Differential gene expression and the importance of regulatory ncRNAs in acidophilic microorganisms

Daniela S. Aliaga Goltsman[1,3#], Loren Hauser[2,4], Mauna Dasari[1,5], Brian C. Thomas[1,6] and Jillian F. Banfield[1#]

[1]University of California Berkeley, Berkeley CA

[2]Oak Ridge National Laboratory, Oak Ridge TN

[3]Current address: Stanford University School of Medicine, Stanford CA

[4]Current address: Digital Infuzion, Inc., Gaithersburg, MD

[5]Current address: University of Notre Dame, Notre Dame, IN

[6]Current address: Metagenomi, Berkeley CA

#Corresponding authors:

Daniela S. Aliaga Goltsman

Stanford University

School of Medicine

Department of Infectious Diseases

3801 Miranda Ave., Bldg. 101, Rm B4-185

Palo Alto, CA 94304

dgolts@stanford.edu

Jillian Banfield

University of California Berkeley

Earth and Planetary Sciences Department

Energy Biosciences Building

2151 Berkeley Way

Berkeley, CA 94720

jbanfield@berkeley.edu

Submitted as a Research Article to the MBio Journal

Running title: “Metatranscriptomic analysis in acidophilic communities”

Abstract word count: 246 and 100. Manuscript word count: 5075.

ABSTRACT

Gene expression profiles provide insight into how microorganisms respond to changing environmental conditions. However, few studies have integrated expression profile analyses of both coding genes and non-coding RNAs (ncRNAs) to characterize the functional activity of microbial community members. Here, we defined gene expression profiles from environmental and laboratory-grown acidophilic biofilms using RNASeq. In total, 15.8 million Illumina reads were mapped to the genomes of 26 acidophilic microorganisms and nine viruses reconstructed from the Richmond Mine at Iron Mountain, California. More than 99% of the genome was transcribed in three *Leptospirillum* species, and > 80% in the archaea G-plasma and *Ferroplasma* Type II. High gene expression by G-plasma and the *Leptospirillum* Group II UBA strain correlated with extremely acidic conditions, whereas high transcriptional expression of *Leptospirillum* Group III and *Leptospirillum* Group II 5way-CG strain occurred under higher pH and lower temperature. While expression of CRISPR Cas genes occurs on the sense strand, expression of the CRISPR loci occurs on the antisense strand in the *Leptospirilli*. A novel riboswitch associated with the biosynthetic pathway for the osmolyte ectoine was upregulated when each specific *Leptospirillum* Group II strain was growing under the conditions most favorable for it. Newly described ncRNAs associated with CO dehydrogenase (CODH) suggest regulation of expression of CODH as a CO sensor in mature biofilms in the *Leptospirilli*. Results reveal the ways in which environmental conditions shape transcriptional profiles of organisms growing in acidophilic microbial communities and highlight the significance of ncRNAs in regulating gene expression.

IMPORTANCE

Microorganisms play important roles in environmental acidification and in metal-recovery based bioleaching processes. Therefore, characterizing how actively growing microbial communities respond to different environments is key to understanding their role in those processes. Microorganisms express their genes, both coding and non-coding, differently depending on environmental factors, thus evaluating community expression profiles inform about the ecology of actively growing microorganisms. Here we used community transcriptomic analyses to characterize gene expression profiles from biofilm communities growing under extremely acidic conditions. Results expand our knowledge of how acidophilic microorganisms respond to changes in their environment and provide insight into possible gene regulation mechanisms.

INTRODUCTION

Extremely acidic environments are usually dominated by relatively few taxa 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 key to understanding the physiology and ecology of microorganisms in acidic environments. The Richmond Mine at Iron Mountain, California, is a well-studied acid mine drainage (AMD) system: for example, deep sequencing of many Richmond Mine biofilms has enabled reconstruction of the genomes of many Bacteria, Archaea, viruses, plasmids, and fungus (e.g. (1-4)). Genome reconstruction analyses, along with high-throughput transcriptomic sequencing and mass spectrometry-based proteomics have the potential to provide information about the community composition and activity in natural ecosystems. Community transcriptomic analyses have been used to describe important metabolic

processes in natural communities such as nitrogen metabolism in marine environments (5), the flow of carbon between organisms in photosynthetic microbial mats (6), and associations of microbial communities with the biogeochemistry of permafrost thawing environments (7). In acidic environments, mass spectrometry-based community proteomics measurements of the dominant member of Richmond Mine AMD biofilms indicated that biofilm maturation affects *Leptospirillum* group II protein expression profiles, with the most protein expression variation found as biofilms transition from early and mid-developmental to the most mature growth stage (8). Similarly, microarray-based community transcriptomics indicated that expression profiles of genes involved in biofilm formation, chemotaxis, motility and quorum-sensing in *Leptospirillum ferrooxidans* were up-regulated when living as part of biofilms vs. in the planktonic fraction of the Rio Tinto AMD system (9). Hua *et al*. recovered the genome of *Ferrovum* spp. from an acid mine drainage system and community gene expression profiles indicated active CO_2 fixation and sulfate reduction pathways (10). In addition, metatranscriptomics studies have highlighted the importance of non-coding RNA (ncRNA) expression in marine systems (11). However, studies of the transcriptional profiles of coding and ncRNAs in acidophilic communities relevant for AMD generation and bioleaching-based metal recovery have not yet been reported.

Many ncRNAs have been identified in Bacteria and Archaea, such as riboswitches, ribozymes, and other regulatory ncRNAs, which play important roles in regulating gene expression (reviewed in (12)). Riboswitches are regions within a messenger RNA (generally located in the 5' untranslated region - UTR) containing ligand-binding sensors that regulate downstream coding sequences (reviewed in (13)). Other regulatory ncRNAs with known function include the tmRNA, which helps unlock stalled ribosomes, and the 6S ncRNA, which acts as mimic of a promoter (reviewed in (14)). In Eukaryotes, a common mechanism of gene regulation involves transcription of long 5' UTRs with secondary structures that can influence

transcription of downstream genes, or containing short upstream open reading frames (uORFs) that attenuate translation of the downstream protein (reviewed in (15)). In addition, high-throughput transcriptomics analyses have enabled identification of ncRNAs that regulate antibiotic resistance in bacteria via 5' UTR transcription and uORF attenuation (reviewed in (16)).

Here we report the analyses of transcriptional profiles of non-ribosomal RNA from biofilms collected from the Richmond Mine at Iron Mountain, California, and from laboratory-grown biofilm communities. Gene expression profiles were evaluated for the whole acidophilic community including Bacteria, Archaea, and viruses, and novel ncRNAs were discovered for many community members. The results greatly expand our understanding of the responses of acidophilic organisms to changes in their environment and provide insight into possible gene regulation mechanisms by ncRNAs in the *Leptospirilli.*

(Part of this article was submitted to an online preprint archive (17))

RESULTS

Transcriptomics reads map across the whole genome of AMD community members.

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 (Table 1). Despite the low efficiency of the rRNA-depletion protocol used, deep sequencing allowed us to detect transcripts from 37 near-complete and partial genomes of Bacteria, Archaea, Fungi, plasmids and viruses (Supplementary Table S1). Transcriptomic data indicate that *Leptospirillum* group II C75 and UBA genotypes dominate early and mid-developmental stage environmental biofilms characterized by extremely low pH and high temperature, whereas *Leptospirillum* group II 5way-CG is the most abundant genotype in the late

94 developmental-stage environmental biofilm (Figure 1A). *Leptospirillum* group III bacteria
95 dominate bioreactors and the A-drift environmental sample. *Leptospirillum ferrooxidans* and
96 *Acidithiobacillus caldus* were detected at very low abundance, as were other organisms and
97 viruses, some of which represent less than 0.001% of the community transcriptome
98 (Supplementary Table S1). Low Archaeal gene expression and high *Acidithiobacillus* sp.
99 expression have been documented on acid mine drainage biofilms growing at higher pH (18).
100 Transcriptomic reads spanned > 95% of the genome of *Leptospirillum* group II and group
101 III, the G-plasma archaeon and one viral genome (Figure 1B). In addition, up to 80% of the
102 genome of *Ferroplasma* Type II, and over 60% of the genomes of *Ferroplasma* Type I and three
103 viruses were identified as transcribed in some samples (Figure 1B). Likely, higher sequencing
104 coverage and/or higher rRNA-removal efficiency would support detecting whole genome
105 transcription in low-abundance community members. The results suggest that for dominant and
106 low-activity organisms and viruses the whole genome is transcribed at some level. This trend has
107 been observed previously in transcriptomics analyses of *Bacillus anthracis* isolates (19).
108
109 **Gene expression profiles from AMD taxa correlate with the environment pH.**
110 Mapping reads to the predicted genes of AMD organisms, plasmids, and viruses were
111 assembled to yield over 26,300 transcripts (Supplementary Materials – R data). Differential
112 expression analysis identified >4,800 significantly differentially expressed genes with a false
113 discovery rate of 0.05 when evaluating environmental vs. bioreactor-grown biofilms
114 (Supplementary Figure S1A). Non-metric multidimensional scaling (NMDS) ordination on the
115 ~4,800 differentially expressed genes indicates two main axes responsible for most of the sample
116 variation (Figure 2, and Supplementary Figure S1B). The NMDS samples plot shows that axis 1
117 separates primarily based on environment type, while axis 2 separates based on pH (Figure 2A).

Gene expression profiles from *Leptospirillum* group II UBA and C75 genotypes appear correlated with negative NMDS2 scores (low pH environments), while expression profiles from *Leptospirillum* group II 5way-CG genotype and group III appear correlated with positive NMDS2 scores (higher pH environments) (Figure 2B-D). Gene expression profiles from Archaea and viruses generally correlated with positive NMDS1 scores, that is, from environmental rather than bioreactor-grown biofilms. Overall, the NMDS genes plots show a gradient of genes from bacteria living at very low pH and high temperature (Figure 2B, bottom right quadrant), transitioning towards genes from Archaea, viruses and other low-abundance organisms (Figure 2C-D), and ending on genes from bacteria living in bioreactors at higher pH and lower temperature (Figure 2B, top left quadrant). Results indicate that environmental gradients directly affect the transcriptional profiles of the community members living along them. Gradients of pH, temperature, and dissolved oxygen, among others, have been suggested as factors influencing community assembly in acidic environments (20, 21).

Environment-specific transcriptional profiles are also evident from hierarchical clustering of transcribed genes with a log2-fold change >5, which separates samples into environmental and bioreactor biofilms, and environmental samples appear to also cluster by developmental stage (Supplementary Figures S1C and S1D). For example, *Leptospirillum* group II 5way-CG and *Leptospirillum* group III genes are overrepresented in the bioreactors and in the A-drift environmental sample (Supplementary Figures S1D). The A-drift biofilm was collected from a very oxidized pool, at higher pH and lower temperature than other environmental samples (see Table 1), conditions that appear to favor growth of these organisms in the laboratory bioreactors. Genes from *Leptospirillum* group II UBA are over-represented in early to mid-growth-stage environmental biofilms relative to bioreactor-grown biofilms (Supplementary Figure S1D). This genotypic group is also well represented in biofilms collected from the C-drift locations, many

of which have reported very low pH and high temperature environments (22-24). Our results support previous proteomic-based studies that suggested distinct ecological adaptation of two *Leptospirillum* group II strains (22). *Leptospirillum* group II-associated AMDV1 phage genes, as well as genes from the AMDV3 virus are overrepresented in the environmental biofilms, while unassigned-viral genes (AMDVIR, likely *Leptospirillum*-associated phage) are overrepresented in bioreactors (Supplementary Figure S1C and Figure 2D). Results indicate that phage/virus activity correlates with their host gene expression.

When looking at the transcriptional profiles of the most abundant community members, hierarchical clustering of expressed genes indicates that genes involved in energy production and conversion, carbon fixation, fatty acid metabolism, transcription and translation factors, and ribosomal proteins from *Leptospirillum* group II UBA and C75 are highly expressed in early to mid-growth-stage environmental biofilms (Figure 3A-B, 3E). These findings suggest rapid growth during early and mid-successional stages. Genes generally overrepresented in bioreactor samples include those 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 from *Leptospirillum* group II 5way-CG and group III (Figure 3C-D and 3E). Bioreactors usually have higher pH and lower temperature than most environmental biofilms from the Richmond Mine (Table 1). These conditions might slow down the activity of *Leptospirillum* group II UBA and C75, and of G-plasma, while enriching for growth and transcriptional activity of *Leptospirillum* group III and group II 5way-CG.

Regulatory motifs, mobile elements and expression of ncRNAs of known function.

Inspection of the upstream regions of transcribed coding genes, tRNAs and non-coding RNAs predicted 207 RpoD promoters in the *Leptospirillum* group II UBA genotype (Supplementary Table S2): 11.5% of them were associated to coding genes or operons transcribed in at least 10 of the 13 biofilm samples. Promoter length varies between 8 and 33 bp and a sequence logo (25) representation of the overall promoter structure indicates that most promoters contain a -10 and -35 motif (Figure 4A).

Transposases in *Leptospirillum* group II UBA and group III are among the most highly expressed genes (Supplementary Materials – R data). In addition, highly expressed multi-copy transposases are enriched in environmental than bioreactor samples (Supplementary Figure S2). Highly abundant transposase expression was reported in community proteomic analyses in acid mine drainage biofilms (26), and in community transcriptomic analyses (27, 28). It is possible that the movement of mobile elements is very important in natural acidophilic biofilms.

A few of the well-characterized ncRNAs in the Rfam database were identified in many genomes of acidophilic microorganisms, some of which are among the most highly expressed genes (Supplementary Table S3 and Figure 4). SsrA (aka transfer-messenger RNA or tmRNA) was detected in the *Leptospirilli* (Figure 4B-C), as well as in G-plasma, a plasmid, and in *Acidithiobacillus caldus,* the later representing only 0.003% of the community transcriptome in the sample in which it is most abundant (Adrift GS0, Supplementary Table S1). The RNAse P ncRNA was the most highly expressed ncRNA in the transcriptomic datasets, and was identified in most AMD organisms, including the fungus, the very low abundance *A. caldus* and *L. ferrooxidans* bacteria, and the archaeon C-plasma (Supplementary Table S3). The signal recognition particle (SRP) RNA was also highly expressed in the *Leptospirilli* (Figure 4B-C), *Ferroplasma* Type I and Type II, the fungus, and in G-, A-, E- and C-plasma. The 6S RNA (SsrS) was identified in the *Leptospirilli* (Figure 4B-C), the *Sulfobacillus*, and *A. caldus*. The cobalamin

189 riboswitch was only identified in *Leptospirillum* group III and contains the expected conserved
190 domains within the core region as described in other organisms (29) (Figure 4C). The cobalamin
191 riboswitch has not yet been described in the genomes of any other member of the *Nitrospira*
192 Phylum.

193 Among the ncRNAs associated to mobile elements detected in the transcriptomes are the
194 Group II introns of *Leptospirillum* groups II and III, *Ferroplasma* Type II, and the
195 Actinobacterial Bin 1 (Supplementary Table S3). In addition, the 5' UTR region of several HNH-
196 endonucleases in *Leptospirillum* groups II, III and IV, and *Ferroplasma* Type II contain the non-
197 coding HEARO RNA, which, along with its associated endonuclease, constitute another type of
198 mobile genetic element (30). The hgcC RNA, a non-coding RNA of unknown function, was
199 detected in *Leptospirillum* group II UBA and group II 5way-CG, and in two viruses
200 (Supplementary Table S3). Previously the hgcC RNA was reported only in Archaea (31),
201 although it is also found in a few viruses in the Rfam database.

202 The small nucleolar RNA (snoRNA), a eukaryotic small RNA with few known homologs
203 in Archaea (32) was detected in most Richmond Mine Archaea A-, C- and G-plasma,
204 *Ferroplasma* Type I and Type II, and the fungus; while the crcB RNA (a fluoride riboswitch,
205 (33)) was detected in the *Leptospirilli*, G-plasma, and in *Ferroplasma* Type II (Supplementary
206 Table S3).

207 The TPP riboswitch binds thiamine pyrophosphate to regulate expression of thiamine
208 biosynthesis and other related genes. These riboswitches have been identified upstream of multi-
209 transmembrane hypothetical proteins in the genomes of *Thermoplasma acidophilum,*
210 *Thermoplasma volcanicum, and Ferroplasma acidarmanus* (34). A-, G-, and E-plasma
211 (members of the *Thermoplasmatales*), I-plasma and *Ferroplasma* Type II express the TPP

212 riboswitch and reads within it have paired-reads to a downstream putative transporter likely
213 regulating its expression (Supplementary Table S3).
214
215 **CRISPR Cas genes and loci expression occurs on both strands of DNA.**
216 Transcripts from CRISPR-associated Cas genes were detected in all samples for
217 *Leptospirillum* groups II and III, the archaeon *Ferroplasma* Type I, and Archaea A- and G-
218 plasma (Supplementary Table S4). Cas genes from *Leptospirillum* group II UBA and G-plasma
219 showed higher expression in bioreactor samples, while those from *Leptospirillum* group III and
220 group II 5way-CG showed higher expression in environmental biofilms (Figure 3E).
221 When surveying the transcriptomic reads using the CRASS software, transcripts
222 containing CRISPR repeats from eight datasets were detected in *Leptospirillum* groups II and III,
223 G-plasma, *Ferroplasma* Type I and Type II, the Actinobacterial bin, and plasmids
224 (Supplementary Tables S4 and S5). Some additional CRISPR transcripts could not be assigned
225 to an AMD genome based on known repeat sequences. The highest diversity of CRISPR repeats
226 expressed was observed in the late growth-stage biofilm (4-way_GS2; Supplementary Table S4).
227 This finding probably reflects the higher richness of late growth-stage biofilms as well as the
228 activity of multiple closely related strains with slightly different CRISPR loci (and thus different
229 phage/viral susceptibility). For *Leptospirillum* group II and G-plasma, the largest number of
230 distinct spacer transcripts was detected in this same biofilm, consistent with a higher diversity of
231 strains in the sample.
232 Transcriptomic reads from precursor CRISPR RNA (pre-crRNAs, as reviewed in (35)) aligned
233 mostly at the trailer end (containing older CRISPR spacers) of the composite CRISPR loci from
234 the assembled genome of *Leptospirillum* group II UBA (Figure 5A), and no repeats and spacer-
235 carrying reads aligned to the CRISPR loci from the assembled *Leptospirillum* group III genome.

These observations are likely due to the composite assemblies not capturing the dynamics of acquired spacers from closely related strains. For example, 407 of the 688 different spacer sequences recovered from *Leptospirillum* group II transcripts, and 5 of the 245 spacers detected in the transcriptome in *Leptospirillum* group III (Supplementary Table S5) had been recovered in a previous analysis of Richmond Mine CRISPR systems (36). Notably, while transcription of Cas genes occurs on the sense strand in which the genes were predicted, transcription of the CRISPR loci (repeat/spacer region) in *Leptospirillum* group II UBA occurs on the antisense strand (Figure 5A). Moreover, 4 of the 5 *Leptospirillum* group III CRISPR spacers and 404 of the 407 spacers in *Leptospirillum* group II UBA recovered here via strand-specific transcriptomics represent the reverse complement of the sequences previously reported in (36) (Supplementary Table S5).

Expression of novel ncRNA.

We observed transcription of long leader sequences upstream of many genes and operons in *Leptospirillum* groups II and III. One example is a transcribed 370 bp 5' UTR of the ectoine biosynthesis operon in both *Leptospirillum* group II UBA and 5way-CG strains (Figure 5B-C). Ectoine is a compatible solute either synthesized or transported from the environment by many organisms during osmotic stress (37, 38). The length of the ncRNA upstream of the ectoine operon, the presence of paired reads from the ncRNA to 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 anti-terminator (-7.03 Kcal/mol) suggest it is likely a riboswitch. The expression of the putative riboswitch in *Leptospirillum* group II UBA in two environmental biofilms is generally higher than that of the operon (Figure 5B, top panel, green and red curves). However, the transcript levels of the putative riboswitch in the bioreactor samples are much lower

than those of the operon (Figure 5B, top panel, blue and magenta curves). These results suggest that the riboswitch may inhibit transcription of the operon in the environmental samples (preferred conditions for *Leptospirillum* group II UBA growth, see Figure 1) and enhance the ectoine operon transcription in bioreactor samples (where growth conditions are more stressful). The expression of the putative riboswitch in *Leptospirillum* group II 5way-CG shows a slight opposite trend than in *Leptospirillum* group II UBA (Figure 5B, bottom panel). The riboswitches of both bacteria share 80% identity at the nucleotide level and their predicted secondary structures look very different (Figure 5C).

Novel carbon monoxide dehydrogenase-associated ncRNAs (referred to here as CODH-ncRNA) were identified in long transcribed leader sequences in *Leptospirillum* group III and in *Leptospirillum* groups II UBA and 5way-CG genotypes (Figure 6 and Supplementary Figure S3). The CODH-ncRNA is expressed at similar abundance and contains paired-end reads that map to its downstream CO-dehydrogenase beta subunit gene (the only subunit identified in the *Leptospirilli*). Although terminators could not be predicted, the dip in the transcriptional expression of the CODH-ncRNAs before the ribosome binding site (RBS) of their associated CO-dehydrogenase (CODH) genes suggest that the CODH-ncRNA may be regulating CODH expression (Figure 6 and Supplementary Figure S3). Although the Archaeon *Ferroplasma* Type II has multiple copies of the full CODH operon and all subunits showed transcriptional expression, a CODH-ncRNA was not identified upstream of the operons (data not shown). Two of the three copies of the CODH-ncRNA in *Leptospirillum* group III and both copies in *Leptospirillum* group II are highly expressed in the late developmental-stage biofilms (Figure 6, left panel; Supplementary Figure S3A-D).

We observed expression of many long leader sequences upstream of other important coding genes and operons in *Leptospirillum* groups II and III. For example, the transcribed 5'

UTR of the pyruvate ferredoxin oxidoreductase (PFOR) operon, cytochrome 572 and cytochrome 579 genes contain putative promoters and untranslated ORFs (uORFs) (Figure 6 and Supplementary Figure S3E-I). Other examples include expression of the 5' UTR of the Uroporphyrin-III C-methyltransferase gene and of the Ribonuclease III gene in *Leptospirillum* group II UBA, with potential folded ncRNAs and predicted promoters (Supplementary Figure S3J-K).

DISCUSSION

Community expression profiles inform about the ecology and physiology of actively growing organisms in their natural environment. Richmond Mine AMD biofilms develop in extremely acidic conditions and their characterization is vital to understanding the process of water acidification. Bioreactor conditions generally resemble environmental conditions from downstream AMD sites at higher pH and lower temperatures, as well as those from metal-release bioleaching systems. Therefore, analyses of environmental and laboratory-grown acidophilic biofilms improve our understanding of the dynamics occurring in acid drainage and bioleaching ecosystems.

Acid mine drain communities have been extensively studied for community membership, and proteomic and transcriptomic studies have been used to infer ecological roles of community members (reviewed in (20)**)**. However, no reports have analyzed the transcriptional profiles of both coding genes and non-coding RNAs (ncRNAs) in acidophilic communities. Here, metatranscriptomic analysis of natural AMD biofilms and of acidophilic biofilms growing in laboratory bioreactors identified patterns in gene expression that correlate with organismal environmental preferences and reveal the potential roles of known and novel ncRNAs.

Differential gene expression analyses indicated that preferred growth conditions for *Leptospirillum* group II UBA and C75 genotypes, and the Archaeon G-plasma are environments at very low pH and high temperatures, while *Leptospirillum* group II 5way-CG and group III prefer higher pH and lower temperature conditions. AMD organisms appear to show base-level activity by expressing the whole genome, and *Leptospirillum*-associated viruses follow transcriptional expression profiles similar to that of their host. In addition, gradients of pH and temperature evident from differentially expressed genes, high expression of transposase genes, and the high diversity of CRISPR loci spacers suggest that AMD communities are highly dynamic ecosystems.

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 (39)). Differential expression of CRISPR Cas genes in environmental vs. bioreactor biofilms point to some level of regulation, where Cas genes are up-regulated in less optimal environments for the organisms carrying them. *Leptospirillum* Group II Cas proteins, at the time considered hypothetical, were highly abundant in the first microbial community proteomic analysis (26) and expression of Cas genes was observed previously in *Leptospirillum* group III (18). When examining the strand-specific nature of the transcriptome, we observed that CRISPR loci (repeats and spacers) transcription occurs on the antisense strand while Cas genes are expressed on the sense strand in the *Leptospirilli*. Co-expression of sense and antisense mRNAs in plants was shown more effective at offering viral immunity than only sense or antisense expression of the relevant genes (40). Therefore, Cas genes and CRISPR loci expression on different strands may be used as a mechanism for effective viral immunity in the *Leptospirilli*. Although, differential expression of Cas genes and CRISPR loci on opposite strands has not yet been reported, bidirectionality of CRISPR loci transcription was observed in *Sulfolobus* sp.,

where the authors suggest the potential need for bidirectionality to neutralize the leader spacer RNAs in the absence of invading elements (41).

RNASeq analyses enable identification of regulatory ncRNAs, leader sequences and transcription start sites (TSS) for coding genes and operons. We confirmed expression of ncRNAs of known function in many high and low abundance organisms, including the RNAse P, SsrA and 6S ncRNAs previously predicted *in silico* in *Acidithiobacillus caldus* (42). In addition, some ncRNAs of unknown function and 5' UTRs show expression levels similar to their neighboring genes and predicted operons, suggesting that expression of upstream non-coding regions might be regulating their neighboring genes. For example, we identified a novel riboswitch associated to the ectoine biosynthetic pathway in *Leptospirillum* group II. The riboswitch appears to inhibit transcription of the ectoine operon in favorable growth conditions while enhancing transcription during stress, suggesting that *Leptospirillum* group II synthesis of compatible solutes to tolerate unfavorable conditions is regulated by the ectoine riboswitch.

Other novel regulatory ncRNAs include a carbon monoxide dehydrogenase-associated ncRNA (CODH-ncRNA) in *Leptospirillum* group II and III. CO dehydrogenase, a nickel iron-sulfur protein, is usually part of a multi-protein complex which metabolizes carbon monoxide (CO) when the cells sense CO in the environment, as well as being used as electron and carbon sources by some bacteria (reviewed by (43)**)**. The CODH beta-subunit, the only subunit identified in *Leptospirillum* groups II and III, has been reported as a single active subunit in metal-reducing *Geobacter* bacteria and in the genomes of two *Chlorobium* phototrophic bacteria (44). It is highly sensitive to oxygen, which might explain the high expression of both CODH and CODH-ncRNA transcripts in late developmental stage biofilms. It is possible that the CODH-ncRNA up-regulates expression of the CO dehydrogenase gene as biofilms mature.

354 Long transcribed 5'UTRs with putative upstream ORFs (uORFs) preceding energy-
355 generation genes (e.g. PFOR and cytochrome c) likely regulate transcription and/or translation
356 of their downstream genes and operons (Figure 6). Analysis of transcription start sites (TSS) in
357 *Clostridium phytofermentans* have shown that the TSS of one *pfor* gene occurs 78 bp upstream
358 of its start codon (45), a length similar to the putative TSS of the *pfor* operons in *Leptospirillum*
359 sp. (Figure 6), while transcripts with long 5' UTRs encoding putative ncRNAs and uORFs have
360 been reported in *Haloferax volcanii* (46) and in *Mycobacterium* sp. (47). Overall, our results
361 point to significant roles of regulatory ncRNAs in acidophilic communities and suggest that gene
362 expression profiles appear to be a factor in ecological diversification.
363
364 **METHODS**
365 Eight biofilms at different developmental stages (growth-stage (GS) 0, 0.5, 1, and 2) were
366 collected from the A-, C-, AB-drift, and 4way locations within the Richmond Mine at Iron
367 Mountain Mines, California (40°40′ 38.42″ N and 122″ 31′ 19.90″ W, elevation of approx.
368 900 m) (Table 1). AMD biofilm developmental stages were estimated visually based on thickness
369 (23). In addition, biofilms were grown in the dark, at pH 1 and 37 °C in laboratory bioreactors,
370 as described in (48) using inocula from within the A-drift location and mine drainage outflow
371 (Table 1). Biofilms were harvested when reaching growth-stage 0, 0.5, and 1, snap-frozen in
372 liquid nitrogen upon collection/harvest and stored at -80 °C.
373 Total RNA was extracted from frozen samples using two acid phenol-chloroform-
374 isoamyl alcohol extractions and immediately purified using the RNEasy MinElute kit (Qiagen).
375 Ribosomal RNA (rRNA) subtraction on 8 of the 13 biofilm samples was done using the
376 MicrobExpress kit (Ambion). Good quality RNA (RIN > 7, assessed by a Bioanalyzer 2100
377 (Agilent Technologies)) from total RNA and rRNA-subtracted RNA were converted to cDNA as

described by (49) in order to keep the strand-specificity of the transcriptome. Briefly, the technique involves adding deoxy-UTP in place of deoxy-TTP during synthesis of the second strand of cDNA. After Illumina library preparation, the second strand of cDNA is selectively digested allowing for sequencing of all molecules in the same direction, and sequences maintain the strand-specificity of the original RNA molecules (49). 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, digestion of the dUTP-containing strand, and sequencing. Five samples were sequenced using the GAIIx platform (75 bp, single end reads), while eight samples were sequenced using the HiSeq 2500 platform (100 bp, paired-end reads) (Table 1).

Low-quality bases were trimmed from the sequencing reads using the fastx_trimmer script (http://hannonlab.cshl.edu/fastx_toolkit/) and the sickle trimmer script with default parameters (https://github.com/najoshi/sickle) and reads < 40 bp in length were discarded. Trimmed reads were mapped to the SSU and LSU rRNA gene Silva database SSURef_102 (50) using Bowtie2 (51) with default parameters to separate ribosomal from non-ribosomal reads.

Non-ribosomal (non-rRNA) reads were mapped using Bowtie2 with default parameters to the available genomes of acidophilic bacteria: *Leptospirillum ferrooxidans* C2-3, *Leptospirillum* group II UBA type (*L. rubarum*), group II '5way-CG' type, group II 'C75' type, group III (*L. ferrodiazotrophum*), group IV UBA BS, *Acidithiobacillus caldus* ATCC 51756, a *Sulfobacillus* bin, and two Actinobacterial bins (2, 3, 22, 52-56); archaea: A, C, D, E, G, and I-plasma, ARMAN-1, -2, -4, and -5, and *Ferroplasma* Type I and Type II (2, 4, 57-59); plasmids (53), a fungus (1); and nine viruses/phage (60). Mapped reads were then assembled into transcript fragments using the Cufflinks pipeline (61). The genome references for mapping are available at: http://genegrabber.berkeley.edu/amd/organisms.

The relative abundance of genomes and genomic fragments from AMD organisms in each sample was estimated as the coverage for all transcripts within a genome (read counts were normalized for genome length and total number of reads per sample) (Figure 1 and Supplementary Table S1). Assembled transcripts from all organisms were also searched vs. the Rfam database (62) for non-coding RNAs (ncRNA) of known function. Additionally, the transcriptional profiles of *Leptospirillum* groups II UBA, group II 5way-CG, group III, and of the archaea A-plasma, C-plasma, G-plasma, and *Ferroplasma* Type II, were visually inspected using Artemis (63). Transcribed regions that did not fall within a coding sequence were evaluated for the presence of non-annotated protein sequences using BlastX (64) vs. the non-redundant NCBI database. 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 intergenic regions that do not appear to encode for protein sequences, based on the above criteria, were labeled as potential non-coding RNA (ncRNA). Bacterial promoters were predicted upstream of manually evaluated ncRNAs using BPROM (65) and rho-independent terminators were evaluated with ARNOLD (66). Riboswitch motifs were predicted using the RibEx webserver (67), and secondary structure prediction of the ncRNAs was done using the RNAfold webserver (68).

Correlation analyses of assembled transcript abundances of non-rRNA reads that mapped to the predicted genes of AMD organisms were done on samples for which total and rRNA-depleted reads were obtained (Supplementary Figure S4). 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 non-rRNA reads from their corresponding total RNA. 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 S4H). 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.

Tables of AMD genes read counts were used for differential expression analyses in R (69). Read counts were normalized for gene length using the function "withinLaneNormalization" from the EDASeq package (70). NMDS ordination was performed on Bray-Curtis distance matrices of variance-stabilizing transformed gene abundance tables with the formula: genes ~ sequencing type + environment type, using the DESeq2 R package (71).

Putative RpoD promoters were predicted upstream of expressed genes in *Leptospirillum* group II UBA genotype by manual inspection of the genome using Artemis (63) along with the visualized transcriptome data. Predicted promoters were listed if they were less than 50 bp upstream of the 5' end of the mapped reads of a transcript and matched 8 out of 12 bp of the standard *E. coli / B. subtilis* RpoD motif (TTGACA_N_{16-19}_TATAAT).

Hierarchical clustering of abundances from predicted genes 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 (72). Clusters and heat maps were visualized using the Java TreeView software (73). Gene trees for *Leptospirillum* spp. transposases were constructed using the MABL website (74). CRISPR loci from transcriptomics reads were reconstructed using the CRASS program (75).

DATA AVAILABILITY

448 Raw sequencing reads were submitted to the NCBI sequence reads archive (SRA) under
449 accession number SRP026490 (76). R code and data tables are available at
450 https://github.com/dgoltsman/acid-mine-drainage.
451
452 **ACKNOWLEDGEMENTS**
453 We thank the late TW Arman, President of Iron Mountain Mines Inc., Mr R Sugarek (US
454 Environmental Protection Agency) for site access, and Mr R Carver for on-site assistance. This
455 project was funded by the U.S. Department of Energy, through the Carbon-Cycling (DE-FG02-
456 10ER64996) program.
457
458

REFERENCES

1. **Mosier AC**, **Miller CS**, **Frischkorn KR**, **Ohm RA**, **Li Z**, **LaButti K**, **Lapidus A**, **Lipzen A**, **Chen C**, **Johnson J**, **Lindquist EA**, **Pan C**, **Hettich RL**, **Grigoriev IV**, **Singer SW**, **Banfield JF**. 2016. Fungi Contribute Critical but Spatially Varying Roles in Nitrogen and Carbon Cycling in Acid Mine Drainage. Front Microbiol **7**:238.

2. **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.

3. **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, 2008 ed. **6**:e177.

4. **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, 2004 ed. **428**:37–43.

5. **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, 2011 ed. **12**:R26.

6. **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. The ISME journal, 2012 ed. **7**:817–829.

7. **Woodcroft BJ**, **Singleton CM**, **Boyd JA**, **Evans PN**, **Emerson JB**, **Zayed AAF**, **Hoelzle RD**, **Lamberton TO**, **McCalley CK**, **Hodgkins SB**, **Wilson RM**, **Purvine SO**, **Nicora CD**, **Li C**, **Frolking S**, **Chanton JP**, **Crill PM**, **Saleska SR**, **Rich VI**, **Tyson GW**. 2018. Genome-centric view of carbon processing in thawing permafrost. Nature **560**:49–54.

8. **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. Environmental microbiology, 2011 ed. **13**:2279–2292.

9. **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, 2010 ed. **11**:404.

10. **Hua ZS**, **Han YJ**, **Chen LX**, **Liu J**, **Hu M**, **Li SJ**, **Kuang JL**, **Chain PS**, **Huang LN**, **Shu WS**. 2014. Ecological roles of dominant and rare prokaryotes in acid mine drainage revealed by metagenomics and metatranscriptomics. The ISME journal.

11. **Shi Y**, **Tyson GW**, **DeLong EF**. 2009. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature, 2009 ed. **459**:266–269.

12. **Wan Y**, **Kertesz M**, **Spitale RC**, **Segal E**, **Chang HY**. 2011. Understanding the transcriptome through RNA structure. Nature reviews Genetics **12**:641–655.

13. **Serganov A**, **Nudler E**. 2013. A decade of riboswitches. Cell, 2013 ed. **152**:17–24.

14. **Wagner EGH**, **Romby P**. 2015. Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. Adv Genet **90**:133–208.

15. **Hinnebusch AG**, **Ivanov IP**, **Sonenberg N**. 2016. Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science **352**:1413–1416.

16. **Dar D**, **Sorek R**. 2017. Regulation of antibiotic-resistance by non-coding RNAs in bacteria. Curr Opin Microbiol **36**:111–117.

17. **Goltsman DS**. 2013. PhD thesis. Chapter 4. University of California, Berkeley. Permalink: https://escholarship.org/uc/item/0zf9f3nx.

18. **Chen L-X**, **Hu M**, **Huang L-N**, **Hua Z-S**, **Kuang J-L**, **Li S-J**, **Shu W-S**. 2015. Comparative metagenomic and metatranscriptomic analyses of microbial communities in acid mine drainage. The ISME journal **9**:1579–1592.

19. **Passalacqua KD**, **Varadarajan A**, **Ondov BD**, **Okou DT**, **Zwick ME**, **Bergman NH**. 2009. Structure and complexity of a bacterial transcriptome. Journal of bacteriology, 2009 ed. **191**:3203–3211.

20. **Quatrini R**, **Johnson DB**. 2018. Microbiomes in extremely acidic environments: functionalities and interactions that allow survival and growth of prokaryotes at low pH. Current Opinion in Microbiology **43**:139–147.

21. **Huang L-N**, **Kuang J-L**, **Shu W-S**. 2016. Microbial Ecology and Evolution in the Acid Mine Drainage Model System. Trends in Microbiology **24**:581–593.

22. **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. Proceedings of the National Academy of Sciences of the United States of America, 2010 ed. **107**:2383–2390.

23. **Wilmes P**, **Remis JP**, **Hwang M**, **Auer M**, **Thelen MP**, **Banfield JF**. 2009. Natural acidophilic biofilm communities reflect distinct organismal and functional organization. The ISME journal, 2008 ed. **3**:266–270.

24. **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.

527 Ecological distribution and population physiology defined by proteomics in a natural microbial
528 community. Mol Syst Biol, 2010 ed. **6**:374.

529 25. **Schneider TD**, **Stephens RM**. 1990. Sequence logos: a new way to display consensus sequences.
530 Nucleic acids research **18**:6097–6100.

531 26. **Ram RJ**, **Verberkmoes NC**, **Thelen MP**, **Tyson GW**, **Baker BJ**, **Blake RC2**, **Shah M**, **Hettich**
532 **RL**, **Banfield JF**. 2005. Community proteomics of a natural microbial biofilm. Science, 2005 ed.
533 **308**:1915–1920.

534 27. **Frias-Lopez J**, **Duran-Pinedo A**. 2012. Effect of periodontal pathogens on the
535 metatranscriptome of a healthy multispecies biofilm model. Journal of bacteriology, 2012 ed.
536 **194**:2082–2095.

537 28. **Hewson I**, **Poretsky RS**, **Beinart RA**, **White AE**, **Shi T**, **Bench SR**, **Moisander PH**, **Paerl RW**,
538 **Tripp HJ**, **Montoya JP**, **Moran MA**, **Zehr JP**. 2009. In situ transcriptomic analysis of the
539 globally important keystone N2-fixing taxon Crocosphaera watsonii. The ISME journal, 2009 ed.
540 **3**:618–631.

541 29. **Peselis A**, **Serganov A**. 2012. Structural insights into ligand binding and gene expression control
542 by an adenosylcobalamin riboswitch. Nature structural & molecular biology, 2012 ed. **19**:1182–
543 1184.

544 30. **Weinberg Z**, **Perreault J**, **Meyer MM**, **Breaker RR**. 2009. Exceptional structured noncoding
545 RNAs revealed by bacterial metagenome analysis. Nature, 2009 ed. **462**:656–659.

546 31. **Klein RJ**, **Misulovin Z**, **Eddy SR**. 2002. Noncoding RNA genes identified in AT-rich
547 hyperthermophiles. Proceedings of the National Academy of Sciences of the United States of
548 America **99**:7542–7547.

549 32. **Omer AD**, **Lowe TM**, **Russell AG**, **Ebhardt H**, **Eddy SR**, **Dennis PP**. 2000. Homologs of small
550 nucleolar RNAs in Archaea. Science **288**:517–522.

551 33. **Breaker RR**. 2011. Prospects for riboswitch discovery and analysis. Molecular cell, 2011 ed.
552 **43**:867–879.

553 34. **Rodionov DA**, **Vitreschak AG**, **Mironov AA**, **Gelfand MS**. 2002. Comparative genomics of
554 thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. The Journal of
555 biological chemistry, 2002nd ed. **277**:48949–48959.

556 35. **Marraffini LA**, **Sontheimer EJ**. 2010. CRISPR interference: RNA-directed adaptive immunity
557 in bacteria and archaea. Nature reviews Genetics **11**:181–190.

558 36. **Sun CL**, **Thomas BC**, **Barrangou R**, **Banfield JF**. 2016. Metagenomic reconstructions of
559 bacterial CRISPR loci constrain population histories. The ISME journal **10**:858–870.

560 37. **Empadinhas N**, **da Costa MS**. 2006. Diversity and biosynthesis of compatible solutes in
561 hyper/thermophiles. International microbiology : the official journal of the Spanish Society for
562 Microbiology, 2006 ed. **9**:199–206.

563 38. **Saum SH**, **Muller V**. 2008. Growth phase-dependent switch in osmolyte strategy in a moderate
564 halophile: ectoine is a minor osmolyte but major stationary phase solute in Halobacillus
565 halophilus. Environmental microbiology, 2007 ed. **10**:716–726.

566 39. **Bhaya D**, **Davison M**, **Barrangou R**. 2011. CRISPR-Cas systems in bacteria and archaea:
567 versatile small RNAs for adaptive defense and regulation. Annual review of genetics, 2011 ed.
568 **45**:273–297.

569 40. **Waterhouse PM**, **Graham MW**, **Wang MB**. 1998. Virus resistance and gene silencing in plants
570 can be induced by simultaneous expression of sense and antisense RNA. Proceedings of the
571 National Academy of Sciences of the United States of America **95**:13959–13964.

41. **Lillestøl RK**, **Shah SA**, **Brügger K**, **Redder P**, **Phan H**, **Christiansen J**, **Garrett RA**. 2009. CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. Molecular microbiology **72**:259–272.

42. **Shmaryahu A**, **Holmes DS**. 2007. Discovery of Small Regulatory RNAs Extends Our Understanding of Gene Regulation in the Acidithiobacillus Genus. Advanced Materials Research **20**:535–538.

43. **Fontecilla-Camps JC**, **Ragsdale SW**. 1999. Nickel–Iron–Sulfur Active Sites: Hydrogenase and Co Dehydrogenase. Advances in Inorganic Chemistry **47**:283–333.

44. **Techtmann S**, **Colman AS**, **Lebedinsky AV**, **Sokolova TG**, **Robb FT**. 2012. Evidence for Horizontal Gene Transfer of Anaerobic Carbon Monoxide Dehydrogenases. Front Microbiol **3**.

45. **Boutard M**, **Ettwiller L**, **Cerisy T**, **Alberti A**, **Labadie K**, **Salanoubat M**, **Schildkraut I**, **Tolonen AC**. 2016. Global repositioning of transcription start sites in a plant-fermenting bacterium. Nature communications **7**:13783.

46. **Babski J**, **Haas KA**, **Näther-Schindler D**, **Pfeiffer F**, **Förstner KU**, **Hammelmann M**, **Hilker R**, **Becker A**, **Sharma CM**, **Marchfelder A**, **Soppa J**. 2016. Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). BMC genomics **17**:629.

47. **Shell SS**, **Wang J**, **Lapierre P**, **Mir M**, **Chase MR**, **Pyle MM**, **Gawande R**, **Ahmad R**, **Sarracino DA**, **Ioerger TR**, **Fortune SM**, **Derbyshire KM**, **Wade JT**, **Gray TA**. 2015. Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. PLoS Genet **11**:e1005641.

593 48. **Belnap CP**, **Pan C**, **Denef VJ**, **Samatova NF**, **Hettich RL**, **Banfield JF**. 2011. Quantitative
594 proteomic analyses of the response of acidophilic microbial communities to different pH
595 conditions. The ISME journal, 2011 ed. **5**:1152–1161.

596 49. **Parkhomchuk D**, **Borodina T**, **Amstislavskiy V**, **Banaru M**, **Hallen L**, **Krobitsch S**, **Lehrach**
597 **H**, **Soldatov A**. 2009. Transcriptome analysis by strand-specific sequencing of complementary
598 DNA. Nucleic acids research, 2009 ed. **37**:e123.

599 50. **Quast C**, **Pruesse E**, **Yilmaz P**, **Gerken J**, **Schweer T**, **Yarza P**, **Peplies J**, **Glöckner FO**. 2013.
600 The SILVA ribosomal RNA gene database project: improved data processing and web-based
601 tools. Nucleic acids research **41**:D590–6.

602 51. **Langmead B**, **Salzberg SL**. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods
603 **9**:357–359.

604 52. **Acuna LG**, **Cardenas JP**, **Covarrubias, P. C.**, **Haristoy JJ**, **Flores R**, **Nunez H**, **Riadi G**,
605 **Shmaryahu A**, **Valdes J**, **Dopson M**, **Rawlings DE**, **Banfield JF**, **Holmes DS**, **Quatrini R**.
606 2013. Architecture and gene repertoire of the flexible genome of the extreme acidophile
607 Acidithiobacillus caldus. PloS one, 2013 ed. **8**:e78237.

608 53. **Goltsman DSA**, **Denef VJ**, **Singer SW**, **VerBerkmoes NC**, **Lefsrud M**, **Mueller RS**, **Dick GJ**,
609 **Sun CL**, **Wheeler KE**, **Zemla A**, **Baker BJ**, **Hauser L**, **Land M**, **Shah MB**, **Thelen MP**,
610 **Hettich RL**, **Banfield JF**. 2009. Community genomic and proteomic analyses of
611 chemoautotrophic iron-oxidizing "Leptospirillum rubarum" (Group II) and “Leptospirillum
612 ferrodiazotrophum” (Group III) bacteria in acid mine drainage biofilms. Applied and
613 Environmental Microbiology, 2009 ed. **75**:4599–4615.

614 54. **Fujimura R**, **Sato Y**, **Nishizawa T**, **Oshima K**, **Kim SW**, **Hattori M**, **Kamijo T**, **Ohta H**. 2012.
615 Complete genome sequence of Leptospirillum ferrooxidans strain C2-3, isolated from a fresh

volcanic ash deposit on the island of Miyake, Japan. Journal of bacteriology, 2012 ed. **194**:4122–4123.

55. **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, 2007 ed. **446**:537–541.

56. **Goltsman DSA**, **Dasari M**, **Thomas BC**, **Shah MB**, **VerBerkmoes NC**, **Hettich RL**, **Banfield JF**. 2013. New group in the Leptospirillum clade: cultivation-independent community genomics, proteomics, and transcriptomics of the new species "Leptospirillum group IV UBA BS". Applied and Environmental Microbiology **79**:5384–5393.

57. **Allen EE**, **Tyson GW**, **Whitaker RJ**, **Detter JC**, **Richardson PM**, **Banfield JF**. 2007. Genome dynamics in a natural archaeal population. Proceedings of the National Academy of Sciences of the United States of America, 2007 ed. **104**:1883–1888.

58. **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. Proceedings of the National Academy of Sciences of the United States of America, 2010 ed. **107**:8806–8811.

59. **Yelton AP**, **Comolli LR**, **Justice NB**, **Castelle C**, **Denef VJ**, **Thomas BC**, **Banfield JF**. 2013. Comparative genomics in acid mine drainage biofilm communities reveals metabolic and structural differentiation of co-occurring archaea. BMC genomics, 2013 ed. **14**:485.

60. **Andersson AF**, **Banfield JF**. 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. Science **320**:1047–1050.

61. **Roberts A**, **Pimentel H**, **Trapnell C**, **Pachter L**. 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics, 2011 ed. **27**:2325–2329.

62. **Griffiths-Jones S**, **Bateman A**, **Marshall M**, **Khanna A**, **Eddy SR**. 2003. Rfam: an RNA family database. Nucleic acids research, 2003rd ed. **31**:439–441.

63. **Rutherford K**, **Parkhill J**, **Crook J**, **Horsnell T**, **Rice P**, **Rajandream MA**, **Barrell B**. 2000. Artemis: sequence visualization and annotation. Bioinformatics, 2000 ed. **16**:944–945.

64. **Altschul SF**, **Gish W**, **Miller W**, **Myers EW**, **Lipman DJ**. 1990. Basic local alignment search tool. J Mol Biol, 1990 ed. **215**:403–410.

65. **Solovyev VV**, **Salamov A**. 2011. Automatic Annotation of Microbial Genomes and Metagenomic Sequences. *In* Li, RW (ed.), Metagenomics and its applications in agriculture, biomedicine, and environmental studies, 1st ed. Nova Science Publishers, Inc.

66. **Lesnik EA**, **Sampath R**, **Levene HB**, **Henderson TJ**, **McNeil JA**, **Ecker DJ**. 2001. Prediction of rho-independent transcriptional terminators in Escherichia coli. Nucleic acids research **29**:3583–3594.

67. **Abreu-Goodger C**, **Merino E**. 2005. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. Nucleic acids research **33**:W690–2.

68. **Hofacker IL**. 2003. Vienna RNA secondary structure server. Nucleic acids research **31**:3429–3431.

69. **Team RC**. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2017.

70. **Risso D**, **Schwartz K**, **Sherlock G**, **Dudoit S**. 2011. GC-content normalization for RNA-Seq data. BMC bioinformatics **12**:480.

71. **Love MI**, **Huber W**, **Anders S**. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **15**:550.

72. **Eisen MB**, **Spellman PT**, **Brown PO**, **Botstein D**. 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America **95**:14863–14868.

73. **Saldanha AJ**. 2004. Java Treeview--extensible visualization of microarray data. Bioinformatics **20**:3246–3248.

74. **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 research, 2008 ed. **36**:W465–9.

75. **Skennerton CT**, **Imelfort M**, **Tyson GW**. 2013. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic acids research, 2013 ed.

76. **Goltsman DSA**. **Richmond Mine Acid Mine Drainage Transcriptome or Gene expression**. NCBI GenBank. (accession no. SRP026490). https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP026490

676 **TABLES**

677 Table 1. Summary statistics and description of samples sequenced. Samples were collected
678 directly from the Richmond Mine (Env) or from bioreactor-grown biofilms (BR). GS: growth
679 stage; Temp: temperature; Type of sequencing: GAII, Illumina GAIIx platform; HiSeq, Illumina
680 HiSeq 2500 platform.

								Total RNA		rRNA-subtracted		
Sample ID	Date collected	Env Type	GS	Location	Temp	pH	Type seq.	No. reads (M)	% rRNA	No. Reads (M)	% rRNA	Non-rRNA Reads (M)
R1_GS0	3/30/10	BR	0	Outflow	37	NA*	GAII	2.56	90.55	4.27	84.85	0.89
R1_GS05	2/19/10	BR	0.5	Outflow	37	1.65	GAII	4.86	88.34	-	-	0.57
R2_GS05	7/20/09	BR	0.5	A drift	37	NA*	GAII	4.79	84.97	-	-	0.72
A-drift	9/17/10	Env	0	A drift	40	1.27	GAII	3.41	86.00	-	-	0.48
C75_GS1	9/17/10	Env	1	C drift	46	0.86	GAII	4.66	90.58	4.59	77.17	1.49
AB10_GS0	11/2/10	Env	0	AB drift	39	0.8	HiSeq	27.87	96.22	53.15	82.09	10.57
AB10_GS1	11/2/10	Env	1	AB drift	39	0.8	HiSeq	30.24	96.46	58.42	81.82	11.69
C10_GS0	11/2/10	Env	0.5	C drift	42	0.8	HiSeq	30.91	94.12	-	-	1.60
C10_GS05	11/2/10	Env	0	C drift	42	0.8	HiSeq	27.12	96.16	51.88	85.39	8.77
C10_GS1	11/2/10	Env	1	C drift	42	0.8	HiSeq	28.60	95.49	50.65	85.66	8.55
G2E1_GS2	6/9/09	Env	2	4-way	39	0.7	HiSeq	26.68	85.73	45.06	68.76	17.88
R3_GS0	9/28/10	BR	0	A drift	37	1.31	HiSeq	31.56	95.00	-	-	1.58
R3_GS1	10/6/10	BR	1	A drift	37	1.74	HiSeq	30.49	94.08	67.98	75.10	18.73

681

682 * NA indicates that no pH measurements were available for these samples.

683
684

685 **FIGURE LEGENDS**

686 Figure 1. (next page) Differential gene and genome expression correlates with environmental
687 conditions in AMD biofilms. A) Relative expression abundance of AMD genomes. Read counts
688 were normalized by the genome length and sample size. B) Proportion of AMD genomes covered
689 by transcriptomic reads. The top 20 most abundant taxa and viruses are shown.

690

691 Figure 2. Non-metric multidimensional scaling (NMDS) ordination on Bray-Curtis distances of
692 differentially expressed genes in AMD communities. A) NMDS samples plot color coded by pH
693 and environment type. B-D) NMDS genes plot colored by: B) the top most abundant
694 *Leptospirillum* species and strains; C) the top most abundant Archaeal species, and D) viral
695 genomes. NMDS genes plots indicate a gradient of points going from the lower right quadrant to
696 the top right quadrant, which then goes to the top left quadrant (horseshoe).

697

698 Figure 3. (previous page) Environmental conditions distinguish expression of groups of genes
699 and functional categories in AMD systems. A-D) Heatmaps of genes with log2-fold change > 5
700 in: A) *Leptospirillum* group II UBA; B) G-plasma; C) *Leptospirillum* group III; and D)
701 *Leptospirillum* group II 5way CG. Yellow: overrepresented, blue: underrepresented. E)
702 Functional categories differ in communities growing in the natural environment vs. laboratory
703 bioreactors. Categories enriched in environmental samples are shown in dark blue, while
704 categories enriched in bioreactor samples are shown in orange. Functional categories in bold
705 within black boxes represent the categories in which the most variation between conditions was
706 observed.

707

Figure 4. Predicted secondary structures of transcribed ncRNAs of known function in *Leptospirillum* sp. A) Sequence logo representation of the promoter structure in *Leptospirillum* group II UBA genotype. The alignment was constructed using the sequences of 207 predicted promoters. B-C) Secondary structures predicted in *Leptospirillum* group II UBA (B, top structures) and in *Leptospirillum* group III (C, bottom structures). Secondary structure prediction was done on the RNAfold webserver.

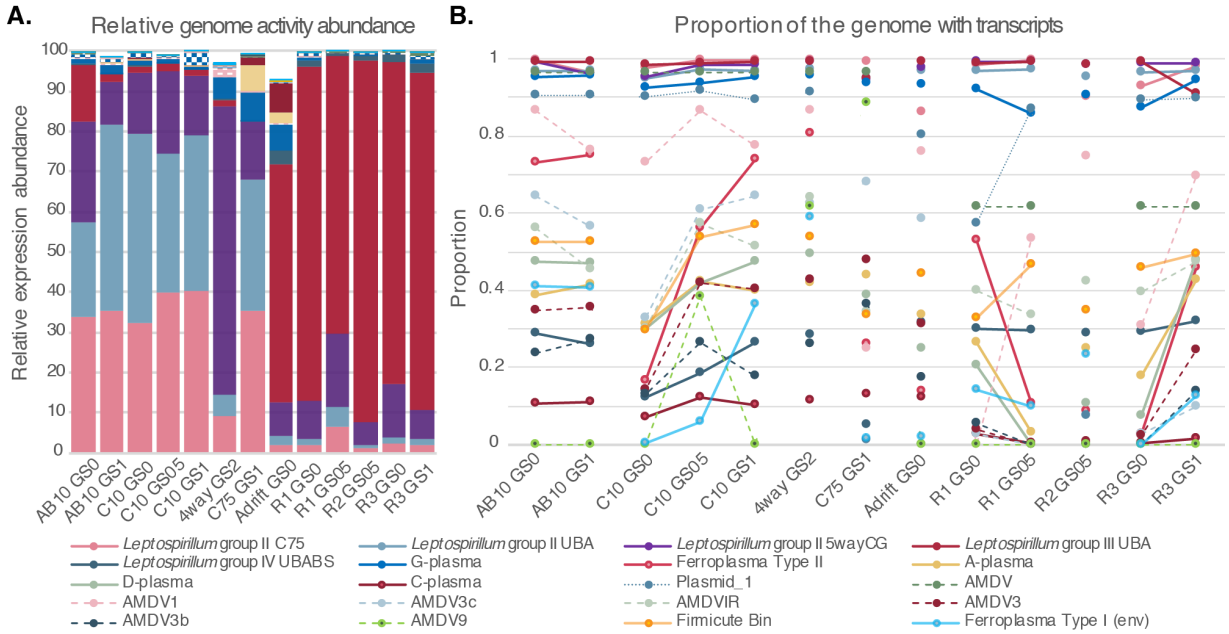
Figure 5. Expression of regulatory ncRNAs in *Leptospirillum* group II. A) Modified Artemis image showing the transcriptional expression of *Leptospirillum* group II UBA CRISPR loci. Expression of Cas genes occurs in the same orientation as the predicted genes on scaffold 8241 (orange line, (-) strand), while pre-crRNAs are *cis*-transcribed on the antisense strand (green line, (+) strand). CPUF: conserved hypothetical protein; TR: transcriptional regulator; TN: transposase. Pre-crRNAs are shown as blue boxes, predicted coding genes are represented by teal boxes, the CRISPR leader sequence (L) is represented by a black box downstream of Cas2 and is followed by the first repeat in the composite assembly (red box). The last repeat of the loci is also show as a red box next to the transcriptional regulator. The composite assembled CRISPR locus, including the Cas genes, is ~ 25 Kb in length. B-C) Novel ectoine riboswitch in *Leptospirillum* group II bacteria. The ectoine operon (teal blue arrows), and the riboswitch (ncRNA, white box) with its promoter (P) and Rho-independent terminator (T) are shown. The modified Artemis image in *Leptospirillum* group II UBA (B, top) and in *Leptospirillum* group II 5way -CG (B, bottom) shows the strand-specific transcriptional expression from two bioreactor samples (blue: R3_GS0; magenta: R3_GS1) and two environmental biofilms (red: AB10_GS0; green: AB10_GS1). The predicted secondary structures of the riboswitches are shown in:

731 *Leptospirillum* group II UBA (C, top) and in *Leptospirillum* group II 5way CG (C, bottom) with
732 their Rho-independent terminator (solid arrow) and an anti-terminator (dashed arrow).
733
734 Figure 6. Expression of regulatory ncRNAs in *Leptospirillum* group III. Panels show the strand-
735 specific expression of energy generation-related genes with their associated regulatory ncRNA.
736 Coding genes are represented by teal arrows, ncRNAs by white arrows, putative uORFs by blue
737 arrows, and promoters by green arrows. Left: CO dehydrogenase (CODH) beta subunit with its
738 transcribed short (70 bp) ncRNA. The predicted secondary structure for the CODH-ncRNA is
739 shown. Purple and green lines: R3_GS1 bioreactor biofilm; blue and teal lines: R3_GS0
740 bioreactor biofilm. Middle: pyruvate:ferredoxin oxidorectase (PFOR) operon with its transcribed
741 5' UTR containing a putative promoter (P, green) and possible regulatory uORFs. Right:
742 Cytochrome c 572 (Cyt572) with its transcribed 5' UTR containing a putative promoter (P) and
743 possible regulatory uORFs. The length of the putative regulatory uORFs ranges from 10 to 40
744 amino acids long.
745

SUPPLEMENTARY MATERIALS LEGENDS

1. Supplementary Table S1. Genome-wide relative abundance of transcripts from AMD genomes.
2. Supplementary Table S2. Predicted promoters upstream of transcribed genes in *Leptospirillum* group II UBA genotype.
3. Supplementary Table S3. Transcripts containing ncRNAs with predicted functions.
4. Supplementary Table S4. Summary of CRISPR spacers recovered from paired-end, strand-specific transcriptomics datasets.
5. Supplementary Table S5. CRISPR CRASS output table and spacer matches to previous CRISPR analyses.
6. Supplementary Figure S1. Differential expression analysis of genes from AMD genomes.
7. Supplementary Figure S2. Gene tree of transposases expressed in *Leptospirillum* sp.
8. Supplementary Figure S3. Examples of expression of leadered transcripts in *Leptospirillum* sp.
9. Supplementary Figure S4. Transcript abundance plots from total vs. rRNA-depleted RNA samples.

765 **FIGURES**



766

767 Figure 1. Transcriptional expression of acid mine drainage genomes. A) Relative genome activity of

768 AMD genomes across samples. Read counts were normalized by the genome length and sample size.

769 B) Proportion of AMD genomes covered by transcriptomic reads. The top 20 most abundant taxa

770 and viruses are shown on both panels.

771

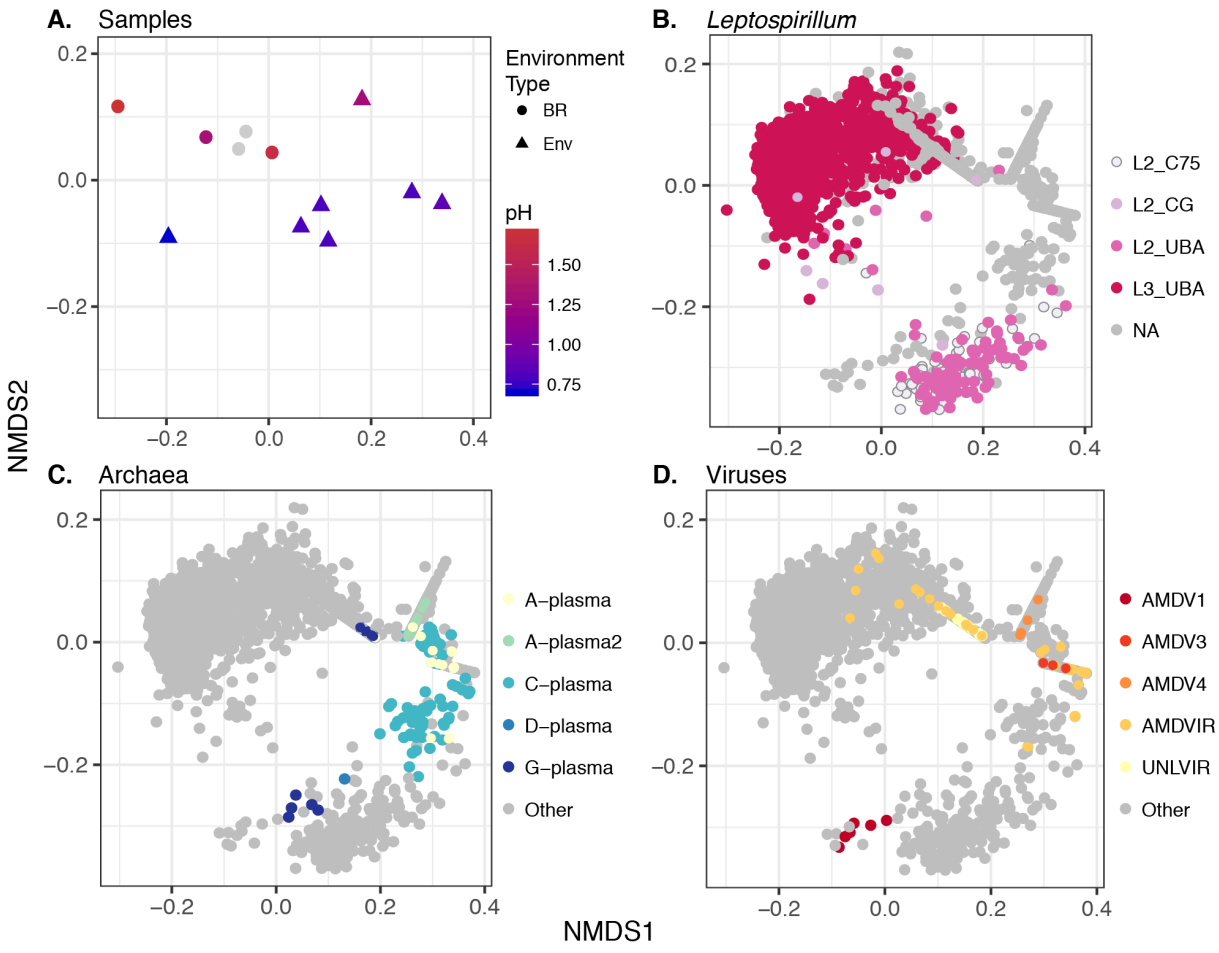
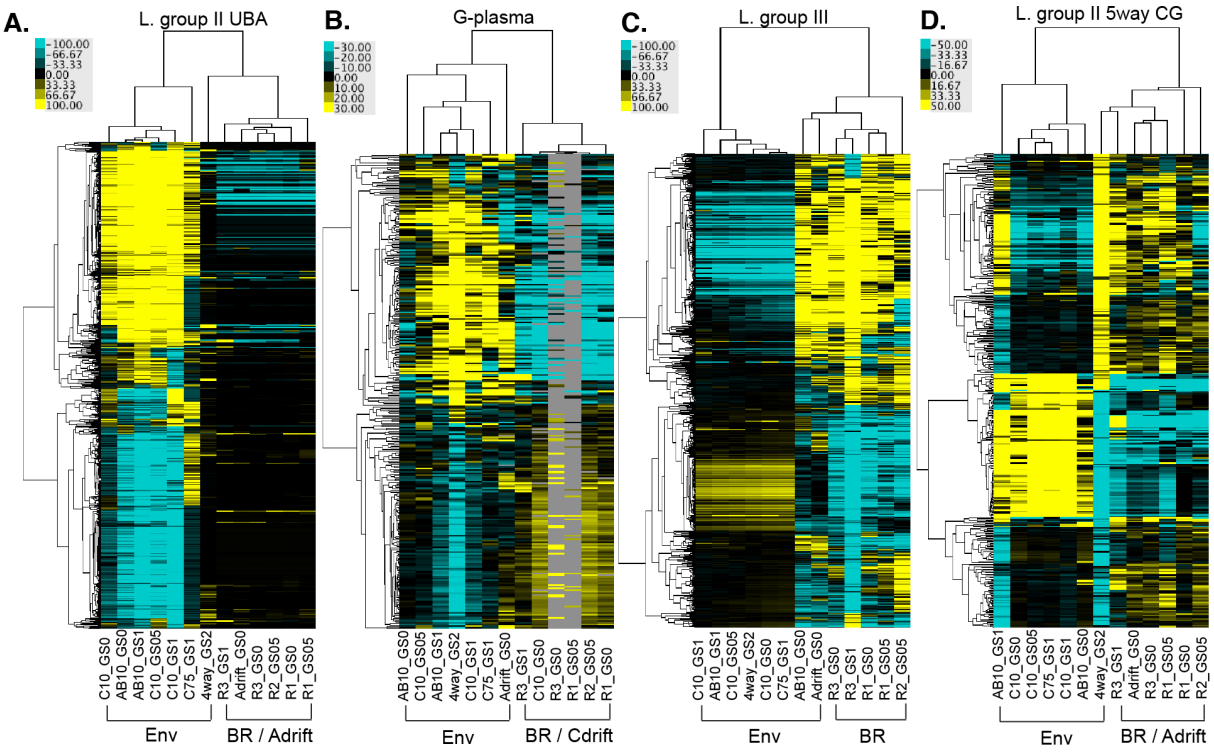
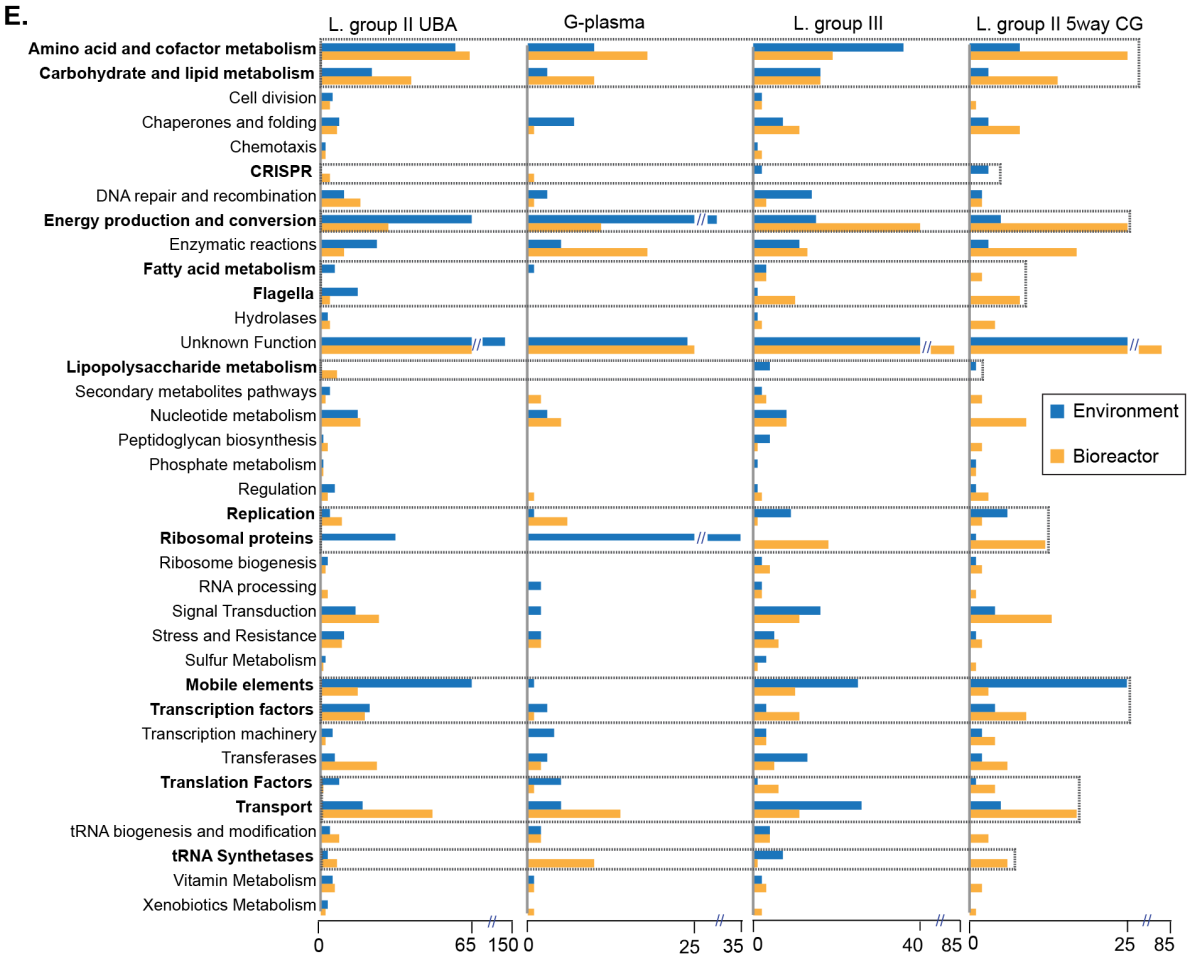


Figure 2. Non-metric multidimensional scaling (NMDS) ordination on Bray-Curtis distances of differentially expressed genes in AMD communities. A) NMDS samples plot coded by pH and environment type. B-D) NMDS genes plot colored by: B) *Leptospirillum* species and strains; C) Archaeal species, and D) viral genomes. NMDS genes plots indicate a gradient of points going from the lower right quadrant towards the top left quadrant (horseshoe).





778

779 Figure 3. (previous page) Environmental conditions distinguish expression of groups of genes and
780 functional categories in AMD systems. A-D) Heatmaps of genes with log2-fold change > 5 in: A)
781 *Leptospirillum* group II UBA; B) G-plasma; C) *Leptospirillum* group III; and D) *Leptospirillum* group
782 II 5way-CG. Yellow: overrepresented, blue: underrepresented. E) Functional categories differ in
783 communities growing in the natural environment vs. laboratory bioreactors. Categories enriched
784 in environmental samples are shown in dark blue, while categories enriched in bioreactor samples
785 are shown in orange. Functional categories in bold within black boxes represent the categories in
786 which the most variation between conditions was observed.

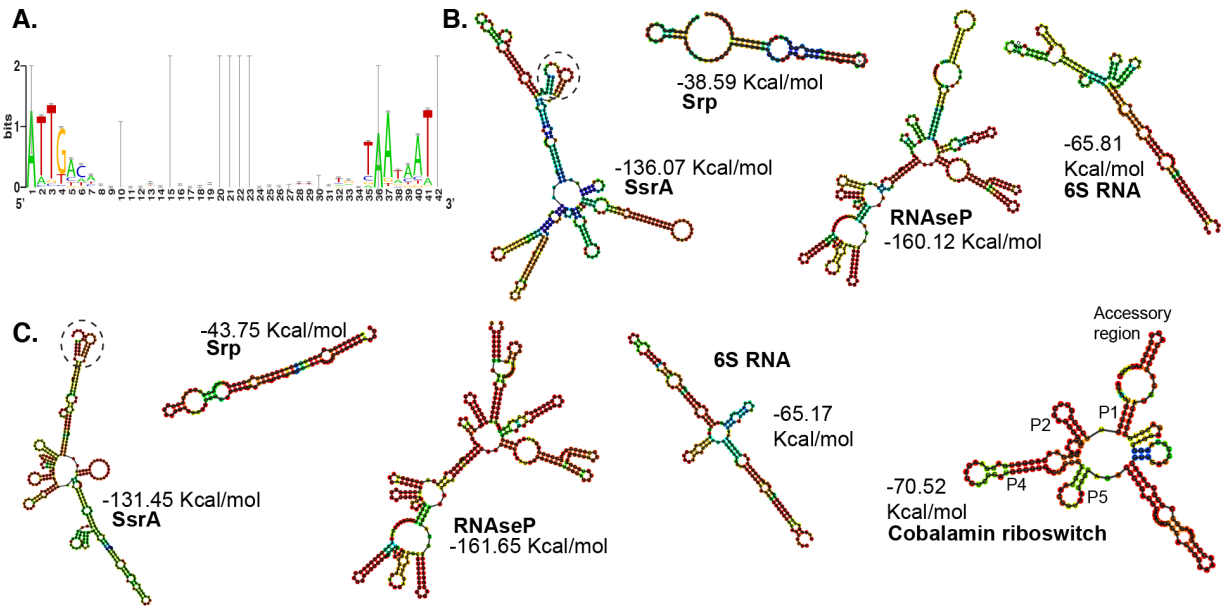


Figure 4. Predicted secondary structures of transcribed ncRNAs of known function in *Leptospirillum* sp. A) Sequence logo representation of the promoter structure in *Leptospirillum* group II UBA genotype. The alignment was constructed using the sequences of 207 predicted promoters. B) structures predicted in *Leptospirillum* group II UBA (Top structures). C) Structures predicted in *Leptospirillum* group III. Secondary structure prediction was done on the RNAfold webserver.

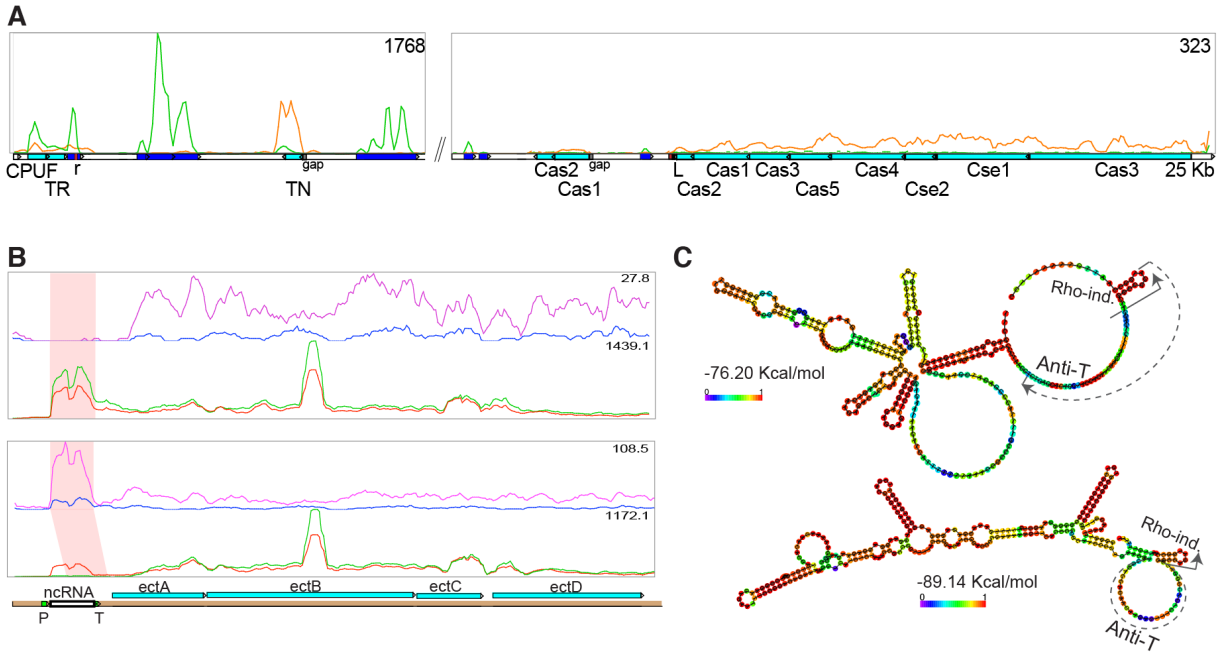
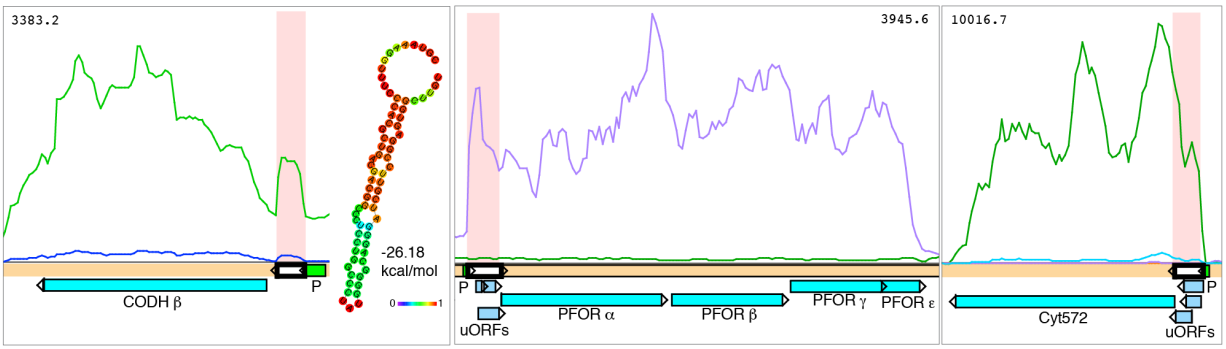


Figure 5. Expression of regulatory ncRNAs in *Leptospirillum* group II. A) Modified Artemis image showing the transcriptional expression of *Leptospirillum* group II UBA CRISPR loci. Expression of Cas genes occurs in the same orientation as the predicted genes on scaffold 8241 (orange line, (-) strand), while pre-crRNAs are *cis*-transcribed on the antisense strand (green line, (+) strand). CPUF: conserved hypothetical protein; TR: transcriptional regulator; TN: transposase. Pre-crRNAs are shown as blue boxes, predicted coding genes are represented by teal boxes, the CRISPR leader sequence (L) is represented by a black box downstream of Cas2 and is followed by the first repeat in the composite assembly (red box). The last repeat of the loci is also show as a red box next to the transcriptional regulator. The composite assembled CRISPR locus, including the Cas genes, is ~ 25 Kb in length. B-C) Novel ectoine riboswitch in *Leptospirillum* group II bacteria. The ectoine operon (teal blue arrows), and the riboswitch (ncRNA, white box) with its promoter (P) and Rho-independent terminator (T) are shown. The modified Artemis image in *Leptospirillum* group II UBA (B, top) and in *Leptospirillum* group II 5way -CG (B, bottom) shows the strand-specific transcriptional expression 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 riboswitches are shown in: *Leptospirillum* group

811 II UBA (C, top) and in *Leptospirillum* group II 5way CG (C, bottom) with their Rho-independent
812 terminator (solid arrow) and an anti-terminator (dashed arrow).



813

814 Figure 6. Expression of regulatory ncRNAs in *Leptospirillum* group III. Panels show the strand-specific
815 expression of energy generation-related genes with their associated regulatory ncRNA. Coding genes
816 are represented by teal arrows, ncRNAs by white arrows, putative uORFs by blue arrows, and
817 promoters by green arrows. Left: CO dehydrogenase (CODH) beta subunit with its transcribed short
818 (70 bp) ncRNA. The predicted secondary structure for the CODH-ncRNA is shown. Purple and green
819 lines: R3_GS1 bioreactor biofilm; blue and teal lines: R3_GS0 bioreactor biofilm. Middle:
820 pyruvate:ferredoxin oxidorectase (PFOR) operon with its transcribed 5' UTR containing a putative
821 promoter (P, green) and possible regulatory uORFs. Right: Cytochrome c 572 (Cyt572) with its
822 transcribed 5' UTR containing a putative promoter (P) and possible regulatory uORFs. The length of
823 the putative regulatory uORFs ranges from 10 to 40 amino acids long.

824

825