# Community structure and metabolism through reconstruction of microbial genomes from the environment

**Gene W. Tyson[1], Jarrod Chapman[3,4], Philip Hugenholtz[1], Eric E. Allen[1], Rachna J. Ram[1], Paul M. Richardson[4], Victor V. Solovyev[4], Edward M. Rubin[4], Daniel S. Rokhsar[3,4] & Jillian F. Banfield[1,2]**

[1]*Department of Environmental Science, Policy and Management,* [2]*Department of Earth and Planetary Sciences, and* [3]*Department of Physics, University of California, Berkeley, California 94720, USA*
[4]*Joint Genome Institute, Walnut Creek, California 94598, USA*

**Microbial communities are vital in the functioning of all ecosystems; however, most microorganisms are uncultivated, and their roles in natural systems are unclear. Here, using random shotgun sequencing of DNA from a natural acidophilic biofilm, we report reconstruction of near-complete genomes of *Leptospirillum* group II and *Ferroplasma* type II, and partial recovery of three other genomes. This was possible because the biofilm was dominated by a small number of species populations and the frequency of genomic rearrangements and gene insertions or deletions was relatively low. Because each sequence read came from a different individual, we could determine that single-nucleotide polymorphisms are the predominant form of heterogeneity at the strain level. The *Leptospirillum* group II genome had remarkably few nucleotide polymorphisms, despite the existence of low-abundance variants. The *Ferroplasma* type II genome seems to be a composite from three ancestral strains that have undergone homologous recombination to form a large population of mosaic genomes. Analysis of the gene complement for each organism revealed the pathways for carbon and nitrogen fixation and energy generation, and provided insights into survival strategies in an extreme environment.**

The study of microbial evolution and ecology has been revolutionized by DNA sequencing and analysis[1–3]. However, isolates have been the main source of sequence data, and only a small fraction of microorganisms have been cultivated[4–6]. Consequently, focus has shifted towards the analysis of uncultivated microorganisms via cloning of conserved genes[5] and genome fragments directly from the environment[7–9]. To date, only a small fraction of genes have been recovered from individual environments, limiting the analysis of microbial communities as networks characterized by symbioses, competition and partitioning of community-essential roles. Comprehensive genomic data would resolve organism-specific pathways and provide insights into population structure, speciation and evolution. So far, sequencing of whole communities has not been practical because most communities comprise hundreds to thousands of species[10].

Acid mine drainage (AMD) is a worldwide environmental problem that arises largely from microbial activity[11]. Here, we focused on a low-complexity AMD microbial biofilm growing hundreds of feet underground within a pyrite ($FeS_2$) ore body[12–15]. This represents a self-contained biogeochemical system characterized by tight coupling between microbial iron oxidation and acidification due to pyrite dissolution[11,16,17]. Random shotgun sequencing of DNA from entire microbial communities is one approach for the recovery of the gene complement of uncultivated organisms, and for determining the degree of variability within populations at the genome level. We used random shotgun sequencing of the biofilm to obtain the first reconstruction of multiple genomes directly from a natural sample. The results provide novel insights into community structure, and reveal the strategies that underpin microbial activity in this environment.

## Initial characterization of the biofilm

Biofilms growing on the surface of flowing AMD in the five-way region of the Richmond mine at Iron Mountain, California[12], were sampled in March 2000. Screening using group-specific[18] fluorescence *in situ* hybridization (FISH) revealed that all biofilms contained mixtures of bacteria (*Leptospirillum*, *Sulfobacillus* and, in a few cases, *Acidimicrobium*) and archaea (*Ferroplasma* and other members of the Thermoplasmatales). The genome of one of these archaea, *Ferroplasma acidarmanus* fer1, isolated from the Richmond mine, has been sequenced previously (http://www.jgi.doe.gov/JGI_microbial/html/ferroplasma/ferro_homepage.html).

A pink biofilm (Fig. 1a) typical of AMD communities was selected for detailed genomic characterization (see Supplementary Information). The biofilm was dominated by *Leptospirillum* species and contained *F. acidarmanus* at a relatively low abundance (Fig. 1b, c). This biofilm was growing in pH 0.83, 42 °C, 317 mM Fe, 14 mM Zn, 4 mM Cu and 2 mM As solution, and was collected from a surface area of approximately 0.05 $m^2$.

A 16S ribosomal RNA gene clone library was constructed from DNA extracted from the pink biofilm, and 384 clones were end-sequenced (see Supplementary Information). Results indicated the presence of three bacterial and three archaeal lineages. The most abundant clones are close relatives of *L. ferriphilum*[19] and belong to *Leptospirillum* group II (ref. 13). Although 94% of the *Leptospirillum* group II clones were identical, 17 minor variants were detected with up to 1.2% 16S rRNA gene-sequence divergence from the dominant type. Tightly defined groups (up to 1% sequence divergence) related to *Leptospirillum* group III (ref. 13), *Sulfobacillus*, *Ferroplasma* (some identical to fer1), 'A-plasma'[15] and 'G-plasma'[15] were also detected. *Leptospirillum* group III, G-plasma and A-plasma have only recently been detected in culture-independent molecular surveys. FISH-based quantification (Fig. 1c; see also Supplementary Information) confirmed the dominance of *Leptospirillum* group II in the biofilm.

## Community genome sequencing and assembly

In conventional shotgun sequencing projects of microbial isolates, all shotgun fragments are derived from clones of the same genome. When using the shotgun sequencing approach on genomes from an

environmental sample, however, variation within each species population might complicate assembly. If intraspecies variation is dominated by limited local polymorphism or homologous recombination, it should be possible to define a composite genome for each species population. Conversely, if the genomic heterogeneity within a species is dominated by large rearrangements, deletions, or insertions, it may be impossible to define composite genomes for species populations from natural communities.

A small insert plasmid library (average insert size 3.2 kilobases (kb)) was constructed from the biofilm DNA for random shotgun sequencing (see Supplementary Information). A total of 76.2 million base pairs (bp) of DNA sequence was generated from 103,462 high-quality reads (averaging 737 bp per read). Analysis of raw shotgun data (Supplementary Figs S1–5) indicated the presence of both bacterial and archaeal genomes at sequence coverages of up to 10×, which would be sufficient to produce a high-quality assembly from a conventional microbial genome project[20,21]. The shotgun data set was assembled with JAZZ, a whole-genome shotgun assembler[22]. Anticipating polymorphisms, we permitted alignment discrepancies beyond those expected from sequencing error if they were consistent with end-pairing constraints. Over 85% of the shotgun reads were assembled into scaffolds longer than 2 kb (a scaffold is a reconstructed genomic region that may contain gaps of a known size range). The combined length of the 1,183 scaffolds is 10.83 megabases (Mb). The assembly is internally self consistent, with 97.2% of end pairs from the same clone assembled with the appropriate orientation and separation, as expected for a low rate of mispairing error (tracking and chimaeric clones).

The first step in assignment of scaffolds to organism types was to separate the scaffolds by average G+C content. These were subsequently subdivided using read depth (coverage). Dinucleotide frequencies did not allow for further subdivision. Notably, separation of scaffolds into low G+C (<43.5%; Supplementary Fig. S3a) and high G+C (≥43.5%) content 'bins' was not significantly compromised by local heterogeneities in G+C content because the scaffolds were binned after assembly. As the scaffolds are typically tens of kilobases long, local fluctuations in G+C content are averaged over the length of each scaffold, allowing, in most cases (>99%), clear assignment to bins of high or low G+C content.

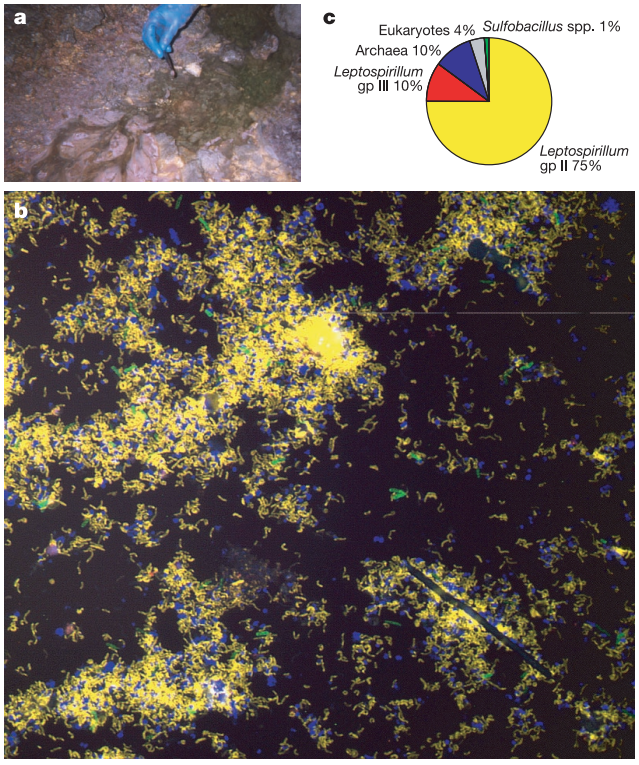


**Figure 1** The pink biofilm. **a**, Photograph of the biofilm in the Richmond mine (hand included for scale). **b**, FISH image of **a**. Probes targeting bacteria (EUBmix; fluorescein isothiocyanate (green)) and archaea (ARC915; Cy5 (blue)) were used in combination with a probe targeting the *Leptospirillum* genus (LF655; Cy3 (red)). Overlap of red and green (yellow) indicates *Leptospirillum* cells and shows the dominance of *Leptospirillum*. **c**, Relative microbial abundances determined using quantitative FISH counts.

The high G+C scaffolds at approximately 10× coverage (70 scaffolds up to 137 kb in length, totalling 2.23 Mb) were identified by the presence of a single 16S rRNA gene as belonging to the genome of a *Leptospirillum* group II species. The average G+C content (55.8%) is comparable to the G+C content (54.9–58%) of *L. ferriphilum*[19]. The total high G+C scaffold length is close to the estimated genome size of *Leptospirillum ferrooxidans*[23] (1.9 Mb). This suggests that essentially the entire *Leptospirillum* group II genome was recovered from the community DNA.

The low G+C scaffolds at approximately 10× coverage were assembled into 59 scaffolds of up to 138 kb in length, totalling 1.82 Mb. The single 16S rRNA gene identified in these scaffolds was 99% identical to that of the fer1 isolate; however, alignment of the scaffolds to the fer1 genome revealed an average of 22% divergence at the nucleotide level (Supplementary Fig. S6). The total scaffold length is close to the genome size of fer1 (1.9 Mb; Allen *et al.*, unpublished data), and local gene order and content are highly conserved (Supplementary Fig. S7). Therefore, these 59 scaffolds represent a nearly complete genome of a previously unknown, uncultured *Ferroplasma* species distinct from fer1. We designate this as *Ferroplasma* type II. The dominance of this organism type was unexpected before the genomic analysis.

We assigned the roughly 3× coverage, high G+C scaffolds to *Leptospirillum* group III on the basis of rRNA markers (474 scaffolds up to 31 kb, totalling 2.66 Mb). Comparison of these scaffolds with those assigned to *Leptospirillum* group II indicates significant sequence divergence and only locally conserved gene order, confirming that the scaffolds belong to a relatively distant relative of *Leptospirillum* group II. A partial 16S rRNA gene sequence from *Sulfobacillus thermosulfidooxidans* was identified in the unassembled reads, suggesting very low coverage of this organism. If any *Sulfobacillus* scaffolds >2 kb were assembled, they would be grouped with the *Leptospirillum* group III scaffolds.

We compared the 3× coverage, low G+C scaffolds (580 scaffolds, 4.12 Mb) to the fer1 genome in order to assign them to organism types (Supplementary Fig. S6). Scaffolds with ≥96% nucleotide identity to fer1 were assigned to an environmental *Ferroplasma* type I genome (170 scaffolds up to 47 kb in length and comprising 1.48 Mb of sequence). The remaining low-coverage, low G+C scaffolds are tentatively assigned to G-plasma. The largest scaffold in this bin (62 kb) contains the G-plasma 16S rRNA gene. The 410 scaffolds assigned to G-plasma comprise 2.65 Mb of sequence. A partial 16S rRNA gene sequence from A-plasma was identified in the unassembled reads, suggesting low coverage of this organism. Any scaffolds from A-plasma >2 kb would be included in the G-plasma bin. Although eukaryotes are present in the AMD system, they were in low abundance in the biofilm studied. So far, no scaffolds from eukaryotes have been detected.

As independent evidence that the *Leptospirillum* group II and *Ferroplasma* type II genomes are nearly complete, we located a full complement of transfer RNA synthetases in each genome data set. An almost complete set of these genes was also recovered from *Leptospirillum* group III. The G-plasma bin contains more than a full set of tRNA synthetases, consistent with inclusion of some A-plasma scaffolds. In addition, we established that the *Leptospirillum* group II, *Leptospirillum* group III, *Ferroplasma* type I, *Ferroplasma* type II and G-plasma bins contained only one set of rRNA genes.

These results, and the agreement between the recovered and anticipated genome sizes, confirm that dividing scaffolds by G+C content, read depth and homology to the fer1 genome is a valid means of sorting most genomes from this community data set. Methods for the analysis of genome signatures currently under development may be required for binning genome data from more complex communities[24,25].

## Population structure and speciation

The biofilm sample contained approximately $10^8$ *Leptospirillum* group II cells. Thus, the roughly 29,000 reads assembled into the *Leptospirillum* group II genome probably all came from different individuals. Despite this, there were no nucleotide polymorphisms in the 16S or 23S rRNA genes, or in their intergenic region. The average nucleotide polymorphism rate in the *Leptospirillum* group II genome is 0.08%; approximately one-third of the polymorphisms cause changes at the protein-coding level. The low incidence of nucleotide polymorphisms indicates that we have sequenced essentially a single strain from the community. Homogeneity within the *Leptospirillum* group II genome may reflect strong recent environmental selection for this genome type or be the result of a founder effect.

There are no nucleotide polymorphisms in the rRNA genes of *Ferroplasma* type II, despite an average polymorphism rate of about 2.2% (see Supplementary Information). Polymorphism-free regions typically a few hundred base pairs long and up to several kilobases in length occur one to tens of kilobases apart on the *Ferroplasma* type II scaffolds (see Supplementary Information). Although these homogenous regions may contain conserved genes, there is no consistency in the function of proteins encoded by them (for example, Fig. 2a; see also Supplementary Fig. S8).

In any given region, typically between one and three distinct patterns of nucleotide polymorphism were observed in the assembled *Ferroplasma* type II composite genome (Fig. 2b). By using sequence read end-pair information, these nucleotide polymorphism patterns could be connected across homogenous regions. We could identify points within individual reads and between end pairs where one distinct nucleotide polymorphism pattern transitioned into another (Fig. 2b; see also Supplementary Fig. S8). The most likely explanation for this, and for the larger homogeneous regions, is that the *Ferroplasma* type II strains have undergone homologous recombination. It is unlikely that the reads with pattern transitions represent variants that arose simply through accumulation of nucleotide polymorphisms, because this

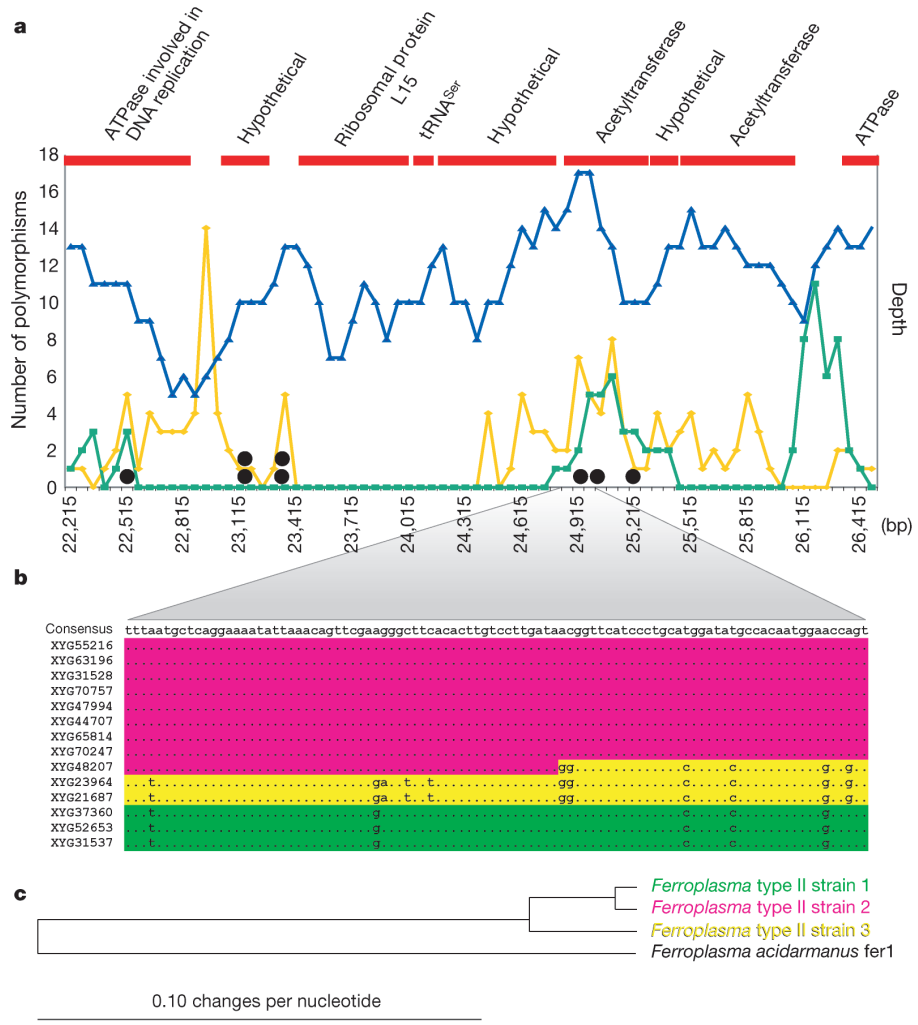


**Figure 2** Segment of the *Ferroplasma* type II composite genome. **a**, A 4.2-kb region showing annotated open reading frames (ORFs) (red), average read depth (blue line), and the number of nucleotide polymorphisms in the 'green' and 'yellow' relative to the 'pink' strain (green and yellow lines) averaged over 60-bp windows. Black dots indicate recombination sites. **b**, Alignment of individual reads (XYG) for a 96-bp region in **a**. Letters indicate nucleotide polymorphisms in the green and yellow strains relative to the pink strain. Note the recombinant sequence (XYG48207). **c**, Evolutionary distance tree inferred from the ancestral strain sequences in **a**.

would require precise selection acting on small degrees of heterogeneity at virtually every locus.

Assembled reads in regions of up to 10 kb chosen at random on *Ferroplasma* type II scaffolds generally contained about 20 transition points, interpreted to be recombination boundaries (see Supplementary Fig. S9). If the characteristics of these regions are typical, then switching of nucleotide polymorphism patterns occurs on average about every 5 kb, and the mosaic genomes of currently existing strains were constructed via at least 400 recombination events. It is impossible to obtain a detailed representation of the strain genome structure or relative abundances of individual types because we cannot directly link polymorphism patterns over long genomic regions with small insert library data.

At the time of sampling, the *Ferroplasma* type II species population seemed to be dominated by strains with mosaic genomes constructed by recombination of three closely related but distinct genome types (pink, green and yellow in Fig. 2b and Supplementary Fig. S9) that we infer correspond to three 'ancestral' strains (Fig. 3). Other ancestral types may have existed, and may have given rise to variants that are too rare to be detected by our analysis. Combinatorial variants such as these have been observed previously in populations of enteric bacteria from disparate locations[26] but they have not been documented in archaea or environmental samples from a single location.

Phylogenetic analysis of genome fragments reconstructed for the three ancestral strains reveals that they derived from a relatively recent common ancestor (Fig. 2c). On the basis of the overall strong internal consistency within the nucleotide polymorphism patterns, we infer that the most recent evolution of the *Ferroplasma* type II population has been dominated by homologous recombination. If the population were to undergo further recombination (a likely scenario), three distinct ancestral strains would still be identifiable at the local scale. Therefore, it is not possible to determine how many episodes of recombination have occurred in the population.

Sequences reconstructed for the three ancestral types (Fig. 2b) were used to calculate relative polymorphism frequencies (Fig. 2a). Most of the proteins encoded by the sequences were slightly different at the amino acid level. In one region (Supplementary Fig. S9) the nucleotide polymorphism rates were 0.6% for the green compared with pink (56% non-synonymous) and 3% for yellow compared with pink (44% non-synonymous). This suggests that recombination involving fragments carrying slightly different protein-coding sequences yields a very large number of genomic combinatorial variants (Fig. 3) with subtly different metabolic characteristics. The existence of mosaic genome types has ecological and evolutionary significance because genome diversity due to extensive recombination would ensure availability of an optimized strain when the system is perturbed, conferring resilience to the species.

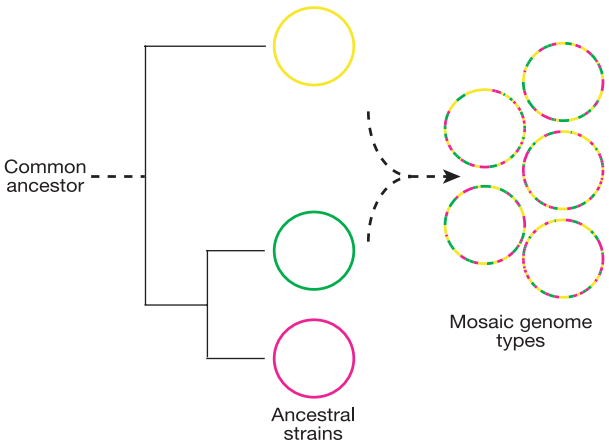


**Figure 3** Schematic diagram illustrating a diversity of mosaic genome types within the *Ferroplasma* type II population that are inferred to have arisen by homologous recombination between three closely related ancestral genome types (pink, yellow and green).

The frequency of recombination between organisms decreases exponentially as they become more divergent[27]. Unlike the *Ferroplasma* type II strains, the *Ferroplasma* type I and II genomes have no anomalous regions of high sequence identity, indicating that they have not undergone recent recombination (at least on length scales smaller than the scaffold size). On the basis of the low recombination rate and the separation of these genomes from each other by assembly, we predict that *Ferroplasma* type I and II are physiologically distinct, and thus are separate species. Recombination and assembly may provide useful genome-based criteria to separate species from strains in cases where one or both organisms are uncultivated.

The combination of a 16S rRNA-based survey with comprehensive genomic sampling provides a snapshot of the population structure. We have not yet determined how stable the community structure is, or the factors responsible for the success of the observed strains. However, an important finding is the dominance of the biofilm by a handful of distinct genome types. A few organism types within a much larger pool of rare types is typical of lognormal abundance distributions in other natural communities[28–30]. This may be attributed to a small number of niches within the AMD system at any one time, possibly because the ecosystem is relatively geochemically simple (for example, the dominant electron donors and acceptors are iron, sulphur and oxygen, and temperature and fluid composition cycle within a relatively narrow range over annual timescales)[12,14].

## Pathways for genetic exchange

Although there is evidence for genetic exchange between *Ferroplasma* type II strains, it is unclear how recombination is achieved. There is also some evidence of ancient gene transfer between the Sulfolobales and members of the Thermoplasmatales, and between bacteria and archaea (data not shown). Transformation by uptake of naked DNA is unlikely, as DNA is not expected to have a significant residence time in the acid solution. There is no evidence for conjugation genes in *Ferroplasma* type I or II, and there is only limited evidence for transduction (some possible phage genes and integrases). We compared the sequences of the probable prophage genes in order to test whether the host range of phage in the AMD system is large enough to provide a mechanism for lateral gene transfer. Identical reverse transcriptases (LambdaSa1) occur in the G-plasma and *Ferroplasma* type II genomes (in very different genomic contexts), suggesting that a single phage type has recently targeted both lineages. Similarly, identical retron-type reverse transcriptases with identical adjacent transposases occur in otherwise different genomic contexts within the *Leptospirillum* group II and III genomes, indicating that a broad host range phage targets both of these groups.

## Metabolic analysis

We recovered near-complete gene inventories for the five dominant members of the biofilm community. The data for *Leptospirillum* group II are particularly notable, as no genome of a *Nitrospira* phylum member had been sequenced previously. Here we focus on the metabolic pathways recovered in *Leptospirillum* group II and *Ferroplasma* type II (Fig. 4; see also annotation files in the Supplementary Information), and on the new insights into the ecological roles of individual members that have changed our understanding of how the community functions.

The acidophilic biofilms are self-sustaining communities that grow in the deep subsurface and receive no significant



NATURE | doi:10.1038/nature02340 | www.nature.com/nature

inputs of fixed carbon or nitrogen from external sources. As with *Leptospirillum* group I, both *Leptospirillum* group II and III have the genes needed to fix carbon by means of the Calvin–Benson–Bassham cycle (using type II ribulose 1,5-bisphosphate carboxylase–oxygenase). All genomes recovered from the AMD system contain formate hydrogenlyase complexes. These, in combination with carbon monoxide dehydrogenase, may be used for carbon fixation via the reductive acetyl coenzyme A (acetyl-CoA) pathway by some, or all, organisms. Given the large number of ABC-type sugar and amino acid transporters encoded in the *Ferroplasma* type

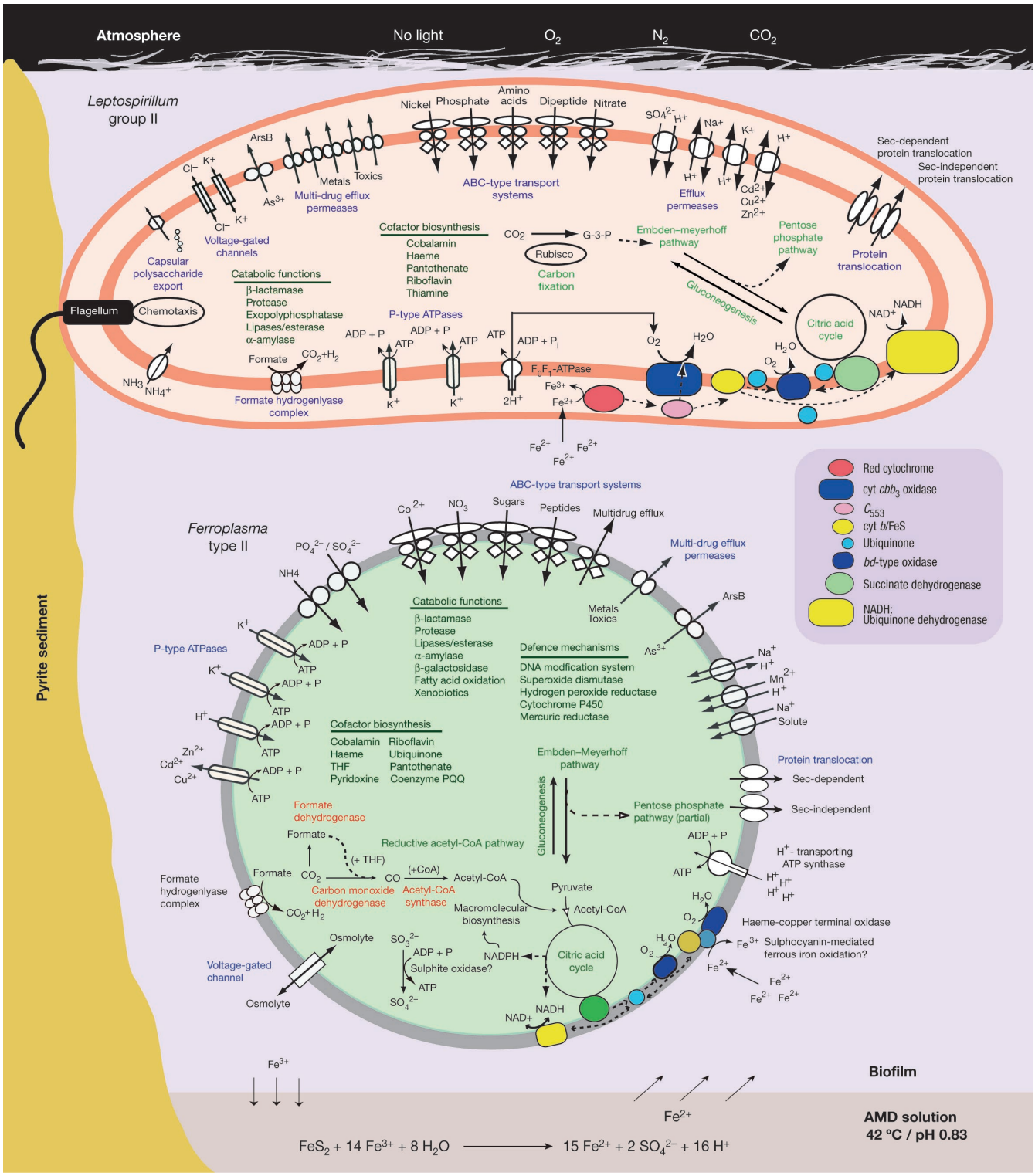


**Figure 4** Cell metabolic cartoons constructed from the annotation of 2,180 ORFs identified in the *Leptospirillum* group II genome (63% with putative assigned function) and 1,931 ORFs in the *Ferroplasma* type II genome (58% with assigned function). The cell cartoons are shown within a biofilm that is attached to the surface of an acid mine drainage stream (viewed in cross-section). Tight coupling between ferrous iron oxidation, pyrite dissolution and acid generation is indicated. Rubisco, ribulose 1,5-bisphosphate carboxylase–oxygenase. THF, tetrahydrofolate.

I and II and G-plasma genomes, these organisms may favour the heterotrophic lifestyle reported for other Thermoplasmatales[31].

Nitrogen fixation is another process essential to the community. Given that *Leptospirillum ferrooxidans* (group I) possesses a complete nitrogen fixation pathway[32] and because *Leptospirillum* group II dominates most biofilms in the AMD system, we suspected that *Leptospirillum* group II would be the primary nitrogen fixer. However, the *Leptospirillum* group II genome does not contain any genes for nitrogen fixation. We detected a complete nitrogen fixation operon with homology to that reported for a *L. ferrooxidans* strain[32] in scaffolds assigned to *Leptospirillum* group III (for example, the *nifH* gene is 84% identical to that of *L. ferrooxidans*). Manual linking of this scaffold to adjacent scaffolds with low coverage supports the assignment of the *nif* operon to *Leptospirillum* group III. In order to confirm this assignment we used primers to amplify *nifH* genes. *nifH* was recovered from the community DNA but not from a mixed culture of *Leptospirillum* group II strains obtained from the site (Tyson *et al.*, unpublished data).

There is no evidence to suggest that *Ferroplasma* type I or II fix nitrogen. It is likely that *Ferroplasma* species obtain nitrogen fixed by other organisms from solution using the numerous amino acid transporters and ammonia permeases (Fig. 4). As the community-essential role of nitrogen fixation is assigned to a relatively low-abundance organism type, *Leptospirillum* group III may be a keystone species in this ecosystem.

The use of ferrous iron oxidation as an energy source directly links microbial and geochemical processes in AMD ecosystems. Iron oxidation has been well studied only in *Acidithiobacillus ferrooxidans*[33–35], despite *Leptospirillum* and *Ferroplasma* being the dominant iron oxidizers in many AMD systems[14]. We have reconstructed a putative electron transport chain for *Leptospirillum* group II (Fig. 4). The genome does not encode the blue copper protein[34] or putative direct ferrous iron oxidase[35,36], which are present in *A. ferrooxidans*; however, a number of novel cytochromes were detected (Ram *et al.*, unpublished data).

All aerobic iron-oxidizing organisms in the AMD system face the challenge of generating energy in microaerophilic solutions. *Leptospirillum* group II and III genomes both encode genes indicative of cytochrome $cbb_3$-type haeme-copper oxidases and cytochrome *bd*-type quinol oxidases, both of which typically have high affinity for oxygen[37,38]. These genes may protect the *Leptospirillum* group III oxygen-sensitive nitrogenase complex during nitrogen fixation, as proposed in other systems[39].

The electron transport chain of *Ferroplasma* type II differs significantly from that in *Leptospirillum* group II (Fig. 4). It contains putative haeme-copper terminal oxidases, cytochrome *b* and associated Rieske iron-sulphur proteins, and a blue copper protein, all of which may be assembled into a terminal oxidase supercomplex similar to SoxM in *Sulfolobus acidocaldarius*[40]. Interestingly, the blue copper protein shares sequence characteristics with both *A. ferrooxidans* rusticyanins and *Sulfolobus* sulphocyanins, thus it may be a component of the electron transport chain during iron oxidation as well as heterotrophic growth. As with *Sulfolobus*, *Ferroplasma* type I and II and G-plasma do not contain coding regions for cytochrome *c*.

The robust macroscopic biofilms provide attachment to the surroundings and allow them to float at the air–water interface, simultaneously optimizing access to $O_2$, $CO_2$, $N_2$ and dissolved ferrous iron. It is unknown which organisms are responsible for polymer production. The *Leptospirillum* group II genome contains an operon of putative cellulose synthase genes. Cellulose has been shown to prevent desiccation and to enable biofilms to float. Numerous glycosyltransferases and polysaccharide export proteins also suggest a role for *Leptospirillum* group II in biofilm formation. In contrast, there is little evidence for genes for extracellular polymer production in the *Ferroplasma* type II genome.

After perturbation, the first colonizers are probably those that are motile. Both *Leptospirillum* group II and III possess a chemotaxis system and flagellar operon, which may be used to respond to ferrous iron[41] and oxygen gradients. There is no evidence for flagellar components or of other genes required for motility in any of the archaeal genomes. This explains the small number of proteins in *Ferroplasma* type II that are assigned to the cell motility and secretion (see Supplementary Fig. S10).

Perhaps the most important challenges facing members of biofilm communities growing in extremely acidic, metal-rich solutions are maintaining a near-neutral cytoplasm and resistance to toxic metals. *Ferroplasma* type II has genes for production of isoprenoid-based lipids, which we presume are largely tetraether linked[42], and thus highly proton impermeable. *Leptospirillum* group II seems to dedicate a comparatively large number (see Supplementary Fig. S10) and variety of genes for cell membrane biosynthesis, suggesting a complex cell wall structure. In addition, there are a variety of proton efflux systems (($H^+$)ATPases, antiporters and symporters). Most community members possess genes for resistance to copper, cobalt, arsenite, mercury, zinc, silver and cadmium (for example, *merA*, *arsB*, *arsR*).

An intriguing finding is the presence of genes for proteins related to DNA photolyases in the *Leptospirillum* group II (SplB and PhrB), III (PhrB) and G-plasma genomes (PhrB). The two putative photolyases in G-plasma are most closely related to photolyases of bacterial origin. These typically light-activated proteins for the repair of ultraviolet-damaged DNA may be present because relatively recent ancestors of these strains populated surface environments.

## Concluding remarks

The culture-independent recovery of two near-complete microbial genomes and partial recovery of three other genomes from an environmental sample is an advance in the study of natural microbial communities. Genome reconstruction was facilitated by the fact that the community was dominated by populations of a few genomically distinct species. The effectiveness of this approach in other environments will be limited by high species richness, heterogeneities in the abundance of community members, as well as by extensive genome rearrangements. However, even in more complex environments, it should be possible to extend the random shotgun sequencing approach to recover the genomes of uncultivated strains and species. These data can then be used to explore the nature of the community metabolic network, to find conditions for cultivating previously uncultivated organisms, to monitor community structure over time, and to construct DNA microarrays to monitor global community gene expression patterns. □

## Methods

The biofilm sample chosen for genomic analysis was characterized using FISH to quantify the abundance of the dominant prokaryotic members. Environmental DNA extracted using an agarose plug method was used for construction of 16S rRNA gene and small insert (approximately 3 kb) pUC18 libraries. Double-ended sequencing reactions were carried out using PE BigDye terminator chemistry (Perkin Elmer) and resolved using an ABI PRISM 3730 (Applied Biosystems) capillary DNA sequencer. Vector and quality trimming of shotgun data was performed to yield 103,462 reads totalling 76.2 Mb (average trimmed read length of 737 bp). After assembly and assignment of scaffolds to organism types, we identified the protein and RNA genes using the Fgenesb_annotator pipeline (Softberry Inc.). The gene prediction algorithm is based on Markov chain models of coding regions and translation and termination sites. We used the tRNAscan-SE package for prediction of tRNA genes. All reads were aligned to the scaffolds using BLASTN for analysis of nucleotide polymorphisms. Details for all methods are provided in the Supplementary Information.

Received 24 November 2003; accepted 19 January 2004; doi:10.1038/nature02340.
Published online 1 February 2004.

1. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51,** 221–271 (1987).
2. Makarova, K. S. & Koonin, E. V. Comparative genomics of archaea: how much have we learned in six years, and what's next? *Genome Biol.* **4,** 115.1–115.16 (2003).
3. Koonin, E. V. & Mushegian, A. R. Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. *Curr. Opin. Genet. Dev.* **6,** 757–762 (1996).
4. Amann, R. I., Ludwig, W. & Schleifer, K. H. Phylogenetic identification and *in-situ* detection of individual microbial-cells without cultivation. *Microbiol. Rev.* **59,** 143–169 (1995).

5. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276,** 734–740 (1997).
6. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* **3,** reviews0003.1–0003.8. (2002).
7. Beja, O. *et al.* Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* **2,** 516–529 (2000).
8. Beja, O. *et al.* Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl. Environ. Microbiol.* **68,** 335–345 (2002).
9. Rondon, M. R. *et al.* Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66,** 2541–2547 (2000).
10. Torsvik, V., Ovreas, L. & Thingstad, T. F. Prokaryotic diversity—magnitude, dynamics, and controlling factors. *Science* **296,** 1064–1066 (2002).
11. Singer, P. C. & Stumm, W. Acidic mine drainage rate-determining step. *Science* **167,** 1121–1127 (1970).
12. Edwards, K. J., Gihring, T. M. & Banfield, J. F. Seasonal variations in microbial populations and environmental conditions in an extreme acid mine drainage environment. *Appl. Environ. Microbiol.* **65,** 3627–3632 (1999).
13. Bond, P. L., Smriga, S. P. & Banfield, J. F. Phylogeny of microorganisms populating a thick, subaerial, predominantly lithotrophic biofilm at an extreme acid mine drainage site. *Appl. Environ. Microbiol.* **66,** 3842–3849 (2000).
14. Bond, P. L., Druschel, G. K. & Banfield, J. F. Comparison of acid mine drainage microbial communities in physically and geochemically distinct ecosystems. *Appl. Environ. Microbiol.* **66,** 4962–4971 (2000).
15. Baker, B. J. & Banfield, J. F. Microbial communities in acid mine drainage. *FEMS Microbiol. Ecol.* **44,** 139–152 (2003).
16. Edwards, K. J. *et al.* Geochemical and biological aspects of sulfide mineral dissolution: lessons from Iron Mountain, California. *Chem. Geol.* **169,** 383–397 (2000).
17. Silverman, M. P. & Ehrlich, H. L. Microbial formation and degradation of minerals. *Adv. Appl. Microbiol.* **6,** 153–206 (1964).
18. Bond, P. L. & Banfield, J. F. Design and performance of rRNA targeted oligonucleotide probes for *in situ* detection and phylogenetic identification of microorganisms inhabiting acid mine drainage environments. *Microb. Ecol.* **41,** 149–161 (2001).
19. Coram, N. J. & Rawlings, D. E. Molecular relationship between two groups of the genus *Leptospirillum* and the finding that *Leptospirillum ferriphilum* sp nov dominates South African commercial biooxidation tanks that operate at 40 °C. *Appl. Environ. Microbiol.* **68,** 838–845 (2002).
20. Fraser, C. M., Eisen, J. A., Nelson, K. E., Paulsen, I. T. & Salzberg, S. L. The value of complete microbial genome sequencing (you get what you pay for). *J. Bacteriol.* **184,** 6403–6405 (2002).
21. Branscomb, E. & Predki, P. On the high value of low standards. *J. Bacteriol.* **184,** 6406–6409 (2002).
22. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297,** 1301–1310 (2002).
23. Amils, R., Irazabal, N., Moreira, D., Abad, J. P. & Marin, I. Genomic organization analysis of acidophilic chemolithotropic bacteria using pulse field gel electrophoretic techniques. *Biochimie* **80,** 911–921 (1998).
24. Sandberg, R. *et al.* Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res.* **11,** 1404–1409 (2001).
25. Abe, T. *et al.* Informatics for unveiling hidden genome signatures. *Genome Res.* **13,** 693–702 (2003).
26. Spratt, B. G., Hanage, W. P. & Feil, E. J. The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Curr. Opin. Microbiol.* **4,** 602–606 (2001).
27. Vulic, M., Dionisio, F., Taddei, F. & Radman, M. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc. Natl Acad. Sci. USA* **94,** 9763–9767 (1997).
28. Magurran, A. E. & Henderson, P. A. Explaining the excess of rare species in natural species abundance distributions. *Nature* **422,** 714–716 (2003).
29. McGill, B. J. A test of the unified neutral theory of biodiversity. *Nature* **422,** 881–885 (2003).
30. Dewdney, A. K. A dynamical model of communities and a new species-abundance distribution. *Biol. Bull.* **198,** 152–165 (2000).
31. Ruepp, A. *et al.* The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* **407,** 508–513 (2000).
32. Parro, V. & Moreno-Paz, M. Gene function analysis in environmental isolates: The nif regulon of the strict iron oxidizing bacterium *Leptospirillum ferrooxidans*. *Proc. Natl Acad. Sci. USA* **100,** 7883–7888 (2003).
33. Blake, R., Shute, E. A., Waskovsky, J. & Harrison, A. P. Respiratory components in acidophilic bacteria that respire on iron. *Geomicrobiol. J.* **10,** 173–192 (1992).
34. Blake, R. & Shute, E. A. Respiratory enzymes of *Thiobacillus ferrooxidans*—kinetic-properties of an acid-stable irron-rusticyanin oxidoreductase. *Biochemistry* **33,** 9220–9228 (1994).
35. Yamanaka, T. & Fukumori, Y. Molecular aspects of the electron transfer system which participates in the oxidation of ferrous ion by *Thiobacillus ferrooxidans*. *FEMS Microbiol. Rev.* **17,** 401–413 (1995).
36. Appia-Ayme, C., Guiliani, N., Ratouchniak, J. & Bonnefoy, V. Characterization of an operon encoding two c-type cytochromes, an aa(3)-type cytochrome oxidase, and rusticyanin in *Thiobacillus ferrooxidans* ATCC 33020. *Appl. Environ. Microbiol.* **65,** 4781–4787 (1999).
37. Preisig, O., Zufferey, R. & Hennecke, H. The *Bradyrhizobium japonicum* fixGHIS genes are required for the formation of the high-affinity cbb(3)-type cytochrome oxidase. *Arch. Microbiol.* **165,** 297–305 (1996).
38. Pitcher, R. S., Brittain, T. & Watmough, N. J. Cytochrome cbb(3) oxidase and bacterial microaerobic metabolism. *Biochem. Soc. Trans.* **30,** 653–658 (2002).
39. Poole, R. K. & Hill, S. Respiratory protection of nitrogenase activity in *Azotobacter vinelandii*—roles of the terminal oxidases. *Biosci. Rep.* **17,** 303–317 (1997).
40. Komorowski, L., Verheyen, W. & Schafer, G. The archaeal respiratory supercomplex SoxM from *S. acidocaldarius combines* features of quinole and cytochrome *c* oxidases. *Biol. Chem.* **383,** 1791–1799 (2002).
41. Acuna, J., Rojas, J., Amaro, A. M., Toledo, H. & Jerez, C. A. Chemotaxis of *Leptospirillum ferrooxidans* and other acidophilic chemolithotrophs—comparison with the *Escherichia coli* chemosensory system. *FEMS Microbiol. Lett.* **96,** 37–42 (1992).
42. Macalady, J. L. *et al.* Tetraether-linked membrane monolayers in *Ferroplasma* spp.: a key to survival in acid. *Extremophiles* (submitted).


**Supplementary Information** accompanies the paper on **www.nature.com/nature**.

**Acknowledgements** This research was funded by the US Department of Energy Microbial Genomics Program and the National Science Foundation Biocomplexity Program. We would like to thank M. Power, W. Getz, R. Blake, J. Handlesman, B. Baker, I. Lo, J. Flanagan, D. Dodds and R. Carver for their contributions to this work. We also thank C. Detter and members of his laboratory at JGI for help with library construction, and T. Arman (Iron Mountain Mines) and R. Sugarek (EPA) for access to the Richmond mine.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to J.F.B. (jill@eps.berkeley.edu). This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accession code AADL00000000. The version described in this paper is the first version, AADL01000000.