each sample. The samples were loaded (~500µg starting material) onto a split-phase column (packed in-house with C18 reverse phase and SCX chromatographic resin) (59) placed behind a 15 cm C18 analytical column (packed in-house). Both were situated in front of a Proxeon nanospray source (Odense, Denmark) on the LTQ. Flow was provided via an Ultimate HPLC pump (LC Packings; a division of Dionex, San Francisco, CA), with an initial flow rate of ~100 µL/min that was split precolumn to obtain a flow rate of ~300 nL/min at the nanospray tip; a voltage of 3.8kV was applied on the waste line. Chromatographic separation of the tryptic peptides was conducted over a 22 h period of increasing (0-500 mM) pulses of ammonium acetate salt followed by a 2 h aqueous to organic solvent gradient. The LTQ was operated in a data-dependent manner with two microscan full scans (400-1700 m/z) and two microscan MS/MS scans (top five most abundant), dynamic exclusion was set at 1 (9, 36).

MS/MS spectra from all individual 24 hr 2D-LC-MS/MS runs were searched using the SEQUEST algorithm (60) against a global database created from proteins predicted from AMD genomic sequences. The database was concatenated with a list of common contaminants (trypsin, keratin, etc.). All searches were run with the following settings: enzyme type, trypsin; Parent Mass Tolerance, 3.0; Fragment Ion Tolerance, 0.5; up to 4 missed cleavages allowed,

and fully tryptic peptides only (no post-translational modifications were considered for this study). The output data files from all searches were filtered and sorted with the DTASelect algorithm (61) using the following parameters: fully tryptic peptides only, with delCN of at least 0.08 and cross-correlation scores (Xcorr) of at least 1.8 (+1), 2.5 (+2), 3.5 (+3). Identification of at least two peptides within the same 24 hr run was required in order for a protein to be deemed identified. From the DTASelect output files, the total numbers of proteins, peptides, spectra and sequence coverage for each protein as well as unique peptides per protein were extracted. The Xcorr values used in the current study have been rigorously tested, and typically give a maximum false positive rate of 1-2% for both bacterial isolates (36) and microbial communities (9, 11, 39). All databases, peptide and protein results, MS/MS spectra and supplementary tables for all database searches are archived and made available as open access via the following link:
http://compbio.ornl.gov/comparative_genomics_proteomics_of_leptospirillum

Proteomic analyses: Given that our goal was to compare *Leptospirillum* Group II to Group III at a functional level, we combined the spectral counts (unique counts for each type + non-unique count shared by the two types) for the two *Leptospirillum* Group II genomic types (UBA and 5-way CG (11, 14)). Similarly, non-unique and unique spectral counts were combined for *Leptospirillum* Group III proteins. Virtually no cross-identification is expected for proteins sharing < 85% sequence identity (40) (*Leptospirillum* Groups II and III share ~55% average sequence identity). Although analysis of similar amounts of protein for each fraction (~ 3mg) could potentially lead to an overrepresentation of proteins in the less abundant extracellular fraction, relative comparisons between samples should not be affected.

For comparison of protein abundance levels for *Leptospirillum* Groups II and III and the extrachromosomal plasmid, proteomics data were normalized using the normalized spectral abundance factors (NSAF) method (62, 63). This method estimates protein abundance by first dividing the spectral count for each protein by the protein length and then dividing this number by the sum of all length-normalized spectral counts for each organism, and multiplying by 100. For each sample we summed the spectral counts from each fraction and combined the spectral counts of each protein over the three technical replicates prior to calculation of the NSAF. The NSAF value for a *Leptospirilllum* Group II protein thus estimates the percentage of the total *Leptospirillum* Group II protein pool that each protein represents.

Circular comparative genomic and proteomic representations were made using the Circos v0.46 software (http://mkweb.bcgsc.ca/circos/). Heat maps of the protein abundance levels were based on $\log_2$-transformed NSAF.

The NSAF values were clustered using Cluster v3.0 (64) (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv) and visualized as heat maps with TreeView software (65). For the clustering analyses, we considered only proteins identified in more than half of the six datasets (two organisms in three environmental samples). To check the robustness of the results, we also performed the analysis (i) considering only proteins identified in five or more of the six datasets, and (ii) considering proteins identified in one or more dataset. For a single organism in a single sample, the median protein NSAF value was determined, and proteins with higher or lower values were classified as over or under represented in the proteome (indicated by shades of red and green, respectively). This is referred to as median centering by organism. In some cases, simultaneously with median centering by organism, the abundance of each protein was compared among the six organism

datasets, the median value for each protein determined, and proteins levels assigned values indicating under or over representation. This is referred to as median centering by protein and organism dataset. For both forms of median centering, the six organism datasets were clustered based on the resulting patterns using average linkage and Kendall's Tau distance matrix.

RESULTS

Genomics statistics and overall proteomic results

We reconstructed a near-complete composite genome for a *Leptospirillum* Group III population closely related (99.8% 16S rRNA gene sequence identity) to the *Leptospirillum ferrodiazotrophum* previously isolated from the site (48) (Figure 2). The *Leptospirillum* Group III genomic dataset comprises 39 scaffolds (average depth ~10X). An isolate with 100% 16S rRNA sequence identity has been recovered and is available for further characterization (strain UBA:5) (Figure 2).

The genome annotations of *Leptospirillum* Group II and *Leptospirillum* Group III are reported in Tables S1 and S2, respectively, and the annotation files have been deposited in Genbank. Table S2 also reports three strain variants contigs within the *Leptospirillum* Group III population. While they share only 92% 16S rRNA gene sequence identity (Figure 2), more than 60% of the genes in *Leptospirillum* Groups II and III are orthologs (55% average amino acid identity) and ~78% of the proteins from each organism identified by proteomics are orthologs (Table 1). Table 1 also summarizes other basic statistics.

Representations of the *Leptospirillum* Groups II and III genomes, illustrating synteny between orthologs (inner ring is color-coded by scaffold), are shown in figures 3 and S2. Proteomic data from ABend and ABfront biofilms are included alongside results for the UBA sample to provide insight into site-to-site variation. Syntenous regions primarily encode core metabolic functions (Figure 3 and tables S1 and S2) and tend to have similar protein abundance patterns across the three biofilms (compare heat map rings within each figure; also Figure S3), and between organisms (Figure 3 and Figure S2). Regions of consistently high protein abundance correspond to ribosomal proteins, RNA polymerase, and proteins involved in energy metabolism. The genomic regions where the gene content shared between *Leptospirillum* Group II and III is lower than average correspond primarily to an integrated plasmid.

Clustering of NSAF values, median centered by organism dataset (Figure S3) and by organism dataset and protein (Figure 4) yielded two clusters, one containing all three datasets for *Leptospirillum* Group II, the other containing all three datasets for *Leptospirillum* Group III. The same results are obtained independent of the filtering level and when clustering was done with four organisms datasets (two organisms in two of the three available environmental samples; data not shown). Many protein sub-clusters are apparent in figure 4. Although ribosomal proteins are highly abundant in all datasets (Figure 3, Figure S2, Figure S3), they are generally less abundant in *Leptospirillum* Group III than Group II (e.g., clusters A and B in Figure 4). Vitamin and cofactor biosynthesis (within cluster A) are generally more abundant in *Leptospirillum* Group II than III, transport and secretion proteins (in cluster B) are overrepresented in all three biofilm samples in *Leptospirillum* Group II, whereas proteins involved in chemotaxis and energy generation (in clusters A and C) are overrepresented in *Leptospirillum* Group III.

Proteins of unknown function in *Leptospirillum* Group II are generally located in genomic regions with consistently high protein abundance, whereas hypothetical proteins not identified by proteomics tend to occur in genomic regions where few proteins are identified. Several proteins of unknown function that are unique to *Leptospirillum* Group II or *Leptospirillum* Group III are highly abundant whereas others show notable inter-sample protein abundance variation (Figure 5).

There are fewer conserved proteins of unknown function than proteins of unknown function in *Leptospirillum* Group II and III, and these appear to be spread across the genomes. Many conserved proteins of unknown function were found in operons with one or more genes with functional predictions, and may have related roles. Examples occur in operons with genes for the proteasome, flagella, transport and secretion, plasmid functions, folate metabolism, and t-RNA synthetases (Tables S1 and S2).

Comparative Genomics

I. Energy metabolism.

A. Electron transport chain. The conserved motif typical of *c*-type cytochromes (CxxCH) was found in 34 predicted *Leptospirillum* Group II proteins. After in-depth analysis, 13 were annotated as *c*-type cytochromes (Table 2). Seven of the putative *Leptospirillum* Group II cytochromes have an ortholog in *Leptospirillum* Group III based on reciprocal best hit. An independent survey of the *Leptospirillum* Group III genome did not uncover any predicted *c*-type cytochromes without a homolog in the *Leptospirillum* Group II genome. Two *Leptospirillum* Group II *c*-type cytochromes, Cytochrome 579 (Cyt_{579}) and Cytochrome 572, isolated directly from Richmond Mine biofilms, were biochemically characterized recently (66, 67). Although the composite *Leptospirillum* Group III genome fragments lack an ortholog of Cyt_{579}, sequence reads not brought into the assembly indicate that *Leptospirillum* Group III has a gene for Cyt_{579}.

Genes encoding a putative bc1 complex have been identified in *Leptospirillum* Groups II and III (Table S3). Cytochrome b/b6 protein is bifurcated and the c1 component contains binding sites for four heme prosthetic groups. Proteins with sequence similarity to two subunits of a cytochrome cbb_3 oxidase were predicted in *Leptospirillum* Groups II and III, and both subunits are duplicated (Table S3). Only *Leptospirillum* Group II cytochrome cbb_3 oxidase gene products were identified by proteomics. Subunits of a cytochrome *bd* oxidase were also predicted in both *Leptospirillum* genomes (Table S3), however peptides for these proteins were not observed in any proteomic datasets.

A conserved cluster of 14 NADH dehydrogenase genes is present in *Leptospirillum* Groups II and III (Table S4), and all of the corresponding proteins were identified by proteomics. In addition, there are several extra copies of different subunits of NADH dehydrogenase, two copies of NADPH:quinone reductase and other energy genes (cytochromes) scattered around the genomes, usually in plasmid/phage regions.

B. CO_2 fixation. Both *Leptospirillum* Groups II and III isolates from the Richmond Mine can be grown without sources of fixed carbon and all previously characterized *Leptospirillum* species only grow chemoautotrophically (2, 11). Therefore, it is very likely that the *Leptospirillum* species in the biofilm utilize CO_2 as their sole carbon source. The most common autotrophic pathway is the Calvin-Benson cycle, however both *Leptospirillum* Group II and III lack ribulose-5-phosphate kinase, a key enzyme in the pathway. *Leptospirillum* Group II and III have two and three copies of a ribulose-bisphosphate carboxylase-like protein

(RuBisCO-like), respectively, but none were predicted to have carboxylase and oxygenase activity based on phylogenetic classification with the Form IV Rubisco-like group (68). Thus, neither *Leptospirillum* Group II nor III appear to use the Calvin-Benson cycle for CO_2 fixation. The RuBisCO-like proteins were identified by proteomics and are likely involved in sulfur metabolism (68).

It is unlikely that *Leptospirillum* Groups II and III fix carbon via the Wood–Ljungdahl pathway, which converts CO_2 to acetyl-CoA through the bifunctional enzyme carbon monoxide dehydrogense/acetyl CoA synthase. Although both have paralogous genes with sequence similarity to CO dehydrogenase (CODH), the predicted proteins lack the active site for CODH and acetyl-CoA synthase based on multiple alignments. It is interesting, however, that one of the candidate CODH genes is in the second cluster of pyruvate:ferredoxin oxidoreductase (PFOR) instead of the epsilon subunit in *Leptospirillum* Group II. All copies of annotated CO dehydrogenase proteins were identified at various levels by proteomics.

The most likely pathway for carbon fixation is via the reductive tricarboxylic acid cycle (rTCA) (Figure 6). Several proteins involved in the rTCA cycle, especially PFOR and isocitrate dehydrogenase, are identified at high levels in proteomic datasets. A key component in the rTCA cycle, phosphoenolpyruvate (PEP) carboxylase, could not be initially located in either species. However, based on a TIGRFam domain and protein structure prediction, we annotated a protein of unknown function as phosphoenolpyruvate carboxylase (Table S1 gene 8241_GENE_507; Table S2 gene 7952_GENE_51; Figure 7). ATP citrate lyase and 2-oxoglutrate:ferredoxin oxidoreductase, were not annotated in *Leptospirillum* Groups II and III. Recent work on *Hydrogenobacter thermophilus* TK-6, an aerobic hydrogen-oxidizing bacterium has demonstrated a novel citrate-cleaving reaction catalyzed by two enzymes, citryl-CoA synthetase and citryl-CoA lyase (69). The first shares similarity with succinyl-CoA synthetase and both *Leptospirillum* species contain two different copies of this enzyme. *H. thermophilus* citryl-CoA lyase shares strong similarity with *Leptospirillum* citrate synthase. Both enzymes were identified by proteomics and may be involved in citrate cleavage. Additionally, both genomes have duplicated operons for PFOR, one of which may carboxylate 2-oxoglutarate instead of pyruvate (70). These results are in agreement with reverse transcriptase PCR results suggesting CO_2 fixation via the rTCA pathway in a strain related to *L. ferriphilum* for which the complete genome sequence has not been deposited (71).

The first part of the rTCA cycle is shared with the pathway for CO_2 incorporation recently described in Archaeum *Igniococcus hospitalis* (72, 73) where acetyl-CoA is carboxylated to pyruvate by PFOR. *I. hospitalis* has been shown to regenerate acetyl-CoA through a complex pathway with 4-hydroxybutyryl-CoA as a central intermediate (73), however *Leptospirillum* Groups II and III lack the genes for this route. If this pathway operates in *Leptospirillum*, a novel pathway to regenerate acetyl-CoA is required.

C. TCA cycle. *Leptospirillum* Group II has genes for most steps in the TCA cycle. The dihydrolipoamide dehydrogenase subunit of the oxoglutarate dehydrogenase complex was predicted and identified by proteomics (intriguingly, three copies are present), but the other two subunits of this complex were not found. The oxoglutarate dehydrogenase complex is similar to the pyruvate dehydrogenase complex, which *Leptosprillum* Group II lacks (likely the carboxylation step catalyzed by PFOR reverses to break down pyruvate to acetyl-CoA). Incomplete TCA cycles have been shown in chemoautotrophs as biosynthetic rather than energy generation pathways (74).

Leptospirillum Group III may have a complete TCA cycle. The uncertain component is also the oxoglutarate dehydrogenase complex. *Leptospirillum* Group III has three subunits that could serve this function or could alternatively be a pyruvate dehydrogenase complex, otherwise lacking in *Leptospirillum* Group III. The key E2 component has a relatively high enzyme-specific profile score for oxoglutarate dehydrogenase (EC 2.3.1.61), while the component E1 could belong to either oxoglutarate or pyruvate dehydrogenase (EC 1.2.4.1). As in *Leptospirillum* Group II, there are three copies of dihydrolipoamide dehydrogenase at different locations in the genome; none are in proximity to the other putative oxoglutarate dehydrogenase complex components. The putative E1, E2 and all dihydrolipoamide dehydrogenase proteins were identified by proteomics.

D. Gluconeogenesis, ***Glycolysis and Sugar Metabolism:*** *Leptospirillum* Groups II and III have all the enzymes needed for gluconeogenesis.

Leptospirillum Groups II and III do not appear to metabolize glucose through the Entner-Doudoroff pathway. Most enzymes in glycolysis (Embden-Meyerhoff) are present in *Leptospirillum* Groups II and III. A key enzyme, phosphofructokinase, has not been identified in either species. A hexokinase is also missing in both organisms, however one of several carbohydrate kinase family proteins may confer this function. Pyruvate kinase, the last energy-generating step in glycolysis, was only found in *Leptospirillum* Group III. Thus, *Leptospirillum* Group III may carry out glycolysis but this function is apparently not possible for Group II. All the proteins identified as potentially involved in gluconeogenesis/glycolysis in *Leptospirillum* Groups II and III were identified by proteomics.

Leptospirillum Group III has two copies of glucoamylase, an enzyme that degrades starch to glucose, and both copies were identified by proteomics. *Leptospirillum* Group II lacks glucoamylase but has genes for degradation of extracellular maltose. It is possible that *Leptospirillum* Group III uses glucoamylase for mobilization through the biofilm.

II. Nitrogen and sulfur metabolism.

A. Nitrogen metabolism genes. *Leptospirillum* Group III carries all the genes for nitrogen fixation (8, 48). These are next to a cluster of molybdenum uptake genes necessary for nitrogen fixation (75). Notably, in our current analysis, the proteins involved in nitrogen fixation were not identified in any of the samples (Table S2).

Leptospirillum Group III has four nitrogen regulatory transduction proteins P-II involved in nitrogen regulation or sensing a-ketoglutarate (76). Two associated with the nitrogen fixation region are not identified by proteomics, thus probably regulate the nitrogen fixation. Of the two identified by proteomics, one is clustered with an ammonium transporter and is likely regulating ammonium uptake and the other is clustered with redox enzymes.

Although *Leptospirillum* Group II does not fix nitrogen, it does harbor various nitrogen metabolism genes. There are three ammonium transporters (all identified by proteomics) clustered with nitrogen regulatory proteins PII. This gene organization is very conserved among other organisms, and suggests that the regulatory proteins are related to ammonium uptake (76, 77). Once inside, the ammonium is assimilated by the glutamine synthase / glutamate synthase pathway (78). Ammonium uptake proteins present in *Leptospirillum* Group II were identified at high levels by proteomics. One copy of nitrogen regulatory protein PII is located close to a transcriptional regulator NifA, Fis family, and shows high protein coverage. Although nitrogen fixation proteins, including NifL, are not present in *Leptospirillum* Group II, NifA may still be involved in nitrogen sensing.

In addition to acquiring ammonium via uptake, *Leptospirillum* Group II may form ammonium from nitrite. A cytochrome c NapC/NirT family protein involved in respiratory nitrite ammonification (79) was found in *Leptospirillum* Group II, but the gene for the catalytic subunit for this route has not been identified. *Leptospirillum* Groups II and III have two genes for nitrite/sulfite reductase (ferredoxin) required in assimilatory nitrite ammonification, and could use these to directly reduce nitrite to ammonium for amino acid biosynthesis (79, 80). Only the gene products for the assimilatory route are identified by proteomics.

The finding of an ammonia monooxygenase subunit, *amoA,* in both *Leptosprillum* Groups II and III genomes is intriguing. AmoA is one of three subunits required to oxidize ammonia and contains the active site for substrate oxidation (81), however the lack of other subunits in *Leptospirillum* prevents us from inferring this functionality. AmoA was not identified by proteomics for either *Leptospirillum* species. Ammonia monooxygenase may also be involved in methane oxidation and hydrocarbon degradation (81).

B. Sulfur metabolism. *Leptospirillum* Group II and *Leptospirillum* Group III have a complete assimilatory pathway for sulfate reduction. Interestingly, APS reductase, which is present as two subunits in *Leptospirillum* species, lacks a conserved motif described by Valdes *et al*. (82). The genes for sulfate assimilation are next to a region that contains Fe-S accessory proteins and a cysteine desulfurase in *Leptospirillum* Group III, but clustered in a plasmid/phage region in *Leptospirillum* Group II.

Leptospirillum Groups II and III could oxidize hydrogen sulfide with a siroheme-like enzyme, rhodanese-like proteins or a sulfide-quinone reductase. The sulfide-quinone reductase is duplicated in both organisms and one is clustered with the cytochrome bd operon, suggesting sulfur oxidation for energy generation. Siroheme-like protein could also be involved in nitrite oxidation (83). Siroheme and rhodanese-like proteins were identified by proteomics only in *Leptospirillum* Group II, whereas sulfide:quinone reductase was identified by proteomics in both species.

III. Biosynthesis and degradation pathways

A. Cofactor biosynthesis. Biotin is a predicted cofactor for several biotin-dependent carboxylases and decarboxylases (84) and both *Leptospirillum* Groups II and Group III have the five genes required for the bacilli-type pathway using BioW. All biotin biosynthesis gene products were identified by proteomics in *Leptospirillum* Group II, but only BioA was identified in *Leptospirillum* Group III.

The biosynthesis of riboflavin and FAD in both *Leptospirillum* Groups II and III is organized in two operons and all of the gene products were identified by proteomics.

All the genes needed for thiamine biosynthesis are present in *Leptospirillum* Groups II and III and all of the gene products were identified by proteomics. Two copies of *thiS* and *thiF* are clustered with genes involved in the biosynthesis of methionine, cysteine and molybdopterin. This gene organization may reflect the common need of sulfur.

Cobalamin biosynthesis can occur via aerobic and anaerobic pathways (85). *Leptospirillum* Groups II and III are probably able to synthesize cobalamin using an anaerobic pathway. They carry 19 of the 20 steps required, including *cbiX*, the second and characteristic step for the anaerobic pathway. Although *Leptospirillum* Groups II and III lack the gene *cbiJ/cobK*, this function could be complemented in very low levels by alternative non-specific reactions as in *Methanococcus maripaludis* (86). It is possible that both organisms could also produce cobalamin through the aerobic pathway; most of the steps are shared between

pathways, and they contain *cobB,* the aerobic alternative to *cbiA*. Most of the gene products putatively involved in cobalamin biosynthesis were identified by proteomics.

B. Fatty acid and lipid biosynthesis. Both species contain the complete pathway for fatty acid biosynthesis. Most of the genes are arranged in clusters and all proteins were identified by proteomics. Only *Leptospirillum* Group II contains a fatty acid desaturase, an enzyme involved in converting saturated bonds to double bonds (87).

In addition to large clusters of genes involved in lipopolysaccharide biosynthesis, there are genes that may indicate production of glycosphingolipids (e.g., ceramide glucosyltransferase) and other membrane components (e.g., squalene/hopene) that may play roles in membrane stabilization and/or acid resistance.

***C. Degradation of aromatic compounds*.** Two putative extradiol (LigB) for cleavage of aromatic rings (part of cathecol dioxygenase) and a carboxymuconolactone decarboxylase were found in *Leptospirillum* Groups II (in a mobile element region) and in *Leptospirillum* Groups III and were identified by proteomics (at low abundance). These enzymes are part of the b-ketoadipate pathway, which degrades protocatechuate, an intermediate product of aromatic compound breakdown (88). However, LigA and other pathway steps were not found.

A 5-carboxymethyl-2-hydroxymuconate d-isomerase, an enzyme that generates an intermediary for the production of oxaloacetate in the benzoate degradation via hydroxylation pathway, is present and identified by proteomics in both species. In addition, carboxymethyl butenolidase (dienelactone hydrolase), an enzyme that generates 2-maleylacetate in the metabolic pathway of chloroaromatic compounds (89), was highly identified by proteomics in *Leptospirillum* Groups II and III. The presence of unique enzymes involved in aromatic compound metabolism suggests that *Leptospirillum* species are degrading aromatic compounds, but sources of the aromatic substrates and the final products remain unclear.

D. Cellulose biosynthesis. *Leptospirillum* Group II has the genes for biosynthesis of cellulose, cellobiose and starch/amylose. The genes for the synthesis of cellulose are clustered. The regulatory subunit is identified at very low levels, while the catalytic subunit and the subunit C were not identified by proteomics. A second copy of cellulose synthase subunit C, in cluster with an endoglucanase and a peptidoglycan glycosyl transferase, was identified by proteomics at high levels. Both copies of cellulose synthase subunit C are much shorter than the sequences with which they share similarity. Although cellulose synthase and cellulase genes were not found in *Leptospirillum* Group III genome, some unassembled reads suggest this function might be present.

E. *Proteasomes*. Within the bacteria, proteasomes involved in proteolysis have previously only been found in Actinobacteria (90), however *Leptospirillum* Group II (90) and *Leptospirillum* Group III contain two gene clusters for this pathway, and all of the gene products are identified by proteomics. Within the community genomic dataset, there are multiple contigs carrying proteasome genes, including an Actinobacterial contig with a cluster of four proteasome genes. The *Leptospirillum*-type, Actinobacterial-type, and other unassigned bacterial proteasomes from the AMD dataset cluster together in a gene tree (Figure S4), suggesting acquisition of proteasomes by via lateral gene transfer.

IV. Signal transduction and information processing.

A. Signal transduction. For most genes encoding signal transduction histidine kinase proteins in *Leptospirillum* Group II we found a syntenous ortholog in *Leptospirillum* Group III. These include chemotaxis – specific (CheA) genes, osmosensitive – specific genes, and genes

for proteins with PAS/PAC sensor domains. Several transcriptional regulators are encoded in *Leptospirillum* Groups II and III, including LysR, ArsR, Fis, LuxR, MerR and other families.

The *Leptospirillum* Group II encodes 29 diguanylate cyclase / phosphodiesterase proteins whereas 39 are encoded by the *Leptospirillum* Group III genome. One of the only four orthologs and 20 of the 25 proteins lacking orthologs were identified by proteomics in *Leptospirillum* Group II. Similarly, in *Leptospirillum* Group III, two of the four orthologs and 20 of the 35 *Leptospirillum* Group III-specific proteins were identified by proteomics.

B. Chemotaxis. *Leptospirillum* Group III lacks *cheB* and *cheR* (Figure 8), which are involved in methylation-dependent adaptation of the receptor in chemotactic and aerotactic sensing pathways (91, 92). Both are present and identified by proteomics in *Leptospirillum* Group II. *Leptospirillum* Group III contains cheV, a methyl-independent adaptation protein that is believed to interact directly with CheA (91), also present in *Leptospirillum* Group II. Overall, however, *Leptospirillum* Group III has many more genes for methyl-accepting chemotaxis sensory transducer-like proteins than *Leptospirillum* Group II, most spread across the genome and identified by proteomics.

C. DNA polymerases. *Leptospirillum* Groups II and III have a gene for DNA polymerase I, and five subunits of DNA polymerase III. These are spread around the genome and have low normalized spectral count values. A DNA polymerase family B is likely the product of lateral transfer from archaea, given its strong similarity only with archaeal proteins. Although detected, this protein is not abundant based on proteomics data.

Leptospirillum Groups II and III contain various genes involved in DNA repair and recombination, including *ruvABC*, organized in an operon and highly expressed (11).

D. RNA polymerase. *Leptospirillum* Groups II and III genomes encode the subunits of the typical bacterial RNA polymerase complex. Both carry sigma-70 (RpoD), sigma-28 (RpoF; in the chemotaxis cluster and close to the flagellar genes) and sigma-54 (RpoN).

Of particular interest is the sigma factor RpoD, which has a fused adenine phosophoribosyltransferase (APRT) domain in *Leptospirillum* Group II (Figure S5). In-depth analysis ruled out an assembly error, missed stop codon or an insertion/deletion that could have generated a frame shift between the two genes. The genes are adjacent to each other in *Leptospirillum* Group III, but not fused. The fused sigma factor shows extensive protein coverage, and at least one peptide maps between normal RpoD and APRT coding regions (Figure S5), confirming that the whole protein is translated. The APRT domain protein may have a regulatory function, perhaps connected to nucleotide synthesis. Another interesting feature of RpoD in both species is the lack of region 1.1, shown to be important in initiation of transcription in *E. coli* (93).

Leptospirillum Groups II and III lack the sigma-32 factor, which responds to extracytoplasmatic stress, for example to induce heat shock genes. All the heat shock gene products are highly abundant based on proteomics, suggesting that the niche for these organisms could require the constitutive expression of otherwise conditionally induced functions. Alternatively, an unknown heat shock sigma factor might play the role.

V. Stress and transport

A. Oxidative stress. *Leptospirillum* Groups II and III have the complete pathway to synthesize phytoene and carotene and most of the gene products were identified. Carotenoids can act as antioxidants (94) and synthesis of carotenoids by the *Leptospirillum* species could be related to radical detoxification.

Leptospirillum Group II contains the genes for rubrerythrin and peroxiredoxin, and both are very highly expressed. *Leptospirillum* Group III carries an alkylhydroperoxidase not found in *Leptospirillum* Group II, the gene products of which were highly identified by proteomics.

B. Ectoine and Trehalose. *Leptospirillum* Groups II and III can synthesize trehalose using three of four known pathways (95). *Leptospirillum* Group II has the complete pathway for ectoine biosynthesis, another compatible solute for tolerance in high salinity and high temperature environments (96). The genes are arranged in an operon (*ectABCD*) and a transporter for ectoine is located upstream the biosynthetic operon; *Leptospirillum* Group III does not have a specific ectoine transporter or genes for the synthesis of ectoine. All of the protein products in both pathways are identified by proteomics. Some of the genes for the ectoine and trehalose biosynthesis pathways were previously documented in *Leptospirillum ferrooxidans* (43), however, this is the first time the complete pathway for the biosynthesis of ectoine and hydroxyectoine is described for an acidophilic bacterium.

C. *Transport and acquisition*. *Leptospirillum* Groups II and III contain several transporters for citrate, potassium, phosphate (Pst in cluster with a phosphate uptake regulator, PhoU), sulfate transporting ATPases, and transporters related to antibiotic resistance. Most secretion proteins are next to metal efflux pumps or membrane efflux proteins. In addition, *Leptospirillum* Group II contains copper translocating ATPases, whereas *Leptospirillum* Group III lacks them. *Leptospirillum* Groups II and III contain an iron permease, many ferric uptake regulators (Fur family proteins) and several TonB-dependent receptors. Most of these proteins were identified by proteomics. Interestingly, figure 3 shows that some *Leptospirillum* Group II transport proteins are overrepresented in all three samples relative to transport proteins in *Leptospirillum* Group III.

D. Metal and antibiotic resistance. Arsenic resistance genes, such as transcriptional regulators (ArsR) and the arsenite transporter (ArsB), are present in *Leptospirillum* Groups II and III. In addition, *Leptospirillum* Group III contains ArsA (an arsenite-activated ATPase), ArsD (arsenical resistance operon trans-acting) and ArsC (arsenate reductase) arranged in an operon and identified by proteomics. These genes may be key to the ability of *Leptospirillum* to thrive in solutions that can contain mM concentrations of arsenic.

Leptospirillum Groups II and III contain a mercuric reductase (MerA) and a mercuric transcriptional regulator in a cluster with an ion channel. A phage/plasmid region in *Leptospirillum* Group II contains a probable mercuric transporter, however this protein was not identified by proteomics. The reduction of mercury is typically favored in anaerobic organisms (97). It is interesting that some strains of *Leptospirillum* Group II have a mercuric reductase frame shifted (insertions/deletions in multiple reads), suggesting that this capability may not be necessary for *Leptospirillum* Group II in the current AMD environment, where mercury levels are very low.

Several antibiotic resistance genes such as beta-lactamase, acriflavin, fusaric acid and glyoxalase family proteins are present in *Leptospirillum* Groups II and III. Most of the gene products were identified by proteomics.

VI. Mobile elements

A. Plasmid regions and extrachromosomal plasmid. Regions of an integrated plasmid are present in *Leptospirillum* Group II (Scaffold 8692) and *Leptospirillum* Group III (Scaffold 4481). The shared gene content and organization, as well as comparable amino acid identities

and GC contents to other non-plasmid regions, indicate that the blocks were acquired long ago, perhaps before the species diverged.

Many conjugal transfer proteins (Type II and IV secretion system components) from the integrated plasmid region share ~30% average amino acid sequence identity with proteins from an extrachromosomal plasmid (Table S5). Consequently, the plasmid may be associated with *Leptospirillum*. The presence of conjugation systems in the integrated and non-integrated plasmids suggests that both *Leptospirillum* types can transfer plasmids. Interestingly, only a few proteins in the integrated plasmid regions of *Leptospirillum* Groups II and III and none of the proteins in the conjugative transfer region were identified by proteomics. In contrast, many of the conjugative transfer proteins in the extrachromosomal plasmid were identified by proteomics (Table S5).

Other mobile regions in *Leptospirillum* Group II encode copper, arsenic, mercury-transporting ATPase and secretion proteins, glycosyltransferases, metal-related transcriptional regulators, and genes from the b-ketoadipate pathway. NADPH:quinone reductase, NADH dehydrogenase subunits and cytochromes are associated with other phage/plasmid regions.

Clusters of toxin-antitoxin system proteins are present in *Leptospirillum* Groups II and III. These systems are known to retain bacterial plasmids during segregation, and some have been suggested to arrest growth during nutritional stress (98). Only one antitoxin protein in *Leptospirillum* Group II was identified by proteomics in the UBA sample.

Leptospirillum Groups II and III contain two copies of reverse transcriptase genes, probably associated to group II introns (99) and all show protein coverage. A gene tree of the reverse transcriptases from *Leptospirillum* Groups II and III and other organisms places them in the chloroplast-like group II intron class (100). Evolutionary mechanisms and function of group II introns are still unknown.

A mobile element on scaffold 8524 in *Leptospirillum* Group II encodes putative defect in organelle trafficking lipoproteins (dot) and intracellular multiplication (Icm) genes. Only some orthologs for Icm and Dot proteins are present in *Leptospirillum* Group III and surrounding genes include many methyl-accepting chemotaxis sensory transducers (all identified by proteomics) without orthologs in *Leptospirillum* Group II. Icm and Dot proteins were not identified by proteomics in any sample.

Two copies of a methyltransferase of the FkbM family occur in a plasmid-like region and another copy is found amongst a large cluster of *Leptospirillum* Group II genes for biosynthesis, export, and reconfiguration of sugar/polysaccharides (only the third of these is identified by proteomics), while *Leptospirillum* Group III lacks these genes.

B. CRISPR. Clustered regularly interspaced short palindromic repeats (CRISPRs) and CRISPR-associated (CAS) genes are involved in a recently described viral and plasmid defense mechanism found in Bacteria and Archaea (28, 101). *Leptospirillum* Group II carries a cluster of Cas proteins (Cas2/1/3/5/4/2/1/3). Orthologs (mostly syntenous) occur in one of the multiple Cas clusters of *Leptospirillum* Group III. Most proteins in the orthologous clusters were identified by proteomics, and Cas protein abundance levels vary significantly amongst biofilms. Another CRISPR region also carrying this repeat occurs at a different genomic locus and in several different mobile elements (without identifiable Cas proteins, data not shown). A second *Leptospirillum* Group III CRISPR-associated cluster encodes Cas and Csm proteins, and transposases interrupt the CRISPR region and Cas1 protein (see Table S2). This CRISPR locus reconstructed from the population genomic dataset is essentially clonal (unlike almost all

other CRISPR loci in the AMD datasets) and none of the spacers match any viral sequences. These observations suggest that this second locus is inactive. However, despite the interruption of Cas1 by a transposase, three of the Csm family proteins were identified by proteomics. This second *Leptospirillum* Group III CRISPR locus is next to an integrase and is interspersed by transposases in a genomic region with many hypothetical proteins, thus it is possibly part of an integrated mobile element. The repeat from this second cluster has a region of similarity to a repeat in a CRISPR locus carried by the extrachromosomal plasmid (locus 13 in scaffold 15659, Table S5). The CRISPR spacer and repeat sequences carried by plasmid-like contigs were reported previously (17).

Spacers from the CRISPRs of the plasmid-like population match plasmid-like contigs, perhaps indicating that CRISPRs spacers are involved in competition among mobile elements. Some spacers only match regions encoding Cas proteins of other plasmids, potentially indicating CRISPR silencing of acquired resistance. Specifically, a spacer from CRISPR locus 13 (on the extrachromosomal plasmid) targets (based on nucleotide identity) the Cas3 helicase from CRISPR locus 19 (contig 15511; Figure 9). Another spacer from CRISPR locus 13 matches a hypothetical protein encoded between the Cas-cys3 and a Cas-cys2 in plasmid-like contig 12113 whereas a spacer from CRISPR locus 11 (plasmid-like contig 11387) targets the Cas_cys3 protein.

Some spacers from plasmid-borne CRISPR loci also target non-Cas genes. For example, a spacer from CRISPR locus 13 matches a hypothetical protein encoded by the same extrachromosomal plasmid (contig 11623). A DNA polymerase encoded by plasmid-like contig 15498 is also targeted by a spacer from CRISPR locus 13.

DISCUSSION

Leptospirillum Groups II and III are the first organisms from the Nitrospirae lineage for which extensive (near complete) genomic data and detailed functional annotations are available. For *Leptospirillum* Group II, there are two composite sequences from the Richmond Mine, the UBA type (11) and 5-way CG type related to *L. ferriphilum* (14). In addition, there are several isolates that are similar to these genomically characterized populations (based on multi-locus sequence typing (11)). We refer to the now extensively characterized UBA subgroup as "*Leptospirillum rubarum*". The genomes of "*L. rubarum*" and *L. ferriphilum* are syntenous, but there are major rearrangements relative to *Leptospirillum* Group III. *Leptospirillum* Groups II and III share most of their core metabolic pathways and the organization of key genes is maintained despite their divergence.

The significant evolutionary distance between the *Leptospirilli* and other organisms for which extensive genomic and biochemical information are available somewhat limits physiological interpretations. Gaps in generally well-known metabolic pathways likely reflect the paucity of information for closely related lineages, although the incomplete nature of both genomic datasets makes any firm conclusion about missing genes impossible. For this reason, proteomic identification of enzymes characteristic of specific pathways provide key indications of the associated functionality. It is important to note that, although there are information gaps, the physiological, ecological, and evolutionary insights obtained here were achieved for natural populations without cultivation.

Studies of the *Leptospirillum* genus support assignment of its members as obligate Fe(II) oxidizers (102). Therefore, Fe(II) oxidation and electron transport are key functions in the maintenance of *Leptospirillum* cellular metabolism. A bifurcated cytochrome b/b6 protein has been found previously only in gram-positive bacteria (103). This unusual structure for a bc1 complex, along with tetraheme cytochrome as the c1 component is also observed in the genome of *Candidatus* "*Kuenenia stuttgartensis"*, a Planctomycete that performs anaerobic ammonium oxidation and generates reductant for autotrophic growth by reverse electron transport (104). The high abundance of now well-studied cytochromes in *Leptospirillum* Group II (and Group III) supports the key role of iron oxidation in growth of these organisms.

For both organisms, ribosomal proteins are generally very abundant, consistent with high levels of activity. Other proteins highly identified in all three proteomic samples in *Leptospirillum* Group II are cold shock proteins and other molecular chaperones, histone-like proteins, cytochromes such as Cyt_{579}, and translation elongation factors. Although it is possible that high levels of cold and heat shock proteins are artifacts of sample preparation, preliminary proteomics data on the same biological sample subjected to different freezing regimes suggest this is not the case (Denef *et al*., unpublished). An alternative explanation is that protein-modifying, RNA-binding and nucleic acid-binding histone-like proteins perform an alternative function, possibly related to cell stabilization in the extreme environment (105).

Some proteins with clear functional annotations were not identified by proteomics. Although the absence of peptide identification could reflect low rather than no abundance or a variety of experimental problems, the failure to detect all predicted peptides from whole groups of proteins from large complexes is notable. For example, *Leptospirillum* species carry formate-hydrogen lyase clusters found in microaerophilic/anaerobic organisms (106), but none of the gene products were identified in the samples studied. This complex may convert formate to hydrogen and CO_2 or play a role in carbon fixation. Its presence in both genomes is consistent with the capability for anaerobic growth. Carbon fixation via the reductive TCA pathway, anaerobic cobalmin pathway genes, and highly abundant PFOR have been found only in known anaerobes, suggesting that *Leptospirillum* Groups II and III may grow in anaerobic as well as microaerophilic and aerobic environments.

Based on functional predictions, proteomic information, and data from studies of isolated species, both organisms are likely capable of carbon fixation and most core metabolic functions. Both face the same environmental challenges, particularly the very low pH and high concentrations of toxic metals. Notable are the identification of multiple pathways for production of compatible solutes that presumably provide a response to surrounding high ionic strength solutions, membrane stabilization molecules, and metal efflux pumps.

Icm proteins may be involved in pathogenesis and conjugation (107), and resistance to eukaryotic predation (108). In *Leptospirillum* Group II, Icm, Dot and FkbM proteins may protect against grazing by protists and from fungi that proliferate in some higher developmental stage biofilms (47, 109). Both organisms appear to have functional CRISPR/Cas loci for viral defense. To date, we have only detected evidence suggestive of silencing of one plasmid's CRISPR locus by another. However, the same approach would be effective to silence the host resistance system (and is likely if lateral transfer from mobile elements is in fact the major source of bacterial and archaeal CRISPR loci).

Differences in gene complement point to important physiological distinctions that may have been key to the likely sympatric divergence of the *Leptospirillum* Groups. *Leptospirillum* Group II is better equipped than *Leptospirillum* Group III to deal with osmotic challenges

associated with the near molar $FeSO_4$ solutions and produce potentially key polymers for establishment of floating biofilms (e.g., cellulose, cellobiose and starch/amylose). *Leptospirillum* Group III is apparently better optimized for energy generation (given a possible complete glycolysis and TCA pathways) and nitrogen fixation. These findings are consistent with the identification of *Leptospirillum* Group II as the early colonist and *Leptospirillum* Group III as a late biofilm developmental stage member (47).

Intriguingly, the complements of signal transduction and chemotaxis genes in *Leptospirillum* Group II and *Leptospirillum* Group III are quite different, as are many regulatory genes, pointing to adaptation to different microenvironments (e.g., with specific levels of oxygen, redox potential, availability of fixed nitrogen). Genomic and proteomic data suggest that signal transduction, motility and chemotaxis are more important in *Leptospirillum* Group III than Group II (Figure 3). Biofilm characterization studies place *Leptospirillum* Group III as dispersed cells and microcolonies in interior regions of biofilms (47), where geochemical gradients are expected to be pronounced. This distribution, in combination with the inferred metabolic characteristics, may point to *Leptospirillum* Group III as a microaerophile that locates in nutrient poor regions of biofilms, where its ability to fix nitrogen may be key. The distribution of *Leptospirillum* Group II at the base of some biofilms, where oxygen availability is almost certainly low, may indicate an optional but as yet incompletely defined anaerobic metabolism (for example, making use of a within-biofilm nitrogen cycle).

Nitrogen fixation proteins of *Leptospirillum* Group III were not identified by proteomics. Both the late arrival of this organism during biofilm development (47) and the near absence of evidence for nitrogen fixation are at odds with the simplest ecological model in which this organism is the keystone species and first colonist. An important observation is that the biofilms studied here form at the confluence of drainage streams with sources throughout the biologically active and probably highly productive subsurface ecosystem. Thus, a significant load of fixed nitrogen in influent solutions is not surprising (Kalnejais *et al*. unpublished). Furthermore, biofilm recycling occurs periodically when biofilms sink. Thus, it is perhaps expected that *Leptospirillum* Group III invest little energy in nitrogen fixation in the biofilms studied here. Likely, this function is important when these bacteria grow in microenvironments not yet studied (e.g., in association with pyrite surfaces in the sediment). Alternatively, nitrogen fixation may occur below detection levels in anaerobic regions of thicker biofilm where fixed nitrogen has been depleted by surrounding organisms.

The identification of an extrachromosomal plasmid with relatively high proteome coverage indicates that the physiology of both *Leptospirillum* cannot be described based on their core metabolic potential alone. The effect of plasmids on the metabolism of these bacteria is difficult to deduce because most proteins identified have no known function. It is interesting that the few proteomically-identified proteins from integrated plasmids (a subset of which are common to both *Leptospirillum* types and may predate lineage divergence) are not involved in conjugation whereas the conjugation apparatus of the extrachromosomal plasmids may contribute to ongoing plasmid transfer.

The genomic and proteomic datasets provide evidence of interesting new biochemical functionalities. For example, the identification of a sigma factor with the fused adenine phosphoribosyltransfersase points to potentially novel aspects of genome regulation. In addition to the indications of functional differentiation noted above, comparative proteogenomic analyses highlight many proteins of unknown function that are unique and, in some cases, highly expressed. These are obvious targets for future functional screening and

crystallography studies (110). Some of these species-specific proteins are relatively abundant in only a subset of biofilm communities, suggesting roles in microenvironmental adaptation. Expression profiles of proteins shared in *Leptospirillum* species strongly cluster by organism rather than environment, suggesting that the organism is more important in determining protein expression pattern than environmental parameters. Proteins shared by both *Leptospirillum* types, but otherwise unique, likely reflect Nitrospirae-lineage adaptations to the low pH, metal-rich environments.

CONCLUSION

This is the first in-depth functional (simultaneous proteomic and genomic) analysis of closely related species as members of communities in the same natural environment, and the first detailed genome-based functional analysis of members of the Nitrospirae lineage. Given that they consistently coexist in acidic, metal-rich environments, *Leptospirillum* group II and III most likely underwent sympatric divergence. Documented differences in their genotypes and protein expression patterns (proteins that cluster by organism not environment) may largely account for different ecological behavior, such as the early predominance of *Leptospirillum* Group II and different partitioning of *Leptospirillum* Groups II and III (47). Specifically, we highlighted important differences in the complement of chemosensory genes and in levels of their protein products, differences in metabolic potential, distinct expression patterns of both orthologous and unique proteins of unknown function, as well as notable differences in complements of signal transduction proteins. This study demonstrates the power of combining comparative, cultivation-independent genomics and community proteomics to study closely related organisms within their natural environment.

ACKNOWLEDGEMENTS

We thank Mr. T. W. Arman, President, Iron Mountain Mines Inc. and R. Sugarek for access to the Richmond Mine and Mr. R. Carver for on-site assistance. E. Watkin provided genotype information for a *Leptospirillum* Group III colony 5 isolate (UBA:5). P. Abraham is thanked for his assistance with UBA proteomic measurements. DNA sequencing was carried out at The Joint Genome Institute. DAG acknowledges support from an NSF Graduate Research Fellowship. Oak Ridge National Laboratory is managed by University of Tennessee-Battelle LLC for the Department of Energy under contract DOE-AC05-00OR22725. This research was supported by The U.S. Department of Energy, Office of Biological and Environmental Research, Genomics: Genomes to Life Program.

CHAPTER 1 FIGURES

Figure 1: Map showing sampling locations.

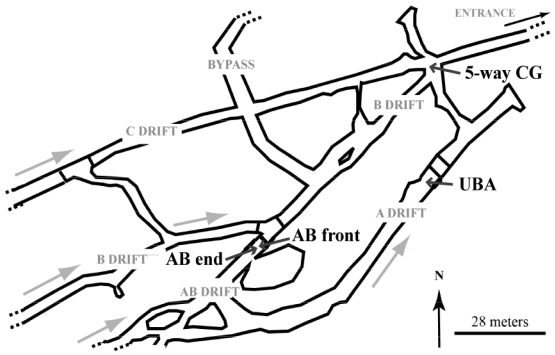


Figure 2. Phylogenetic tree based on 16S rRNA genes of *Leptospirillum* spp. (maximum likelihood method). Statistically supported bootstrap values are labeled at the nodes. Scale bar is 0.10 changes per site or 10%. Filled squares indicate isolates while filled circles indicate composite genomes.

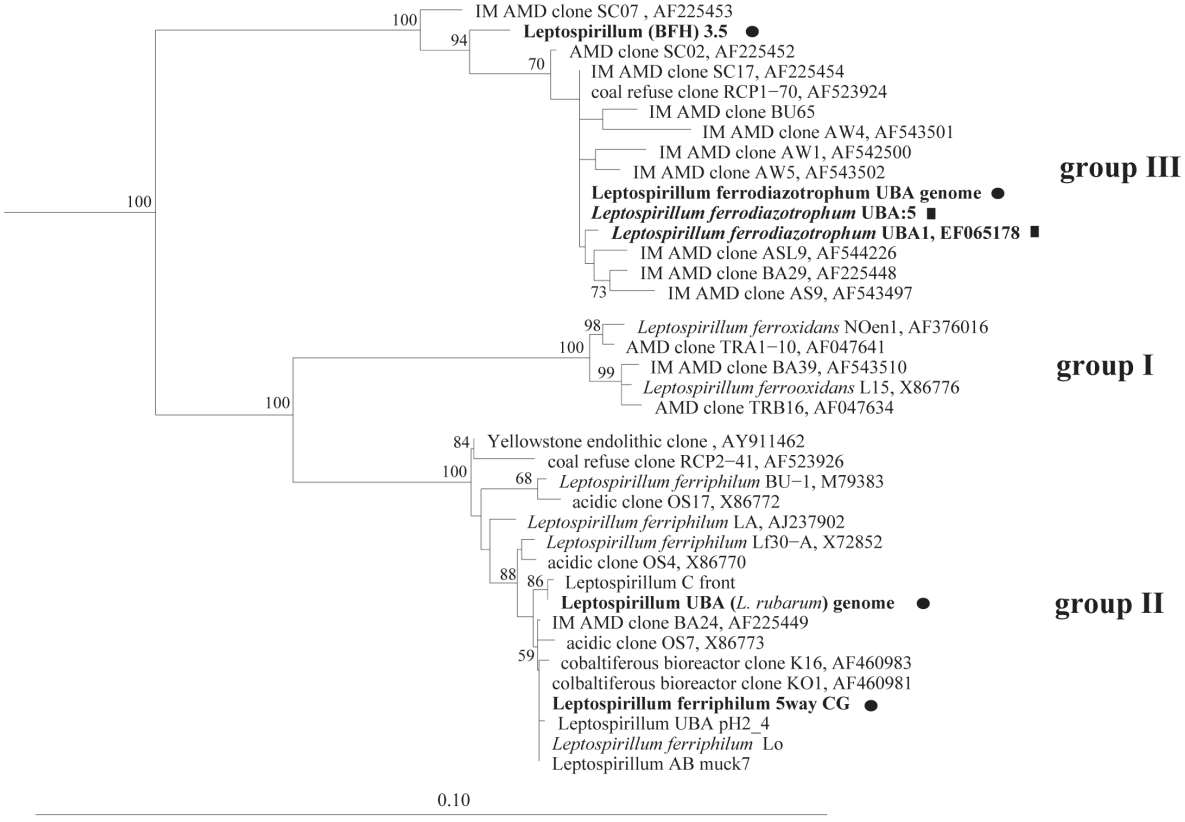


Figure 3. Diagram of the genome of *Leptospirillum* Group II (outer circle) and orthologs in *Leptospirillum* Group III (inner circle, color-coded by scaffold). Histogram of percent identity between orthologs is black bars on inner ring. Heat maps of protein identification values (NSAF) are given by six middle rings (grey: no expression).

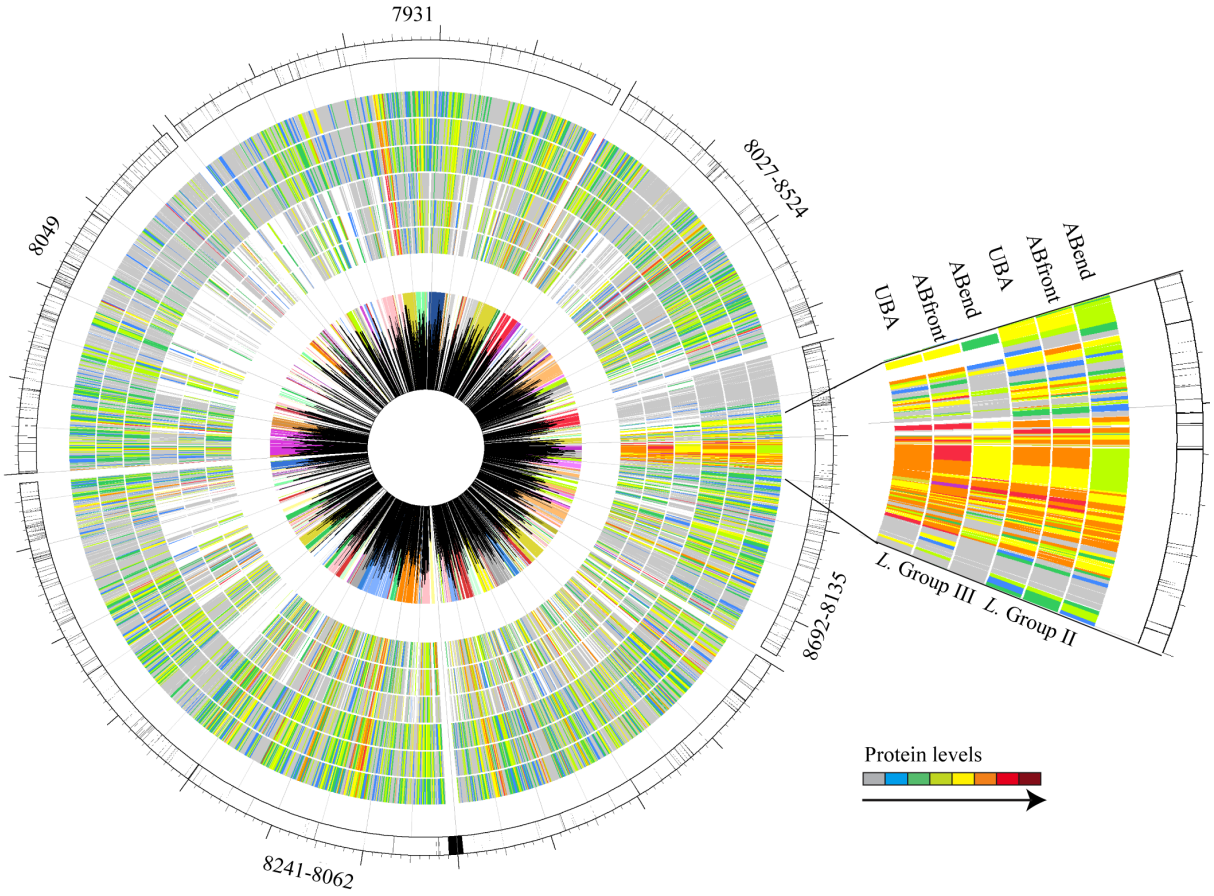


Figure 4. Protein abundance values (NSAF) for *Leptospirillum* Groups II and III orthologs in samples UBA, ABfront (AB-F) and ABend (AB-E). Red: overrepresented, green: underrepresented, black: median, grey: no identification. Functional categories for the most abundant proteins in three clusters (number of examples in parentheses): Cluster A (only part shown): Transcription, translation, ribosomal structure and biogenesis (7), coenzyme transport and metabolism (6), transport and secretion (5), energy production and conversion (3), others (23). Cluster B: Transport and secretion (10), translation, ribosomal structure and biogenesis (5), posttranslational modification, protein turnover, chaperones (4), lipid transport and metabolism (2), others (25). Cluster C: Energy production and conversion (10), cell motility (10), amino acid transport and metabolism (6), transcription, translation, ribosomal structure and biogenesis (6), others (24)

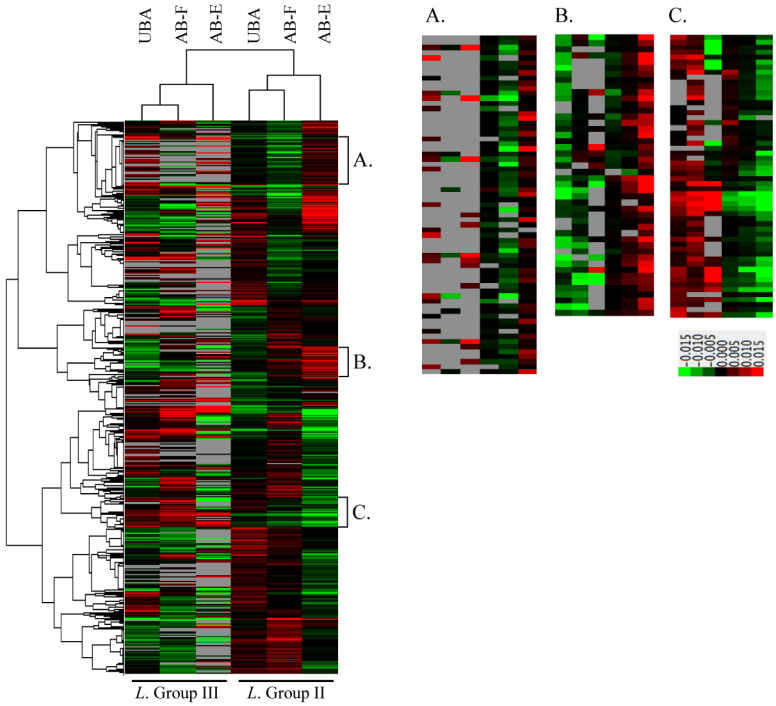


Figure 5. Inferred abundances (NSAF) of proteins of unknown function in *Leptospirillum* groups II and III. A) orthologs, B) unique to *Leptospirillum* Group II, and C) unique to *Leptospirillum* Group III. Red: overrepresented, green: underrepresented, black: median, grey: no expression.

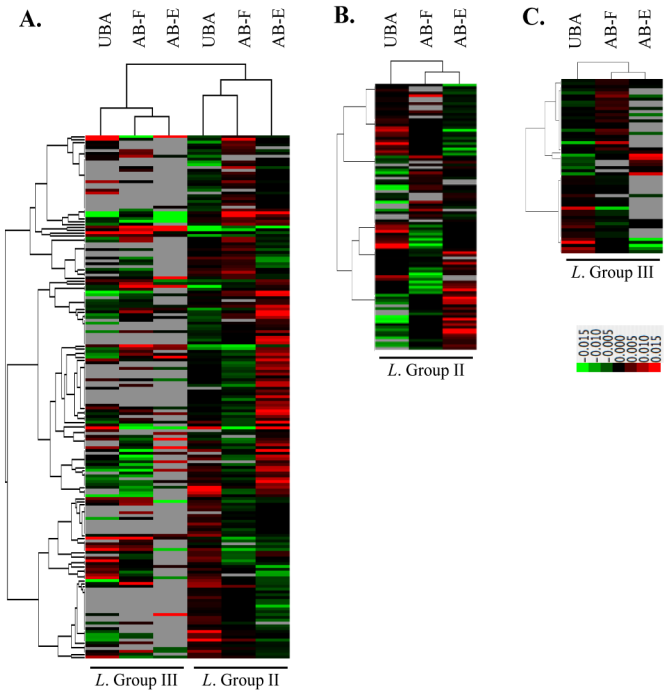


Figure 6. Proposed CO_2 fixation pathway (rTCA) for *Leptospirillum* Groups II and III. 1. Pyruvate:ferredoxin oxidoreductase (PFOR), 2. Phosphoenolpyruvate (PEP) synthase, 3. Phosphoenolpyruvate carboxylase (PEPC), 4. malate dehydrogenase, 5. fumarate hydratase, 6. Fumarate reductase, 7. Succinyl-CoA synthetase, 8. PFOR (second copy), 9. Isocitrate dehydrogenase, 10. Aconitate hydratase, 11. Succinyl-CoA synthetase (second copy) and citrate synthase.

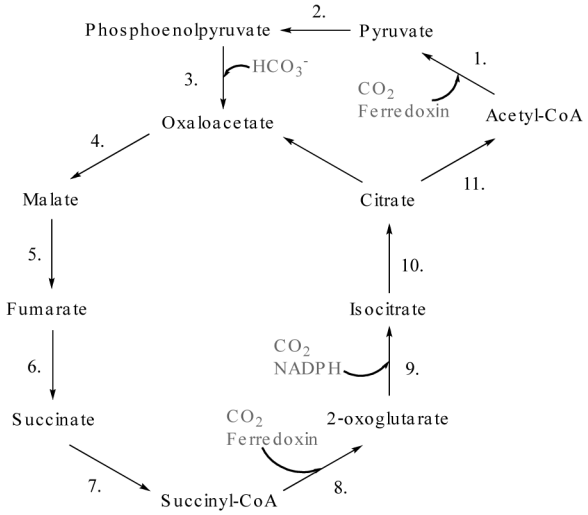


Figure 7. A) Alignment of PEP carboxylase regions containing conserved active site residues (in red) from Maize and *E. coli* [Matsumura, 2002], and *Leptospirillum* group II. Other identical residues in purple. B) Predicted protein structure of PEP carboxylase in *Leptospirillum* group II with active site residues shown in red.

A.

```
L. Group II      R104 VFLTFRLPNIW   E191 VIPLIEGVPQL   D229 FIARSDPALNAG
E. coli          R396 VRIDIRQESTR   E506 VAPLFETLDDL   D543 MIGYSDSAKDAG
Maize            R456 VKLDIRQESER   E566 VVPLFERLADL   D603 MVGYSDSGKDAG

L. Group II      R273 GSLPFRGGLNP    R382 IGLFGYSRG-IGQKRLPRAISFTGA R391
E. Coli          R587 GGSIGRGGAPA    R703 LGSRPAKRRPTGGVESLRAIPWIFA R713
Maize            R647 GGTVGRGGGPT    R763 IGSRPAKRRPGGGITTLRAIPWIFS R773
```

B.

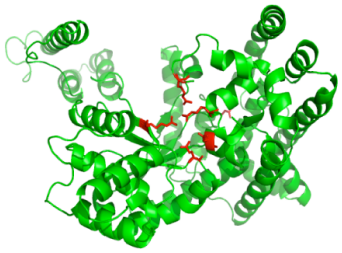


Figure 8. A) Diagram of the chemotaxis gene cluster in *Leptospirillum* Groups II and III. Orthologs shown in grey, unique proteins are shown in black pattern. MCP: methyl-accepting chemotaxis sensory transducer. Cartoon of predicted methyl-dependent (B) and methyl-independent (C) chemotaxis systems in *Leptospirillum* Groups II and III, respectively. Adaptation chemotaxis proteins: R, CheR; B, CheB; and V, CheV.

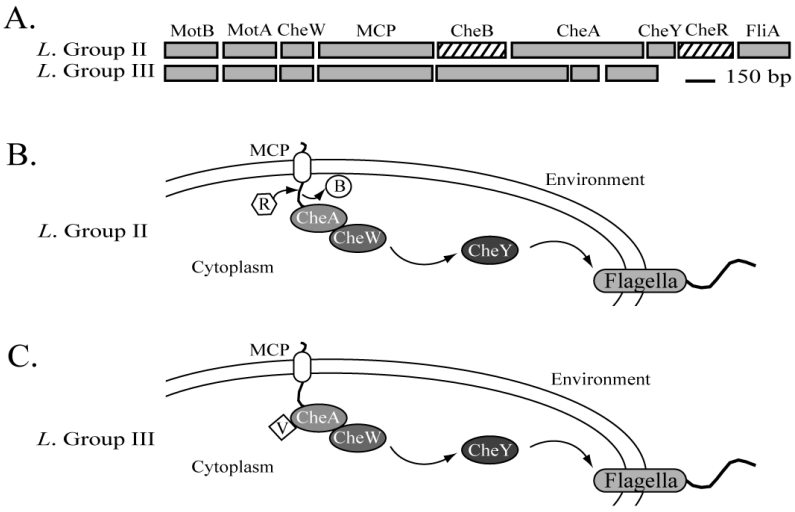
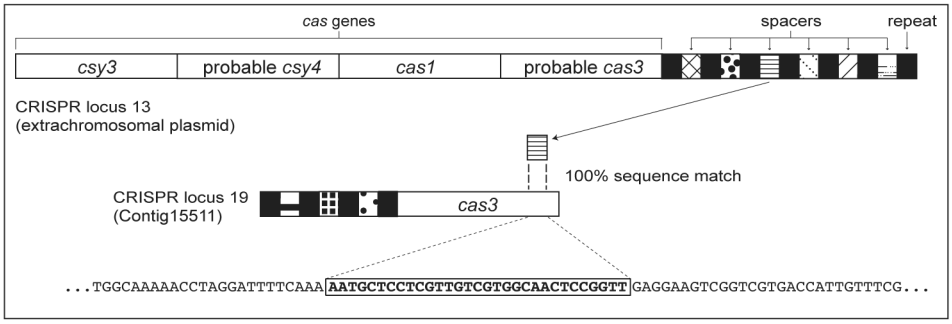


Figure 9. Diagram of CRISPR/CAS loci associated with the extrachromosomal plasmid. Black bars show repeat sequences, while bars between the repeats represent spacer sequences. (Note: While *cas* genes are displayed accurately, the sets of spacers are shown schematically). A spacer from CRISPR locus 13 (extrachromosomal plasmid) targets a *cas*3 gene from CRISPR locus 19 (plasmid-like Contig 15511). The inset displays a portion of the cas3 gene targeted by the spacer shown bold.



CHAPTER 2

A new group in the *Leptospirillum* clade: cultivation-independent community genomics, proteomics and transcriptomics of the new species *Leptospirillum* group IV UBA BS.

ABSTRACT

Leptospirillum spp. are widespread members of acidophilic microbial communities that catalyze ferrous iron oxidation, thereby increasing sulfide mineral dissolution rates. These bacteria play important roles in environmental acidification and are harnessed for bioleaching-based metal recovery. Known members of the *Leptospirillum* clade of the *Nitrospira* phylum are *L. ferrooxidans* (group I), *L. ferriphilum* and "*L. rubarum*" (group II), and *L. ferrodiazotrophum* (s). In the Richmond Mine acid mine drainage (AMD) system, biofilm formation is initiated by "*L. rubarum*"; *L. ferrodiazotrophum* appears in later developmental stages. Here we used community metagenomic data from unusual, thick floating biofilms to identify distinguishing metabolic traits in a rare and uncultivated community member, *Leptospirillum* group IV UBA BS. These biofilms typically also contain a variety of Archaea, *Actinobacteria*, and few other *Leptospirillum* spp. *Leptospirillum* group IV UBA BS shares 98% 16S rRNA sequence identity and 70% average amino acid identity between orthologs with its closest relative, *L. ferrodiazotrophum*. The presence of nitrogen fixation and reverse TCA cycle proteins suggest an autotrophic metabolism similar to that of *L. ferrodiazotrophum*, while hydrogenase proteins suggest anaerobic metabolism. Community transcriptomic and proteomic analyses demonstrate expression of a multicopper oxidase unique to this species, as well as hydrogenases, and core metabolic genes. Results suggest that *Leptospirillum* group IV UBA BS might play important roles in carbon fixation, nitrogen fixation, hydrogen metabolism, and iron oxidation in some acidic environments.

INTRODUCTION

Some members of acidophilic microbial communities catalyze ferrous iron oxidation, accelerating ferric iron-mediated oxidative dissolution of sulfide minerals and thus formation of acid mine drainage (AMD) (4, 111). Because of this capacity, acidophilic microbial communities are harnessed to release metals from sulfide minerals in biomining (reviewed in (112)). Microbial communities, especially those adapted to very low pH conditions (< pH 2), are often dominated by *Leptospirillum* bacteria of the phylum *Nitrospira* (3, 42, 113). To date, there are three recognized groups within the clade *Leptospirillum* based on 16S rRNA phylogeny: group I (*L. ferrooxidans*), group II (*L. ferriphilum* and "*L. rubarum*"), and group III (*L. ferrodiazotrophum*) (44, 48, 114). All three *Leptospirillum* groups have been observed in 16S rRNA gene surveys and metagenomic studies from acidic and bioleaching environments worldwide (e.g. 3, 46, 115-117). Based on isolate characterization studies, all are iron-oxidizing chemoautotrophs and two groups (I and III) are reported to be capable of nitrogen fixation (48, 118). Near complete genomes for "*Leptospirillum rubarum*" (UBA type), *Leptospirillum* group II 5wayCG type, and *L. ferrodiazotrophum* have been recovered from community genomic datasets (11, 12, 14) and the complete genomes of *L. ferrooxidans* and *L. ferriphilum* isolates are now available (119, 120). Much research has focused on microns- to hundred micron-thick floating biofilms sampled from the Richmond Mine at Iron Mountain, California, that are typically dominated by *Leptospirillum* group II (e.g. 47, 121). Community genomics, proteomics and transcriptomics have allowed for the cultivation-independent study of generally abundant members of these communities (e.g. 8, 10, 12, 19, 122), however the roles of lower abundance community members have not been well studied. Here we describe unusual, near

centimeter-thick floating biofilm communities that grow within the Richmond Mine and report the partial genome of a new group in the *Leptospirillum* clade: *Leptospirillum* group IV UBA BS. These bacteria comprise less than 3% of the sequenced community, so this study demonstrates the power of community genomics for achieving insight into the physiology of relatively low abundance community members.

MATERIALS AND METHODS

Biofilm samples, which we refer to as "UBA BS", were obtained from the A-drift tunnel from within the Richmond Mine, at Iron Mountain, CA (40°40′ 38.42″ N and 122″ 31′ 19.90″ W, elevation of ~900 m). Samples were collected in November 2005 (Nov05), August 2007 (Aug07), November 2007 (Island 2 and 3), June 2008 (Jun08), and December 2011 (Dec11). Sample Nov05 was subject to Sanger sequencing, as previously described (8, 11). Additionally, community proteomic data were obtained for Nov05 and Aug07, as described earlier (9, 12). Briefly, proteins were released from biofilms via sonication, fractionated based on cellular location, denatured and reduced with 6M Guanidine/10mM DTT, digested using sequencing grade trypsin (Promega, Madison, WI), desalted and analyzed via two-dimensional nano LC-ES-MS/MS (linear ion trap Orbitrap Thermo Fischer Scientific). For the Nov05 samples, two protein extraction methods were used as described in (9): the first using an acidic buffer referred to as M2 buffer, and the second using a 0.1 M sodium acetate (pH 5.0) buffer (S buffer). The resultant MS/MS spectra from individual runs were then used to search a database of predicted proteins from AMD genomic sequences as well as common contaminants (trypsin/keratin) with SEQUEST and filtered with DTASelect (9). For comparison of protein abundance levels for *Leptospirillum ferrodiazotrophum* and group IV UBA BS, proteomics data were analyzed using clustered normalized spectral abundance factors (NSAF), as described earlier (18, 62).

Fluorescence *in situ* hybridization (FISH) was done on samples Aug07, Island 2 and 3, Jun08, and Dec11 as described in (123) using the following probes: Eubmix (general bacteria), Arc415 (general archaea), Lf1252 (*Leptospirillum* group III specific), and Lf288CG (*Leptospirillum* group II 5wayCG type specific probe). We have designed two probes that target the *Leptospirillum* group IV UBA BS 23S rRNA gene: LIV307 (5'-CCCTCTTTGGCGGACCTTTC-3'); and LIV1191 (5'- CACTCCAGGCCGAACGCTCC-3'). FISH was done using 40% formamide concentration.

Community genomics data obtained from sample UBA (11) and the Nov05 biofilm (UBA BS) were used to assemble the partial genome of *Leptospirillum* group IV UBA BS. Briefly, reads belonging to "*L. rubarum*" and *L. ferrodiazorophum* were removed from both datasets and the remaining reads co-assembled using Phred/Phrap/consed as described previously (17). Contigs were binned using ESOM (13), coverage and sequence similarity to *L. ferrodiazotrophum*. Manual curation of the assembly was done using methods reported previously (12). To confirm the accuracy of the binning, reads belonging to "*L. rubarum*" and *L. ferrodiazotrophum* were removed from an additional community genomic dataset (5way CG, (8)), and the remaining reads were mapped onto the assembled genome of *Leptospirillum* group IV UBA BS using gsMapper (Roche/454) with parameters 90% minimum sequence identity and 40 bp minimum overlap. Automated annotation was done using an in-house pipeline and gene calls were manually curated using BlastX (56) against the Swissprot database.

To assess abundance of organisms in the genomically characterized UBA BS Nov05 biofilm sample, reads were mapped to the available genomes of AMD organisms using gsMapper with parameters 99% sequence identity and 40 bp minimum overlap. Relative abundance of each organism was estimated based on coverage statistics. To calculate coverage, the number of reads mapping to all scaffolds binned to each organism was multiplied by the average read length (800 bp) and the result divided by the genome size (cumulative length of scaffolds in each bin).

Biofilms for transcriptomic exploration were grown in laboratory bioreactors, as described in (124), in the dark, at pH 1 and 37 °C, and harvested at early and mid stages of development. Briefly, bioreactors consist of a Teflon channel (30 cm long x 5 cm wide x 3 cm deep), which allows the acidic modified 9K medium to flow at a constant rate (124). The medium was recycled through the reactor until oxidized (turning from bright green to red color), at which point spent medium was replaced by fresh 9K medium.

Community transcriptomic data were obtained from eight environmental and five bioreactor grown samples. Biofilms were lysed using the MirVana lysis buffer (Ambion) and bead-beating. Total RNA was extracted using acid phenol:chloroform:isoamyl alcohol (Ambion), pelleted with cold isopropanol for about 1 hour and immediately purified using the RNEasy MinElute kit (Qiagen). Integrity of the RNA was confirmed using a Bioanalyzer 2100 (Agilent Technologies). An aliquot of good quality RNA (RNA integrity number >7) from six environmental and three bioreactor samples underwent ribosomal RNA depletion using the MicrobExpress kit (Ambion). Good quality total and rRNA-depleted RNA was converted to cDNA using Superscript III (Invitrogen) as described in (125), and the cDNA was fragmented with a Covaris S-system (Covaris, Inc.) to an average fragment size of 200 bp. Fragmented cDNA was sent to the University of California Davis Sequencing facility for Illumina genomic library preparation and sequencing. Samples were indexed to sequence multiple samples in an Illumina lane.

To separate ribosomal from non-ribosomal reads, transcriptomic reads were mapped to a modified Silva database containing ribosomal RNA genes from AMD bacteria and archaea (126) using bowtie (127) with parameters –v 1 –best –y. Non-ribosomal reads from the nine rRNA-depleted samples were pooled with non-ribosomal reads obtained from the corresponding total RNA. Non-ribosomal reads from all 13 samples were then mapped to the partial *Leptospirillum* group IV UBA BS genome using bowtie with parameters –v 1 –best –y. Transcript abundance per gene was normalized by dividing the read counts by the gene length, and the resulting value was divided by the total sum of the length-normalized values in each sample.

The 16S rRNA phylogenetic tree was built using ARB (128) with 1000 bootstraps. Protein model predictions were done using the Phyre website (129) and visualized using Pymol (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC). Ligand binding was predicted using the 3DLigandSite webserver (130). Clustering of NSAF values was done using the Cluster v3.0 for Mac OSX, centering genes and samples by the median, using the Spearman Rank Correlation similarity matrix, and Average Linkage as clustering method (64). The heatmaps were visualized with the TreeView software (65).

The genome annotation and proteomics datasets are already publically available via the open knowledgebase ggKbase website (http://genegrabber.berkeley.edu/amd/organisms/506)

RESULTS

Diversity of UBA BS biofilms

The UBA BS biofilms occur as near-centimeter thick aggregates of fuzzy strips that float on the surface of slowly draining AMD ponds (Supplementary Figure S1). The thickness of these biofilms is remarkable, given that previously described AMD biofilms are rarely more than 200 μm thick (47). Biofilm Nov05 was collected from an AMD pool in the A-drift of the Richmond Mine in Northern California (see methods). The temperature of the pool was 38 °C and pH 0.89. DNA and proteins were recovered from this sample for metagenomic sequencing and mass spectrometry-based proteomic analysis. Community proteomic data were also obtained for biofilm Aug07, which was collected from approximately the same site (36 °C and pH 1.94). In addition, five UBA BS biofilms, including Aug07, were analyzed with fluorescence *in situ* hybridization (FISH) for organism identification (Table 1, Figure 1, and Supplementary Figure S2). Cell counts obtained from FISH data show that archaea appear to dominate these biofilms (average 68% ± 9%), and *Leptospirillum* bacteria represent an average 34% of the organisms detected by FISH (Figure 1A and Table 1). Only one UBA BS biofilm appeared to be dominated by *Leptospirillum* group III, but also contained *Leptospirillum* group II 5way-CG at low abundance (Sample Dec11, Supplementary Figure S2). Most bacteria that bound the general *Leptospirillum* FISH probe (data not shown) were not labeled by groups II- and III-specific FISH probes, suggesting the presence of an unknown *Leptospirillum* type in these samples (e.g., Figure 1C (Jun08), Figure 1D (Aug07), and Figure 1E (Island 2)). *Leptospirillum* group II 5wayCG type was detected at low abundance (Table 1, Figure 1D, and Supplementary Figure S2), but "*L. rubarum*" (group II UBA type) was not identified in any of the biofilms. Protists were present in some biofilms, however fungi were not observed with FISH (data not shown).

An average 84% of the reads from the Nov05 community genomics dataset mapped to assembled genome fragments from bacteria or archaea, including fragments binned to known organisms (8, 11-13, 16), one bacteriophage population, three viruses of archaea (17), and one *Leptospirillum* mega-plasmid population (Figure 2A). The remaining reads must be associated with community members whose abundance is low enough to preclude assembly. Based on coverage, *Thermoplasmatales* lineage archaea E-plasma, A-plasma and C-plasma, a related archaeon, I-plasma, and ARMAN nanoarchaea dominate the sample (61% of the reads), whereas *Leptospirillum* bacteria comprise ~ 4%, other bacteria and fungi represent 7%, and viruses and plasmids represent 28% of the sequenced community (Figure 2A).

We also evaluated community composition using proteomic data. Interestingly, community proteomic analysis of the Nov05 biofilm and of the Aug07 sample shows that *L. ferrodiazotrophum* proteins are the most abundant, and only 28% of the proteins are archaeal (Figure 2B). The discrepancy between DNA and protein measures could be attributed to differences in activity levels for bacteria vs. archaea (16) and/or differences in cell size (e.g., see Figure 1A). However, biases in DNA vs. protein extraction could also be a factor.

Identification and "omic" sampling of *Leptospirillum* group IV UBA BS

The partial genome of a new *Leptospirillum* bacterium was assembled primarily from reads from the Nov05 community genomics dataset (average depth of 3.3X). Assembled genome fragments > 1.6 kb were assigned to this genome in part based on some sequence similarity to the genome of *Leptospirillum ferrodiazotrophum*. Highly dissimilar sequences

belonging to the new bacterium may have been missed in this approach. In addition, sequences were classified as belonging to this organism based on tetranucleotide sequence signatures analyzed in the context of an emergent self-organizing map (13). The partial genome contains 1,923 predicted protein-encoding genes, one ribosomal RNA operon, and 18 tRNAs on 295 scaffolds representing 1.48 Mb of sequence (Table 2). The genomes of *L. ferrooxidans, L. ferriphilum,* "*L. rubarum*", *Leptospirillum* group II 5wayCG and *L. ferrodiazotrophum* are on average 2.6 ± 0.16 Mb in length. Thus, we estimate that ~ 56% of the newly assembled genome was recovered. The presence of 67.5% of the single copy genes selected to measure genome completeness supports this estimation (Supplementary Figure S3). 2.57 % of the Nov05 reads mapped uniquely to the new *Leptospirillum* genome; hence, partial genome reconstruction was achieved for a low abundance organism (Figure 2A and Supplementary Table S1).

A 16S rRNA phylogenetic tree of the genus *Leptospirillum*, which includes the partial (>1000 bp) sequence from the newly assembled *Leptospirillum* genome and two other closely related cloned sequences from other acidic environments, identifies a new group IV clade (Figure 3). We refer to the new clade as *Leptospirillum* group IV. The phylogenetic distance and percent nucleotide identity of the 23S rRNA gene, and amino acid identity of the RpoB protein, ribosomal protein S3, and ribosomal protein L5 support the placement of *Leptospirillum* group IV UBA BS in a separate clade: the distance between the *Leptospirillum* group IV UBA BS and *Leptospirillum* group III sequences is comparable to those of *Leptospirillum* groups I and II (Supplementary Figure S4). We designed 23S rRNA FISH probes to specifically target ribosomes of *Leptospirillum* group IV UBA BS (see methods). When applied to UBA BS biofilms, FISH analyses indicate that the *Leptospirillum* cells that did not bind the group II- and III-specific probes are *Leptospirillum* group IV UBA BS (Figure 1B (Aug07) and Figure 1F (Island 2)).

Comparison of *Leptospirillum* group IV UBA BS with other *Leptospirillum* spp.:

Table 2 reports a comparison of the *Leptospirillum* group IV UBA BS genome with the genomes of other *Leptospirillum* species. About 75% of the predicted genes in *Leptospirillum* group IV UBA BS have an ortholog in the genomes of the other *Leptospirillum* spp., with an average amino acid identity of 57 %.

The *Leptospirillum* group IV UBA BS genome contains several subunits for a pyruvate:ferredoxin oxidoreductase (PFOR) cluster identified by transcriptomics and proteomics (Supplementary Table S2). Cytochromes Cyt_{572} and Cyt_{579}, which were reported in "*L. rubarum*" and *L. ferrodiazotrophum* (12), were also identified (Supplementary Table S2). Two tetra-heme *c*-type cytochromes were predicted in the *Leptospirillum* group IV UBA BS genome, one of them belonging to the NapC/NirT family (UBABSL4_11800G0003 and UBABSL4_8194G0002a, Supplementary Table S2).

A multicopper oxidase gene with no orthologs in the other *Leptospirillum* bacteria is encoded in the genome of *Leptospirillum* group IV UBA BS. Multicopper oxidases (MCO) are involved in the oxidation of ferrous iron, copper and manganese (131). The sequence contains two predicted MCO domains, including all copper centers found in other two-domain MCOs (Figure 4).

As in *Leptospirillum ferrodiazotrophum* and *L. ferrooxidans* (48, 118), *Leptospirillum* group IV UBA BS contains a nitrogen (N_2) fixation pathway. *nif* transcripts were not observed for *Leptospirillum* group IV UBA BS, and only a few transcriptomics reads were observed for *L. ferrodiazotrophum* in one environmental sample (Supplementary Table S2). The operon in

Leptospirillum group IV UBA BS and *L. ferrodiazotrophum* contain paralogous *nifZ* (Figure 5). The genomic assemblies were manually curated to verify the existence of paralogous *nifZ* genes, which share only 39 % sequence identity. This duplication is not part of the *L. ferrooxidans* Nif operon (118, 120).

The *Leptospirillum* group IV UBA BS genome contains genes for osmoprotection (ectoine operon and trehalose synthase), synthesis of vitamins and cofactors, amino acid biosynthesis, polar flagella, chemotaxis, nitrite/sulfite reductase and sulfur assimilation, ATP synthase, 20S proteasome cluster, an arsenic resistance operon, phosphate transport, and formate hydrogenlyase cluster (Supplementary Table S2), all of which are present in "*L. rubarum*" and *L. ferrodiazotrophum* (12). Similar to "*L. rubarum*", *Leptospirillum* group IV UBA BS contains a cellulose synthase gene, suggesting that it too may be involved in biofilm formation. A CRISPR system type I (132) implicated in phage/plasmid resistance (28) has also been identified in *Leptospirillum* group IV UBA BS, and all CRISPR-associated Cas genes were detected by transcriptomics generally at low level (Supplementary Table S2).

The genomes of *Leptospirillum ferriphilum, L. ferrooxidans,* "*L. rubarum*" and *L. ferrodiazotrophum* contain two mercury (Hg) resistance genes in cluster: *merA/merR*, and the first three also encode the Hg transporter *merT*. *Leptospirillum* group IV UBA BS encodes in cluster the genes *merR*, *merA*, and the mercury transporter, *merC*, which contains four transmembrane domains and a conserved metal-binding motif MxCxxC at the C-terminus. The operon was identified by transcriptomics in four biofilm samples (Supplementary Table S2). At low concentrations, Hg can enter the cell where it is volatilized by the mercuric reductase MerA, but MerC may be necessary when high concentrations of Hg are encountered and efficient reduction is required (reviewed in (133)).

Leptospirillum group IV UBA BS contains genes for a respiratory [NiFe]-hydrogenase. HydB is the large subunit of the complex, which catalyzes the reversible oxidation of H_2 (reviewed in (134)). The predicted structure model for HydB based on the structure of *Escherichia coli* HydB contains perfectly conserved active site residues (cysteine 75 and 78, which bind nickel; and glutamate 56 and alanine 528 which bind ferrous iron and magnesium) (Figure 6), and transmembrane helix prediction suggests that it is membrane bound. The small subunit, HydA, is encoded in the cluster with HydB. It contains the conserved motif (S/T)RRxFxK within the signal peptide, and shares high sequence identity over the full length with a membrane-bound hydrogenase small subunit from *Ralstonia eutropha* (data not shown). Hydrogenase maturation proteins HypBFCDE are encoded as a cluster with a nickel transporter. Hydrogenases have not yet been described in the *Leptospirillum* spp., but the recently deposited genome of a *Leptospirillum ferrooxidans* isolate contains a single cluster of hydrogenase and maturation genes (120), which share an average 66% amino acid identity with the *Leptospirillum* group IV UBA BS orthologs.

A second [NiFe]-hydrogenase large subunit (likely HoxH) and a hydrogenase maturation protein encoded in the genome of *Leptospirillum* group IV UBA BS have homologs in *Acidithiobacillus ferrooxidans,* sharing an average 55% identity. Ligand binding prediction shows that HoxH is also able to bind nickel through four conserved cysteines (Cys64, Cys67, Cys415, and Cys418) as well as magnesium and ferrous iron (Glu45, Ile370, and His421). However no transmembrane helices were predicted for this hydrogenase, suggesting that it is a cytoplasmic hydrogenase involved in H_2 uptake during N_2 fixation (reviewed in (134)).

HydB was identified by proteomics in one UBA BS biofilm, and genes *hydA*, *hydB, hoxH,* maturation proteins and the nickel transporter were identified by transcriptomics only in

the sample in which the nitrogen fixation operon in *L. ferrodiazotrophum* was transcribed (Supplementary Table S2).

Proteomics and transcriptomics of *Leptospirillum* group IV UBA BS

A total of 22.9% of *Leptospirillum* group IV UBA BS predicted proteins were identified by mass spectrometry based proteomics (see methods) in three UBA BS biofilms (Supplementary Table S2). In order to compare the functional profiles of *L. ferrodiazotrophum* and *Leptospirillum* group IV UBA BS, we clustered NSAF values for orthologous proteins identified in these three biofilms (Figure 7). All ribosomal proteins, the ATPase and PFOR clusters, chemotaxis and flagella, and phosphate uptake proteins were largely overrepresented in *Leptospirillum* group IV UBA BS (Figure 7, cluster 2), whereas lipid and fatty acid biosynthesis, and peptide processing proteins were overrepresented in *L. ferrodiazotrophum*. Proteins involved in carbohydrate metabolism, amino acid, vitamin and cofactor, and nucleotide biosynthesis were overrepresented in both.

To more broadly evaluate the activity of *Leptospirillum* group IV UBA BS in AMD biofilms, we obtained community transcriptomics data from eight relatively thin floating biofilms from a range of locations in the Richmond Mine, and five laboratory-grown thin floating biofilms. *Leptospirillum* group IV UBA BS accounts for less than 1% of the total community RNA on average in all of these transcriptomic samples (data not shown). Despite the low abundance, up to 46% of its predicted genes were identified by at least one Illumina read (Supplementary Table S2). Clustering of normalized transcriptomic values shows that *Leptospirillum* group IV UBA BS transcript abundance is generally higher in bioreactor-grown than environmental samples (Supplementary Figure S5). Bioreactor solutions tend to have a higher pH and higher concentration of ferric iron compared to natural AMD solutions (Shufen Ma, personal communication). Interestingly, *Leptospirillum* group IV UBA BS genes encoded in mobile elements (transposases, hypothetical proteins, secretion proteins and transporters) are overrepresented in bioreactor transcriptomic datasets (Supplementary Figure S5). Core metabolism genes are generally similarly transcribed in natural and bioreactor-grown biofilms (Supplementary Table S2).

DISCUSSION

Bacteria generally dominate Richmond Mine AMD biofilms, and only recently Archaea were reported to dominate sunken biofilms degrading under microaerophilic and anaerobic conditions (135). Here we described an unusual Archaeal-dominated acidophilic biofilm community that contains a novel bacterial species, *Leptospirillum* group IV UBA BS. The identity of its 16S rRNA gene sequence compared to sequences of the other *Leptospirillum* spp. is sufficiently low to warrant its designation as a distinct species (< 98.7% identity, as recommended by Stackebrandt in (136)). Enrichment of this species has not been observed in cultures from the Richmond Mine, possibly due to its habitat within an extraordinarily thick polymer environment and its association with uncultivated Archaea. Notably, *Leptospirillum* group IV UBA BS is most abundant in biofilms that contain only very low abundances of other cultivated *Leptospirillum* spp. Consistent with its designation as a distinct species, among the Leptospirilli, *Leptospirillum* group IV UBA BS is unique in its capacity for mercury resistance,

the presence of a multicopper oxidase, and many unique hypothetical proteins. As in *Leptospirillum ferrooxidans*, it has the potential for hydrogen metabolism.

Its genomic content suggests that cells are likely motile, capable of both carbon and nitrogen fixation. It has been suggested that motility allows *L. ferrodiazotrophum* to redistribute into micro-colonies (12). Motility may be particularly important for *Leptospirillum* group IV UBA BS growing in the gel-like environment of the thick floating biofilms, allowing it to find a suitable habitat as conditions change during biofilm development. *Leptospirillum* group IV UBA BS might be capable of anaerobic growth using H_2 as electron donor, as shown for *Acidithiobacillus ferrooxidans* (137), although experimental validation is required to confirm this function. Hydrogen metabolism could be associated with growth under anaerobic conditions predicted within AMD biofilms thicker than a few microns (138).

Leptospirillum bacteria are known iron oxidizers, likely using cytochrome 572 (Cyt_{572}) to take electrons from Fe(II) (67, 139). Cytochrome 579 (Cyt_{579}) is involved in electron transport (66, 140). The presence of a multicopper oxidase, as well as Cyt_{572} and Cyt_{579}, whose biochemical function was verified in *Leptospirillum* Group II from the Richmond mine (66, 67), provides strong evidence supporting the role of *Leptospirillum* group IV UBA BS in iron oxidation. Likely, *Leptospirillum* group IV UBA BS fixes CO_2 via the reverse TCA cycle using pyruvate:ferredoxin oxidoreductase (PFOR), as shown in other *Leptospirillum* spp. (reviewed in (141)).

Community transcriptomics shows that the fraction of *Leptospirillum* group IV UBA BS genes for which a transcript was detected was highest in the environmental sample A-drift GS0 (Supplementary Table S2). This early developmental stage biofilm, dominated by *L. ferrodiazotrophum,* was collected from the A-drift tunnel in September 2010. The temperature at that location was 40 °C, the pH was 1.27, and the solution contained unusually high concentrations of ferric iron (S. Ma, S. E. Spaulding, D. S. Aliaga Goltsman, M. Dasari, and J. F. Banfield, submitted for publication). The sample in which the lowest fraction of genes for which a transcript was detected was the environmental sample C75 GS1 (Supplementary Table S2). This mid-developmental stage biofilm, collected from a pool in the C-drift tunnel with pH 0.86 and temperature 46 °C, was dominated by "*Leptospirillum rubarum*" (group II, UBA-type) and *L. ferrodiazotrophum* was detected at low abundance (121). Therefore, both physiologically and in terms of certain environmental preferences, *Leptospirillum* group IV UBA BS is most similar to *L. ferrodiazotrophum* (group III), to which it is most closely related based on phylogenetic analysis. It is notable that expression of hydrogenases in *Leptospirillum* group IV UBA BS was only detected in the sample in which the nitrogen fixation operon is expressed in *L. ferrodiazotrophum.* H_2 is a byproduct of nitrogen fixation, but there is no evidence that it can be used by *L. ferrodiazotrophum.* Thus, consumption of H_2 may be the basis for cooperative interactions between *Leptospirillum* group IV UBA BS and *L. ferrodiazotrophum*.

ACKNOWLEDGEMENTS

We thank Mr. TW Arman, President, Iron Mountain Mines Inc., Mr R Sugarek (US Environmental Protection Agency) for site access, and Mr R Carver for on-site assistance. DSAG acknowledges support from an NSF GRFP fellowship. We thank Christine Sun for help with ruby scripts, and Edward Ralston for access to, and help with, the Covaris system. DNA

sequencing was carried out at The Joint Genome Institute. Transcriptomic sequencing was done at the University of California Davis. This research was supported by the U.S. Department of Energy through the Genomic Sciences (DE-FG02-05ER64134) and Carbon-Cycling (DE-FG02-10ER64996) programs.

TABLES

Table 1. Relative abundance (%) of organisms estimated from FISH in four UBA BS biofilms. LeptoIII: *Leptospirillum ferrodiazotrophum*; LeptoIV: *Leptospirillum* group IV UBA BS; LeptoII CG: *Leptospirillum* group II 5way CG type.

				% Total		**% *Leptospirillum***		
Sample	**Date collected**	**pH**	**T (°C)**	**Archaea**	**Bacteria**	**LeptoIII**	**LeptoIV**	**LeptoII CG**
Aug07	8/22/07	1.94	36	72.87	27.13	3.09	11.10	0.00
Island 2	11/7/07	ND*	ND*	57.31	42.69	5.10	26.34	1.09
Island 3	11/7/07	ND*	ND*	77.29	22.71	5.92	41.01	1.34
Jun08	6/26/08	1.08	37	64.36	35.64	27.17	14.74	0.00
Average				**67.96**	**32.04**	**10.32**	**23.30**	**0.61**

*ND: no data available

Table 2. Statistics for the *Leptospirillum* group IV UBA BS genome and comparison with other *Leptospirillum* spp. genomes. LeptoIV: *Leptospirillum* group IV UBA BS; LeptoIII: *L. ferrodiazotrophum*; LeptoII UBA: "*L. rubarum*"; LeptoII CG: *Leptospirillum* group II 5way CG type.

	LeptoIV	**LeptoIII**	**LeptoII UBA**	**LeptoII CG**	***L. ferriphilum* ML-04**	***L. ferrooxidans* C2-3**
Group classification	IV	III	II	II	II	I
Reference	this study	(12)	(11)	(14)	(119)	(120)
Genome length (Mbp)	1.48	2.84	2.65	2.72	2.41	2.56
Assembled scaffolds	295	25	10	77	1	1
GC content (%)	58.98	58.51	54.7	54.28	54.55	50.1
No. predicted proteins	1923	2654	2625	2584	2471	2421
No. tRNAs	18	46	48	47	48	51
Compared to *L.* UBA BS						
No. orthologs	-	1463	1382	1387	1310	1346
% id orthologs	-	69.62	55.20	55.35	54.66	47.66
% id 16S rRNA	-	98.46	92.43	92.2	92.3	92.25
% id 23S rRNA	-	96.3	91.39	91.53	91.39	89.92

CHAPTER 2 FIGURES

Fig. 1. (next page) Fluorescence *in situ* hybridization images of UBA BS biofilms. Scale bar = 1 μm. A) Sample Aug07: DNA-stain (blue, most small dots are Archaea) and general bacteria (green). B) Sample Aug07: general bacteria (green) and *Leptospirillum* group IV UBA BS (yellow). C) Sample Jun08: general bacteria (green) and *L. ferrodiazotrophum* (red). D) Sample Aug07: general bacteria (green), *L. ferrodiazotrophum* (red), and *Leptospirillum* group II 5wayCG-type (blue). E) Sample Island 2: general bacteria (blue), *L. ferrodiazotrophum* (green/light blue). F) Sample Island 2: general bacteria (blue) and *Leptospirillum* group IV UBA BS (purple/pink). Images A and B were taken from the same field of view, as were images E and F. Arrows indicate *Leptospirillum* group III cells.

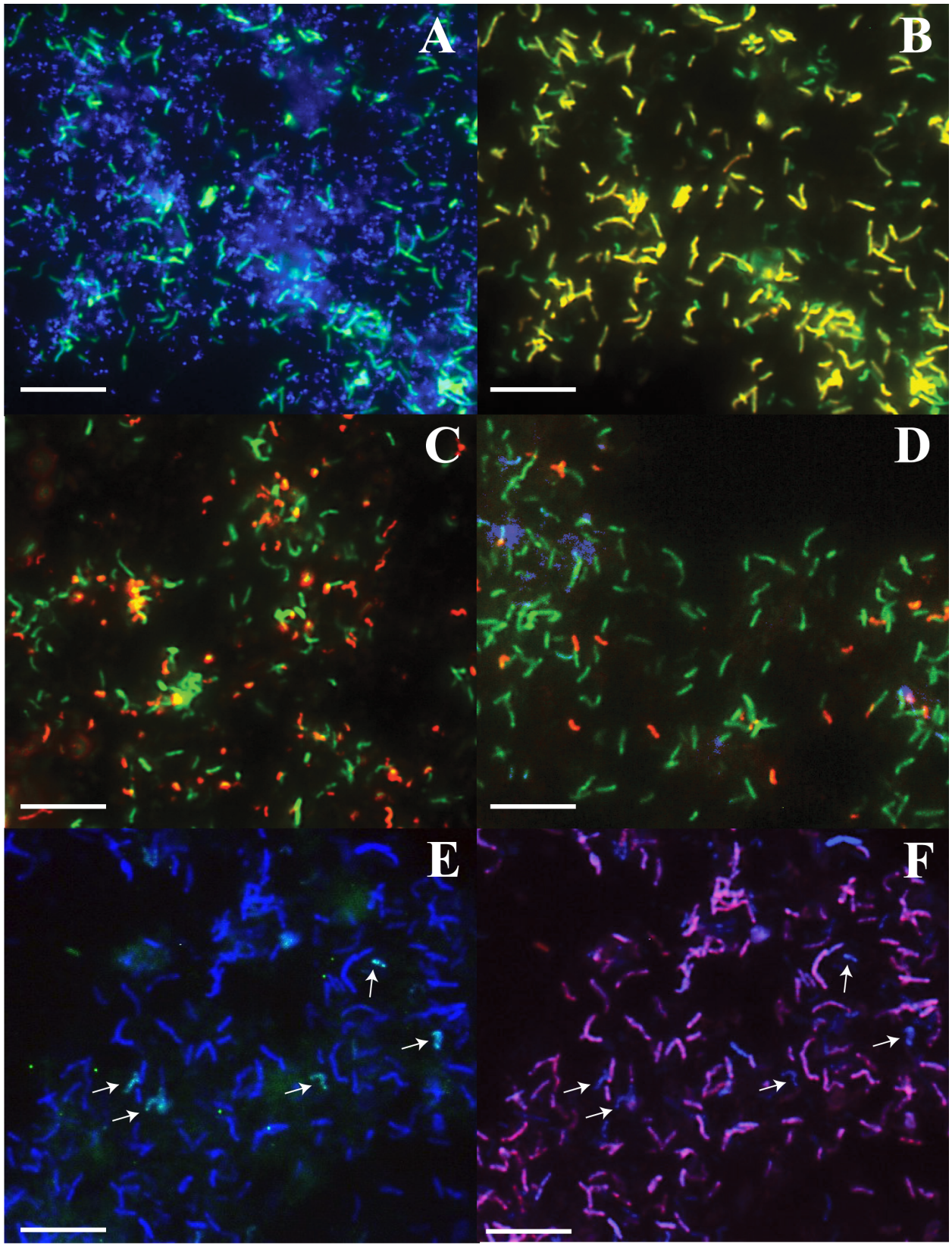

A
B
C
D
E
F

Fig. 2 A) Percent of uniquely mapping community genomics reads from the Nov05 sample to assembled genomes from AMD organisms. B) Percent NSAF values in three community proteomics datasets (III and IV refer to *L. ferrodiazotrophum* (group III) and *Leptospirillum* group IV UBA BS protein abundances, respectively).

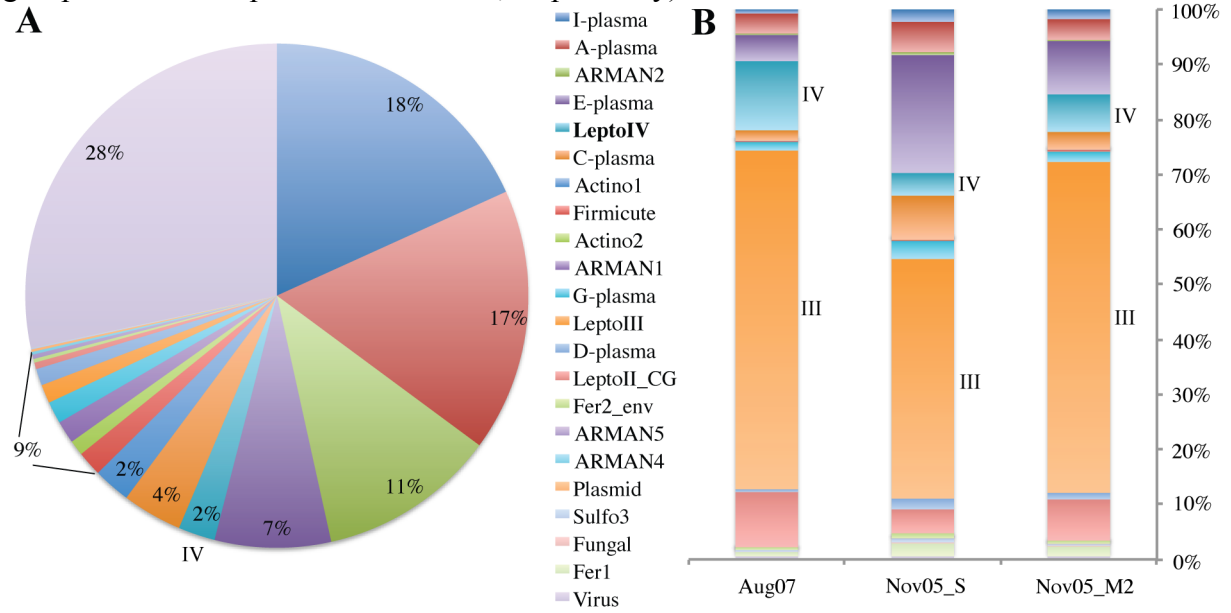


Fig. 3. Neighbor Joining 16S rRNA phylogenetic tree of *Leptospirillum* spp. (1000 bootstraps). Percent bootstrap is shown at the nodes, scale bar indicates 10% sequence divergence. The sequence of a Delta Proteobacterium was used as outgroup.

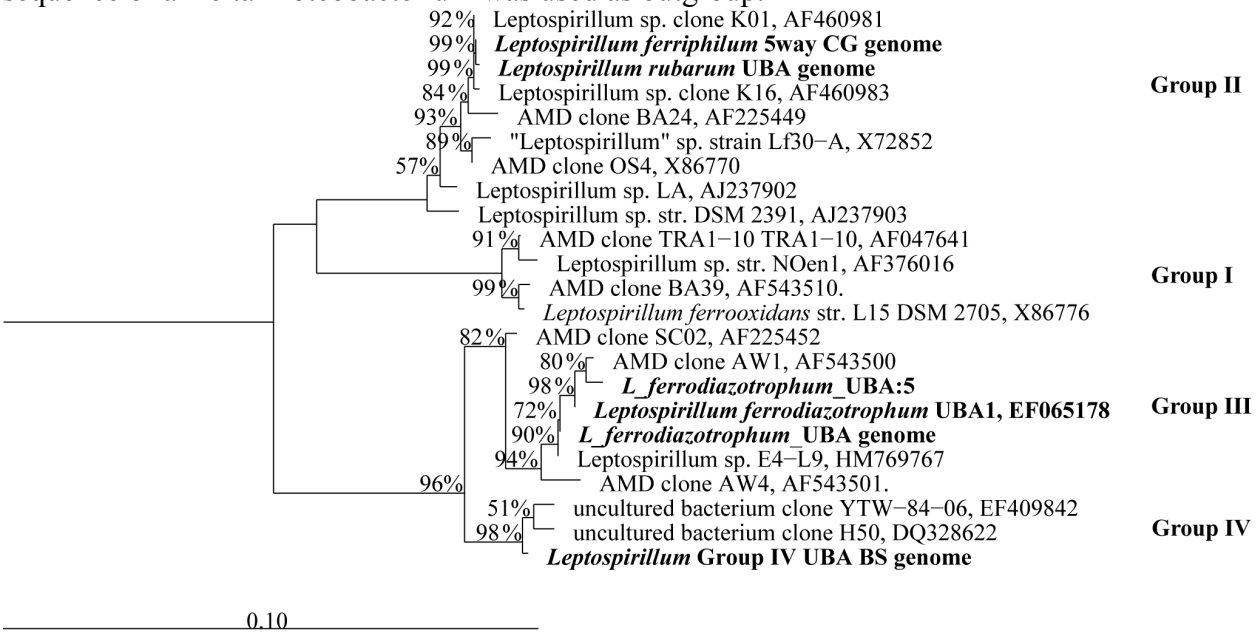


Fig. 4. Multiple sequence alignment of MCOs, showing the conserved copper center residues: T1 (*); and T2/T3 (◆). Alignment includes sequences from: 1) *Leptospirillum* Group IV UBA BS (dotted box), 2) *Nitrosomonas europaea*, 3) *Nitrosococcus halophilus*, 4) *Herminiimonas arsenicoxydans*, 5) *Shewanella woodyi* ATCC 51908, 6) *Deinococcus maricopensis* DMS 21211, and 6) *Flavobacterium frigoris* PS1.

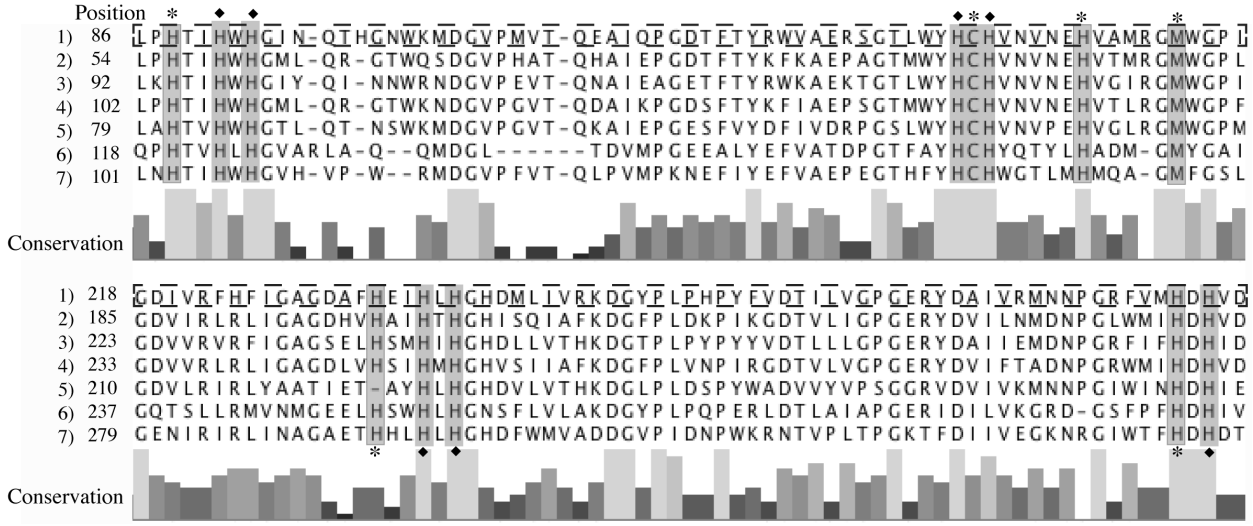


Fig. 5. Diagram of the nitrogen fixation operon in *Leptospirillum* Group IV UBA BS, *L. ferrodiazotrophum,* and *L. ferrooxidans*. NifX-assoc: NifX-associated protein; Fdx: ferredoxin; [FeS]: FeS cluster assembly protein. Gaps (brackets) in the *Leptospirillum* Group IV UBA BS operon correspond to genes: a) two hypothetical proteins, *nifS*, aminotransferase; and b) *nifU*, *nifT*, two hypothetical proteins, ferredoxin, *nifW*, and two final nitrogen regulatory proteins P-II.

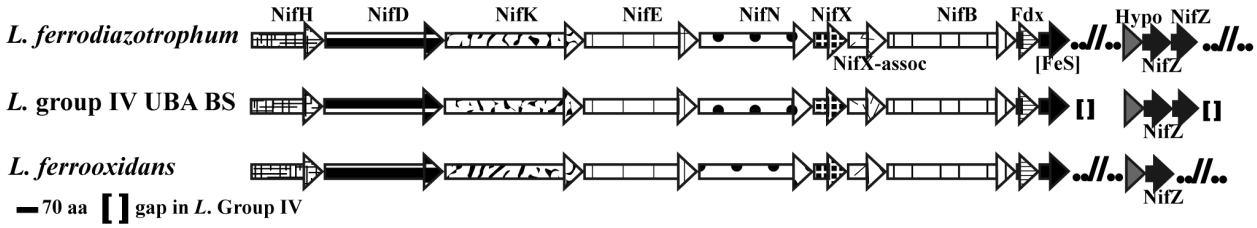


Fig. 6. 3D structure representation of the hydrogenase HydB in *Leptospirillum* Group IV modeled on the *Escherichia coli* protein: A) HydB full model, and B) Zoomed image of the active site highlighting catalytic residues (NiFe-binding: red; and Fe2-binding: orange). Conserved structural regions are shown in teal blue, while less conserved regions are shown in magenta. C) Multiple alignment screenshot showing catalytic residues (Cys75 and Cys78 – red; Glu56 – orange) and other perfectly conserved residues around the active site (green). Alignment includes hydrogenase sequences from: 1) *Leptospirillum* Group IV UBA BS; 2-3) *E. coli,* 4) *Allocromatium vinosum;* 5) *Bradirhizobium japonicum*; 6) *Desulfovibrio vulgaris*; 7) *Wolinella succinogenes*; and 8) *Desulfovibrio baculatus*.

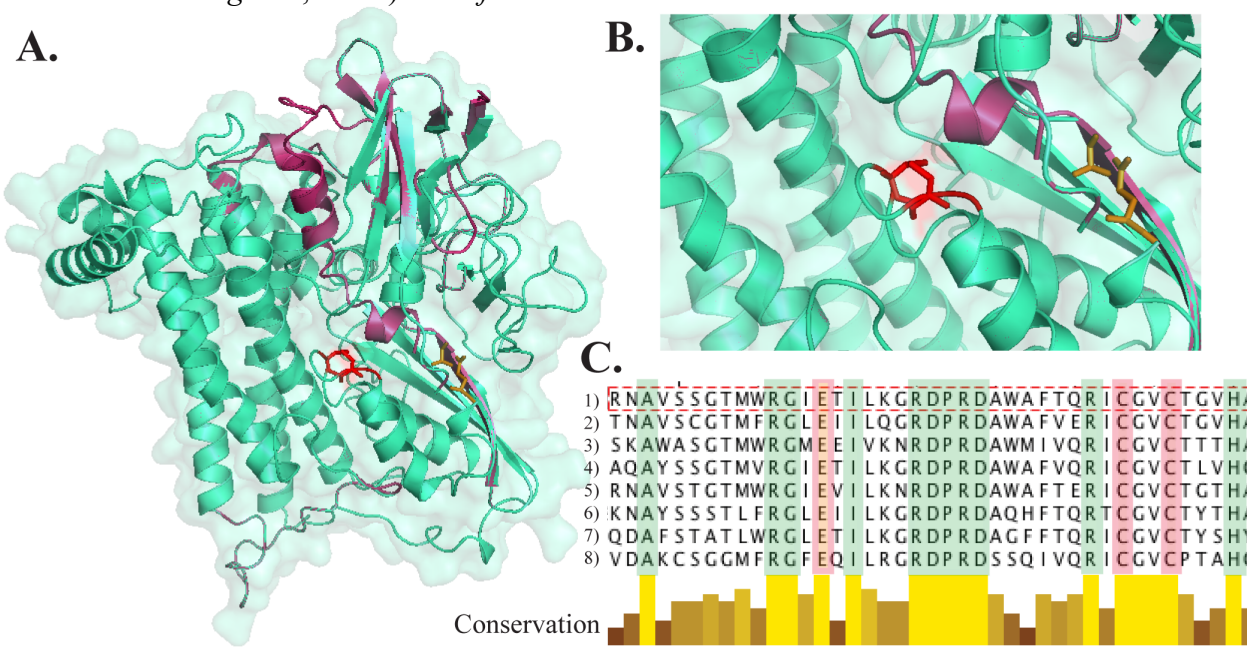
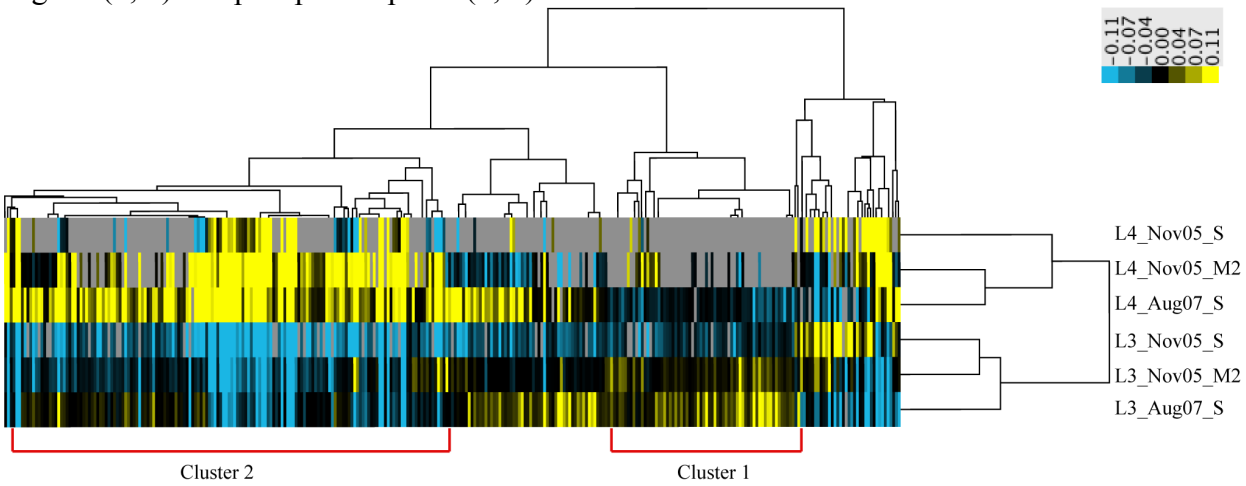


Fig. 7. NSAF clustering in *Leptospirillum* Group IV UBA BS compared to *L. ferrodiazotrophum* in three UBA BS biofilms. Cluster 1: overrepresented in *L. ferrodiazotrophum*. Cluster2: overrepresented in *L.* Group IV UBA BS (Cluster 1; cluster 2): energy related (8; 17), carbohydrate metabolism (10; 20), nucleotide metabolism (2; 5), amino acid metabolism (10; 16), lipid and fatty acid metabolism (4; 0), vitamin and cofactor biosynthesis (5; 9), ribosomal proteins (0; 20), peptide processing (7; 4), chemotaxis and flagella (1; 9) and phosphate uptake (0; 3).



CHAPTER 3

Community rRNA Gene Transcriptomics Reveals Unexpected High Microbial Diversity in Acidophilic Biofilm Communities

ABSTRACT

A fundamental question in microbial ecology relates to community structure, and how this varies across environment types. It is widely believed that some environments, such as those at very low pH, host communities that are simple based on the low number of taxa. However, so far, analyses of species richness have relied on cultivation-independent methods that provide relatively low sampling depth. Here we used community transcriptomic sequencing to analyze the microbial diversity of natural acid mine drainage biofilms from the Richmond Mine at Iron Mountain, California. Our analyses target deep pools of ribosomal RNA gene transcripts recovered from both natural and laboratory-grown biofilms across varying developmental stages. 91.8% of the ~254 million Illumina reads mapped to rRNA genes represented in the Silva database. Remarkably, up to 159 different taxa, including Bacteria, Archaea and Eukaryotes, were identified. The results indicate that the primary characteristic that has enabled prior extensive cultivation-independent "omic analyses is not "simplicity" but rather the high dominance by a few taxa. Diversity measures, ordination and hierarchical clustering separate environmental from laboratory-grown biofilms. In part, this is due to the much larger number of rare members in the environmental biofilms. Although *Leptospirillum* bacteria generally dominate biofilms, we detect a wide variety of other *Nitrospirae* organisms present at very low abundance (up to ~ 1 % of the community). Bacteria from the *Chloroflexi*, *Deferribacteres* and other phyla without cultivated representatives were also detected at low abundance. The findings greatly expand our knowledge of the range of organisms adapted to extremely low pH conditions and demonstrate that communities once considered "simple" can harbor a striking amount of hidden complexity when analyzed with next generation sequencing technology.

INTRODUCTION

Microbial ecology and evolution studies in model natural ecosystems can greatly advance understanding the ecology of more complex natural environments. The acid mine drainage (AMD) environment has been established as a model ecosystem, due in part to microbial compositional characteristics that make the biofilms tractable for cultivation-independent molecular analyses (reviewed in (7)). AMD and other acidic environments have also been extensively studied systems due to their importance in acid and metal contamination, and application in the biomining industry (reviewed in (112, 142)).

AMD systems are generally dominated by few acidophilic Bacteria belonging to the *Nitrospira* and *Proteobacteria* phyla, whereas lower abundance members include *Firmicutes*, *Actinobacteria*, and *Acidobacteria* phyla, Archaea and Eukaryotes (e.g., 113, 143-146, and reviewed in (6)). Cultivation-independent techniques such as community genomics, proteomics, and microarrays have been successfully applied to study the physiology and ecology of acidophilic organisms in their natural environments. Results have provided new insight into the functional capacities of many previously unstudied organisms (e.g. 12, 122), and Yelton *et al*., in review) and the environmental and biological factors that influence community structure (10, 19). For example, protein expression of the most abundant bacteria in AMD biofilms from the Richmond Mine at Iron Mountain, California, is influenced by the community composition (19), whereas expression in lower abundance members may be

controlled by abiotic factors (18). In addition, environmental factors such as pH and ferrous iron concentrations promote changes in the dominant community members in AMD biofilms (Shufen Ma, personal communication). Furthermore, fluorescence microscopy has shown that the community composition and organization of AMD biofilms change as the biofilm matures (47). All analyses to date have focused exclusively on relatively abundant organisms (more than a few percent of each community). Lack of information about less abundant organisms has limited our understanding of the biology of extremely acidic environments and factors that impact community structure.

Studies of the total RNA pool using deep-sequencing technologies to reveal the diversity of acidophilic communities have the potential to deeply explore microbial community composition, and provide extensive datasets suitable for diversity analyses. The most widely used indices for exploring microbial diversity are the species richness (which gives equal weight to all phylotypes present in a community), the Shannon-Wiener index (which measures the information content of a community) and the Simpson's index (which measures the probability that two phylotypes are the same) (reviewed in (147)). In addition, Hill's diversity indices summarize different properties of a community, depending on the presence and relative abundance of taxa. Here we apply these indices to transcriptomic data to evaluate changes in biofilm diversity with developmental stage, location-dependent environmental factors, and growth in laboratory bioreactors vs. the natural system. Results greatly expand our understanding of the diversity of AMD biofilms, verifying that these are truly complex ecosystems. The findings clarify that it is dominance by a few taxa, not lack of complexity (29) that makes AMD environments good model systems for studying community physiology and ecology.

MATERIALS AND METHODS

Eight biofilms were collected from the A-drift, C-drift, AB-drift, and 4-way locations within the Richmond Mine at Iron Mountain Mines, California (40°40′ 38.42″ N and 122″ 31′ 19.90″ W, elevation of ~900 m) (Figure 1 and Table 1). In addition, biofilms were grown at pH 1 and 37 °C in laboratory bioreactors using inoculum from two locations within the A-drift, and mine outflow, as previously described (124). Biofilms were snap-frozen in liquid nitrogen upon collection and stored at -80 °C.

Community transcriptomic data were obtained from all frozen samples. Briefly, total RNA was obtained using two acid phenol:chloroform:isoamyl alcohol (Ambion) extractions, and immediately purified using the RNEasy MinElute kit (Qiagen), as described earlier (15) and Chapter 2, methods section). Integrity and concentration of the RNA was assessed using a Bioanalyzer 2100 (Agilent Technologies). Good quality (RNA integrity number > 7) total RNA was converted to cDNA as described by (125) in order to keep the strand-specificity of the transcriptome. Resulting cDNA was fragmented using a Covaris S-system (Covaris, Inc.) to an average fragment size of 200 bp. Fragmented cDNA was sent to the University of California Davis for library preparation and sequencing. Illumina single-end GAII sequencing was obtained for five samples and Illumina paired-end HiSeq sequencing for the other eight biofilms (referred to as GAII and HiSeq, respectively, through the manuscript; Table 1). Data from the two rounds of sequencing were analyzed separately.

Low-quality bases were trimmed from the sequencing reads using the fastx_trimmer script (http://hannonlab.cshl.edu/fastx_toolkit/) or the sickle trimmer script with default

parameters (https://github.com/najoshi/sickle). Reads > 40 bp were kept for further analyses. Trimmed reads were mapped to the small and large subunit (SSU and LSU) rRNA gene Silva databases (148) using bowtie (127) with default parameters to separate ribosomal from non-ribosomal reads. Ribosomal reads were then mapped using bowtie with default parameter to the SSU_Ref_108 rRNA Silva database. Mapped reads were assembled using Cufflinks (149), and assembled transcripts were clustered at 97% identity using Uclust (150). Abundance measures (normalized count values generated by Cufflinks) are reported in the Supplementary Table 1.

Small subunit (SSU) rRNA gene sequences were aligned using the SINA Aligner (151). Phylogenetic tree construction was done using FastTree with parameters –gtr –nt –gamma (152), and trees were visualized using the iTol website (153). The *Nitrospira* Phylum phylogenetic tree was built in ARB, with 1000 bootstraps (128).

Rank abundance curves, and non-metric multidimensional scaling (NMDS) analyses were done using R (154) and the R packages Picante (155), and BiodiversityR (156). NRI and NTI values were estimated using the vegan R package (157), and principal coordinates analyses (PCoA) were estimated using the Fast Unifrac website (158). Diversity profiles were calculated as presented by (159).

Reads mapping to *Leptospirillum* groups II and III rRNA genes were removed from the ribosomal RNA fastq files using bowtie. Full-length SSU rRNA gene sequences were then reconstructed from these *Leptospirillum*-subtracted reads using EMIRGE (126) with parameters –l 101 –i 300 –s 100 –phred33 and run until 40 iterations. The sequences obtained were clustered at 98% identity with CD-HIT (160), and searched against the Silva SSU_Ref_108 for classification using Blast (56).

Assembled SSU rRNA sequences analyzed in this work are available as supplementary materials.

RESULTS AND DISCUSSION

Community transcriptomics

Biofilms at increasing stages of development were obtained from eight locations within the Richmond Mine for community transcriptomic analyses (Table 1 and Figure 1). Biofilms were also grown in laboratory bioreactors in order to compare the diversity of environmental and bioreactor transcriptomes. An average 91.8% of the total sequencing reads mapped to 16S/18S and 23S/28S ribosomal RNA genes from the Silva database, and Archaea, Bacteria, and Eukaryotes were identified (Figure 2 and Table 2). Members of the phylum *Nitrospira* represent more than 85% of the community in most biofilms with the exception of biofilm A-drift, which contain a large proportion of Eukaryotes (Table 2). A-drift is an unusual, very early developmental stage biofilm sampled from a site with relatively high pH (pH 1.27). Other Bacteria include *Proteobacteria* (of the classes Alpha, Delta, Epsilon, Gamma, and T18), *Actinobacteria* (generally of the *Acidimicrobiales* class), *Acidobacteria, Firmicutes* (class *Clostridia*), *Deferribacteres* and *Chloroflexi* (*Anaerolineae* and *Caldilineae* classes) (Figure 2 and Supplementary table 1). Protists and fungi have been observed in fluorescence microscopy studies of acid mine drainage biofilms (144), and represent the majority of the Eukaryotic hits in this study. *Thermoplasmatales* Archaea A, E, and G-plasma, *Ferroplasma* Type I and the Nanoarchaea ARMAN-1 and ARMAN-4 were observed in all biofilms, but were most abundant in high developmental-stage environmental biofilms (Supplementary table 1).

Community diversity

A total 1773 SSU rRNA transcripts were assembled with Cufflinks and were clustered at 97% identity using Uclust into 425 OTUs. Of these, there were 159 OTUs longer than 130 bp, which were present in at least two transcriptomics datasets, were classified by The Ribosome Database Project (161), as well as by blast search against the SILVA database (Supplementary Table 1). Community diversity analyses were done with these 159 OTUs.

Rank abundance curves show that biofilms sequenced using the HiSeq platform reach a mean species richness of less than 100 taxa and organismal abundance curves fall quickly, indicating that samples are largely dominated by a few OTUs. For the GAII platform data, the calculated mean species richness goes up to 139, but curves indicate a more even distribution (Figure 3). The sequencing technology-dependence of taxonomic richness is likely due to the paired-end vs. single-end nature of the datasets: HiSeq paired-end data assembles more accurately, whereas GAII single-end data likely over-estimates richness by misplacing reads. For both data types, C-drift curves indicate lower richness and evenness compared to other biofilms (Figure 3). C-drift environmental biofilms were growing on the surface of solutions with slightly harsher conditions (low pH and highest temperature) than at other locations or in the bioreactors (Table 1). Biofilms grown in bioreactors have been reported to resemble those found naturally in the acidic natural system (124), therefore, it is expected that bioreactor biofilms show comparable richness and evenness to environmental samples (Figure 3).

Taxonomy-based Shannon-Wiener diversity index and Inverse Simpson's index of diversity show much higher diversity in GAII biofilms, likely due to an over-estimation of taxonomic richness (Figure 4A). However, indices for biofilms from the same site but of different developmental stage show similar diversity values (R1, R3, AB10, and C10 samples), indicating that diversity varies more with location than with developmental stage. Again, C-drift samples (C75, and C10) show the lowest diversity indices (within GAII and HiSeq datasets), consistent with the lower richness observed in rank abundance curves (Figure 3 and Figure 4A).

The net relatedness index (NRI) and nearest taxon index (NTI) are measures of phylogenetic clustering or overdispersion closer to the root (NRI) and at the tips (NTI) of a phylogeny (162). Most NRI values are negative in HiSeq-sequenced datasets, indicating over-dispersion closer to the root of the tree (i.e., taxa are more distantly related to each other at higher taxonomic levels; Figure 4B). The exceptions are the C10 GS0 biofilm and the bioreactor sample R3 GS0, which are early developmental-stage biofilms that show the lowest taxonomic richness and evenness (Figure 3). Environmental stability has been suggested to contribute to negative NRI values observed at one Rio Tinto location (145). High NRI values for GAII datasets indicate phylogenetic clustering at the root of the phylogeny, e.g. taxa are more closely related at the root (Figure 4B). The high number of closely related sequences in GAII-sequenced samples (as observed in rank abundance curves) likely contributes to the over-estimation of taxonomic richness and leads to high NRI values. Positive NTI values show phylogenetic clustering in all samples, indicating that taxa are more closely related at the tips of the phylogeny (Figure 4B). This trend was also observed in most Rio Tinto samples (145). Overall, we conclude that the AMD system is characterized by high diversity at the phylum level, but with relatively few closely related members of each phylum.

We evaluated differences between microbial communities using weighted Unifrac distances (Figure 5). PCoA analyses separate bioreactor from environmental biofilms along the

first component (P1), but places the sample A-drift GS0 away from other samples (Figure 5A). This result is likely due to the higher pH and the unusually large Eukaryotic and Archaeal richness observed for the A-drift sample (Table 1). Component P2 separates late from early to mid-growth stage environmental biofilms (Figure 5A). Hierarchical clustering confirms the separation shown by component P1 of the PCoA analyses, and shows that samples from the same site (but of differing developmental stage) cluster together, as observed in Figure 4A. Non-Metric Multidimensional Scaling (NMDS) analyses were calculated using Bray-Curtis distance matrix and three dimensions were selected (a significant reduction in stress was observed from one to two and three dimensions - Supplementary Figure 1). NMDS ordination separates GAII from HiSeq samples along dimension 1 and environmental from bioreactor samples along dimension 2, while dimension 3 separates early to mid-growth stage biofilms from late growth-stage biofilms (Figure 6). The results support the findings of the PCoA and hierarchical cluster analyses, and suggest that microbial community composition in the AMD system is driven, first, by environmental factors, and, to a lesser extent, by the developmental-stage of the biofilm. Similar results have been observed in community proteomic analyses of AMD biofilms (18).

Diversity profiles (159) were used to explore differences in community diversity while taking into account taxonomy or phylogenetic similarity (Figure 7). In diversity profile plots, the Y-axis represents a calculated effective diversity value and the X-axis represents a sensitivity parameter "q", where smaller "q" gives higher weight to rare taxa and this weight decreases with increasing "q" (159). Overall, one community is considered more diverse than another if its diversity profile curve lies above the other curve. When considering only taxonomic identity (i.e., just OTUs), C-drift samples are, once again, the least diverse (Figure 7A). The A-drift biofilm is more diverse that the other samples sequenced by GAII (presumably due to the higher pH).

Unexpectedly, the bioreactors appear to be more diverse than HiSeq-environmental biofilms, especially if emphasis is placed on the more highly abundant taxa (higher "q") (Figure 7B). However, when adding phylogenetic similarity to the profiles, it becomes evident that the environmental samples are more diverse than bioreactor biofilms, and more mature biofilms are more diverse than early biofilms (Figure 7C and 7D). The "open" nature and constant change in conditions of environmental samples likely contribute to higher community diversity. Additionally, larger community membership, as observed on microscopy-based analyses that considered relatively abundant taxa (47), likely increases diversity as biofilms mature.

Nitrospira phylum and low abundance organisms.

There were 43 OTUs identified as belonging to the *Nitrospira* Phylum and 21 were confidently classified as *Leptospirillum* groups I – IV by the Ribosome Database Project, blast to the SILVA database, and ARB phylogenetic tree construction (Figure 8 and Supplementary Table 1). *Leptospirillum* Groups II generally dominates environmental biofilms, whereas *Leptospirillum* Group III dominates bioreactor grown samples (in this study), as well as biofilm A-drift (Table 3). A change in dominance from *Leptospirillum* group II to group III has been observed in biofilms exposed to low Fe^{+2}/Fe^{+3} ratios and higher pH (Shufen Ma, personal communication). *Leptospirillum* groups I and IV are usually at low abundance in all samples, and group IV is more abundant in biofilms where *Leptospirillum* group III is the dominant member (Table 3). Other *Nitrospira* Phylum OTUs observed in most transcriptomics samples

include sequences related to the *Magnetobacterium* and *Nitrospira* genera, as well as other uncultured and unclassified *Nitrospiraceae* (Figure 8 and Supplementary table 1). Sequences belonging to *Magnetobacterium* spp. have not yet been recovered from metagenomic datasets from AMD biofilms. In fact, magnetotactic bacteria from the *Nitrospira* phylum are typically found in freshwater lakes, lake sediments, and hot springs (163). Using cryogenic transmission electron microscopy (cryo-TEM), we have observed long vibrio-shaped cells with intracellular magnetosome-like structures in AMD biofilms (Figure 9). Bullet-shaped chains of magnetite, like those in Figure 9, have been reported in the magnetotactic, sulfate-reducing Delta-Proteobacterium *Desulfovibrio magneticum* (reviewed by Komeili *et al.* (164)). Therefore, the cryo-TEM images confirm the presence of magnetotactic bacteria in AMD systems and support the transcriptomics findings.

Sequences longer than 1000 bp related to *Acidithiobacillus caldus* and other uncultured *Acidithiobacilli* were assembled from the data, and are mostly observed in bioreactor samples (Supplementary table 1). These bacteria are often encountered in higher pH and lower temperature environments such as downstream of AMD sites and bioleaching systems (reviewed in (6)). Among the *Actinobacteria*, the full sequence of *Ferromicrobium* sp. Mc9KL−1−9 and sequences longer than 900 bp related to uncultured *Acidimicrobium* bacteria were recovered from bioreactors and late growth-stage biofilms. Among the *Firmicutes*, sequences of ~ 1500 bp closely related to *Sulfobacillus thermosulfidooxidans* and *Alicyclobacillus disulfidooxidans* were observed in all biofilms, and in HiSeq bioreactor samples, respectively (Supplementary table 1). Both of these organisms have been detected by cultivation-based and/or cultivation-independent methods in the AMD system (unpublished). Other organisms for which there is transcriptomic evidence include members of the Chloroflexi, Deferribacteres, and the TM6 phyla.

Small subunit (SSU) rRNA gene reconstruction

Using EMIRGE and a dataset of transcriptomics reads after subtraction of those mapping to *Leptospirillum* groups II and III, we reconstructed SSU rRNA gene sequences for 662 operational taxonomic units (OTU), which range from ~500 bp to ~1900 bp in sequence length. Reconstructed sequences were clustered using CD-HIT at 97% identity into 336 clusters and based on percent identity alignment to SSU rRNA genes from the Silva database, indicate the presence of a diversity of Archaea, Bacteria and Eukaryotes (Supplementary Table 2). *Leptospirillum* group II and the Archaea *Ferroplasma* Type II, G-plasma, and A-plasma were most abundant in environmental samples, whereas *Leptospirillum* group III and *Actinobacteria* were most abundant in bioreactor biofilms (Supplementary Table 2). Nearly half of the OTUs reconstructed were Eukaryotic, including fungi, protists and red algae, and sequence lengths range from ~ 460 bp to 1900 bp. Full-length 18S and 28S rRNA genes from *Acidomyces richmondensis*, a fungus previously observed in Richmond Mine biofilms (3), could be reconstructed and was found increasingly abundant in mature biofilms. Mitochondrial and chloroplast ribosomal sequences were only observed in the most mature 4way GS2 biofilm (Supplementary Table 2). The results support the findings observed by the Cufflinks assembly.

An important consideration for sequence-based “omic” studies is that the depth of sampling is large enough that organisms at very low abundance levels can be detected. However, some organisms identified at low abundance may be contaminants. We went to considerable length to prevent contamination during RNA extraction or library preparation, but the introduction of externally-derived DNA at any of several steps cannot be completely ruled

out. It is notable that we do not detect a wide variety of organisms typically encountered in the laboratory environment (e.g., typical human or room microbiome organisms, such as skin-associated bacteria) or sequences that would suggest contamination from an isolate or other metagenomic project. These considerations give us confidence that rare community members, detected in both laboratory-grown bioreactors and field-collected samples, were intrinsic to those samples.

CONCLUSION

Although both sequencing methods allowed for the detection of rare taxa in all transcriptomics biofilms, HiSeq paired-end datasets were more accurate for the assembling SSU rRNA transcripts and relative abundance estimation. Despite the likely over-estimation of taxonomic richness in GAII-sequenced biofilms, the evaluation of community diversity patterns agree with those observed for HiSeq data, especially when phylogenetic analyses are included (Figures 5-7). Phylogenetic similarity-based diversity analyses are an appropriate method to compare levels of diversity between different samples and developmental stages, and are an important addition to the common metrics used in diversity studies.

Bacteria of the *Magnetobacterium* and *Nitrospira* genera and the Chloroflexi, Deferribacteres, and Gemmatimonadetes phyla have not been observed previously in community genomics or microscopy-based analyses of the well studied Richmond Mine AMD model system. Our analyses indicate that, despite the low pH, elevated temperature, and very high metal concentrations, AMD systems can be much more diverse than previously thought. It is the dominance of a few taxa that makes acidophilic communities well suited for studies ecology, physiology, and diversity.

ACKNOWLEDGEMENTS

We thank Mr. TW Arman, President, Iron Mountain Mines Inc., and Mr R Sugarek (US Environmental Protection Agency) for site access, and Mr R Carver for on-site assistance. DSAG acknowledges funding from the Department of Environmental Science, Policy, and Management at UC Berkeley. We thank The Dimensions of Biodiversity Distributed Graduate Seminar (DBDGS, funded by NSF project 1050680) for useful discussion. We thank Chris Miller for providing purified RNA from the biofilm 4way. David Armitage is thanked for his help with R scripts. Transcriptomic sequencing was done at the University of California Davis. This project was funded by the U.S. Department of Energy, through the Carbon-Cycling (DE-FG02-10ER64996) program.

TABLES

Table 1. Description of samples. Samples 1 to 8 were collected from the natural environment, samples 9 to 13 were grown in laboratory bioreactors. Env: environmental; BR: bioreactor; GS: growth stage.

Sample	Date collected	Location	GS	Type	T (C°)	pH	Type of Seq	No. Reads (M)	% rRNA
1) A-drift	07.17.10	A drift	0	Env	40	1.27	GAII	3.41	80.55
2) C75	09.17.10	C drift	1	Env	46	0.86	GAII	4.66	87.79
3) AB10	11.02.10	AB drift	0	Env	39	0.8	HiSeq	27.87	93.43
4) AB10	11.02.10	AB drift	1	Env	39	0.8	HiSeq	30.24	93.73
5) C10	11.02.10	C drift	0	Env	42	0.8	HiSeq	27.12	91.47
6) C10	11.02.10	C drift	0.5	Env	42	0.8	HiSeq	30.91	93.55
7) C10	11.02.10	C drift	1	Env	42	0.8	HiSeq	28.60	92.54
8) 4-way	07.15.11	4-way	2	Env	39	0.7	HiSeq	26.68	81.88
9) R1	03.30.10	Outflow*	0	BR	37	ND**	GAII	2.56	87.96
10) R1	02.19.10	Outflow*	0.5	BR	37	ND**	GAII	4.86	85.91
11) R2	07.20.09	A drift*	0.5	BR	37	ND**	GAII	4.79	82.60
12) R3	09.28.10	A drift*	0	BR	37	1.31	HiSeq	31.56	92.01
13) R3	10.06.10	A drift*	1	BR	37	1.74	HiSeq	30.49	91.15

*For bioreactor-grown biofilms, location refers to the place where the inoculum was obtained from within the Richmond Mine.

**ND: no data available

Table 2. Relative abundance of SSU rRNA genes (%).

	AB10 GS0	**AB10 GS1**	**C10 GS0**	**C10 GS05**	**C10 GS1**	**4-way GS2**	**Adrift GS0**	**C75 GS1**	**R1 GS0**	**R1 GS05**	**R2 GS05**	**R3 GS0**	**R3 GS1**
Nitrospira	91.81	85.05	91.08	90.47	90.76	85.78	76.67	87.07	94.83	93.18	89.85	95.50	94.90
Bacteria	7.24	12.61	8.25	8.62	7.84	5.35	9.35	4.70	4.42	6.73	6.16	4.44	4.44
Archaea	0.82	1.32	0.66	0.81	0.80	5.30	7.18	8.23	0.75	0.08	0.55	0.06	0.51
Eukaryotes	0.13	1.02	0.01	0.11	0.60	3.58	6.80	0.00	0.00	0.00	3.44	0.00	0.15

Table 3. Relative abundance (%) of *Leptospirillum* spp. and other Nitrospira phylum SSU rRNA genes.

	AB10 GS0	**AB10 GS1**	**C10 GS0**	**C10 GS05**	**C10 GS1**	**4-way GS2**	**Adrift GS0**	**C75 GS1**	**R1 GS0**	**R1 GS05**	**R2 GS05**	**R3 GS0**	**R3 GS1**
Group I	2.53	5.22	3.07	2.62	2.88	2.62	2.59	10.58	2.40	3.41	1.58	0.39	0.31
Group II	74.28	80.43	88.28	88.90	87.94	83.85	14.06	77.34	10.99	25.78	7.36	11.76	7.10
Group III	18.43	13.80	8.22	8.18	8.63	12.71	65.08	11.36	66.68	56.11	71.53	63.12	64.79
Group IV	4.76	0.54	0.44	0.29	0.55	0.82	17.39	0.70	19.01	14.21	18.70	24.58	27.51
Non-Leptos.	0.11	0.03	0.02	0.02	0.02	0.03	4.78	2.77	4.65	3.38	4.20	0.60	1.02

CHAPTER 3 FIGURES

Figure 1. A) Map of the Richmond Mine shows the locations from where biofilms and inoculum for bioreactors were collected.

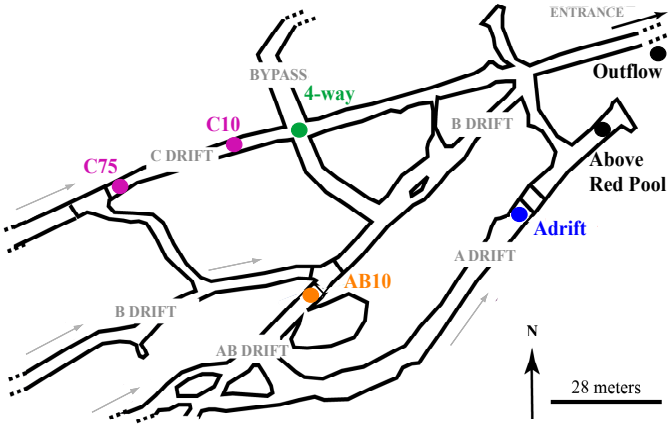


Figure 2. A) Phylogenetic tree of the SSU rRNA genes identified in transcriptomics samples. Assembled SSU rRNA sequences were aligned using the SINA aligner, and phylogenetic tree reconstruction was done using FastTree.

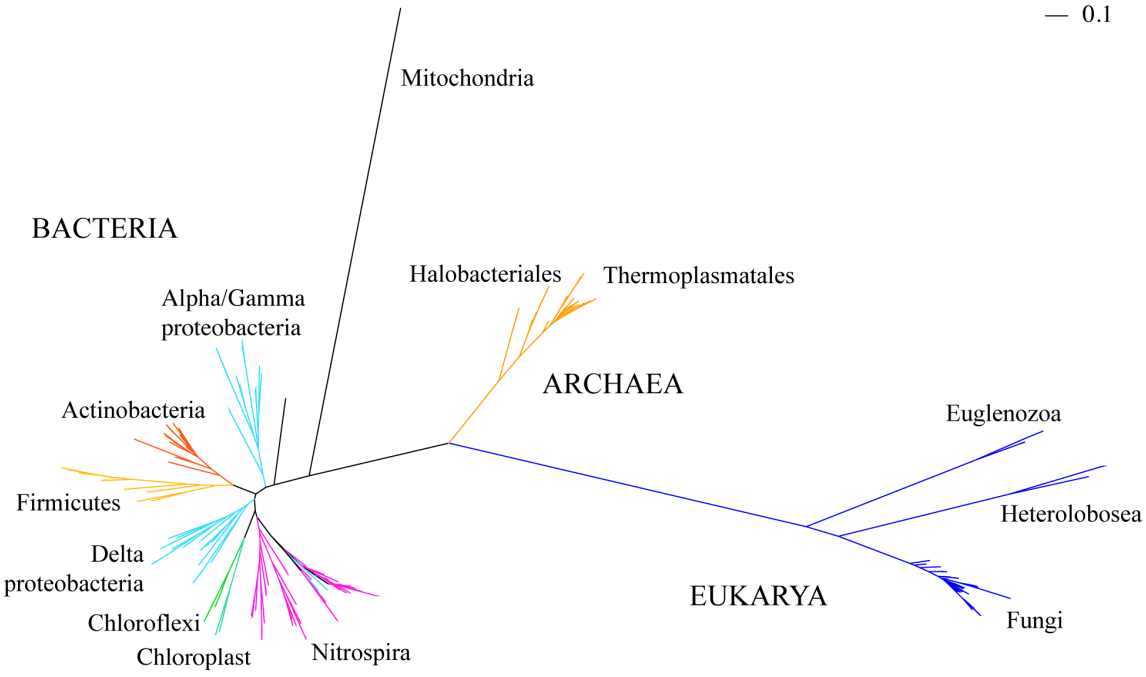


Figure 3. A) Rank abundance curves of assembled SSU rRNA sequences categorized by growth stage and location. Y-axis is represented in logarithmic scale.

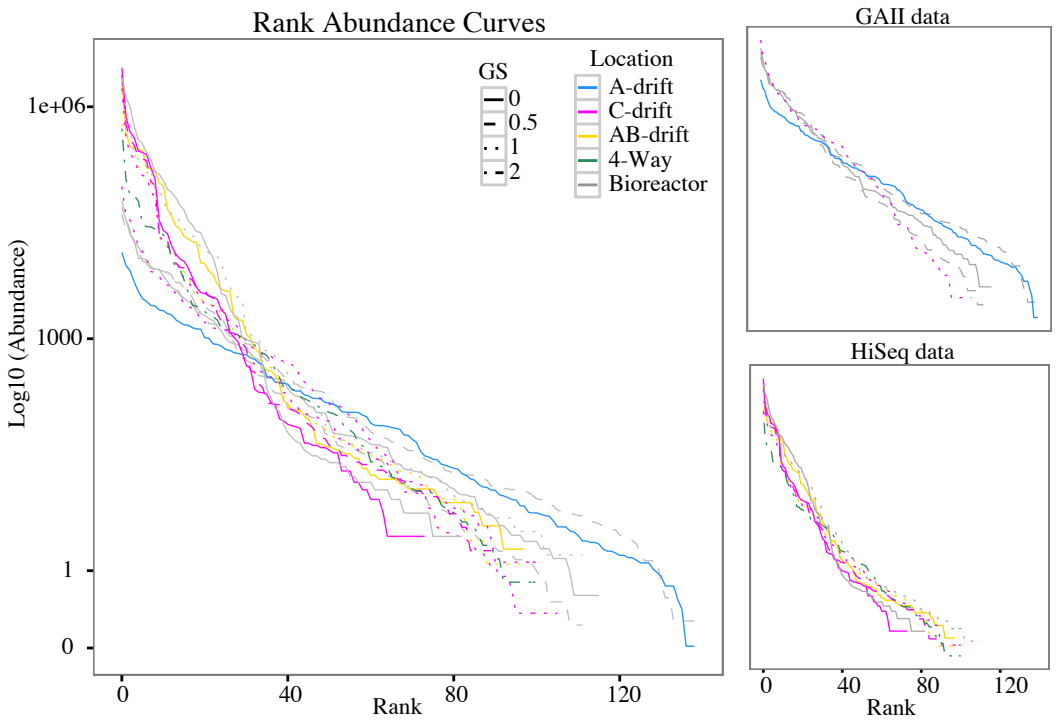


Figure 4. (A) Shannon-Wiener's diversity index (right Y-axis), and Simpson Index of Diversity (left Y-axis). (B) NRI and NTI.

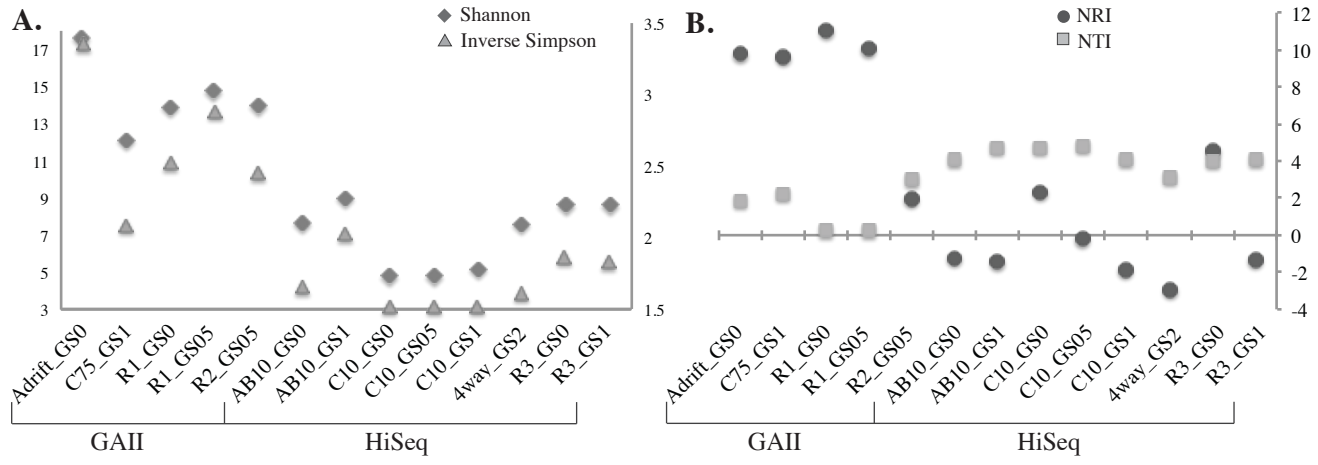


Figure 5. A) PCoA analyses of microbial communities. B) Hierarchical clustering of samples. BR: bioreactor samples, Env: environmental samples. Adrift GS0 is an unusual environmental biofilm sampled from a high pH and low temperature site.

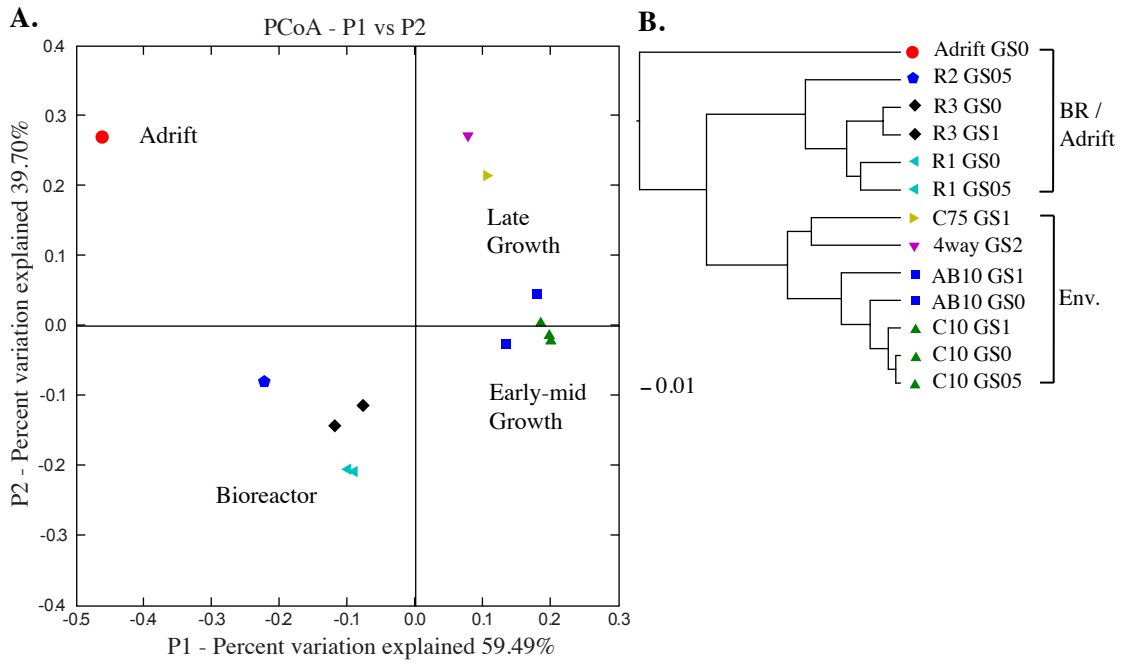


Figure 6. NMDS analyses of microbial communities: A) dimensions 1 and 2. B) dimensions 1 and 3. Bioreactor samples: R1-GS0, R1-GS05, R2-GS05, R3-GS0, and R3-GS1.

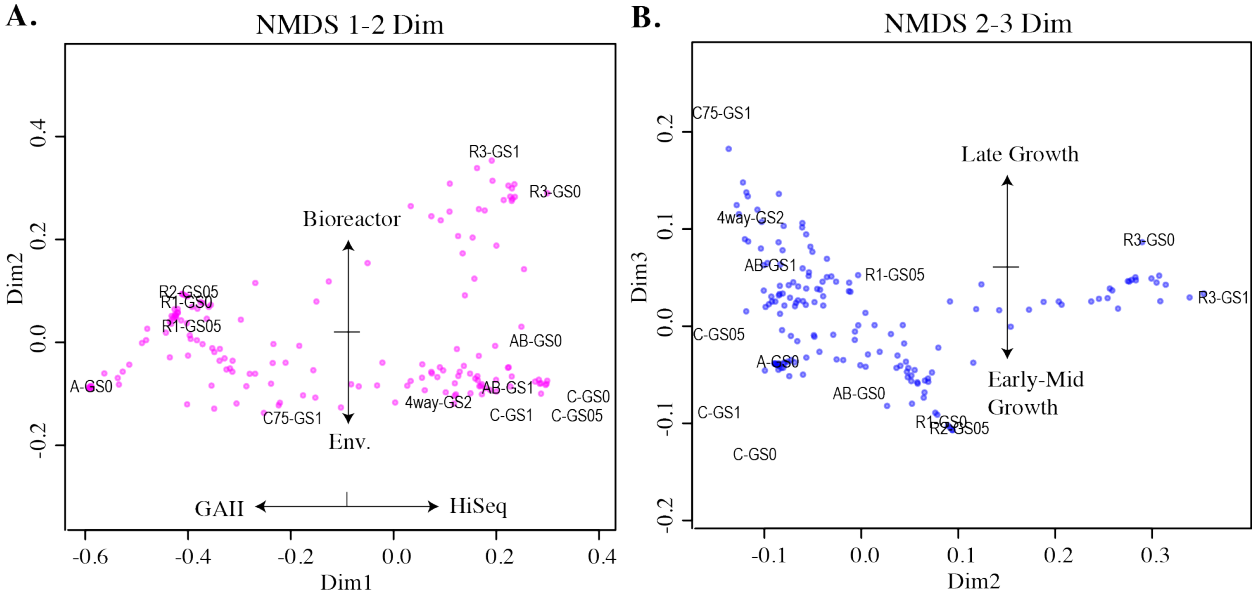


Figure 7. Diversity profiles categorized by growth stage (GS) and location. A and B) taxonomy-based; C and D) phylogenetic similarity-based.

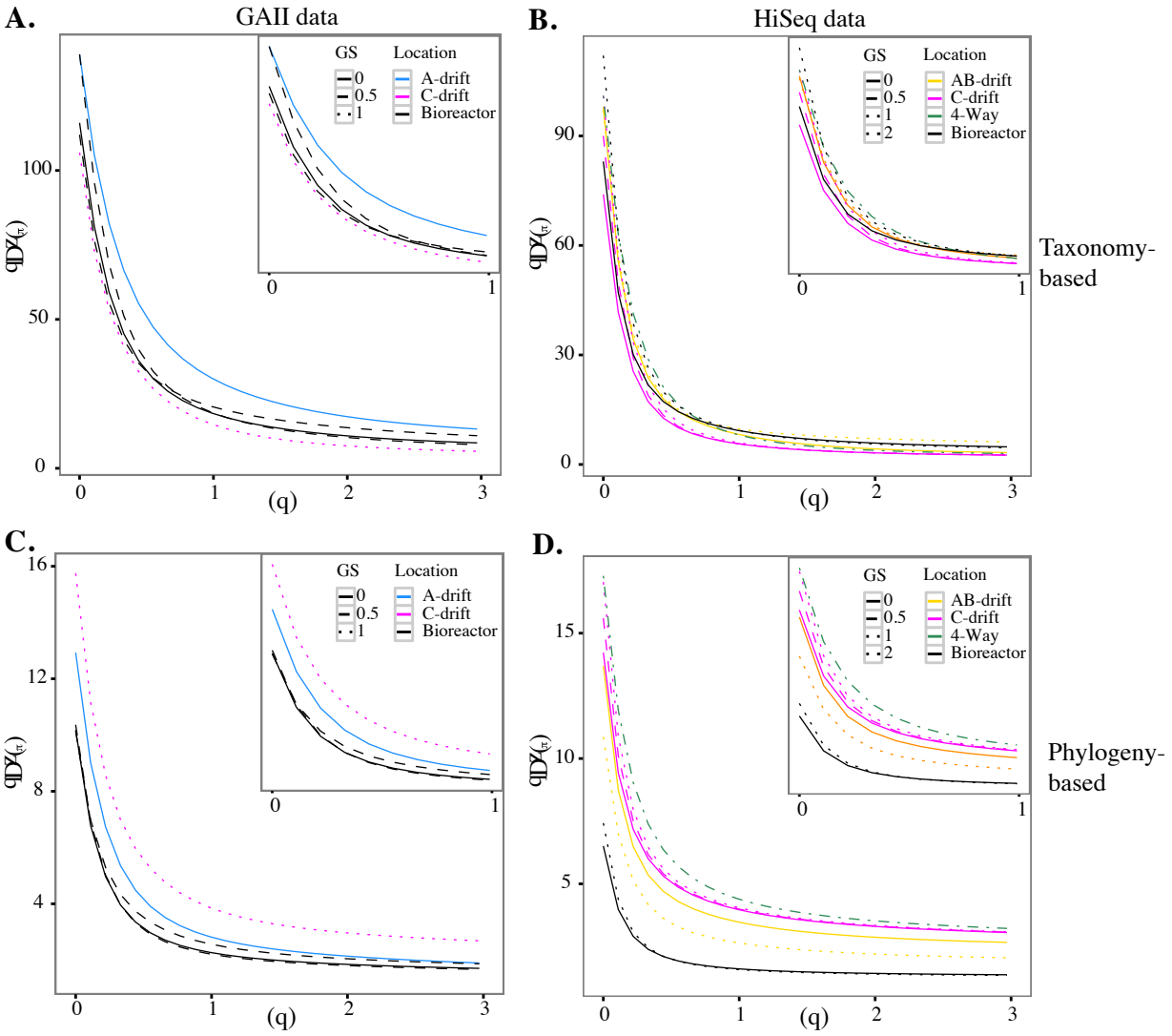


Figure 8. Neighbor-Joining phylogenetic tree of members of the phylum Nitrospira. The sequence of a Delta Proteobacterium was used as outgroup.

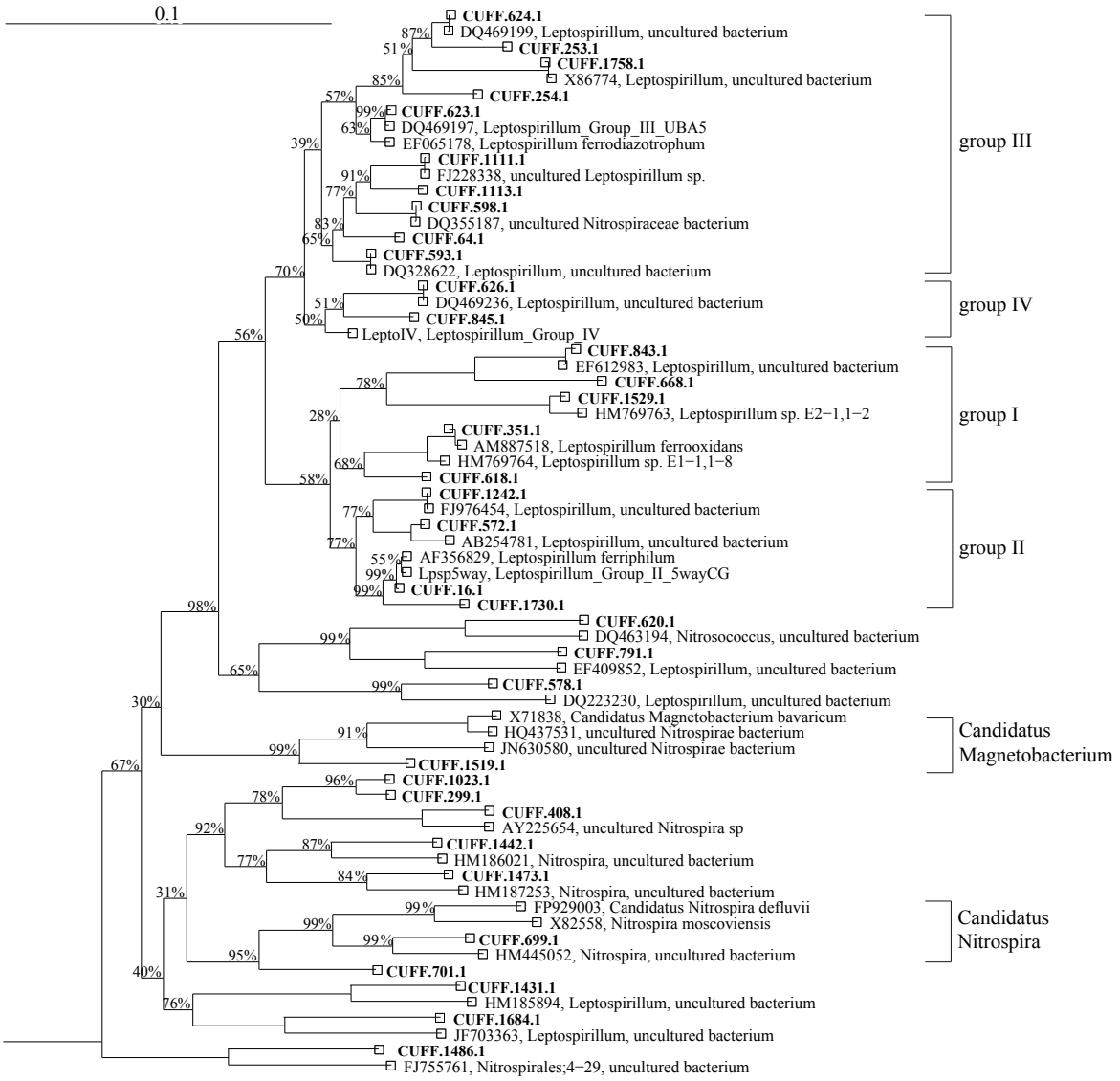
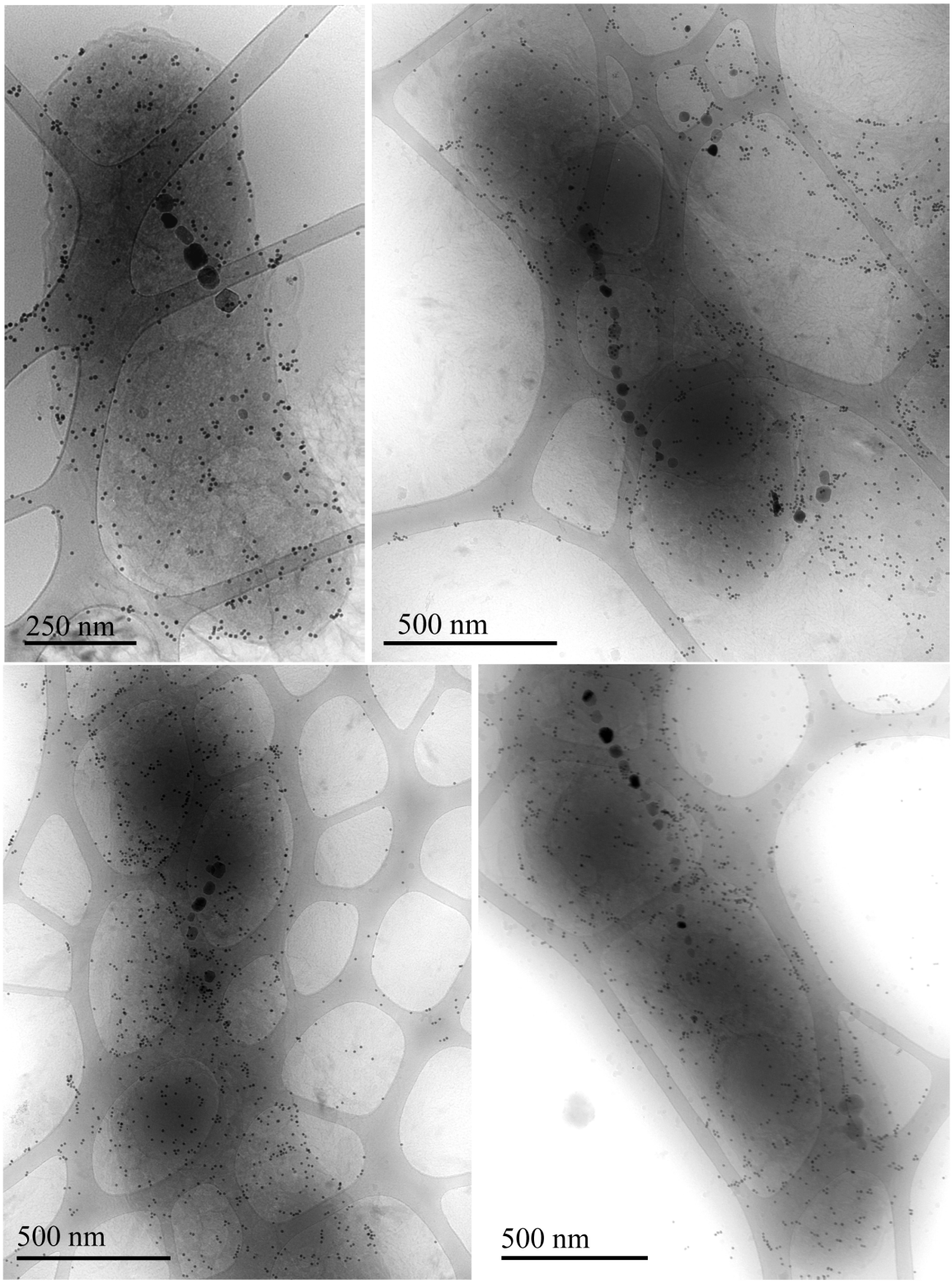


Figure 9. Cryo-TEM images of magenetosome-like containing cells recovered from AMD biofilms.



CHAPTER 4

Community Transcriptomics Provides New Insights into the Ecology of Acidophiles and the Regulation of Their Genes in Acid Mine Drainage Biofilm Communities

ABSTRACT

Gene expression profile studies reveal how organismal responses change with community composition and environmental conditions. To date, community transcriptomic analyses have allowed for studies of gene expression of the most abundant members in a few natural environments. Even less common are integrated studies of the expression of genes and non-coding RNAs in microbial communities. Acid mine drainage (AMD) biofilms from the Richmond Mine, Iron Mountain, California, have served as a model system for "omics"-based microbial community research. Here, we sequenced total and rRNA-depleted RNA from eight AMD biofilms, and from five biofilms grown in the laboratory. A total of 15.8 million non-ribosomal reads were mapped to 20 available genomes of AMD microorganisms and five viruses/phage, and gene expression profiles were obtained for the most abundant biofilm community members. More than 95% of the genes in three bacteria of the genus *Leptospirillum*, and nearly half of the genes in the archaea G-plasma and *Ferroplasma* Type II, and in the *Leptospirillum*–associated virus AMDV1 were detected. Gene expression profiles reflect different environmental preferences of two closely related *Leptospirillum* Group II genotypes, *Leptospirillum* group II UBA and group II 5way CG. Direct evidence for distinct mechanisms of genome regulation in these bacteria likely contributes to this adaptation. An example is a novel riboswitch associated to the biosynthetic pathway of the compatible solute ectoine, which was predicted in each of the *Leptospirillum* group II genotypes and its expression was confirmed by transciptomics. Functions were also assigned for some non-coding RNAs identified previously in other bacteria. The gene expression profile of G-plasma is similar to that of *Leptospirillum* Group II UBA, whereas the expression profile of *Leptospirillum* Group III resembles that of *Leptospirillum* Group II 5way CG. Transposases, cytochromes, and some non-coding RNAs were among the most highly expressed genes in most samples. CRISPR Cas genes were detected at low abundance, as were phage/viral transcripts. Overall results highlight the importance of gene expression profiling in understanding functioning and adaptation of acidophilic biofilm communities.

INTRODUCTION

Studies of gene expression provide clues into the ecology and physiology of organisms in their natural environments. Community transcriptomic analyses have been used to describe important metabolic processes in communities such as nitrogen metabolism in marine environments (22), the flow of carbon between organisms in photosynthetic microbial mats (23), and the modes of life of the most abundant member in an acid mine drainage (AMD) system (122). In addition, metatranscriptomics studies have highlighted the importance of non-coding RNA (ncRNA) expression in marine systems (25).

Extremely acidic environments are usually dominated by relatively few taxa (reviewed in (6)) making them good model systems for ecology and physiology studies. Because of the roles acidophilic microorganisms play in environmental acidification and in metal-recovery based bioleaching processes, gene expression studies are necessary to understanding the ecology and physiology of microorganisms in acidic environments. The Richmond Mine at Iron Mountain, California, is a well-studied acid mine drainage (AMD) system. Deep sequencing of many Richmond Mine biofilms has allowed for the reconstruction of the genomes of many bacteria (8, 11-15, Justice *et al*., in preparation, and unpublished data),

archaea (8, 13, 16, Yelton *et al.*, in review), viruses (17), plasmids (12) and fungi (Miller *et al*. in preparation). Mass spectrometry-based community proteomics measurements have allowed for studies of the physiology of the dominant members of Richmond Mine AMD organisms (e.g. 10, 18, 19, 165), and microarray-based community transcriptomics has been used to study expression profiles of *Leptospirillum* bacteria living in biofilms and in the planktonic fraction of the Rio Tinto AMD system (122). However, studies of the total transcriptome of acidophilic communities have not yet been performed. Here we sequenced total and rRNA-depleted RNA from biofilms collected from the Richmond Mine, and from laboratory bioreactors. Gene expression profiles were obtained for the more abundant AMD organisms, and non-coding RNAs were evaluated in the most abundant bacteria: two closely related *Leptospirillum* group II species that share 99.7 % 16S rRNA sequence identity (*Leptospirillum* group II UBA and group II 5way CG types), and *Leptospirillum* group III. The results greatly expand our knowledge of the responses of AMD organisms to changes in their environment, and provide insight into some gene regulation mechanisms in the *Leptospirilli.*

METHODS

Eight biofilms at different developmental stages (growth-stage (GS) 0, 0.5, 1, and 2) were collected from the A-, C-, AB-drift, and 4way locations within the Richmond Mine at Iron Mountain Mines, California (40°40′ 38.42″ N and 122″ 31′ 19.90″ W, elevation of ~900 m), as described previously (Chapter 3 of this dissertation and (15)). In addition, biofilms were grown in the dark, at pH 1 and 37 °C in laboratory bioreactors using inoculum from within the A-drift, and mine outflow. Biofilms were snap-frozen in liquid nitrogen upon collection and stored at -80 °C.

Total RNA was extracted from all frozen samples using an acid phenol-chloroform-isoamyl alcohol extraction method and purified using the RNEasy MinElute kit (Qiagen), as previously reported (15). RNA integrity was determined by looking at the ratio of 23S to 16S peaks, and at the RNA integrity number (RIN) using a Bioanalyzer 2100. Most samples had an RIN > 8, as suggested in the Illumina mRNA sequencing sample preparation guide, however some samples with RIN > 7 were also considered. Ribosomal RNA-subtraction on 11 of the samples was done with the MicrobExpress kit (Ambion). Total RNA and rRNA-subtracted RNA were converted to cDNA as described by (125) in order to keep the strand-specificity of the transcriptome. Resulting cDNA was fragmented using a Covaris S-system (Covaris, Inc.) to an average fragment size of 200 bp and sent to the University of California Davis for Illumina library preparation and sequencing. Five samples were sequenced using the GAII platform, while eight samples were sequenced using the HiSeq platform.

Low-quality bases were trimmed from the sequencing reads using the fastx_trimmer script (http://hannonlab.cshl.edu/fastx_toolkit/) or the sickle trimmer script with default parameters (https://github.com/najoshi/sickle), and reads < 40 bp in length were filtered out. Trimmed reads were mapped to the SSU and LSU rRNA gene Silva databases (148) using bowtie (127) with default parameters to separate ribosomal from non-ribosomal reads (Figure 1).

Non-ribosomal reads were mapped using bowtie with parameters –n 1 –best –y (127) to predicted genes from the genomes of eight bacteria: *Leptospirillum* group II UBA, group II 5way-CG, group II C75, group III, group IV, a *Sulfobacillus* bin, and two Actinobacterial bins

(8, 11-14, Justice *et al*., in preparation, and unpublished data), nine archaea (A, E, G, and I-plasmas, ARMAN 1, 2, and 4, and *Ferroplasma* Type I and Type II) (8, 13, 16, Yelton *et al.*, in review), and five viruses/phage (17), and unpublished data. Mapped reads were then assembled into transcript fragments using the Cufflinks pipeline and relative abundance measures (FPKM) were obtained (149).

Non-ribosomal reads were also mapped to the genomes of *Leptospirillum* groups II UBA, group II 5way CG, and group III, and the mapping results visually inspected using Artemis (166). Transcribed regions that did not fall within a coding sequence were evaluated for the presence of non-annotated protein sequences using BlastX (56) against the non-redundant NCBI database. In addition, we scanned these regions for the presence of possible ribosome binding sites and start codons that could hint to hypothetical proteins not yet identified in the public databases. Transcribed regions that do not encode for protein sequences, based on the above criteria, were then labeled as non-coding RNA (ncRNA). The ncRNAs were searched against the Rfam database (167), riboswitch motifs were predicted using the RibEx webserver (168), and secondary structure prediction was done using the RNAfold webserver (169).

Hierarchical clustering of FPKM values was done using the software Cluster 3.0 for Mac OSX, centering genes and samples by the median, using the Spearman Rank Correlation similarity matrix, and Average Linkage as clustering method (64). Clusters and heat maps were visualized using the Java TreeView software (65). Gene trees for *Leptospirillum* spp. transposases were constructed using the MABL website (170). CRISPR loci were reconstructed using the CRASS algorithm (171). Reads containing CRISPR repeats and the associated spacers in each sample studied are available in the supplementary materials (CRASS_output).

RESULTS AND DISCUSSION

Transcriptomics statistics.

On average 91.82% ± 4.35% of the transcriptomic reads in total RNA samples and 79.80% ± 9.59% in the rRNA-subtracted samples mapped to rRNA genes from the SSU and LSU Silva databases (Figure 1). The low efficiency of rRNA-depletion kits is a common problem, and methods to accurately remove bacterial, archaeal and eukaryotic rRNA are being developed (172-174). Ribosomal RNA reads were separated from non-rRNA reads, the non-rRNAreads were mapped to the genes of AMD organisms, and correlation analyses of assembled transcript abundance (FPKM values) were done on samples for which total and rRNA-depleted reads were obtained. Transcript abundances correlated well for most samples (R^2 values range from 0.74 to 0.95), hence, non-rRNA reads from rRNA-depleted RNA were pooled with those obtained from their corresponding total RNA (Figure 1). No correlation could be found in one case (A-drift), therefore, only reads obtained from total RNA were considered for further analyses (Supplementary Figure 1H). The correlation between rRNA-depleted and total RNA transcripts from an early developmental stage bioreactor sample, R1 GS0, was low (and dispersion of points was high), likely due to the much shorter fragments assembled from the total RNA sample (Supplementary Figure 2). Given that the correlation between transcript abundances was slightly positive, and that transcript length improved in the rRNA-depleted sample, reads from total and rRNA-depleted RNA were also pooled for R1 GS0.

After mapping reads to the predicted genes of available genomes of AMD organisms, 16,917 transcripts were assembled. Up to 95% of the predicted annotated genes in *Leptospirillum* group II and ~ 98% in *Leptospirillum* group III were detected (Table 1), suggesting that the whole genome is transcribed at some level. This trend has been observed in transcriptomics analyses of isolated bacteria (175), but has not been reported for bacteria growing in microbial communities. In addition, more than half of the predicted genes were transcribed in the archaea G-plasma and *Ferroplasma* Type II, and in the *Leptospirillum* spp.-associated virus AMDV1 (Table 1). Assembled transcripts were also observed for *Actinobacteria*, *Firmicutes*, other alphabet plasmas and viruses at low abundance. Despite the low efficiency of the ribosomal RNA-depletion protocol used, deep sampling allowed us to detect the expression of genes from many community members, including very low abundance organisms.

Gene expression profiles.

Assembled transcripts were tested for relative abundance expression changes between samples using the log2-fold change obtained from the Cufflinks pipeline (see methods). 96.8% of the assembled genes showed a log2-fold change > 2, while 14,749 transcripts had a log2-fold change > 5, and 604 of these had log2-fold change values larger than 10. Hierarchical clustering of these 604 genes separates samples into early to mid growth-stage environmental biofilms, bioreactor samples plus biofilm A-drift, and a cluster represented by a late growth stage biofilm (4way GS2, Figure 2, Supplementary Table S1). *Leptospirillum* group II 5way CG and *Leptospirillum* group III genes are overrepresented in both the bioreactor and A-drift biofilm. The results may reflect similarity between the bioreactor and A-drift environments. Specifically, the A-drift biofilm was collected from a very oxidized pool, at higher pH and lower temperature than other environmental samples, conditions that appear to favor growth of *Leptospirillum* group II 5way CG and *Leptospirillum* III in the laboratory. This finding is consistent with previous reports that *Leptospirillum ferriphilum*, to which *Leptospirillum* group II 5way CG is most closely related, is present in environments at pH > 1.5 (e.g. 122).

Genes from *Leptospirillum* group II UBA are over-expressed in early to mid growth-stage environmental biofilms relative to bioreactor-grown biofilms (Figure 2). *Leptospirillum* group II UBA appears to prefer low pH and high temperature environments, whereas *Leptospirillum* group II 5way CG and *Leptospirillum* group III prefer environments with high pH and low temperatures (Figure 2, Supplementary Table S1). Microscopy and proteomic-based studies have reported that *Leptospirillum* group II UBA generally dominates early growth-stage biofilms, as well as biofilms collected from the C-drift location, a generally low pH and higher temperature environment (10, 19, 47). Our results agree with previous studies that indicate distinct ecological adaptation of the two *Leptospirillum* group II types (10).

Clustering of genes with log2-fold change > 5 in *Leptospirillum* group II UBA and in the archaeon G-plasma indicate that genes involved in energy production and conversion, carbon fixation, fatty acid metabolism, transcription and translation factors, and ribosomal proteins are generally over-expressed in early to mid growth-stage environmental biofilms (Figure 3A, 3B and 3E, Supplementary Table S1). These findings suggest rapid growth during early and mid successional stages. Genes generally over-represented in bioreactor samples include genes involved in amino acid and cofactor metabolism, carbohydrate and lipid metabolism, DNA repair and recombination, lipopolysaccharide metabolism, nucleic acid metabolism, signal transduction, tRNA synthetases and transport genes. Bioreactors usually

have higher pH and lower temperature than those observed in environmental biofilms (See Table 1 in Chapter 3 of this dissertation). These results point to altered (likely slower) growth behavior under bioreactor conditions. The findings confirm the preference of the *Leptospirillum* group II UBA and G-plasma for growth in environmental biofilms over conditions observed in bioreactors. The opposite trend is observed for *Leptospirillum* group III and group II 5way CG: gene expression profiles indicate that these bacteria appear to prefer conditions of higher pH and lower temperature, and therefore, grow best in laboratory bioreactors (Figure 3C, 3D and 3E).

Clustering of viral genes with log2-fold change > 2 also separate environmental from bioreactor samples (Figure 4). *Leptospirillum* spp.-associated AMDV1 phage genes are more highly expressed in one mid-developmental stage bioreactor biofilm compared to all other samples, and Archaeal and unassigned-viral genes are overrepresented in all other samples.

Mobile elements.

Interestingly, transposases in *Leptospirillum* group II UBA and group III are amongst the most highly expressed genes in environmental biofilms. In addition, multi-copy transposases appear to be more highly expressed, and only few single-copy transposases are highly abundant in bioreactor samples (Figure 5). Community proteomic analyses also reported highly abundant transposase proteins in acid mine drainage biofilms (9), and highly abundant transposase genes have been observed previously in community transcriptomic analyses (176, 177). It is possible that the movement of mobile elements is very important in the natural environment, where community membership and environmental conditions change constantly. It has also been suggested that shared habitats promote lateral transfer of transposases among organisms (178).

Transcripts from CRISPR-associated Cas genes were detected in all samples at low abundance for *Leptospirillum* groups II and III, the archaeon *Ferroplasma* Type I, and *Thermoplasmatales* A- and G-plasma (Figure 3, and data not shown). Cas genes from *Leptospirillum* group II UBA and G-plasma were over-represented in bioreactor samples, while those from *Leptospirillum* group II 5way CG and group III were over-represented in environmental biofilms (Figure 3E). Our findings point to some level of regulation of Cas genes.

It has been suggested that Cas proteins and CRISPR loci primary transcripts are constitutively expressed, and their expression levels might be induced as invasion occurs (reviewed in Bhaya (179)). We recovered CRISPR transcripts from the eight datasets, and assigned these to *Leptospirillum* groups II and III, G-plasma, *Ferroplasma* Type 1 and Type 2, *Actinobacteria*, and plasmids (Table 2). Some additional CRISPR transcripts could not be assigned to an organism based on a known repeat sequence. CRISPR transcripts were not abundant in any of the AMD biofilm communities. However, the largest number of distinct *Leptospirillum* group III spacer transcripts was identified in a mid-growth stage bioreactor sample (R3 GS1; Table 2), where the AMDV1 genes were relatively highly expressed. This finding points to diversification rather than up-regulation of transcription of a specific CRISPR locus type as an important response to viral/phage proliferation.

The highest diversity of CRISPR repeats and was observed in a late growth-stage biofilm (4way GS2). This finding suggests the activity of multiple closely related strains with slightly different CRISPR loci (and thus different phage/viral susceptibility) in the more complex, mature biofilms. In *Leptospirillum* group II and G-plasma, the largest number of

distinct spacer transcripts was detected in the same biofilm, consistent with a higher diversity of strains of these species in this sample.

Non-coding RNA (ncRNA).

641 non-coding regions in *Leptospirillum* groups II and III were transcribed: 578 of these showed a fold-change > 2, and 12 have log2-fold change larger than 10 (Supplementary Table 1). Some ncRNA regions show comparable expression levels as their neighboring genes or operons, and many contain reads whose mate-pair falls within coding genes. These results suggest that ncRNAs might play a role in the regulation of their neighboring genes. One example is a transcribed 370 bp 5' UTR region of the ectoine operon in both *Leptospirillum* group II UBA and 5way CG gentypes (Figure 6). Ectoine is a compatible solute synthesized or transported from the environment by many organisms during osmotic stress (95, 180). The length of the ectoine operon associated ncRNA, as well as the presence of mate-paired reads within the transcribed operon, and the presence of a putative promoter (TTGACA-N17-(A)A(A)A(C)T), a Rho-independent terminator (-6.30 Kcal/mol) and an antiterminator (-7.03 Kcal/mol), suggest it is likely a riboswitch. Riboswitches are regions within an mRNA (generally located in the 5' UTR) containing ligand-binding sensors, and regulate the downstream coding sequences (reviewed in (27)). The expression of the riboswitch in *Leptospirillum* group II UBA in two environmental biofilms is higher than that of the operon (Figure 6A, green and red curves): thus, the riboswitch is probably inhibiting transcription of the operon. In contrast, the transcript levels for the bioreactor samples in *Leptospirillum* group II UBA (Figure 6A, blue and magenta curves) are much lower than those of the operon, so transcription of the operon appears enhanced. The expression of the riboswitch in *Leptospirillum* group II 5way CG shows a slight opposite trend to that observed in *Leptospirillum* group II UBA (Figure 6B). The riboswitches of both bacteria share 80% identity at the nucleotide level, and their predicted secondary structures are very different (Figure 6C). These results suggest that *Leptospirillum* group II UBA needs to synthesize compatible solutes to tolerate the environmental conditions present in bioreactor samples, while the opposite is true for *Leptospirillum* group II 5way CG. Therefore, the putative ectoine riboswitch might function by altering the expression of the ectoine operon to synthesize compatible solutes in response to changing environmental conditions. This finding is consistent with a prior suggestion that *Leptospirillum* group II UBA type invests in production of compatible solutes (e.g., trehalose) that become an energy resource for the *Leptospirillum* group II 5way CG genotype (10). To date, there are no reports in the literature of riboswitches associated the ectoine operon.

We identified a few of the well-characterized ncRNAs in the genomes of *Leptospirillum* group II and III, some of which are among the most highly expressed genes (Figure 7). SsrA (or tmRNA) is involved in the rescue of stalled ribosomes during translation (181), and both *Leptospirillum* spp. secondary structures appear to have the conserved tRNA-like domain (Figure 7, dashed circle). RNAseP is the RNA component of a ribonucleoprotein with endonuclease activity whose function is, among others, the processing of precursor tRNAs (182). The *Leptospirillum* spp. RNAseP RNA secondary structures show a high degree of conservation (Figure 7), and both are encoded next to a transposase. The signal recognition particle (SRP) RNA is part of a ribonucleoprotein complex involved in the secretion of proteins (183). The *Leptospirillum* spp. SRP RNA sequences share 78.3% identity between them, and their predicted secondary structures are highly conserved (Figure 7). In addition, the secondary

structures of the 6S RNA show some degree of conservation despite sharing only 56.9% identity at the nucleotide level between the *Leptospirillum* group II UBA and group III sequences. The 6S RNA interacts with the Sigma-70 factor of the RNA polymerase and helps regulate transcription (184). The cobalamin riboswitch of was only identified in *Leptospirillum* group III and contains the conserved domains found within the core region (Figure 7) (185). Many other transcribed ncRNAs with unknown function are present in the genomes of *Leptospirillum* group II-IV at high abundance. The lack of conservation between some ncRNAs, such as riboswitches, makes their structural and functional prediction difficult. Some riboswitches, for example, appear to be specific to a single species (reviewed in Breaker (186)). Overall, the results point to significant roles of regulatory RNAs in *Leptospirillum* spp. Regulatory RNA control of differential genome expression appears to be a factor in species divergence and ecological diversification.

ACKNOWLEDGEMENTS

We thank Mr. TW Arman, President, Iron Mountain Mines Inc., and Mr R Sugarek (US Environmental Protection Agency) for site access, and Mr R Carver for on-site assistance. We thank Mr. Loren Hauser for help with promoter prediction. Transcriptomic sequencing was done at the University of California Davis. This project was funded by the U.S. Department of Energy, through the Carbon-Cycling (DE-FG02-10ER64996) program.

TABLES

Table 1. Percent of genes transcribed in AMD genomes.

	AB10 GS0	AB10 GS1	C10 GS1	C10 GS0	C10 GS05	4way GS2	R3 GS0	R3 GS1	Adrift GS0	C75 GS1	R1 GS0	R1 GS05	R2 GS05	Reference
A-plasma	0.79	5.13	1.27	0.04	6.05	4.65	-	8.29	0.35	29.52	0.04	-	0.13	(13)
Actino1	-	-	-	-	-	0.08	-	6.36	0.08	0.04	0.43	-	-	Unpublished
Actino2	0.23	0.23	0.23	0.11	0.23	0.23	-	0.57	0.45	0.28	0.06	0.06	0.11	Unpublished
AMDV1	36.00	2.00	2.00	2.00	34.00	44.00	-	-	-	2.00	-	-	-	(17)
AMDV3	0.88	0.88	0.88	-	0.88	3.96	-	0.44	-	9.69	-	-	-	(17)
AMDV4	-	-	-	-	-	-	-	-	0.65	-	-	-	-	(17)
AMDVIR	1.35	1.16	2.32	0.58	1.54	2.12	0.97	4.25	0.19	9.07	0.19	0.58	0.39	Unpublished
ARMAN1	-	-	-	-	-	0.10	-	-	-	1.10	-	-	-	(16)
ARMAN2	-	-	-	-	-	-	-	-	-	0.11	-	-	-	(16)
ARMAN4	-	-	-	-	-	0.47	-	-	-	0.35	-	-	-	(16)
C-plasma	0.06	0.35	0.06	-	1.40	0.47	-	-	1.10	7.15	-	-	-	Unpublished
D-plasma	0.17	0.22	0.11	-	0.11	0.77	-	0.28	-	0.55	-	-	-	Unpublished
E-plasma	0.18	0.30	0.18	-	0.18	1.08	-	0.36	1.26	0.36	0.06	-	-	(13)
G-plasma	27.67	31.98	22.46	0.78	27.56	76.55	-	35.62	2.29	40.30	2.81	-	0.21	(13)
I-plasma	0.18	0.29	0.12	-	0.24	0.94	-	0.24	-	0.53	-	-	-	(13)
FER1 CG	0.06	-	0.06	-	-	23.19	-	-	-	-	-	-	-	(187)
FER1 Isolate	0.05	0.10	-	-	0.05	21.84	-	-	-	-	-	-	-	(187)
FER2 CG	0.80	2.56	0.84	-	0.15	50.29	-	0.15	-	0.48	-	-	-	(32)
Firmicute bin	0.50	-	-	-	0.43	0.71	-	0.14	-	0.07	0.07	0.14	-	Unpublished
Lepto2 C75	78.52	75.99	78.12	57.62	77.09	47.59	3.52	18.60	0.32	80.69	5.77	16.19	1.70	(121)
Lepto2 CG	69.12	21.83	26.90	10.26	21.48	92.80	28.83	76.78	2.13	24.30	41.87	62.58	16.91	(14)
Lepto2 UBA	71.77	72.25	74.41	55.57	79.66	34.82	1.61	11.63	0.62	62.78	2.88	8.44	1.96	(11)
Lepto3	68.20	13.40	14.37	0.62	11.97	53.40	72.44	90.37	23.63	0.28	85.29	81.33	85.01	(12)
Lepto4	2.55	0.26	0.10	0.05	0.21	2.13	2.50	6.82	1.40	0.05	3.85	2.76	3.28	(15)
Sulfobacillus 1	-	-	-	-	-	0.33	-	-	-	-	-	-	-	Unpublished
UNLVIR	2.22	1.67	1.11	0.56	2.22	2.78	1.11	8.89	-	3.89	-	0.56	0.56	Unpublished

Table 2. CRISPR loci recovered from paired-end transcriptomics datasets. Columns 3-10 indicate the number of spacers present in each locus and the coverage (No. reads).

Repeat	Organism	AB10 GS0	AB10 GS1	C10 GS0	C10 GS05	C10 GS1	4way GS2	R3 GS0	R3 GS1	Reference
CGGTTCATCCCCACGAACGTGGGGAATAC	*L.* group II	144 (6094)	98 (5580)	139 (976)	200 (11531)	113 (4722)	338 (1567)	9 (35)	67 (222)	CRISPR #5 (17)
CGGTTCATCCCCGCGGGCGCGGGGAACAC	*L.* group II								4 (8)	Variant of CRISPR #5 (17)
ACAGGATTCACGCAAGGCTCACCCTAGAGTGATTTCG	*L.* group II		3 (4)							This study
ACCTATGCTACACAGGTTATCCTAGCACCC	*L.* group II						3 (16)			This study
AGAAACACCCCCGCGCATGCGGGGAAAACAG	*L.* group IV		2 (5)							CRISPR #11 (17)
(A/G)GAAACACCCCCACGGGCGTGGGGAAGAC	*L.* group III	11 (87)					8 (60)	6 (60)	220 (2829)	CRISPR #7 (17)
(A)GTGTACCGCCGCAAGGCGGCTTAGAAA	Plasmid		2 (9)	16 (59)	19 (50)		10 (30)		15 (48)	CRISPR #19 (17)
(A)GTTATCCACGACATATGTGGCTTAGAAA(T)	Plasmid	7 (23)	13 (45)	2 (7)	5 (15)		4 (25)		2 (5)	CRISPR #13 (17)
GTTTTTATTTGCCATAGGCAGTAATAATTTGTTTGATAATTCA	Plasmid		3 (6)		4 (6)			2 (5)	6 (18)	CRISPR #17 (17)
ATTTCAGAAAAACTAGTTAGTATGGAAG	G-plasma		3 (19)		2 (17)		21 (88)		1 (5)	CRISPR #22 (17)
CTTTCAATCCTATCAAGGTTCTATTTTTAC	G-plasma						20 (125)			CRISPR #9 (17)
CTTTGAAACTTTCTAAATAAGATTCTAAC	G-plasma	2 (12)					23 (113)			CRISPR #6 (17)
CTTTCAATCCTATTAAGGTTCGATTTTAAC	G-plasma						16 (78)			variant of CRISPR #3 (17)
ACTTCCATACTACCTAGTGCTTTTTAAAC	*Ferroplasma* Type 2						9 (36)			This study
ATTTCAAACCCTTATAGATAGACTAAACAC	*Ferroplasma* Type 1						9 (63)			This study
CTCCTCCCCGCACACGCGGGGGTCATCCC	Actinobacteria								6 (9)	This study
ATTGTTAGAATACCTATAAGGACTTGAAAC	Unassigned						22 (74)			Unplaced CRISPR #12 (17)

CHAPTER 4 FIGURES

Figure 1. Flow diagram of the transcriptomics protocol used.

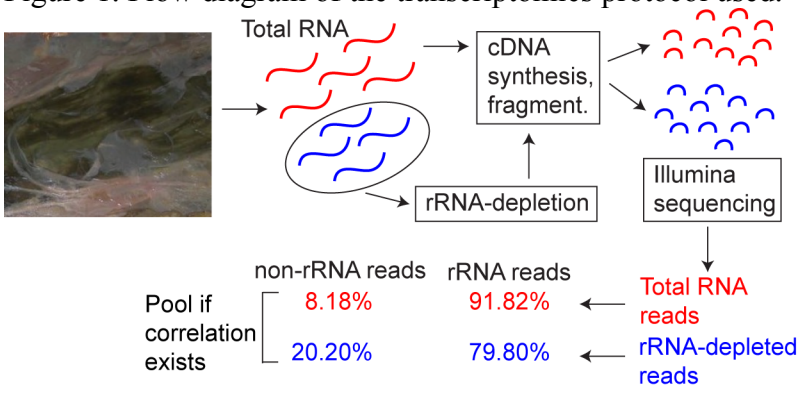


Figure 2. Heatmap of genes with log2-fold change > 10. Samples cluster in environmental or bioreactor samples, some clustering is observed based on growth stage.

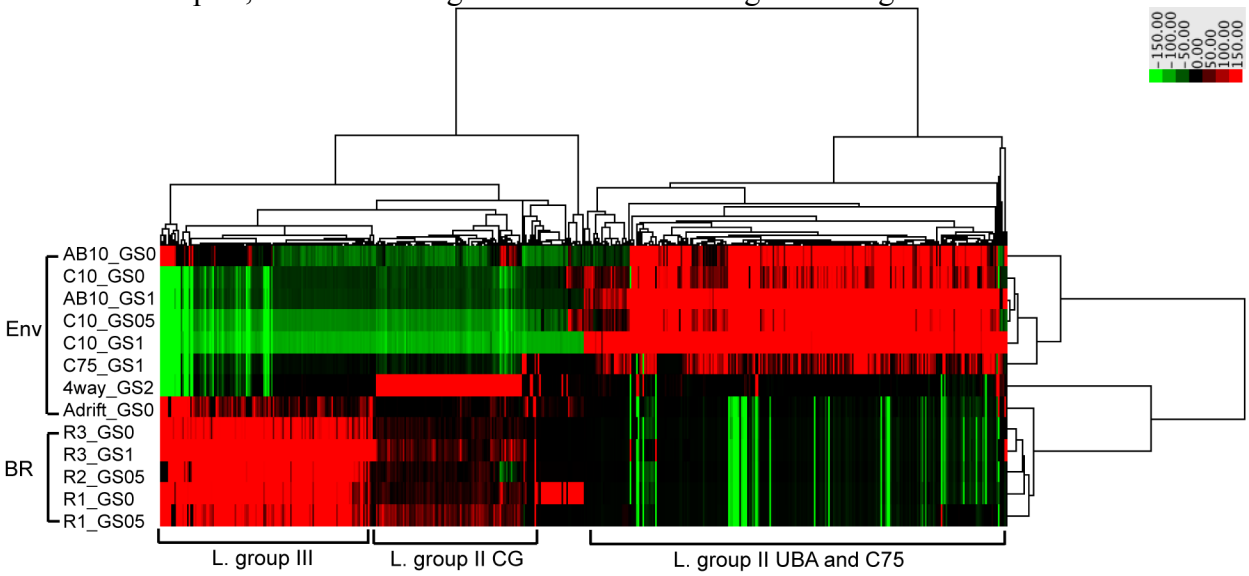


Figure 3. (next page) Heatmap of genes with log2-fold change > 5 in: A) *Leptospirillum* group II UBA; B) G-plasma; C) *Leptospirillum* group III; and D) *Leptospirillum* group II 5way CG. Yellow: overrepresented, blue: underrepresented. E) Distribution of the number of genes in functional categories highly expressed in environmental and bioreactor samples. The x-axis represents the number of genes detected by transcriptomics in each functional category.

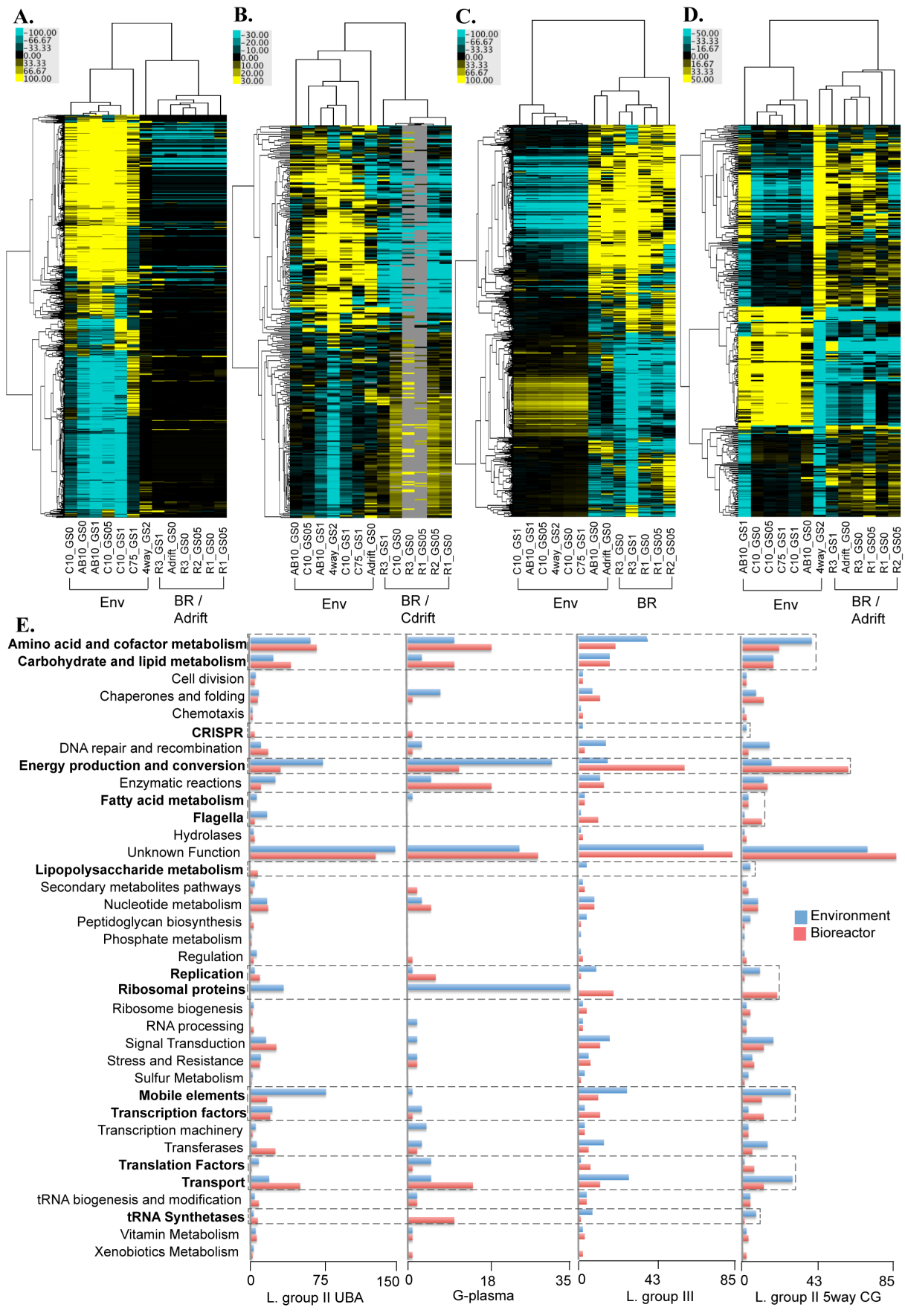

A.
B.
C.
D.
E.
Env
BR / Adrift
BR / Cdrift
BR
Amino acid and cofactor metabolism
Carbohydrate and lipid metabolism
Cell division
Chaperones and folding
Chemotaxis
CRISPR
DNA repair and recombination
Energy production and conversion
Enzymatic reactions
Fatty acid metabolism
Flagella
Hydrolases
Unknown Function
Lipopolysaccharide metabolism
Secondary metabolites pathways
Nucleotide metabolism
Peptidoglycan biosynthesis
Phosphate metabolism
Regulation
Replication
Ribosomal proteins
Ribosome biogenesis
RNA processing
Signal Transduction
Stress and Resistance
Sulfur Metabolism
Mobile elements
Transcription factors
Transcription machinery
Transferases
Translation Factors
Transport
tRNA biogenesis and modification
tRNA Synthetases
Vitamin Metabolism
Xenobiotics Metabolism
Environment
Bioreactor
L. group II UBA
G-plasma
L. group III
L. group II 5way CG

Figure 4. Heatmap of viral genes with log2-fold change > 2: Cluster 1) Unassigned viral genes (11), and Archaeal-associated phage genes (AMDVIR, 25). Cluster 2) *Leptospirillum*-associated phage genes (AMDV1, 23), and Archaeal-associated phage genes (AMDV3, 16; AMDVIR, 24). Red: overrepresented, green: underrepresented.

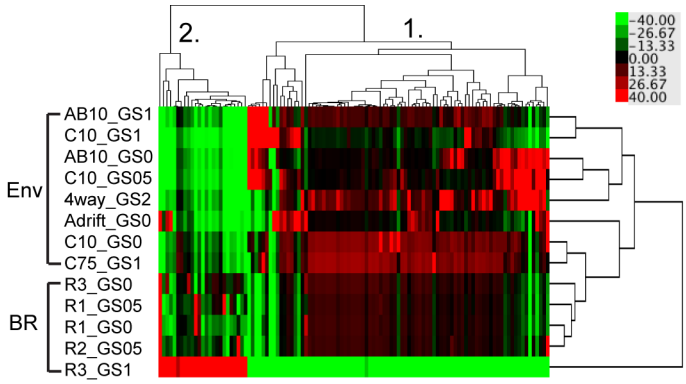


Figure 5. Gene tree of transposases in: A) *Leptospirillum* group II UBA; and B) *Leptospirillum* group III. Blue squares represent transposases highly expressed in environmental samples, while red squares represent transposases highly expressed in bioreactors.

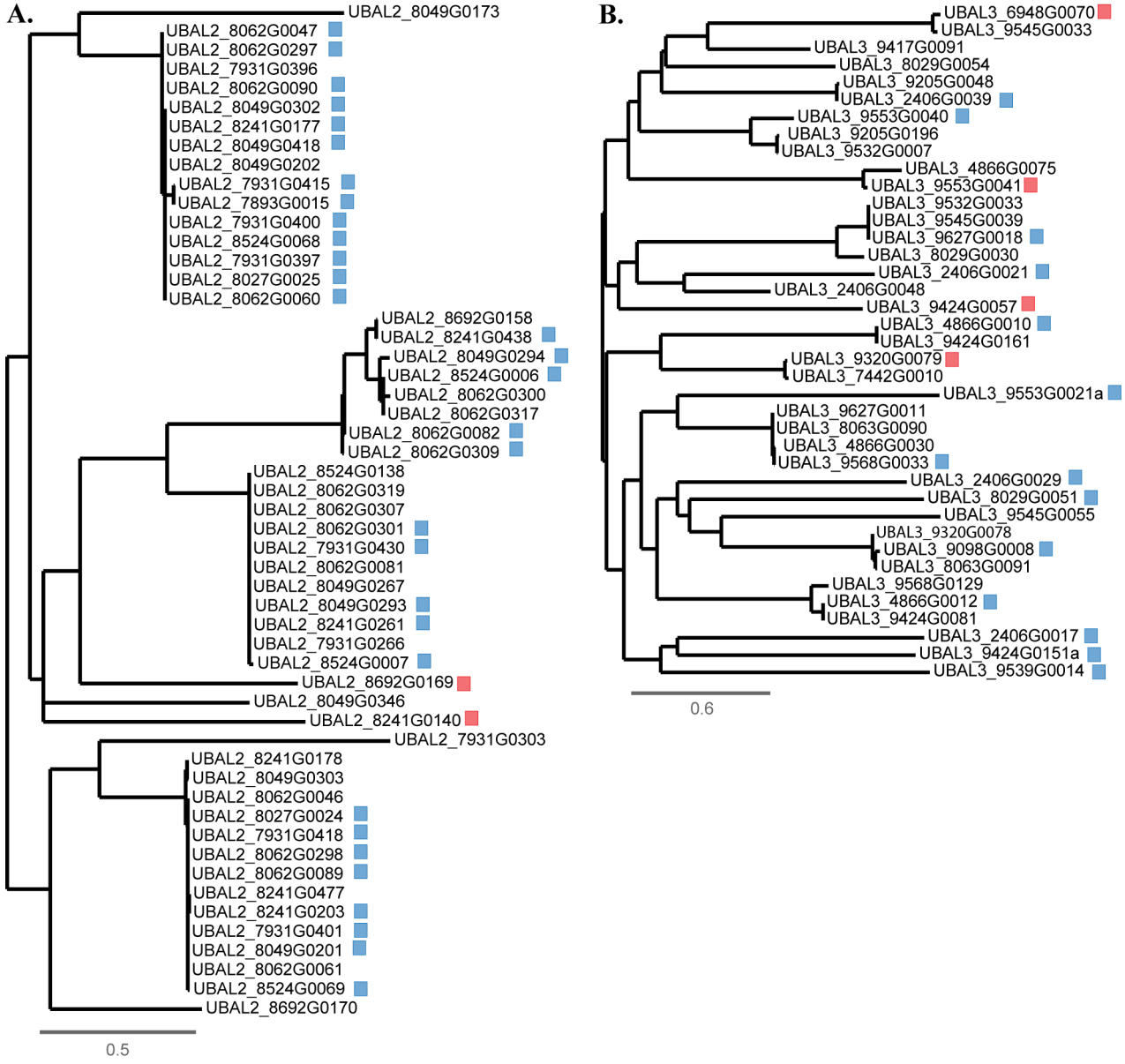


Figure 6. Predicted ectoine riboswitch in *Leptospirillum* group II bacteria. The ectoine operon (teal blue arrows), and the riboswitch (ncRNA) with its promoter (P) and Rho-independent terminator (T) are shown. The modified Artemis screenshot in *Leptospirillum* group II UBA (A) and in *Leptospirillum* group II 5way CG (B) shows the strand-specific transcriptomic reads distribution from two bioreactor samples (blue: R3_GS0; magenta: R3_GS1) and two environmental biofilms (red: AB10_GS0; green: AB10_GS1). The predicted secondary structure of the riboswitch in *Leptospirillum* group II UBA (C) and in *Leptospirillum* group II 5way CG (D) show the predicted Rho-independent terminator (Solid arrow) and the antiterminator (dashed arrow). Colors indicate base-pairing probabilities.

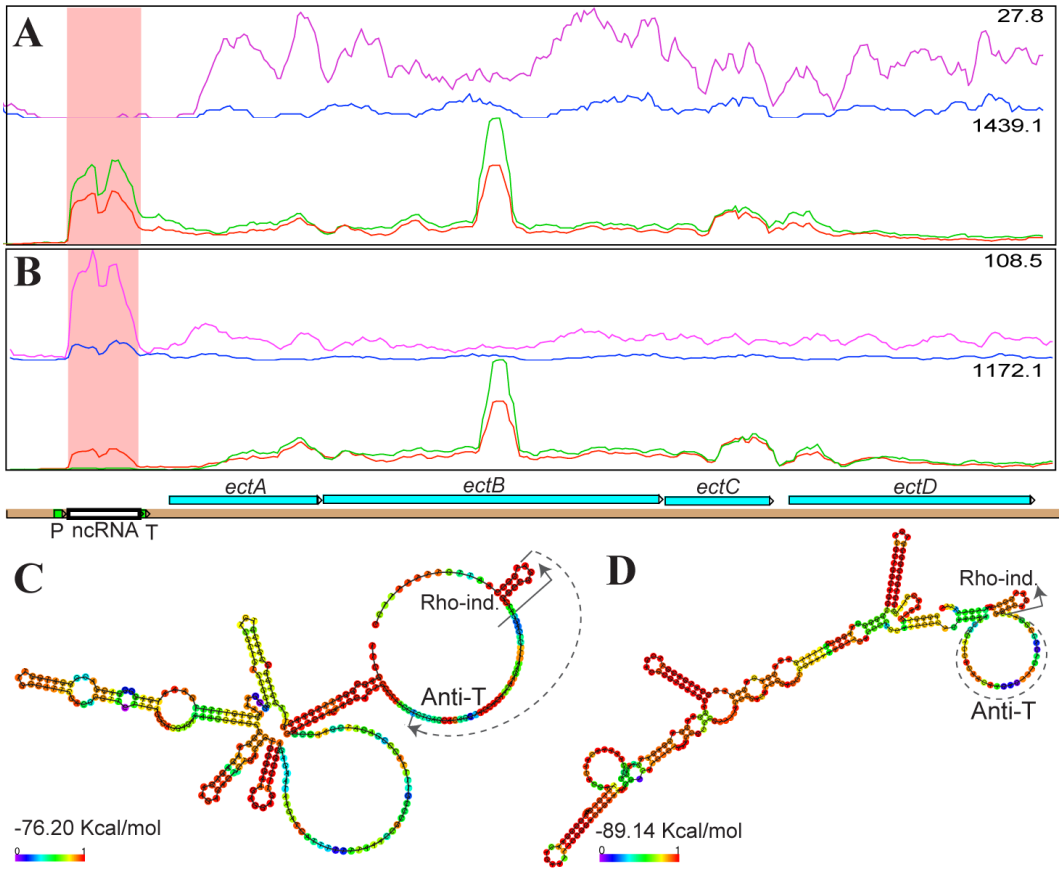
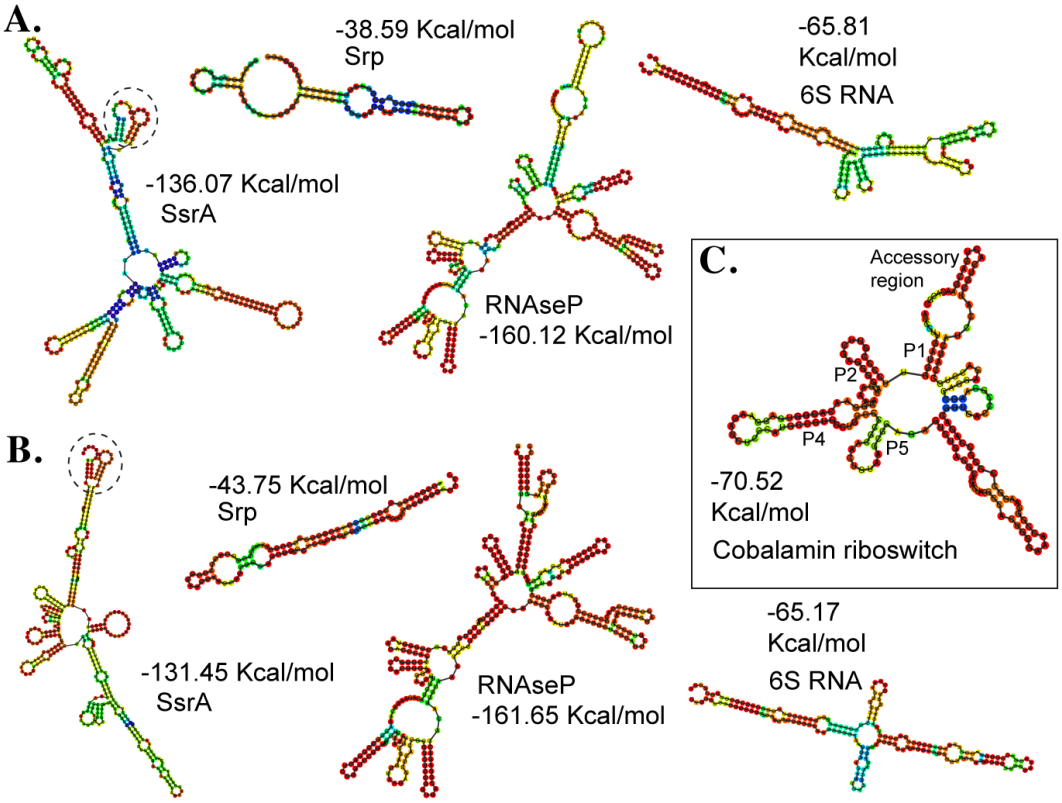


Figure 7. Predicted secondary structures of transcribed ncRNAs in: A) *Leptospirillum* group II UBA, and B-C) *Leptospirillum* group III. Structures were folded using the RNAfold webserver.



REFERENCES

1. **Domic E.** 2007. A Review of the Development and Current Status of Copper Bioleaching Operations in Chile: 25 Years of Successful Commercial Implementation, p. 81-95, Biomining. Springer Berlin Heidelberg.
2. **Rohwerder T, Gehrke T, Kinzler K, Sand W.** 2003. Bioleaching review part A: progress in bioleaching: fundamentals and mechanisms of bacterial metal sulfide oxidation. Appl Microbiol Biotechnol **63:**239-248.
3. **Baker BJ, Banfield JF.** 2003. Microbial communities in acid mine drainage. FEMS Microbiol Ecol **44:**139-152.
4. **Edwards KJ, Bond PL, Druschel GK, McGuire MM, Hamers RJ, Banfield JF.** 2000. Geochemical and biological aspects of sulfide mineral dissolution: lessons from Iron Mountain, California. Chemical Geology **169:**383-397.
5. **Schrenk MO, Edwards KJ, Goodman RM, Hamers RJ, Banfield JF.** 1998. Distribution of thiobacillus ferrooxidans and leptospirillum ferrooxidans: implications for generation of acid mine drainage. Science **279:**1519-1522.
6. **Johnson DB.** 2012. Geomicrobiology of extremely acidic subsurface environments. FEMS Microbiol Ecol **81:**2-12.
7. **Denef VJ, Mueller RS, Banfield JF.** 2010. AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. ISME J **4:**599-610.
8. **Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428:**37-43.
9. **Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake RC, 2nd, Shah M, Hettich RL, Banfield JF.** 2005. Community proteomics of a natural microbial biofilm. Science **308:**1915-1920.
10. **Denef VJ, Kalnejais LH, Mueller RS, Wilmes P, Baker BJ, Thomas BC, VerBerkmoes NC, Hettich RL, Banfield JF.** 2010. Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. Proc Natl Acad Sci U S A **107:**2383-2390.
11. **Lo I, Denef VJ, Verberkmoes NC, Shah MB, Goltsman D, DiBartolo G, Tyson GW, Allen EE, Ram RJ, Detter JC, Richardson P, Thelen MP, Hettich RL, Banfield JF.** 2007. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. Nature **446:**537-541.
12. **Aliaga Goltsman DS, Denef VJ, Singer SW, VerBerkmoes NC, Lefsrud M, Mueller RS, Dick GJ, Sun CL, Wheeler KE, Zemla A, Baker BJ, Hauser L, Land M, Shah MB, Thelen MP, Hettich RL, Banfield JF.** 2009. 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 **75:**4599-4615.
13. **Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF.** 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biol **10:**R85.

14. **Simmons SL, Dibartolo G, Denef VJ, Goltsman DS, Thelen MP, Banfield JF.** 2008. Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. PLoS Biol **6:**e177.
15. **Aliaga Goltsman DS, Dasari M, Thomas BT, Shah MB, VerBerkmoes NC, Hettich RL, Banfield JF.** 2013. A new group in the Leptospirillum clade: cultivation-independent community genomics, proteomics and transcriptomics of the new species Leptospirillum group IV UBA BS. Appl Environ Microbiol**:**In press.
16. **Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, Land ML, Verberkmoes NC, Hettich RL, Banfield JF.** 2010. Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci U S A **107:**8806-8811.
17. **Andersson AF, Banfield JF.** 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. Science **320:**1047-1050.
18. **Mueller RS, Denef VJ, Kalnejais LH, Suttle KB, Thomas BC, Wilmes P, Smith RL, Nordstrom DK, McCleskey RB, Shah MB, Verberkmoes NC, Hettich RL, Banfield JF.** 2010. Ecological distribution and population physiology defined by proteomics in a natural microbial community. Mol Syst Biol **6:**374.
19. **Mueller RS, Dill BD, Pan C, Belnap CP, Thomas BC, VerBerkmoes NC, Hettich RL, Banfield JF.** 2011. Proteome changes in the initial bacterial colonist during ecological succession in an acid mine drainage biofilm community. Environ Microbiol **13:**2279-2292.
20. **Denef VJ, Verberkmoes NC, Shah MB, Abraham P, Lefsrud M, Hettich RL, Banfield JF.** 2008. Proteomics-inferred genome typing (PIGT) demonstrates inter-population recombination as a strategy for environmental adaptation. Environ Microbiol.
21. **Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, Delong EF.** 2008. Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A **105:**3805-3810.
22. **Stewart FJ, Sharma AK, Bryant JA, Eppley JM, DeLong EF.** 2011. Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. Genome Biol **12:**R26.
23. **Burow LC, Woebken D, Marshall IP, Lindquist EA, Bebout BM, Prufert-Bebout L, Hoehler TM, Tringe SG, Pett-Ridge J, Weber PK, Spormann AM, Singer SW.** 2013. Anoxic carbon flux in photosynthetic microbial mats as revealed by metatranscriptomics. ISME J **7:**817-829.
24. **VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF.** 2009. Systems biology: Functional analysis of natural microbial consortia using community proteomics. Nat Rev Microbiol **7:**196-205.
25. **Shi Y, Tyson GW, DeLong EF.** 2009. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature **459:**266-269.
26. **Gottesman S, Storz G.** 2011. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harbor perspectives in biology **3**.
27. **Serganov A, Nudler E.** 2013. A decade of riboswitches. Cell **152:**17-24.
28. **Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P.** 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science **315:**1709-1712.

29. **Schloss PD, Handelsman J.** 2005. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol **6:**229.
30. **Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ, Szeto E, Kyrpides NC, Mussmann M, Amann R, Bergin C, Ruehland C, Rubin EM, Dubilier N.** 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature **443:**950-955.
31. **Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E, Dalin E, Putnam NH, Shapiro HJ, Pangilinan JL, Rigoutsos I, Kyrpides NC, Blackall LL, McMahon KD, Hugenholtz P.** 2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotechnol **24:**1263-1269.
32. **Eppley JM, Tyson GW, Getz WM, Banfield JF.** 2007. Genetic exchange across a species boundary in the archaeal genus ferroplasma. Genetics **177:**407-416.
33. **Washburn MP, Wolters D, Yates JR, 3rd.** 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol **19:**242-247.
34. **Lipton MS, Pasa-Tolic L, Anderson GA, Anderson DJ, Auberry DL, Battista JR, Daly MJ, Fredrickson J, Hixson KK, Kostandarithes H, Masselon C, Markillie LM, Moore RJ, Romine MF, Shen Y, Stritmatter E, Tolic N, Udseth HR, Venkateswaran A, Wong K-K, Zhao R, Smith RD.** 2002. From the Cover: Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. Proceedings of the National Academy of Sciences **99:**11049-11054.
35. **VerBerkmoes NC, Shah MB, Lankford PK, Pelletier DA, Strader MB, Tabb DL, McDonald WH, Barton JW, Hurst GB, Hauser L, Davison BH, Beatty JT, Harwood CS, Tabita FR, Hettich RL, Larimer FW.** 2006. Determination and comparison of the baseline proteomes of the versatile microbe Rhodopseudomonas palustris under its major metabolic states. J Proteome Res **5:**287-298.
36. **Brown SD, Thompson MR, Verberkmoes NC, Chourey K, Shah M, Zhou J, Hettich RL, Thompson DK.** 2006. Molecular dynamics of the Shewanella oneidensis response to chromate stress. Mol Cell Proteomics **5:**1054-1071.
37. **Callister SJ, Dominguez MA, Nicora CD, Zeng X, Tavano CL, Kaplan S, Donohue TJ, Smith RD, Lipton MS.** 2006. Application of the accurate mass and time tag approach to the proteome analysis of sub-cellular fractions obtained from Rhodobacter sphaeroides 2.4.1. Aerobic and photosynthetic cell cultures. J Proteome Res **5:**1940-1947.
38. **Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF, Carlson CA, Smith RD, Giovanonni SJ.** 2008. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. ISME J.
39. **Wilmes P, Andersson AF, Lefsrud MG, Wexler M, Shah M, Zhang B, Hettich RL, Bond PL, VerBerkmoes NC, Banfield JF.** 2008. Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. ISME J **2:**853-864.
40. **Denef VJ, Shah MB, Verberkmoes NC, Hettich RL, Banfield JF.** 2007. Implications of strain- and species-level sequence divergence for community and isolate shotgun proteomic analysis. J Proteome Res **6:**3152-3161.

41. **Banfield JF, Verberkmoes NC, Hettich RL, Thelen MP.** 2005. Proteogenomic approaches for the molecular characterization of natural microbial communities. OMICS **9:**301-333.
42. **Bond PL, Druschel GK, Banfield JF.** 2000. Comparison of acid mine drainage microbial communities in physically and geochemically distinct ecosystems. Appl Environ Microbiol **66:**4962-4971.
43. **Parro V, Moreno-Paz M, Gonzalez-Toril E.** 2007. Analysis of environmental transcriptomes by DNA microarrays. Environ Microbiol **9:**453-464.
44. **Coram NJ, Rawlings DE.** 2002. Molecular relationship between two groups of the genus Leptospirillum and the finding that Leptospirillum ferriphilum sp. nov. dominates South African commercial biooxidation tanks that operate at 40 degrees C. Appl Environ Microbiol **68:**838-845.
45. **Norris PR.** 2006. Acidophile Diversity in Mineral Sulfide Oxidation p. 199-216. *In* Rawlings DE, Johnson, D. B. (ed.), Biomining. Springer Berlin Heidelberg.
46. **Xie X, Xiao S, He Z, Liu J, Qiu G.** 2007. Microbial populations in acid mineral bioleaching systems of Tong Shankou Copper Mine, China. J Appl Microbiol **103:**1227-1238.
47. **Wilmes P, Remis JP, Hwang M, Auer M, Thelen MP, Banfield JF.** 2009. Natural acidophilic biofilm communities reflect distinct organismal and functional organization. ISME J **3:**266-270.
48. **Tyson GW, Lo I, Baker BJ, Allen EE, Hugenholtz P, Banfield JF.** 2005. Genome-directed isolation of the key nitrogen fixer Leptospirillum ferrodiazotrophum sp. nov. from an acidophilic microbial community. Appl Environ Microbiol **71:**6319-6324.
49. **Gordon D.** 2003. Viewing and editing assembled sequences using Consed. Curr Protoc Bioinformatics **Chapter 11:**Unit11 12.
50. **Konstantinidis KT, Tiedje JM.** 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A **102:**2567-2572.
51. **Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T.** 2003. Informatics for unveiling hidden genome signatures. Genome Res **13:**693-702.
52. **Badger JH, Olsen GJ.** 1999. CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol **16:**512-524.
53. **Delcher AL, Harmon D, Kasif S, White O, Salzberg SL.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res **27:**4636-4641.
54. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:**955-964.
55. **Eddy SR.** 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics **3:**18.
56. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J Mol Biol **215:**403-410.
57. **Zemla A, Zhou CE, Slezak T, Kuczmarski T, Rama D, Torres C, Sawicka D, Barsky D.** 2005. AS2TS system for protein structure modeling and analysis. Nucleic Acids Res **33:**W111-115.
58. **Wheeler KE, Erickson BK, Mueller R, Singer SW, Verberkmoes NC, Hwang M, Thelen MP, Hettich RL.** 2012. Metal affinity enrichment increases the range and depth of proteome identification for extracellular microbial proteins. J Proteome Res **11:**861-870.

59. **McDonald WH, Ohi R, Miyamoto DT, Mitchison TJ, Yates JR.** 2002. Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. International Journal of Mass Spectrometry **219:**245-251.
60. **Eng JK, McCormack AL, Yates JR.** 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry **5:**976-989.
61. **Tabb DL, McDonald WH, Yates JR, 3rd.** 2002. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J Proteome Res **1:**21-26.
62. **Florens L, Carozza MJ, Swanson SK, Fournier M, Coleman MK, Workman JL, Washburn MP.** 2006. Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. Methods **40:**303-311.
63. **Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP.** 2006. Statistical Analysis of Membrane Proteome Expression Changes in Saccharomyces cerevisiae. Journal of Proteome Research **5:**2339-2347.
64. **Eisen MB, Spellman PT, Brown PO, Botstein D.** 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A **95:**14863-14868.
65. **Saldanha AJ.** 2004. Java Treeview--extensible visualization of microarray data. Bioinformatics **20:**3246-3248.
66. **Singer SW, Chan CS, Zemla A, VerBerkmoes NC, Hwang M, Hettich RL, Banfield JF, Thelen MP.** 2008. Characterization of cytochrome 579, an unusual cytochrome isolated from an iron-oxidizing microbial community. Appl Environ Microbiol **74:**4454-4462.
67. **Jeans C, Singer SW, Chan CS, Verberkmoes NC, Shah M, Hettich RL, Banfield JF, Thelen MP.** 2008. Cytochrome 572 is a conspicuous membrane protein with iron oxidation activity purified directly from a natural acidophilic microbial community. ISME J **2:**542-550.
68. **Ashida H, Danchin A, Yokota A.** 2005. Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulfur metabolism? Res Microbiol **156:**611-618.
69. **Aoshima M, Ishii M, Igarashi Y.** 2004. A novel enzyme, citryl-CoA lyase, catalysing the second step of the citrate cleavage reaction in Hydrogenobacter thermophilus TK-6. Mol Microbiol **52:**763-770.
70. **Aoshima M, Igarashi Y.** 2006. A novel oxalosuccinate-forming enzyme involved in the reductive carboxylation of 2-oxoglutarate in Hydrogenobacter thermophilus TK-6. Mol Microbiol **62:**748-759.
71. **Levican G, Ugalde JA, Ehrenfeld N, Maass A, Parada P.** 2008. Comparative genomic analysis of carbon and nitrogen assimilation mechanisms in three indigenous bioleaching bacteria: predictions and validations. BMC Genomics **9:**581.
72. **Jahn U, Huber H, Eisenreich W, Hugler M, Fuchs G.** 2007. Insights into the autotrophic CO2 fixation pathway of the archaeon Ignicoccus hospitalis: comprehensive analysis of the central carbon metabolism. J Bacteriol **189:**4108-4119.
73. **Huber H, Gallenberger M, Jahn U, Eylert E, Berg IA, Kockelkorn D, Eisenreich W, Fuchs G.** 2008. A dicarboxylate/4-hydroxybutyrate autotrophic carbon assimilation

cycle in the hyperthermophilic Archaeum Ignicoccus hospitalis. Proc Natl Acad Sci U S A **105:**7851-7856.
74. **Lengeler JW, Drews G, Schlegel HG.** 1999. Biology of the prokaryotes. Thieme ; Distributed in the USA by Blackwell Science, Stuttgart ; New York Malden, MA.
75. **Delgado MJ, Tresierra-Ayala A, Talbi C, Bedmar EJ.** 2006. Functional characterization of the Bradyrhizobium japonicum modA and modB genes involved in molybdenum transport. Microbiology **152:**199-207.
76. **Ninfa AJ, Jiang P.** 2005. PII signal transduction proteins: sensors of alpha-ketoglutarate that regulate nitrogen metabolism. Curr Opin Microbiol **8:**168-173.
77. **Conroy MJ, Durand A, Lupo D, Li XD, Bullough PA, Winkler FK, Merrick M.** 2007. The crystal structure of the Escherichia coli AmtB-GlnK complex reveals how GlnK regulates the ammonia channel. Proc Natl Acad Sci U S A **104:**1213-1218.
78. **Zhang Y, Wolfe DM, Pohlmann EL, Conrad MC, Roberts GP.** 2006. Effect of AmtB homologues on the post-translational regulation of nitrogenase activity in response to ammonium and energy signals in Rhodospirillum rubrum. Microbiology **152:**2075-2089.
79. **Simon J.** 2002. Enzymology and bioenergetics of respiratory nitrite ammonification. FEMS Microbiol Rev **26:**285-309.
80. **Curdt I, Singh BB, Jakoby M, Hachtel W, Bohme H.** 2000. Identification of amino acid residues of nitrite reductase from Anabaena sp. PCC 7120 involved in ferredoxin binding. Biochim Biophys Acta **1543:**60-68.
81. **Arp DJ, Stein LY.** 2003. Metabolism of inorganic N compounds by ammonia-oxidizing bacteria. Crit Rev Biochem Mol Biol **38:**471-495.
82. **Valdes J, Veloso F, Jedlicki E, Holmes D.** 2003. Metabolic reconstruction of sulfur assimilation in the extremophile Acidithiobacillus ferrooxidans based on genome analysis. BMC Genomics **4:**51.
83. **Murphy MJ, Siegel LM, Tove SR, Kamin H.** 1974. Siroheme: a new prosthetic group participating in six-electron reduction reactions catalyzed by both sulfite and nitrite reductases. Proc Natl Acad Sci U S A **71:**612-616.
84. **Rodionov DA, Mironov AA, Gelfand MS.** 2002. Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. Genome Res **12:**1507-1516.
85. **Raux E, Schubert HL, Warren MJ.** 2000. Biosynthesis of cobalamin (vitamin B12): a bacterial conundrum. Cell Mol Life Sci **57:**1880-1893.
86. **Kim W, Major TA, Whitman WB.** 2005. Role of the precorrin 6-X reductase gene in cobamide biosynthesis in Methanococcus maripaludis. Archaea **1:**375-384.
87. **Aguilar PS, de Mendoza D.** 2006. Control of fatty acid desaturation: a mechanism conserved from bacteria to humans. Mol Microbiol **62:**1507-1514.
88. **Masai E, Katayama Y, Fukuda M.** 2007. Genetic and biochemical investigations on bacterial catabolic pathways for lignin-derived aromatic compounds. Biosci Biotechnol Biochem **71:**1-15.
89. **Nikodem P, Hecht V, Schlomann M, Pieper DH.** 2003. New bacterial pathway for 4- and 5-chlorosalicylate degradation via 4-chlorocatechol and maleylacetate in Pseudomonas sp. strain MT1. J Bacteriol **185:**6790-6800.
90. **De Mot R.** 2007. Actinomycete-like proteasomes in a Gram-negative bacterium. Trends in Microbiology **15:**335-338.

91. **Szurmant H, Ordal GW.** 2004. Diversity in chemotaxis mechanisms among the bacteria and archaea. Microbiol Mol Biol Rev **68:**301-319.
92. **Stephens BB, Loar SN, Alexandre G.** 2006. Role of CheB and CheR in the complex chemotactic and aerotactic pathway of Azospirillum brasilense. J Bacteriol **188:**4759-4768.
93. **Wilson C, Dombroski AJ.** 1997. Region 1 of [sigma]70 is required for efficient isomerization and initiation of transcription by Escherichia coli RNA polymerase. Journal of Molecular Biology **267:**60-74.
94. **Umeno D, Tobias AV, Arnold FH.** 2005. Diversifying carotenoid biosynthetic pathways by directed evolution. Microbiol Mol Biol Rev **69:**51-78.
95. **Empadinhas N, da Costa MS.** 2006. Diversity and biosynthesis of compatible solutes in hyper/thermophiles. Int Microbiol **9:**199-206.
96. **Garcia-Estepa R, Argandona M, Reina-Bueno M, Capote N, Iglesias-Guerra F, Nieto JJ, Vargas C.** 2006. The ectD gene, which is involved in the synthesis of the compatible solute hydroxyectoine, is essential for thermoprotection of the halophilic bacterium Chromohalobacter salexigens. J Bacteriol **188:**3774-3784.
97. **NÌ Chadhain SM, J. K. Schaefer, S. Crane, G. J. Zylstra, Barkay T.** 2006. Analysis of mercuric reductase (merA) gene diversity in an anaerobic mercury-contaminated sediment enrichment. Environmental Microbiology **8:**1746-1752.
98. **Arcus VL, Rainey PB, Turner SJ.** 2005. The PIN-domain toxin-antitoxin array in mycobacteria. Trends Microbiol **13:**360-365.
99. **Robart AR, Zimmerly S.** 2005. Group II intron retroelements: function and diversity. Cytogenet Genome Res **110:**589-597.
100. **Toor N, Hausner G, Zimmerly S.** 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. RNA **7:**1142-1152.
101. **Marraffini LA, Sontheimer EJ.** 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science **322:**1843-1845.
102. **Sand W, Rohde K, Sobotke B, Zenneck C.** 1992. Evaluation of Leptospirillum-Ferrooxidans for Leaching. Applied and Environmental Microbiology **58:**85-92.
103. **Sone N, Sawa G, Sone T, Noguchi S.** 1995. Thermophilic bacilli have split cytochrome b genes for cytochrome b6 and subunit IV. First cloning of cytochrome b from a gram-positive bacterium (Bacillus stearothermophilus). J Biol Chem **270:**10612-10617.
104. **Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, Taylor MW, Horn M, Daims H, Bartol-Mavel D, Wincker P, Barbe V, Fonknechten N, Vallenet D, Segurens B, Schenowitz-Truong C, Medigue C, Collingro A, Snel B, Dutilh BE, Op den Camp HJ, van der Drift C, Cirpus I, van de Pas-Schoonen KT, Harhangi HR, van Niftrik L, Schmid M, Keltjens J, van de Vossenberg J, Kartal B, Meier H, Frishman D, Huynen MA, Mewes HW, Weissenbach J, Jetten MS, Wagner M, Le Paslier D.** 2006. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature **440:**790-794.
105. **van der Oost J, de Vos WM, Antranikian G.** 1996. Extremophiles. Trends in Biotechnology **14:**415-417.
106. **Sawers RG.** 2005. Formate and its role in hydrogen production in Escherichia coli. Biochem. Soc. Trans. **33:**42-46.

107. **Segal G, Feldman M, Zusman T.** 2005. The Icm/Dot type-IV secretion systems of Legionella pneumophila and Coxiella burnetii. FEMS Microbiol Rev **29:**65-81.
108. **Pukatzki S, Ma AT, Sturtevant D, Krastins B, Sarracino D, Nelson WC, Heidelberg JF, Mekalanos JJ.** 2006. Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system. Proc Natl Acad Sci U S A **103:**1528-1533.
109. **Baker BJ, Lutz MA, Dawson SC, Bond PL, Banfield JF.** 2004. Metabolically active eukaryotic communities in extremely acidic mine drainage. Appl Environ Microbiol **70:**6264-6271.
110. **Kuznetsova E, Proudfoot M, Sanders SA, Reinking J, Savchenko A, Arrowsmith CH, Edwards AM, Yakunin AF.** 2005. Enzyme genomics: Application of general enzymatic screens to discover new enzymes. FEMS Microbiol Rev **29:**263-279.
111. **Rawlings DE, Dew D, du Plessis C.** 2003. Biomineralization of metal-containing ores and concentrates. Trends Biotechnol **21:**38-44.
112. **Rawlings DE, Johnson DB.** 2007. The microbiology of biomining: development and optimization of mineral-oxidizing microbial consortia. Microbiology **153:**315-324.
113. **Druschel G, Baker B, Gihring T, Banfield J.** 2004. Acid mine drainage biogeochemistry at Iron Mountain, California. Geochemical Transactions **5:**13.
114. **Hippe H.** 2000. Leptospirillum gen. nov. (ex Markosyan 1972), nom. rev., including Leptospirillum ferrooxidans sp. nov. (ex Markosyan 1972), nom. rev. and Leptospirillum thermoferrooxidans sp. nov. (Golovacheva et al. 1992). International journal of systematic and evolutionary microbiology **50 Pt 2:**501-503.
115. **Diaby N, Dold B, Pfeifer HR, Holliger C, Johnson DB, Hallberg KB.** 2007. Microbial communities in a porphyry copper tailings impoundment and their impact on the geochemical dynamics of the mine waste. Environ Microbiol **9:**298-307.
116. **Garcia-Moyano A, Gonzalez-Toril E, Aguilera A, Amils R.** 2007. Prokaryotic community composition and ecology of floating macroscopic filaments from an extreme acidic environment, Rio Tinto (SW, Spain). Syst Appl Microbiol **30:**601-614.
117. **Galleguillos PA, Hallberg KB, Johnson DB.** 2009. Microbial Diversity and Genetic Response to Stress Conditions of Extremophilic Bacteria Isolated from the Escondida Copper Mine. Advanced Materials Research **71 - 73:**55-58.
118. **Parro V, Moreno-Paz M.** 2004. Nitrogen fixation in acidophile iron-oxidizing bacteria: the nif regulon of Leptospirillum ferrooxidans. Res Microbiol **155:**703-709.
119. **Mi S, Song J, Lin J, Che Y, Zheng H, Lin J.** 2011. Complete genome of Leptospirillum ferriphilum ML-04 provides insight into its physiology and environmental adaptation. J Microbiol **49:**890-901.
120. **Fujimura R, Sato Y, Nishizawa T, Oshima K, Kim SW, Hattori M, Kamijo T, Ohta H.** 2012. Complete genome sequence of Leptospirillum ferrooxidans strain C2-3, isolated from a fresh volcanic ash deposit on the island of Miyake, Japan. J Bacteriol **194:**4122-4123.
121. **Denef VJ, Banfield JF.** 2012. In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. Science **336:**462-466.
122. **Moreno-Paz M, Gomez MJ, Arcas A, Parro V.** 2010. Environmental transcriptome analysis reveals physiological differences between biofilm and planktonic modes of life of the iron oxidizing bacteria Leptospirillum spp. in their natural microbial community. BMC Genomics **11:**404.

123. **Hugenholtz P, Tyson GW, Blackall LL.** 2001. Design and Evaluation of 16S rRNA-Targeted Oligonucleotide Probes for Fluorescence In Situ Hybridization, p. 29-42, Gene Probes, vol. 179.
124. **Belnap CP, Pan C, VerBerkmoes NC, Power ME, Samatova NF, Carver RL, Hettich RL, Banfield JF.** 2010. Cultivation and quantitative proteomic analyses of acidophilic microbial communities. ISME J **4:**520-530.
125. **Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A.** 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res **37:**e123.
126. **Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF.** 2011. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol **12:**R44.
127. **Langmead B.** 2010. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics **Chapter 11:**Unit 11 17.
128. **Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, Forster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, Konig A, Liss T, Lussmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer KH.** 2004. ARB: a software environment for sequence data. Nucleic Acids Res **32:**1363-1371.
129. **Kelley LA, Sternberg MJ.** 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc **4:**363-371.
130. **Wass MN, Kelley LA, Sternberg MJ.** 2010. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res **38:**W469-473.
131. **Kosman DJ.** 2009. Multicopper oxidases: a workshop on copper coordination chemistry, electron transfer, and metallophysiology. J Biol Inorg Chem **15:**15-28.
132. **Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV.** 2011. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol **9:**467-477.
133. **Barkay T, Miller SM, Summers AO.** 2003. Bacterial mercury resistance from atoms to ecosystems. FEMS Microbiol Rev **27:**355-384.
134. **Vignais PM, Colbeau A.** 2004. Molecular biology of microbial hydrogenases. Curr Issues Mol Biol **6:**159-188.
135. **Justice NB, Pan C, Mueller R, Spaulding SE, Shah V, Sun CL, Yelton AP, Miller CS, Thomas BC, Shah M, VerBerkmoes N, Hettich R, Banfield JF.** 2012. Heterotrophic archaea contribute to carbon cycling in low-pH, suboxic biofilm communities. Appl Environ Microbiol **78:**8321-8330.
136. **Stackebrandt E, Ebers J.** 2006. Taxonomic parameters revisited: tarnished gold standards. Microbiology Today **33:**152-155.
137. **Ohmura N, Sasaki K, Matsumoto N, Saiki H.** 2002. Anaerobic respiration using Fe(3+), S(0), and H(2) in the chemolithoautotrophic bacterium Acidithiobacillus ferrooxidans. J Bacteriol **184:**2081-2087.
138. **Ma S, Banfield JF.** 2011. Micron-scale Fe2+/Fe3+, intermediate sulfurspecies and O2 gradients across the biofilm–solution–sediment interface control biofilm organization. Geochimica et Cosmochimica Acta **75:**3568–3580.

139. **Bonnefoy V, Holmes DS.** 2012. Genomic insights into microbial iron oxidation and iron uptake strategies in extremely acidic environments. Environ Microbiol **14:**1597-1611.
140. **Blake RC, 2nd, Griff MN.** 2012. In situ Spectroscopy on Intact Leptospirillum ferrooxidans Reveals that Reduced Cytochrome 579 is an Obligatory Intermediate in the Aerobic Iron Respiratory Chain. Front Microbiol **3:**136.
141. **Cardenas JP, Valdes J, Quatrini R, Duarte F, Holmes DS.** 2010. Lessons from the genomes of extremely acidophilic bacteria and archaea with special emphasis on bioleaching microorganisms. Appl Microbiol Biotechnol **88:**605-620.
142. **Gadd GM.** 2010. Metals, minerals and microbes: geomicrobiology and bioremediation. Microbiology **156:**609-643.
143. **Golyshina OV, Timmis KN.** 2005. Ferroplasma and relatives, recently discovered cell wall-lacking archaea making a living in extremely acid, heavy metal-rich environments. Environ Microbiol **7:**1277-1288.
144. **Baker BJ, Tyson GW, Goosherst L, Banfield JF.** 2009. Insights into the diversity of eukaryotes in acid mine drainage biofilm communities. Appl Environ Microbiol **75:**2192-2199.
145. **Amaral-Zettler LA, Zettler ER, Theroux SM, Palacios C, Aguilera A, Amils R.** 2011. Microbial community structure across the tree of life in the extreme Rio Tinto. ISME J **5:**42-50.
146. **Garcia-Moyano A, Gonzalez-Toril E, Aguilera A, Amils R.** 2012. Comparative microbial ecology study of the sediments and the water column of the Rio Tinto, an extreme acidic environment. FEMS Microbiol Ecol **81:**303-314.
147. **Bent SJ, Forney LJ.** 2008. The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. ISME J **2:**689-695.
148. **Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO.** 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res **41:**D590-596.
149. **Roberts A, Pimentel H, Trapnell C, Pachter L.** 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics **27:**2325-2329.
150. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26:**2460-2461.
151. **Pruesse E, Peplies J, Glockner FO.** 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics **28:**1823-1829.
152. **Price MN, Dehal PS, Arkin AP.** 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol **26:**1641-1650.
153. **Letunic I, Bork P.** 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23:**127-128.
154. **Team RDC.** 2008. R: A language and environment for statistical computing.
155. **Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO.** 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics **26:**1463-1464.
156. **Kindt R, Coe R.** 2005. Tree diversity analysis. A manual and software for common statistical methods for ecological and biodiversity studies. World Agroforestry Centre (ICRAF), Nairobi (Kenya).

157. **Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H.** 2013. vegan: Community Ecology Package. R package version 2.0-6.
158. **Hamady M, Lozupone C, Knight R.** 2010. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J **4:**17-27.
159. **Leinster T, Cobbold CA.** 2012. Measuring diversity: the importance of species similarity. Ecology **93:**477-489.
160. **Li W, Godzik A.** 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22:**1658-1659.
161. **Wang Q, Garrity GM, Tiedje JM, Cole JR.** 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol **73:**5261-5267.
162. **Vamosi SM, Heard SB, Vamosi JC, Webb CO.** 2009. Emerging patterns in the comparative analysis of phylogenetic community structure. Molecular ecology **18:**572-592.
163. **Amann R, Peplies J, Schüler D.** 2007. Diversity and Taxonomy of Magnetotactic Bacteria, p. 25-36, Magnetoreception and Magnetosomes in Bacteria. Springer Berlin Heidelberg.
164. **Komeili A.** 2012. Molecular mechanisms of compartmentalization and biomineralization in magnetotactic bacteria. FEMS Microbiol Rev **36:**232-255.
165. **Belnap CP, Pan C, Denef VJ, Samatova NF, Hettich RL, Banfield JF.** 2011. Quantitative proteomic analyses of the response of acidophilic microbial communities to different pH conditions. ISME J **5:**1152-1161.
166. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944-945.
167. **Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR.** 2003. Rfam: an RNA family database. Nucleic Acids Res **31:**439-441.
168. **Abreu-Goodger C, Merino E.** 2005. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. Nucleic Acids Res **33:**W690-692.
169. **Hofacker IL.** 2003. Vienna RNA secondary structure server. Nucleic Acids Res **31:**3429-3431.
170. **Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O.** 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res **36:**W465-469.
171. **Skennerton CT, Imelfort M, Tyson GW.** 2013. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res.
172. **Archer MJ, Lin B.** 2011. Development of a single-step subtraction method for eukaryotic 18S and 28S ribonucleic acids. Journal of biomedicine & biotechnology **2011:**910369.
173. **Kumar N, Creasy T, Sun Y, Flowers M, Tallon LJ, Dunning Hotopp JC.** 2012. Efficient subtraction of insect rRNA prior to transcriptome analysis of Wolbachia-Drosophila lateral gene transfer. BMC research notes **5:**230.
174. **Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, Levin JZ, Livny J, Earl AM, Gevers D, Ward DV, Nusbaum C, Birren BW, Gnirke A.** 2012. Efficient and

robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biol **13:**R23.
175. **Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, Zwick ME, Bergman NH.** 2009. Structure and complexity of a bacterial transcriptome. J Bacteriol **191:**3203-3211.
176. **Hewson I, Poretsky RS, Beinart RA, White AE, Shi T, Bench SR, Moisander PH, Paerl RW, Tripp HJ, Montoya JP, Moran MA, Zehr JP.** 2009. In situ transcriptomic analysis of the globally important keystone N2-fixing taxon Crocosphaera watsonii. ISME J **3:**618-631.
177. **Frias-Lopez J, Duran-Pinedo A.** 2012. Effect of periodontal pathogens on the metatranscriptome of a healthy multispecies biofilm model. J Bacteriol **194:**2082-2095.
178. **Hooper SD, Mavromatis K, Kyrpides NC.** 2009. Microbial co-habitation and lateral gene transfer: what transposases can tell us. Genome Biol **10:**R45.
179. **Bhaya D, Davison M, Barrangou R.** 2011. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annual review of genetics **45:**273-297.
180. **Saum SH, Muller V.** 2008. Regulation of osmoadaptation in the moderate halophile Halobacillus halophilus: chloride, glutamate and switching osmolyte strategies. Saline systems **4:**4.
181. **Dulebohn D, Choy J, Sundermeier T, Okan N, Karzai AW.** 2007. Trans-translation: the tmRNA-mediated surveillance mechanism for ribosome rescue, directed protein degradation, and nonstop mRNA decay. Biochemistry **46:**4681-4693.
182. **Esakova O, Krasilnikov AS.** 2010. Of proteins and RNA: the RNase P/MRP family. RNA **16:**1725-1747.
183. **Herskovits AA, Bochkareva ES, Bibi E.** 2000. New prospects in studying the bacterial signal recognition particle pathway. Mol Microbiol **38:**927-939.
184. **Wassarman KM.** 2007. 6S RNA: a small RNA regulator of transcription. Curr Opin Microbiol **10:**164-168.
185. **Peselis A, Serganov A.** 2012. Structural insights into ligand binding and gene expression control by an adenosylcobalamin riboswitch. Nature structural & molecular biology **19:**1182-1184.
186. **Breaker RR.** 2011. Prospects for riboswitch discovery and analysis. Molecular cell **43:**867-879.
187. **Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF.** 2007. Genome dynamics in a natural archaeal population. Proc Natl Acad Sci U S A **104:**1883-1888.

APPENDIX 1. Supplementary Materials Information

Chapter 1:

- Supplementary figures S1, S2, S3, S4, and S5
- Supplementary Table S1
- Supplementary Table S2
- Supplementary Table S3
- Supplementary Table S4
- Supplementary Table S5
- References

Chapter 2:

- Supplementary figures S1, S2, S3, S4, and S5
- Supplementary Table S1
- Supplementary Table S2
- References
- Fluorescence *in situ* hybridization protocol

Chapter 3:

- Supplementary figure S1
- Supplementary Table S1
- Supplementary Table S2
- SSU rRNA gene sequences assembled by Cufflinks (fasta format)

Chapter 4:

- Supplementary figures S1 and S2
- Supplementary Table S1
- Non-coding RNA sequences predicted by Rfam and RibEx (fasta format)
- Reads containing CRISPR loci (identified via CRASS assembler)

APPENDIX 2. List of publications

- **Aliaga Goltsman DS**, Dasari M, Thomas BC, Shah MB, VerBerkmoes NC, Hettich RL, and Banfield JF. New group in the *Leptospirillum* clade: cultivation-independent community genomics, proteomics and transcriptomics of the new species *Leptospirillum* group IV UBA BS. Appl Environ Microbiol. 2013. In press.

- Wheeler K, Zemla A, Jiao Y, **Aliaga Goltsman DS**, Singer S, Banfield J and Thelen MP. Functional insights from computational modeling of orphan proteins expressed in a microbial community. J Proteomics Bioinform. 2010 Sep; 3: 266-274.

- **Aliaga Goltsman DS**, Denef VJ, Singer SW, VerBerkmoes NC, Lefsrud M, Mueller RS, Dick GJ, Sun CL, Wheeler KE, Zemla A, Baker BJ, Hauser L, Land M, Shah MB, Thelen MP, Hettich RL, and Banfield JF. 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. Jul;75(13):4599-615.

- Simmons SL, Dibartolo G, Denef VJ, **Aliaga Goltsman DS**, Thelen MP, Banfield JF. Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. PLoS Biol. 2008 Jul 22; 6(7): e177

- Lo I, Denef VJ, Verberkmoes NC, Shah MB, **Goltsman D**, DiBartolo G, Tyson GW, Allen EE, Ram RJ, Detter JC, Richardson P, Thelen MP, Hettich RL, Banfield JF. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. Nature. 2007 Mar 29; 446(7135): 537-41.

- Veloso F, Riadi G, **Aliaga D**, Lieph R and Holmes DS. Large Scale, Multi-Genome Analysis of Alternate Open Reading Frames in Bacteria and Archaea. OMICS. Mar 2005, Vol. 9, No. 1: 91-105.