Enigmatic, ultrasmall, uncultivated Archaea

Brett J. Baker[a], Luis R. Comolli[b], Gregory J. Dick[a,1], Loren J. Hauser[c], Doug Hyatt[c], Brian D. Dill[d], Miriam L. Land[c], Nathan C. VerBerkmoes[d], Robert L. Hettich[d], and Jillian F. Banfield[a,e,2]

[a]Department of Earth and Planetary Science and [e]Environmental Science, Policy, and Management, University of California, Berkeley, CA 94720; [b]Lawrence Berkeley National Laboratories, Berkeley, CA 94720; and [c]Biosciences and [d]Chemical Sciences Divisions, Oak Ridge National Laboratory, Oak Ridge, TN 37831

Edited by Norman R. Pace, University of Colorado, Boulder, CO, and approved March 30, 2010 (received for review December 16, 2009)

Metagenomics has provided access to genomes of as yet uncultivated microorganisms in natural environments, yet there are gaps in our knowledge—particularly for Archaea—that occur at relatively low abundance and in extreme environments. Ultrasmall cells (<500 nm in diameter) from lineages without cultivated representatives that branch near the crenarchaeal/euryarchaeal divide have been detected in a variety of acidic ecosystems. We reconstructed composite, near-complete ~1-Mb genomes for three lineages, referred to as ARMAN (archaeal Richmond Mine acidophilic nanoorganisms), from environmental samples and a biofilm filtrate. Genes of two lineages are among the smallest yet described, enabling a 10% higher coding density than found genomes of the same size, and there are noncontiguous genes. No biological function could be inferred for up to 45% of genes and no more than 63% of the predicted proteins could be assigned to a revised set of archaeal clusters of orthologous groups. Some core metabolic genes are more common in *Crenarchaeota* than *Euryarchaeota*, up to 21% of genes have the highest sequence identity to bacterial genes, and 12 belong to clusters of orthologous groups that were previously exclusive to bacteria. A small subset of 3D cryo-electron tomographic reconstructions clearly show penetration of the ARMAN cell wall and cytoplasmic membranes by protuberances extended from cells of the archaeal order *Thermoplasmatales*. Interspecies interactions, the presence of a unique internal tubular organelle [Comolli, et al. (2009) *ISME J* 3:159–167], and many genes previously only affiliated with Crenarchaea or Bacteria indicate extensive unique physiology in organisms that branched close to the time that *Cren*- and *Euryarchaeotal* lineages diverged.

acid mine drainage | archaeal Richmond Mine acidophilic nanoorganisms | metagenomics | phylogeny | microbial ecology

Metagenomics is providing new insights into the physiological capacities and evolutionary histories of microorganisms from a wide range of environments [reviewed recently by Wilmes et al. (1)]. Many datasets provide fragmentary glimpses into genetic diversity (2–4) and a few have reported near-complete genomic sequences for uncultivated organisms (5–8). In most cases where extensive reconstruction has been possible, insights have been restricted to relatively dominant members. Furthermore, it has been difficult to use genomic information to infer the nature of interorganism interactions, although these are likely to be very important aspects of microbial community functioning. The need for topological and organizational information to place genomic data in context motivates the combination of cultivation-independent genomics and 3D cryogenic transmission electron microscope-based ultrastructural analyses of microbial communities.

Despite the importance of cellular interactions (symbiosis and parasitism), most of what we know about microorganismal associations is from cultivation-based studies (9–11). However, sequencing of the genomes of several endosymbiotic and parasitic Bacteria has revealed reduction in gene and genome sizes, reflecting evolved dependence of the endosymbiont or parasite on its host (12, 13). The ultrasmall archaeal parasite *Nanoarchaeum equitans* has only 552 genes and requires a connection to its archaeal host, *Ignicoccus hopstialis*, to survive (10). Recently, it was shown that this interaction involves contact between outer membranes (14). Given the vast diversity of microbial life (15), it is likely that other unusual relationships critical to survival of organisms and communities remain to be discovered. In the present study, we explored the biology of three unique, uncultivated lineages of ultrasmall Archaea by combining metagenomics, community proteomics, and 3D tomographic analysis of cells and cell-to-cell interactions in natural biofilms. Using these complementary cultivation-independent methods, we report several unexpected metabolic features that illustrate unique facets of microbial biology and ecology.

Chemoautotrophic biofilms grow in acidic, metal-rich solutions (millimolar to molar Fe, Zn, Cu, As) within Richmond Mine, at Iron Mountain, California. Despite the low species richness, biofilms are complex enough to capture important evolutionary and ecological processes, making them an ideal model system for community ecological studies (16). The biofilms are dominated by *Leptospirillum* spp. Bacteria and contain Archaea of the order *Thermoplasmatales*, such as *Ferroplasma* sp (17), and uncultivated *Thermoplasmatales*, such as G-plasma (7, 18) and A-plasma (19). 16S rRNA gene sequences recovered from metagenomic data revealed the existence of unique low abundance organisms from the ARMAN lineages (archaeal Richmond Mine acidophilic nanoorganisms) that had been missed in prior PCR-based surveys because of mismatches with common primers (20). The lineages have been detected in other acidic environments (21, 22). ARMAN cells have volumes of 0.009 μm^3 to 0.04 μm^3, close to the theoretical lower limit for life, and almost all cells contain a mysterious tubular organelle that is roughly 200 nm long and 60 nm wide (23).

Results and Discussion

Recovery of ARMAN Genomes. One hundred megabases of sequence (mate paired ~700-bp Sanger reads from 3-kb clones) was acquired from the ultra-back A drift blanket strips (UBA-BS) biofilm (*SI Materials and Methods*) shown by fluorescence in situ hybridization to contain more ARMAN-2 cells than were detected in biofilms studied previously. These sequences were combined with other unclassified sequences from a second biofilm (UBA

Author contributions: B.J.B., L.R.C., and J.F.B. designed research; B.J.B., L.R.C., and N.C.V. performed research; B.J.B.; G.J.D., L.J.H., D.H., B.D.D., M.L.L., N.C.V., R.L.H., and J.F.B. contributed new reagents/analytic tools; B.J.B., L.R.C., G.J.D., L.J.H., D.H., B.D.D., M.L.L., N.C.V., and J.F.B. analyzed data; and B.J.B., L.R.C., and J.F.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: *Candidatus* Micrarchaeum acidiphilum ARMAN-2 has been deposited at DDBJ/EMBL/GenBank under the project accession ACVJ00000000 (scaffolds with annotation are GG697234–GG697241). The filtrate library has been deposited under NCBI Genome project ID #36661. The *Candidatus* Parvarchaeum acidiphilum ARMAN-4 Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession ADCE00000000. The version described in this article is the first version, ADCE01000000 (scaffolds with annotation are ADCE01000001–ADCE01000045). The *Candidatus* Parvarchaeum acidophilus ARMAN-5 Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession ADHF00000000 (scaffolds with annotation are ADHF01000001–ADHF01000073).

[1]Present address: Department of Geological Sciences, University of Michigan, Ann Arbor, MI 48109.

[2]To whom correspondence should be addressed. E-mail: jbanfield@berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.0914470107/-/DCSupplemental.

dataset, ~117 Mb) (24). The resulting 229,084 reads were assembled into 13,256 contigs containing at least two reads and totaling 34.1 Mb of sequence. 16S rRNA genes and other phylogenetic markers indicated that the dataset included the near complete (15× coverage) genome of an ARMAN-2 population (ARMAN-2 reads represent ~7% of the sequences from the two biofilms).

ARMAN cells from a third biofilm (UBA) were enriched by filtration (targeting cells <450 nm in diameter) to enable genomic analysis. DNA concentrations in the filtrate were increased by multiple displacement amplification (MDA) before library construction. From ~60 Mb of random shotgun sequence, we assembled 4,825 fragments totaling 13.6 Mb. Two fragments encoded 16S rRNA genes from previously unidentified archaeal groups, here named ARMAN-4 and ARMAN-5. These share 92% pair-wise identity, but only 72% identity with the 16S rRNA gene of ARMAN-2. The organisms with closest 16S rRNA gene sequences to ARMAN-4 and -5 are uncultivated (77% sequence identity) (Fig. S1). We constructed trees using ribosomal proteins, elongation factors EF-1 and EF-2, a membrane secretion protein (SecY), and a DNA repair and recombination protein (RadA) to elucidate the phylogenetic position of the ARMAN groups. The majority of the statistically supported analyses for these genes indicate that ARMAN branch near the root of the *Euryarchaeota* (Fig. S2). Using probes that target ARMAN-2, -4, and -5, we showed that ARMAN-4 and -5 are much less abundant that ARMAN-2 in the seven biofilm communities surveyed (*SI Materials and Methods*).

To identify ARMAN genome fragments, scaffolds >1.5 kb were binned using a variety of methods (*SI Materials and Methods*), including clustering of tetranucleotide frequencies within emergent self-organizing maps. Tetranucleotide mapping created distinct clusters, consistently binned ARMAN fragments of known affiliation, and confirmed that essentially all ARMAN fragments were included in the genomic reconstructions (5). A subset of low coverage (3×) ARMAN fragments could not be firmly assigned to either group and were designated "ARMAN-unassigned" (430,607 bp of sequence). These fragments may belong to variant genotypes.

To further evaluate genome completeness, we checked for the presence of 40 genes expected in all genomes (25). ARMAN-2 has a single copy of each; ARMAN-4 and -5 lack nine and four genes, respectively. Thus, we estimate that ARMAN-2 is ~99%, ARMAN-4 is 80%, and ARMAN-5 90% complete, with average genome sizes of ~1 Mb. The 552 predicted orthologous proteins of ARMAN-4 and -5 share 71% average amino acid identity. The genome of ARMAN-2 is in three fragments and ARMAN-4 and -5 are in 46 and 75 fragments, respectively. Genome statistics are summarized in Table 1. For group two, we propose the name *Candidatus* Micrarchaeum acidiphilum ARMAN-2. We suggest the names *Candidatus* Parvarchaeum acidiphilum for ARMAN-4 and *Candidatus* Parvarchaeum acidophilus for ARMAN-5.

Comparative Genomics. ARMAN proteins have an unusually large number of top BLAST hits to bacterial proteins (Fig. 1 and Fig. S3) and 12 only have bacterial orthologs in the current clusters of orthologous groups (COG) database (26) (Table S1). Other notable features of the ARMAN genomes are the anomalously small fraction of genes that can be assigned to archaeal COGs (arCOGs) (<66%) compared with all other Archaea except *Crenarchaeum symbiosum* (6) (58%) and the preponderance of crenarchaeal over euryarchaeal arCOGs, despite the robust phylogenetic placement within the *Euryarchaeota* (Figs. S1 and S2 and Table S2). In addition, ARMAN-4 and -5 have a TAA motif upstream of many genes that is likely involved in transcriptional control, a feature seen previously in crenarchaea. ARMAN-2 lacks 18, ARMAN-4 lacks 24, and ARMAN-5 lacks 14 of the 166 universal arCOGs (Table S3). Many of the missing arCOGs are involved in translation or are ribosomal. We infer that proteins missing from within otherwise complete ribosomal operons are very likely not present in the genome. We were able to assign a biochemical function or general biological activity to between 58 and 65% of ARMAN genes. A large number of genes (25% in ARMAN-2, 35% in ARMAN-4, 38% in ARMAN-5) have no matches (e-value $>10^{-5}$) to sequences in public databases (Fig. S3).

Table 1. Summary of genomic information for ARMAN

	ARMAN-2	ARMAN-4	ARMAN-5	Unassigned
Total bp	999,043	800,887	921,220	430,607
G+C content %	47.2	34.8	34.1	36.5
Number of fragments	3	44	73	145
Average contig size (bp)	333,014	17,546	12,381	2,935
Predicted protein coding genes (CDS/ORF)	1,033	916	1046	NA
Percent of genome protein coding	90.7	90.1	89.5	NA
Number of KEGG matches	735	569	687	NA
Number of arCOG matches	685	587	687	NA
Average CDS/ORF length (bp)	892	774	788	NA
Biological function assigned	668	539	561	NA
Conserved hypotheticals	152	139	159	NA
tRNAs	35	39	42	6

We examined the taxonomic distributions of the top hits of all genes of ARMAN, Euryarchaea, Crenarchaea, and two separate basal Archaea (*N. equitans* and *Candidatus* Korarchaeum cryptofilum) to the KEGG database (Fig. 1). The ARMANs have a larger percentage of their genes, with best matches to crenarchaeal organisms than any other Euryarchaea. They also have many genes with no matches (e-value of 10^{-5}) in the KEGG database. The taxonomic match-based profiles of the ARMANs are most similar to the deeply branched *N. equitans* and *Ca.* K. cryptofilum.

Proteomic Analysis of ARMAN. ARMAN proteins were identified from lysates of seven biofilm communities via shotgun proteomics based on assignment of tandem mass spectra to peptides predicted from the genomic datasets. The database used for peptide identification included predicted proteins from four bacterial and six *Thermoplasmatales* archaeal species and from ARMAN-2, -4, and -5. Most identified ARMAN peptides were uniquely assigned to proteins of a single ARMAN type. Searches yielded 173 proteins from ARMAN-2 (17% of predicted proteins), 32 from ARMAN-4 (3% of predicted proteins), and 27 from ARMAN-5 (3% of predicted proteins) (Tables S4–S6). The most abundant ARMAN-2, -4, and -5 proteins were involved in energy production, translation, protein modification, amino acid metabolism, or had unknown functions (Tables S4–S6). H^+-transporting subunits of ATP synthase are the most highly expressed proteins with assigned function from ARMAN-4, and -5, and account for 8, 31, and 40% of the total spectral counts for proteins from ARMAN-2, -4, and -5, respectively).

A protein with homology to the alpha subunit of a multisubunit archaeal thermosome (ARMAN-2, UNLARM2_0780), likely a group II chaperonin (27), is also among the most highly expressed ARMAN proteins with known function. Chaperonins are generally highly expressed in the acid mine drainage (AMD) organisms (28) and may play a role in protein refolding or stabilization. However, this thermosome protein alternatively may be a component of the intracellular ARMAN tube, given that ar-

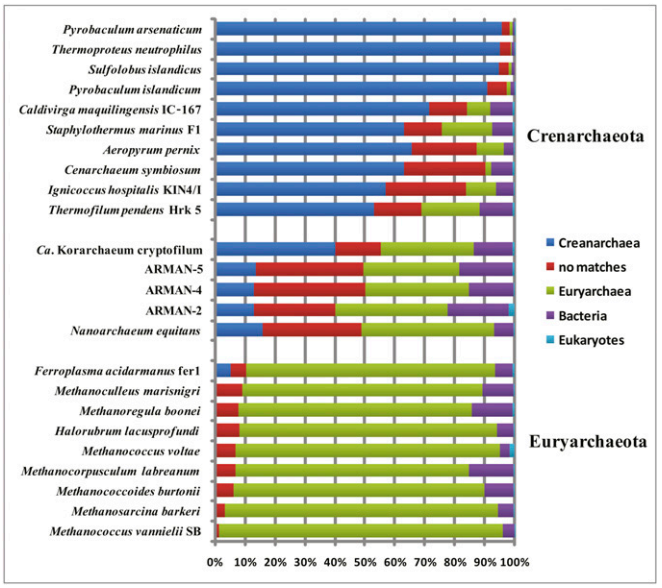

Fig. 1. The taxonomic affliations of the top blast hit for all proteins from each genome to the KEGG genome database. *Crenarchaeota* are shown on top and *Euryarchaeota* on the bottom of the diagram. The cutoff for the "no match" category was an e-value of 10^{-5}.

chaeal chaperonins have previously been shown to form hetero-oligomeric complex cylinders (29). The ARMAN-4 homolog of the archaeal thermosome protein has been detected by proteomics as well (BJBARM4_0177 in two datasets).

Roughly 25% of the ARMAN proteins identified were annotated as hypothetical or conserved hypothetical and, across all samples, 32% of spectra assigned to ARMAN derived from hypothetical proteins. The most abundant protein of ARMAN-2 (UNLARM2_0493) has no assigned function. Interestingly, this protein was not identified in some community proteomic datasets, suggesting that it may be involved in a specific environmental response.

Metabolism. Despite few genes for glycolysis, ARMAN-2 has genes for breakdown of fatty acids via beta oxidation and all three ARMANs have complete or near complete tricarboxylic acid (TCA) cycles. ARMAN-4 and -5 have near complete glycolytic pathways, use the pentose phosphate pathway for carbohydrate metabolism, and have genes required for glycerol utilization. Five proteins of the TCA cycle were detected in the proteomic data and high levels of succinate dehydrogenase were detected from all three ARMAN groups (two homologs in ARMAN-2). Therefore, we know that aerobic respiration is active in the ARMANs in the biofilm. Aconitate hydratase was also detected in the proteomic data from all three ARMAN groups (two homologs in ARMAN-5), but succinate-CoA ligase and fumarate lyase were detected only from ARMAN-4. All ARMAN genomes encode a single superoxide dismutase, two peroxiredoxin-like genes, and alkyl hydroperoxide reductase to respond to oxidative stress, consistent with an aerobic lifestyle, although cytochrome *c*-oxidase genes have not been found. All three genomes encode components required for oxidative phosphorylation and ARMAN-2 has heme synthesis genes needed to produce cytochromes. ARMAN-2 has an alternative thymidylate synthase (ThyX) for thymine synthesis (30). We have not identified the ribulose monophosphate pathway that is common to Archaea in ARMAN-2. For CoA synthesis, all three groups are missing the bifunctional phosphopantothenoylcysteine synthetase/decarboxylase that is found in all other Archaea except *N. equitans*, *Ca.* K. cryptofilum (31), and *Thermofilum pendens* (32).

There are many pathways for which genes were not identified, but this could be because of genome incompleteness or high levels of divergence from characterized genes. For example, we identified only two genes for cobalamin (vitamin B12) biosynthesis, cobalamin adenosyltransferase, and precorrin-3B synthetase (*cobZ*). CobZ has only been reported previously in bacteria and is a key enzyme in a poorly understood cobalamin biosynthetic pathway in photosynthetic bacteria (33). Searches of the UBA-BS and UBA datasets revealed that this is the only CobZ in either community. Interestingly, the ARMAN-2 gene UNLARM2_0870, encoding a protein of unknown function, contains a cobalamin binding site.

DNA Replication and Cell Cycle. All three ARMAN groups have orc1/cdc6 replication initiation protein homologs, indicating that they have circular chromosomes. ARMAN-2 has primases, *priS*, and multiple DNA polymerases. Additionally, the ARMAN-2 genome encodes a sliding clamp proliferating cell nuclear antigen (PcnA) protein (UNLARM2_0836), and this protein was identified in the UBA-BS sample. All three groups have a clamp-loader complex that encodes the large and small subunit of replication factor C and ARMAN-4 and -5 have a replication factor A homolog. Chromatin-associated genes include *Alba* (in ARMAN-5) and A3 archaeal-type histone (just in ARMAN-2). ARMAN-2 has a MinD protein involved in chromosome partitioning, as well as two structural maintenance of chromosomes-like proteins for chromosomal segregation, both of which were identified by proteomics. It also has a HerA (FtsK-like) previously only found in Crenarchaea. ARMAN-2 has four cell-division proteins (FtsZ), whereas ARMAN-4 and -5 have two. One ARMAN-2 cell-division FtsZ protein was identified (UNLARM2_0213) in two datasets (UBA and UBA-BS). These observations indicate that ARMAN have at least a fraction of the replication and cell-cycle pathways common to Archaea, whereas the proteomic data imply that the cells were active and dividing in the samples characterized by proteomics.

Transcription and Translation. ARMAN-2 has a total of 53 ribosomal proteins, 30 large ribosomal subunit (LSU) and 23 small ribosomal subunit (SSU), only 18 of which were identified by proteomics. Ribosomal genes typically have conserved local context. One large block of ribosomal genes shares conserved gene order with many other genomes (e.g., S17, L14, L24, S4e, L5, S14, S8, L6, L32, L19, L18, S5). However, many ribosomal proteins are scattered around the genomes. ARMAN-4 and -5 have 43 and 44 ribosomal proteins (ARMAN-4 has 21 LSU, 22 SSU and ARMAN-5 has 24 LSU, 20 SSU). Only one ribosomal protein (L3) from ARMAN-4 was identified by proteomics, suggesting its low abundance and activity in the biofilm communities.

ARMAN-2 does not have a Shine-Dalgarno motif in its translation initiation regions. Instead it has a strong GGTG motif with normal upstream spacing (5–10 bp). Over 80% of the genes are expected to have three or four of the consecutive nucleotides in the motif GGTG. GGTG is a common motif in Archaea, occurring in *Aeropyrum pernix*, where numerous start sites have been verified experimentally (34). In ARMAN-4 and -5, a small percentage of genes use the GAGG end of the Shine-Dalgarno motif. However, more common is the use of a TAA motif. This has previously only been seen in Bacteria (35), Crenarchaea (*Sulfolobus tokodaii* TAAA at 13–15 bp, *Nitrosopumilus maritimus* AATAA at 13–15 bp, and *Caldivirga maquilingensis* TAA at 13–15 bp, weak motif), and *N. equitans* (TAAAA at 5–10 bp and TATAA 13–15 bp weak motif). The TAA motifs were found 3 to 15 bp upstream of genes in ARMAN-4 but only 3 to 10 bp upstream in ARMAN-5.

ARMAN-2 has a full complement of archaeal DNA-dependent RNA polymerase subunits. The *rpoA* and *rpoB* genes, encoding the largest subunits, are present in a single operon along with *rpoH*. Several transcriptional archaeal and bacterial-type regulators

including *asnC/lrp*, *trmB*, *arsR*, *padR*, and *marR* were identified in the genome of ARMAN-2.

In the ARMAN-2 genome we identified a minimal set of aminoacyl-tRNA synthetases for 20 amino acids, including Asn-tRNA (Asn) and Gln-tRNA (Gln) synthetases. We did not identify a Sep-tRNA (Sec) synthetase. Aminoacyl-tRNA synthetases were one of the largest classes of proteins detected by proteomics (Tables S4–S6). tRNAscan-SE (36) identified 35 tRNA genes in the ARMAN-2 genome. Thirteen ARMAN-2 tRNA genes contain introns, but introns are only found in six and five ARMAN-4 and -5 tRNA genes, respectively. Interestingly, no genes were identified for tRNA(Cys) in any of the ARMAN genomes. Although it is possible that the variants of this gene from all three ARMANs were among the portions of the genomes that are currently unsampled, it is more likely that tRNA(Cys) variants were missed because their sequences are too divergent to be identified by current search methods. Alternatively, ARMAN may obtain cysteine from other organisms.

Genomic Characteristics Hint at the Importance of Interspecies Interactions. The average lengths of ARMAN genes are small (Table 1), compared with the average gene length for most Archaea of 924 bp. This finding is notable, given general conservation of gene length in most Bacteria and Archaea (37). To rule out effects because of genome fragmentation, we verified the small gene size across several large syntenic fragments. Furthermore, comparison of 293 orthologs between ARMAN-5 and *Pyrococcus abyssi* (GE5) (excluding incomplete ORFs) revealed that the ARMAN-5 genes are 12% smaller. Most of the genes that are short in ARMAN have similar sized homologs in other genomes. Thus, we infer that ARMAN has not evolved unusually short genes but rather, evolution has enriched the genomes of ARMAN in short gene variants. Additionally, ARMAN-4 and -5 appear to have a high percentage of overlapping genes (ARMAN-4 ~18% and ARMAN-5 ~19%). These phenomena contribute to the observed higher-than-average coding density (Fig. 2). The only sequenced genome with a shorter average gene length is *Anaplasma phagocytophilum* (9) (775 bp), but this genome has the normal coding density for Bacteria and Archaea.

Interestingly, other organisms with small average gene sizes, *N. equitans* (824 bp) (10) and *Neorickettsia sennetsu* (804 bp) (9), also have small cell volumes. ARMAN-2 has three split genes, with the halves in different parts of the genome, as does *N. equitans* (10). We found three noncontiguous split genes in ARMAN-2: transcription initiation factor TFIID (TATA box binding protein) (UNLARM2_0715 and UNLARM2_0193), threonyl-tRNA synthetase (the first 137 aa is in UNLARM2_0834 and the rest is in UNLARM2_0226) and tryptophanyl-tRNA synthetase (UNLARM2_0333 and UNLARM2_0034). The Threonyl-tRNA synthetase protein (UNLARM2_0226) was detected in the UBA-BS proteome (two spectral counts). These genes are not split in ARMAN-4 and -5. All ARMAN have an unusual $tRNA^{Ile}$ with an UAU anticodon predicted to code for ATA (38), as well as CAU, commonly found in Archaea. This also occurs in *N. equitans* and *Ca*. K. crytopfilum. However, unlike in *N. equitans*, no split tRNA genes were identified (39, 40).

Characteristics of the ARMAN genomes, such as split genes, low GC content, frequent overlapping genes, a large number of hypothetical proteins, small genome size, and unusually short genes (13), are shared by host-associated/symbiotic microbes. These features suggest that ARMAN may depend upon another community member for some fraction of its resources.

Interactions with Other Community Members. We examined organismal associations in biofilm samples using 3D cryo-transmission electron microscopy. ARMAN cells are readily identifiable because they are small and have a distinct cell wall (21, 23). The vast majority of cells (over 500 imaged) were not directly interacting with other organism types (Fig. 3). However, in some 3D tomographic reconstructions, the ARMAN cell wall is penetrated by cell wall-less Archaea of the *Thermoplasmatales* lineage (Fig. 4 and additional examples in Fig. S4, and Movies S1, S2, and S3). This interaction could involve injection of nutrients from *Thermoplasmatales* cells into ARMAN, parasitism of ARMAN, or exchange of molecules between them. Notable features distinguishing ARMAN–*Thermoplasmatales* interactions from those reported previously for *N. equitans* and *I. hospitalis* (10, 11, 14) include the rarity of cell-to-cell interactions and the cytoplasmic connection between ARMAN and the associated cell.

No CRISPR loci were found in any of the genomes, but ARMAN-4 has one possible CRISPR-associated protein (BJBARM5_1007). The lack of this viral defense system (41) may explain why ARMAN cells are often infected with one or two morpho-types of viruses (4). A 11.3

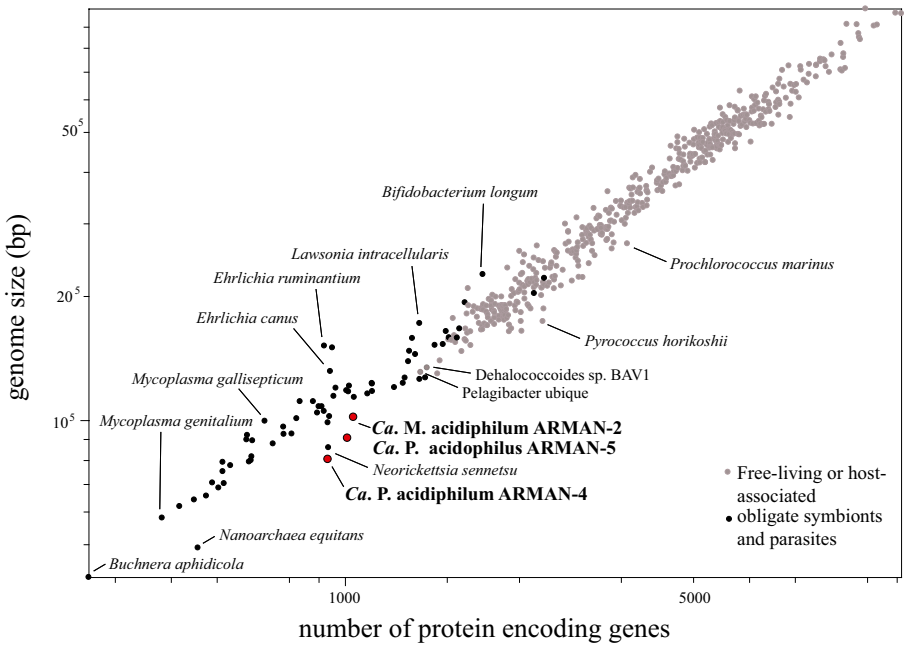

Fig. 2. Plot of archaeal and bacterial genomes (from National Center for Biotechnology Information database) sizes versus the number of protein encoding genes per genome.

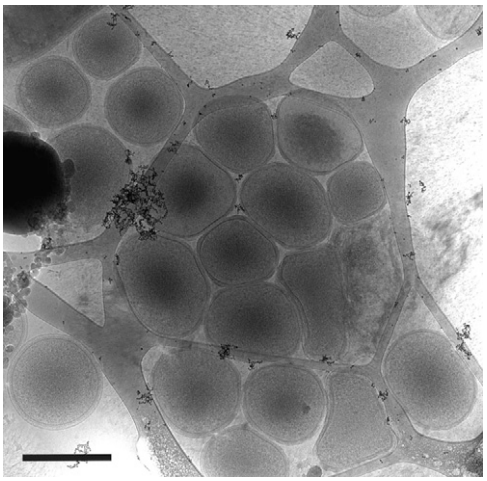

Fig. 3. A cryo-electron micrograph of the biofilm. Notice that the ARMAN cells are not connected to *Thermoplasmatales* cells present in this area. (Scale bar, 500 nm.)

kb provirus is encoded on an ARMAN-5 contig (1668). The provirus includes a predicted protein with homologs in *Ferroplasma acidarmanus*, *Sulfolobus islandicus*, and in a *S. islandicus* rod-shape virus.

All ARMAN genomes have several genes inferred to play a role in pilus assembly. The biosynthesis of pili by ARMAN has been identified by cryo-electron microscopy (23), although no ARMAN plasmids have been documented to date.

Community Proteomic Analyses of ARMAN Activity Levels. We identified a relatively small number of proteins from ARMAN by proteomics compared with other microbes in the biofilm (Table S7). As two of proteomic datasets were acquired from the same biofilms from which genomic data derived (UBA and UBA-BS), this observation cannot be explained by sequence divergence relative to the reference database. The most ARMAN-2 proteins were identified in the UBA-BS biofilm, consistent with the higher abundance of this organism in the genomic dataset from that sample. ARMAN protein identification was not obscured by an anomalously low or high number of tryptic cleavage sites (lysine and arginine), pI, or lengths. Another factor that could lower proteome representation is extensive posttranslational modification, but there is no reason to expect higher-than-normal levels of modifications in these compared to other organisms. Assuming that experimental factors do not explain the low protein identification rates and taking the approximately five times smaller cell volume of ARMAN compared with Bacteria and other Archaea into consideration, the very low proteomic (0.7–1.8% of peptides identified in the UBA-BS proteome) relative to genomic representation (~17% of cells represented in the UBA-BS genomic dataset) implies low activity levels of ARMAN relative to other organisms in the community. The notably low abundance of ribosomal proteins relative to other proteins (Tables S4 – S6) is consistent with the documented small number of ribosomes per cell (23) and also supports this inference.

Conclusions

Over the past decade, surveys of genes and gene fragments have indicated the existence of vast microbial diversity in natural ecosystems (15), but there are still many lineages about which we know very little. Deeply divergent lineages are of particular interest, as they provide insight into the form and pattern of biological evolution. In the present study, we acquired genomic and proteomic information for three such archaeal lineages. The data further expand the number and variety of known genes, some of which must be involved in newly described interorganism interactions and formation of large internal organelles of unknown function. Interestingly, these Euryarchaea have many genes with

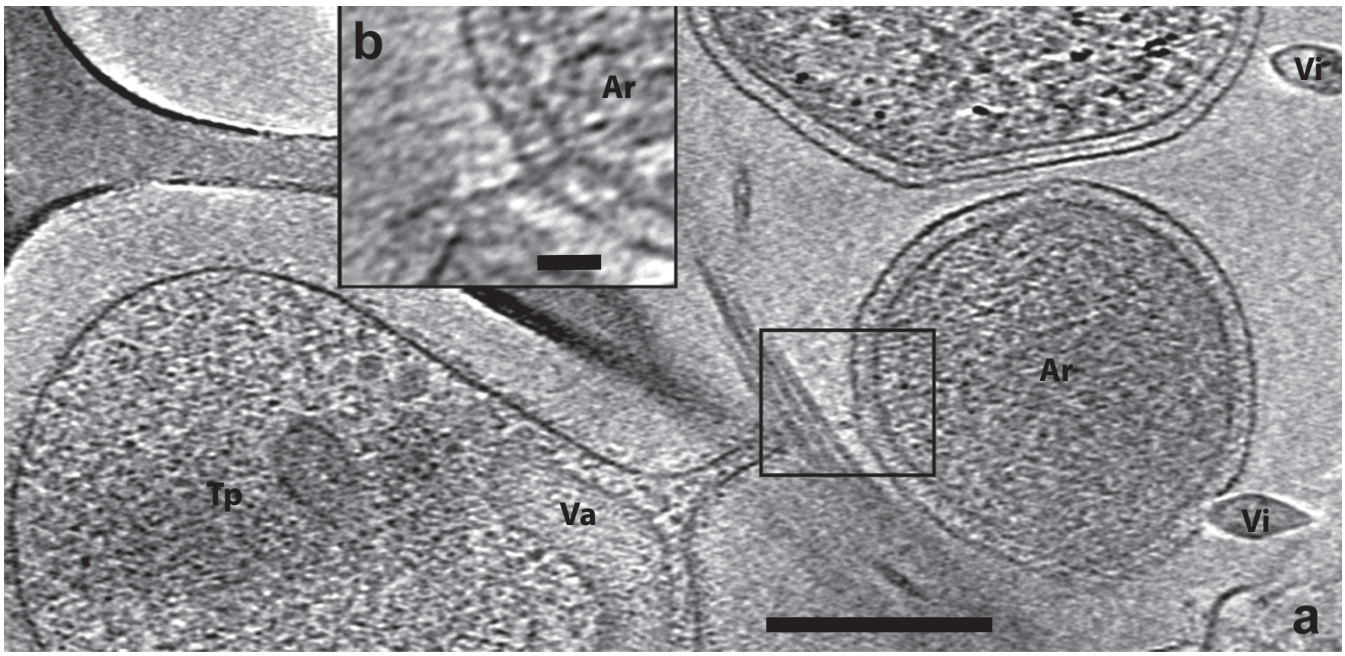

Fig. 4. Image illustrating interaction between ARMAN and a *Thermoplasmatales* lineage archaeon in the community. (*A*) A 1-pixel-thick slice through a cryo-electron tomographic reconstruction documenting interaction between an ARMAN cell on the right and a *Thermoplasmatales* lineage cell on the left. Also shown is the previously reported association of viruses and ARMAN cells (*Upper* and *Lower*, *Right*) and the presence of vacuoles in the single membrane-bounded *Thermoplasmatales* cell. A different slice through the reconstruction (2 pixels thick) shown in *B* illustrates the penetration of the ARMAN wall by the *Thermoplasmatales* cell. For a 3D movie of this field of view, see Movie S1. Also see Movie S2 and Movie S3 for other 3D examples of these connections. Ar, ARMAN; Tp, *Thermoplasmatales* lineage cell; Va, vacuole; VI, virus. (Scale bar in *A*, 300 nm; in *B*, 50 nm.)

closest similarity to genes from Crenarchaea and Bacteria, consistent with their early divergence from their common ancestor.

Materials and Methods

Sample Collection, DNA Extraction, and Genome Reconstruction. All samples for DNA extraction were collected from the A drift of the Richmond Mine (24) at Iron Mountain in northern California. The UBA-BS sample collected in November 2005 (38 °C) was an ≈1-cm-thick, pink, floating biofilm with a gelatinous texture. DNA extraction, library construction, and sequencing follow methods reported previously (24). For details of filtrate preparation from the UBA biofilm collected at the base of the UBA waterfall (March 2005), see Baker et al. (21), and for DNA extraction details see Lo et al. (24). MDA amplifications of genomic DNA from the filtrate used GenomiPhi kit V2 (GE Healthcare). MDA amplification products were screened by PCR using archaeal- (23F and 1492R), bacterial- (27F and 1492R), and ARMAN-specific (ARM979F 5′-TATTACCAGAAGCGACGGC-3′ and ARM1365R 5′-AGGGACGTATTCACCGCTCG-3′) primer sets (21). The product with the least-visible amplification with archaeal and bacterial primers and the most ARMAN PCR product visible on an agarose gel was chosen for small-insert cloning. For details about genome assembly and annotation, see *SI Materials and Methods*.

Proteomics. For global analysis of ARMAN protein expression, we analyzed previously published proteomic data from seven AMD biofilms: ABend (January 2004), ABfront (June 2004), UBA (June 2005), ABmuck (November 2006), ABmuck Friable (November 2006), UBA-BS (November 2005), and UBA-BS2 (August 2007). Comprehensive genomic data area available for the UBA and UBA-BS samples (5, 24). Samples were collected on site, frozen on dry ice, and transported back to the laboratory; cells were lysed and fractionated into extracellular, whole-cell, soluble and membrane fractions [see Ram et al. (28) for details]. Either a neutral buffer (M2) or acidic buffer (S buffer) was used during lysis; in most cases, both buffers were used (generating additional datasets). Technical triplicates were analyzed via shotgun proteomics with nano-LC-MS/MS on either a linear ion trap (LTQ) or high resolution LTQ-Orbitrap (both Thermo-Fisher Scientific). All datasets were searched with SEQUEST (42) against a composite community database AMD_CoreDB_04232008; a subset of datasets from samples known to have relatively high ARMAN abundance (UBA, UBA-BS, and UBA-BS2) was additionally searched against a database containing ARMAN-4 and -5 sequences called in multiple frames (amdv1allfrm_arman_AMD_CoreDB_04232008). All data analyses methods are as previously described and all datasets were filtered with at least two peptides per protein and have been shown to have a very low false-positive rate, generally 1 to 2% [described in detail previously (16, 24, 28)].

Cryo-Electron Microscope Specimen Preparation. For cryo-transmission electron microscope characterization, aliquots of 5 μL were taken directly from fresh biofilm samples and placed onto lacey carbon grids (Ted Pella 01881) that were pretreated by glow-discharge. The support grids were preloaded with 10-nm colloidal gold particles that serve as reference points for 3D reconstruction. The Formvar support was not removed from the lacey carbon. The grids were manually blotted and plunged into liquid ethane by a compressed air piston, then stored in liquid nitrogen. For details on cryo-electron micrscope imaging, see *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Gene Tyson and Eric Allen for assistance with filtration and genomic DNA preparations, Mr. Ted Arman (President, Iron Mountain Mines), Mr. Rudy Carver, and Dr. Richard Sugarek for site access and other assistance, Dr. Sussanah Tringe for sequencing logistics, Manesh Shah for assistance with proteomic analyses, and Dr. Hans Truper for the naming of the ARMAN groups. This work was funded by the US Department of Energy's Office of Science, Biological and Environmental Research Program (DOE Genomics:GTL project Grant DE-FG02-05ER64134), the National Aeronautics and Space Administration Astrobiology Institute, and by Laboratory Directed Research and Development support from the University of California, Lawrence Berkeley National Laboratory. This work was also supported by the Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under Contract DE-AC02-05CH11231. The sequencing was provided through the Community Sequencing Program at the Department of Energy Joint Genome Institute.

1. Wilmes P, Simmons SL, Denef VJ, Banfield JF (2009) The dynamic genetic repertoire of microbial communities. *FEMS Microbiol Rev* 33:109–132.
2. Béjà O, et al. (2000) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* 2:516–529.
3. Rusch DB, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5:e77.
4. Venter JC, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74.
5. Dick GJ, et al. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol* 10:R85.
6. Hallam SJ, et al. (2006) Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc Natl Acad Sci USA* 103:18296–18301.
7. Tyson GW, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43.
8. Allen EE, et al. (2007) Genome dynamics in a natural archaeal population. *Proc Natl Acad Sci USA* 104:1883–1888.
9. Hotopp JC, et al. (2006) Comparative genomics of emerging human ehrlichiosis agents. *PLoS Genet* 2:e21.
10. Waters E, et al. (2003) The genome of *Nanoarchaeum equitans*: Insights into early archaeal evolution and derived parasitism. *Proc Natl Acad Sci USA* 100:12984–12988.
11. Podar M, et al. (2008) A genomic analysis of the archaeal system *Ignicoccus hospitalis-Nanoarchaeum equitans*. *Genome Biol* 9:R158.
12. Ochman H (2005) Genomes on the shrink. *Proc Natl Acad Sci USA* 102:11959–11960.
13. Nakabachi A, et al. (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314:267.
14. Junglas B, et al. (2008) *Ignicoccus hospitalis* and *Nanoarchaeum equitans*: Ultrastructure, cell-cell interaction, and 3D reconstruction from serial sections of freeze-substituted cells and by electron cryotomography. *Arch Microbiol* 190:395–408.
15. DeLong EF, Pace NR (2001) Environmental diversity of bacteria and archaea. *Syst Biol* 50:470–478.
16. Denef VJ, et al. (2009) Proteomics-inferred genome typing (PIGT) demonstrates inter-population recombination as a strategy for environmental adaptation. *Environ Microbiol* 11:313–325.
17. Edwards KJ, Bond PL, Gihring TM, Banfield JF (2000) An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. *Science* 287:1796–1799.
18. Druschel GK, Baker BJ, Gihring TM, Banfield JF (2004) Acid mine drainage biogeochemistry at Iron Mountain, California. *Geochem Trans* 5:13–32.
19. Baker BJ, Banfield JF (2003) Microbial communities in acid mine drainage. *FEMS Microbiol Ecol* 44:139–152.
20. Baker BJ, et al. (2006) Lineages of acidophilic archaea revealed by community genomic analysis. *Science* 314:1933–1935.
21. Juottonen H, Tuittila E-S, Juutinen S, Fritze H, Yrjälä K (2008) Seasonality of rDNA- and rRNA-derived archaeal communities and methanogenic potential in a boreal mire. *ISME J* 2:1157–1168.
22. Sunna A, Bergquist PL (2003) A gene encoding a novel extremely thermostable 1,4-beta-xylanase isolated directly from an environmental DNA sample. *Extremophiles* 7:63–70.
23. Comolli LR, Baker BJ, Downing KH, Siegerist CE, Banfield JF (2009) Three-dimensional analysis of the structure and ecology of a novel, ultra-small archaeon. *ISME J* 3:159–167.
24. Lo I, et al. (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* 446:537–541.
25. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol* 8:R10.
26. Tatusov RL, et al. (2003) The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4:41.
27. Ditzel L, et al. (1998) Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT. *Cell* 93:125–138.
28. Ram RJ, et al. (2005) Community proteomics of a natural microbial biofilm. *Science* 308:1915–1920.
29. Bigotti MG, Clarke AR (2005) Cooperativity in the thermosome. *J Mol Biol* 348:13–26.
30. Myllykallio H, et al. (2002) An alternative flavin-dependent mechanism for thymidylate synthesis. *Science* 297:105–107.
31. Elkins JG, et al. (2008) A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc Natl Acad Sci USA* 105:8102–8107.
32. Anderson I, et al. (2008) Genome sequence of *Thermofilum pendens* reveals an exceptional loss of biosynthetic pathways without genome reduction. *J Bacteriol* 190:2957–2965.
33. McGoldrick HM, et al. (2005) Identification and characterization of a novel vitamin B12 (cobalamin) biosynthetic enzyme (CobZ) from *Rhodobacter capsulatus*, containing flavin, heme, and Fe-S cofactors. *J Biol Chem* 280:1086–1094.
34. Yamazaki S, et al. (2006) Proteome analysis of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1. *Mol Cell Proteomics* 5:811–823.
35. Chen S, Bagdasarian M, Kaufman MG, Walker ED (2007) Characterization of strong promoters from an environmental *Flavobacterium hibernum* strain by using a green fluorescent protein-based reporter system. *Appl Environ Microbiol* 73:1089–1100.
36. Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 33:W686–W689.
37. Xu L, et al. (2006) Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Mol Biol Evol* 23:1107–1108.
38. Marck C, Grosjean H (2002) tRNomics: Analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA* 8:1189–1232.
39. Randau L, Münch R, Hohn MJ, Jahn D, Söll D (2005) *Nanoarchaeum equitans* creates functional tRNAs from separate genes for their 5′- and 3′-halves. *Nature* 433:537–541.
40. Huber H, et al. (2002) A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* 417:63–67.
41. Barrangou R, et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315:1709–1712.
42. Eng JK, McCormack AL, Yates I, John R (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5:976–989.