Genome dynamics in a natural archaeal population

Eric E. Allen*[†], Gene W. Tyson*[‡], Rachel J. Whitaker*[§], John C. Detter[¶], Paul M. Richardson[¶], and Jillian F. Banfield*[‖]**

Departments of *Environmental Science, Policy, and Management and [‖]Earth and Planetary Science, University of California, Berkeley, CA 94720; and [¶]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598

Edited by James M. Tiedje, Michigan State University, East Lansing, MI, and approved November 30, 2006 (received for review June 9, 2006)

Evolutionary processes that give rise to, and limit, diversification within strain populations can be deduced from the form and distribution of genomic heterogeneity. The extent of genomic change that distinguishes the acidophilic archaeon *Ferroplasma acidarmanus* fer1 from an environmental population of the same species from the same site, fer1(env), was determined by comparing the 1.94-megabase (Mb) genome sequence of the isolate with that reconstructed from 8 Mb of environmental sequence data. The fer1(env) composite sequence sampled ≈92% of the isolate genome. Environmental sequence data were also analyzed to reveal genomic heterogeneity within the coexisting, coevolving fer1(env) population. Analyses revealed that transposase movement and the insertion and loss of blocks of novel genes of probable phage origin occur rapidly enough to give rise to heterogeneity in gene content within the local population. Because the environmental DNA was derived from many closely related individuals, it was possible to quantify gene sequence variability within the population. All but a few gene variants show evidence of strong purifying selection. Based on the small number of distinct sequence types and their distribution, we infer that the population is undergoing frequent genetic recombination, resulting in a mosaic genome pool that is shaped by selection. The larger genetic potential of the population relative to individuals within it and the combinatorial process that results in many closely related genome types may provide the basis for adaptation to environmental fluctuations.

Archaea | *Ferroplasma* | population genomics | recombination

The sequencing of microbial isolate genomes has enabled comparison of gene complements from a broad selection of bacterial and archaeal species. Comparative genomic explorations lend insight into the metabolic and ecological diversity of microbial taxa and provide us with the raw materials for inferring evolutionary history among lineages. However, in most cases only a single representative of a particular lineage has been analyzed genomically. Where multiple strain genomes are available, the organisms are typically pathogens, and the strains are derived from distinct localities (1–5). Lacking are comparative population genomic studies that include analysis of an isolate versus an environmental population of the same species obtained from the same site. Such data would enable evaluation of how representative of the population an individual is and would reveal the form and distribution of genetic variability that can give rise to the emergence of new species.

Genetic techniques such as multilocus sequence typing (MLST) (6, 7) of multiple individuals derived from a microbial population can provide a glimpse into the extent of genetic diversity within a natural population and insight into important processes that shape population structure. However, MLST requires isolation of the organisms under investigation and typically surveys only a small (6 to 10) number of loci. Consequently, MLST approaches are limited in the extent to which they can be applied to a wide variety of organisms, and extrapolation of such data to the whole-genome scale may be limited. Similarly, the utility of DNA microarray technologies to survey genetic diversity among closely related strains or environmental populations is inherently limited owing to the fact that only genes shared between reference and test strain(s) can be surveyed (8). Missing from such analyses are genes unique to the interrogated strains and the inability to extract syntenic information.

An alternative route to assess genetic diversity and reveal microevolutionary processes within natural microbial populations involves the cultivation-independent analysis of sequences from many individuals obtained from environmentally derived shotgun sequence data (9, 10). At the population level, comparison of reconstructed genome fragments and the associated raw sequence data can reveal genomic regions that are invariant as well as those regions that possess heterogeneity.

In the present study, we reconstructed the genome sequence of an isolate, *Ferroplasma acidarmanus* fer1, and compared it with ≈103 megabases (Mb) of environmental sequence data obtained directly from a natural biofilm community from the same site. Both the isolate (collected in July 1997) (11) and the environmental biofilm (≈1-cm^3 sample collected in March 2002) (9) were recovered from the five-way region of the Richmond Mine at Iron Mountain in Northern California. *F. acidarmanus* fer1 is an extremely acidophilic ferrous iron [Fe(II)]-oxidizing Euryarchaeote (order *Thermoplasmatales*) implicated in acid mine drainage (AMD) generation (11). Isolation of fer1 and description of its habitat have been reported (11). The fer1 strain grows optimally at 42°C, pH 1.2, and is capable of growth at pH 0. It is closely related to *Ferroplasma acidiphilum* (12, 13), based on comparison of 16S rRNA gene sequences (99.8% nucleotide similarity). Detailed phenotypic characterization of *Ferroplasma* species, including *F. acidarmanus* fer1, has been reported (12). To date, fer1 represents the only archaeal isolate recovered from the Richmond Mine.

Here, we investigate the structure and genomic diversity of an environmental population of *F. acidarmanus*, herein referred to as fer1(env), making use of the fer1 isolate genome sequence as a reference. Although both fer1 and the fer1(env) population were obtained from the same site, it cannot be claimed that the fer1 isolate is a member of the fer1(env) population because it was collected >4 years earlier. However, because both fer1 and the fer1(env) population had a very recent common ancestor, we can use the fer1 isolate genome in a comparative approach to quantify genome-wide heterogeneity and deduce the relative rates of the

Author contributions: E.E.A. and J.F.B. designed research; E.E.A., G.W.T., J.C.D., and P.M.R. performed research; E.E.A., G.W.T., R.J.W., and J.F.B. analyzed data; and E.E.A. and J.F.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS direct submission.

Abbreviations: MLST, multilocus sequence typing; Mb, megabase; IS, insertion sequence; R/M, restriction-modification; dN/Ds ratio, ratio of nonsynonymous to synonymous nucleotide substitutions.

Data deposition: Accession numbers for the community genomic data are AADL01000000 and AADL00000000. The *Ferroplasma acidarmanus* fer1 genome data associated with this study have been deposited in the GenBank database (accession no. AABC05000000).

[†]Present address: Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA 92093.

[‡]Present address: Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139.

[§]Present address: Department of Microbiology, University of Illinois, Urbana, IL 61801.

**To whom correspondence should be addressed. E-mail: jill@eps.berkeley.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0604851104/DC1.

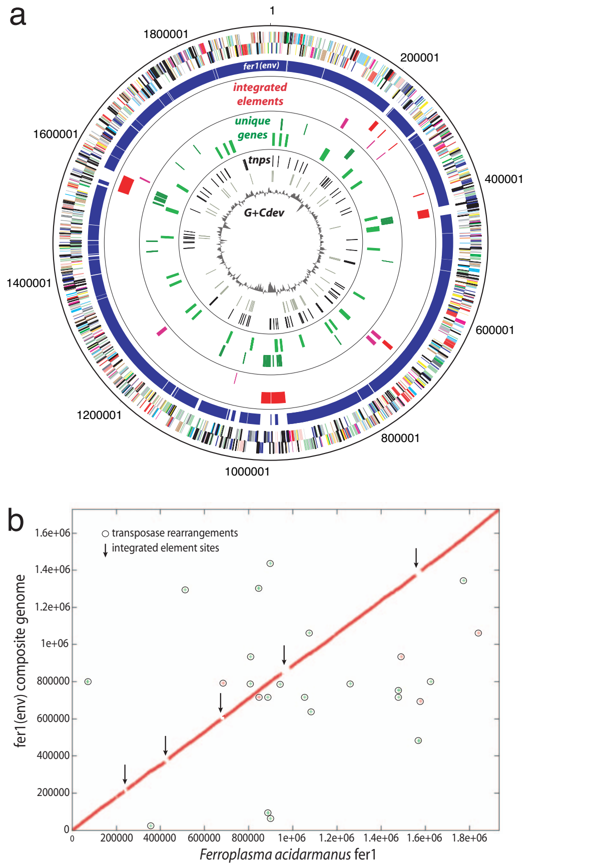


Fig. 1. Isolate versus environmental population genome comparison. (*a*) Circular diagram of the *F. acidarmanus* fer1 genome. Progressing inward: double line shows predicted gene sequences color-coded according to functional category (see SI Fig. 6); fer1(env) genome reconstruction (blue circle); putative regions impacted by integrated elements in fer1 and fer1(env) (red and pink, respectively); unique fer1 and fer1(env) genes (dark and light green, respectively); transposases (tnps) in fer1 and fer1(env) (dark and light gray, respectively). Inner circle shows % G+C content in fer1 (window size 8 kb). (*b*) Dot plot showing the shared synteny between assembled genomic fragments of the fer1(env) composite genome sequence and the fer1 isolate genome. Circled regions show the 25 rearranged transposase genes (opposite strand orientation in green) and arrows denote regions of fer1-specific phage elements.

various modes of genome evolution. The approach taken here, in which an isolate genome sequence is used as a reference to which a sequence from a closely related environmental population is compared, provides an innovative route to address population-level ecological and evolutionary questions.

Results and Discussion

Isolate Versus Population Composite Genome Analysis. The fer1 isolate genome sequence was assembled from data obtained from small and large insert libraries. General features of the 1.94-Mb genome, classification of the 1,963 ORFs by functional category, and a metabolic reconstruction are reported in supporting information (SI) Table 3 and SI Figs. 6 and 7.

Composite scaffolds and contigs from the fer1(env) population were assembled independently of the fer1 isolate sequence. Assembled fragments were ordered and aligned by using the fer1 genome as a reference sequence to reconstruct ≈92% (1,792 genes) of the fer1 isolate genome (Fig. 1*a*), with an average read depth of 4.5×. The colinearity plot (Fig. 1*b*) illustrates a very high degree of conserved synteny between fer1 and fer1(env).

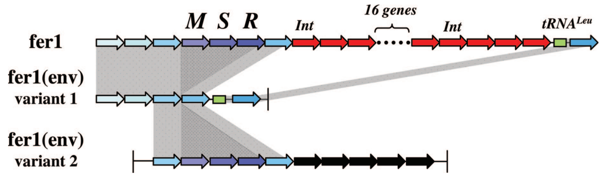


Fig. 2. Example of a heteromorphic fer1(env) genomic region that includes a type I restriction-modification system (M, methylation; S, specificity; R, restriction) that is present in only a subset of fer1(env) members, and is distinct from the comparable region in fer1. The 24-gene insertion in fer1 (shown in red) includes a phage integrase, four hypothetical genes, a phage-related DNA primase, 13 hypothetical proteins, another phage integrase, a conserved hypothetical membrane protein, a site-specific DNA methylase, and a hypothetical protein. The insertion region is 3′ delimited by a tRNA^{Leu}.

Fig. 1 highlights the differences between the fer1 and the composite fer1(env) genomes. There are 45 genes in fer1(env) not detected in the fer1 isolate and 152 genes in fer1 not present in fer1(env). These gene products are predominantly hypothetical and conserved hypothetical proteins, as well as transposases, integrases, transport-related proteins, products involved in restriction/modification and DNA repair, and glycosyltransferases (see SI Table 4). No missing genes were found in functional categories involved in core energy metabolism, lipid biosynthesis, cofactor biosynthesis, or transcription. Of the unique fer1 isolate genes, 79% occur in distinct genomic blocks (the largest comprising 24 genes; Fig. 2 and SI Fig. 8). Such regions often posses anomalous G+C contents (Fig. 1*a*) and are often associated with phage-type integrase genes. Thus, we infer that the blocks of unique genes are the result of prophage insertion. The localization of integrated elements at tRNA genes and the presence of partitioned integrase gene fragments flanking the insertion is consistent with a mechanism of site-specific integration typified by the SSV-type archaeal integrated elements (14).

Movement of insertion sequence (IS) elements also occurs fast enough to give rise to strain-level differences (Fig. 1*a*). There are 94 transposase insertions in fer1 and 51 in fer1(env) representing 10 families of ISs (SI Table 5). The much higher incidence of ISs in fer1 may indicate that these mobile elements proliferated while the organism was in culture. Only 18 of the 43 transposases in common to fer1 and fer1(env) occur in the same location (the 25 rearranged elements are shown in Fig. 1*b*). A few notably proliferous transposases have been copied into multiple locations, particularly in fer1 (e.g., seven identical copies of one novel transposase). As sources of strain-level diversity, together with the effects IS elements can elicit on gene expression and genome topography (15–20), heterogeneous IS dynamics may alter the adaptive fitness of population members population over short time scales.

Genomic heterogeneity between fer1 and the composite fer1(env) sequence was quantified by determining the number of differences that occur across comparable genomic regions (Table 1). The average nucleotide-level divergence between fer1 and fer1(env) is ≈3%. However, ≈50% of proteins encoded by the composite fer1(env) sequence are identical to the orthologous isolate sequences. The average amino acid-level similarity among orthologs is 99.16%. Overall, blocks of strain-specific genes localized to integrated elements represent the largest input of novel genetic potential within the population. The gain or loss of single or a few genes (mostly hypothetical) is the second most significant source of strain heterogeneity. In contrast, gene duplications (paralogous sequences) are very infrequent events and gene order is almost completely preserved.

Environmental Genotypic Heterogeneity. Comparison among composite genome scaffolds revealed that coexisting members of the fer1(env) population exhibit differences in gene content. When a subset of individuals contain genes not present in all population

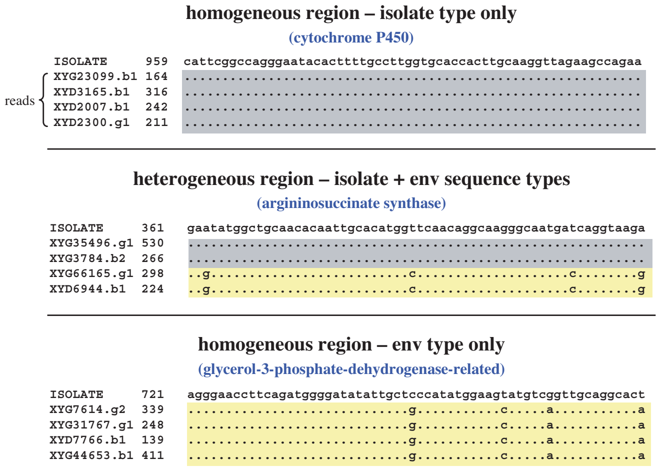
Table 1. Incidence of the documented forms of genome heterogeneity

Genomic feature	Events*	Incidence fer1, %	Incidence fer1 (env), %
Acquisition of phage-like genes	103 and 35	5	1.9
Nucleotide polymorphisms	≈58,200	3	3
Gene insertion/deletion	47 and 38	2.4	2
Syntenic deviations (shared 1,792 genes)	32	1.8	1.8
Transposase rearrangements	25	1.13	1.1
Transposase duplication	43 and 10	2	0.5
Transposase acquisition	8	0.40	0.40
Amino acid changes	≈5,570	0.97	0.97
Gene rearrangements (excluding transposases)	7	0.40	0.40
Gene duplications	2	0.10	0.10

*Defined as occurrence per 1,963 genes or 1.94 Mb in fer1 compared with 1,837 genes or 1.79 Mb in the composite fer1(env) genome. Events refer to incidence in fer1 and composite fer1(env) genomes, respectively.

members, the same genomic region will be assembled into two or more discrete fragments (10). The genes most commonly responsible for this phenomenon are transposases. In addition to documenting their presence/absence in composite genome sequences, the insertion of transposases into a subset of population members is evidenced by the localized "stretching" of mate-pairs (paired end reads) across genomic regions.

Integrated elements also contribute to genetic heterogeneity between fer1(env) members. Fig. 2 shows that fer1(env) variant 1 is missing three genes comprising a complete type I restriction-modification (R/M) system. In fer1, this region is adjacent to a tRNA^{Leu} and includes an insertion of 24 putative prophage-associated genes flanked by a partitioned integrase gene. Interestingly, a distinct block of possibly phage-derived genes is also present in the same region in fer1(env) variant 2, suggesting that the same genomic region has been targeted by disparate phage. Consequently, the form of heterogeneity that distinguishes individuals from one another within the fer1(env) population is similar to that which distinguishes fer1 from fer1(env). Whereas the importance of phage as sources of novel genes that differentiate species and strains has been documented (5, 14, 21, 22), our results indicate that the insertion and loss of blocks of genes of probable phage origin occurs fast enough to differentiate the gene content of members of a coexisting population and may give rise to phenotypic differences. These results extend previous findings documenting extensive genomic variation within natural coexisting bacterial populations (23), including the significance of phenotypic differentiation mediated by phage (24), and serve to highlight the dynamic nature of microbial population structure.



homogeneous region – isolate type only
(cytochrome P450)

```
         ISOLATE     959  cattcggccagggaatacacttttgccttggtgcaccacttgcaaggttagaagccagaa
       ( XYG23099.b1 164  ............................................................
reads <  XYD3165.b1  316  ............................................................
       ( XYD2007.b1  242  ............................................................
         XYD2300.g1  211  ............................................................
```

heterogeneous region – isolate + env sequence types
(argininosuccinate synthase)

```
ISOLATE     361  gaatatggctgcaacacaattgcacatggttcaacaggcaagggcaatgatcaggtaaga
XYG35496.g1 530  ............................................................
XYG3784.b2  266  ............................................................
XYG66165.g1 298  ..g..........................c....................c........g
XYD6944.b1  224  ..g..........................c....................c........g
```

homogeneous region – env type only
(glycerol-3-phosphate-dehydrogenase-related)

```
ISOLATE     721  agggaaccttcagatggggatatattgctcccatatggaagtatgtcggttgcaggcact
XYG7614.g2  339  .............................g...........c.....a...........a
XYG31767.g1 248  .............................g...........c.....a...........a
XYD7766.b1  139  .............................g...........c.....a...........a
XYG44653.b1 411  .............................g...........c.....a...........a
```

Fig. 3. Range of sequence types observed in the fer1(env) environmental population. Isolate-type sequences are shown in gray; env-type sequences are shown in yellow. Sequence types were quantified for all fer1(env) loci; see also Fig. 4 and SI Fig. 9.

Analyses presented above rely upon comparison between the fer1 isolate and composite fer1(env) sequences. However, greater resolution into the fer1(env) population structure can be achieved because each environmentally derived sequencing read likely originated from a different individual within the natural population (9). We evaluated sequence variability in the population on a gene-by-gene basis by aligning all reads constituting the fer1(env) population with the nucleotide sequence of fer1 (average coverage at each locus is ≈4.5×). Overall, 56% of the genes/gene fragments encoded on fer1(env) reads (out of a total of ≈14,000 reads) have the fer1 sequence type (defined as having ≥99.9% nucleotide-level similarity to the fer1 ortholog). At least one read has the fer1 sequence type at ≈80% of loci (see gene content dendogram, SI Fig. 9).

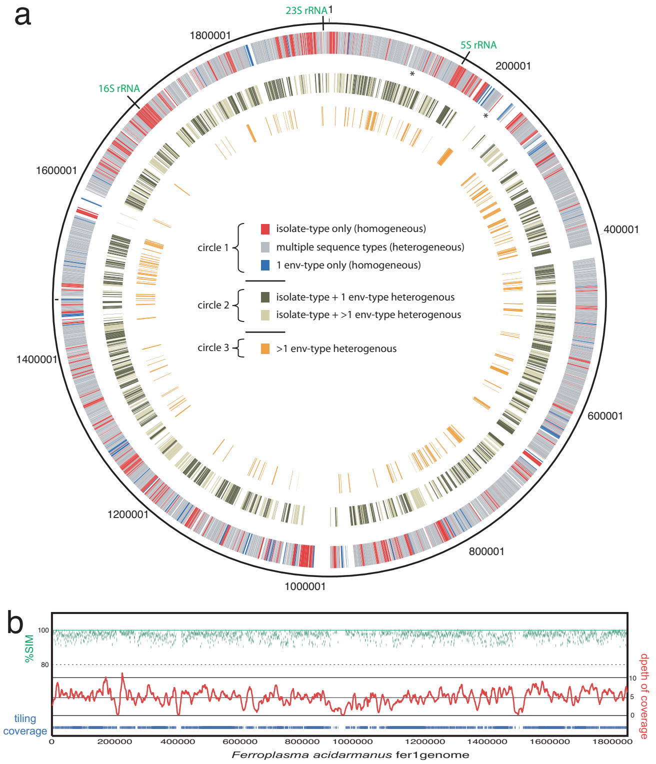


Fig. 4. Form and distribution of sequence-type variation within the fer1(env) population. (*a*) Circular diagram showing the genome-wide distribution of sequence variation in the fer1(env) population. Data are derived from the gene content dendogram (SI Fig. 9). Each fer1(env) locus was queried with respect to sequence type present (isolate-type only, isolate plus env-type, or env-type only) by alignment of environmental sequencing reads constituting the fer1(env) population to the orthologous fer1 locus. Isolate-type-only sequences were defined as having an environmental nucleotide identity ≥99.9% to the respective fer1 sequence. Genes <99.9% similar to the fer1 sequence were defined as having env-type sequence. Homogeneous loci represent those possessing a single sequence type; isolate-type only (red) or 1 env-type only (blue). Cumulative heterogeneous loci (gray) possess both isolate-type plus env-type sequences or multiple (>1) env-type sequences, which can be further separated by the sequence composition of those loci (brown, light brown, or orange). The locations of 23S, 16S, and 5S rRNA genes are shown for reference. Segment spanning the two asterisks (*) shows region depicted in Fig. 5. (*b*) Percentage identity (green), depth of coverage (10-kb window size) (red), and tiling coverage (blue) of fer1(env) reads across the fer1 genome.

Table 2. Genome-wide distribution of gene sequence types in the fer1(env) population

Loci	%
Homogeneous isolate type	24
Homogeneous non-isolate type	7
Heterogeneous, isolate type present	56
Heterogeneous, no isolate type	13

We categorized all gene types in the fer1(env) population as either homogeneous (one sequence type only, 30.4%) or heterogeneous (multiple variants at one locus, 69.6%) (Figs. 3 and 4). Of the homogeneous group, 78.3% of genes are isolate type (427 genes in total), indicating that a significant number of genes possessing the isolate sequence type have persisted over the >4 years since the isolate was obtained. The genome-wide distribution of sequence variation in the fer1(env) population is summarized in Table 2.

To quantify and evaluate the distribution of gene variants within the fer1(env) population, we evaluated several >100-kb regions in detail, verifying that mate-pairs carrying environmental variant sequences anchor them into the same genomic context (see 109-gene region in Fig. 5). Of the fer1(env) heterogeneous loci, 55% possess two variants and 45% possess three or more variants (these products share on average ≈99% amino acid level identity with fer1 isolate orthologs).

Given the ≈4.5× sequencing depth across the fer1(env) population, we cannot determine whether regions are homogeneous because variants were absent or simply present in low abundance. However, our results do show the dominance of sequence types present. In the future, PCR-based analyses together with population genomic sampling will provide the sequence depth necessary to more fully resolve the abundance distribution of gene types in the population.

Mate-pair information (Fig. 5*b*; see also SI Fig. 10) and direct examination of nucleotide polymorphism patterns within the reads confirm that the population does not consist of a few dominant clonal genome types that have been sampled to different depths of coverage. Several kilobase long regions characterized by only isolate-type genes transition into regions with several gene variants at each locus (Figs. 4 and 5). Multiple linkage patterns between isolate-type and non-isolate type genes can be identified in single regions and some reads capture the transitions directly (Fig. 5 and SI Fig. 11). Thus, it is evident that the population exhibits a mosaic genome structure inferred to have arisen via recombination between distinct yet closely related strain variant sequence types. Frequently, but not exclusively, the site of recombination occurs at the same nucleotide position in multiple reads indicating that the resulting recombinant sequence has been amplified in the natural community, either by natural selection or by genetic drift.

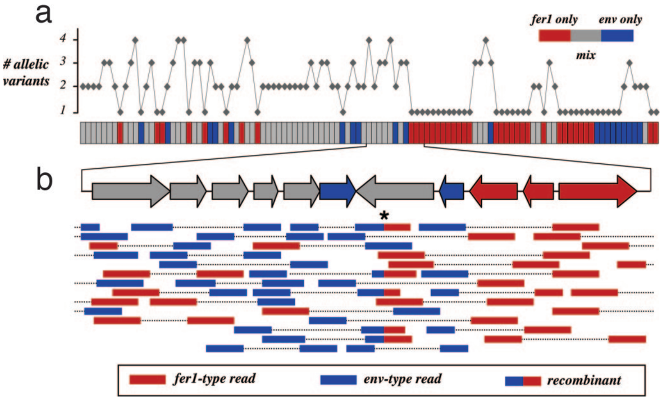


Fig. 5. Detailed view of fer1(env) population structure. (*a*) Genomic region (109 genes) showing the distribution of sequence variability and mosaic structure of the fer1(env) population. The majority of loci possess more than one sequence type, and the fer1 isolate sequence type is present at most loci (red plus gray). Note that env-only type genes (blue) may be homogeneous or heterogeneous depending on the number of allelic variants present. (*b*) An enlargement of a region in a illustrating the sampling of environmental sequence types across 11 genes. Individual sequencing reads constituting the region are shown. Note that the sequence at one end of a paired-end clone may be isolate type whereas the other end may be non-isolate type. Example of recombination site (env- to isolate-type sequence) occurring within reads is noted (*).

Extensive recombination has been documented in *Ferroplasma* type II (9). Subsequently, MLST-based analyses provided evidence for genetic recombination in the halophilic archaeon *Halorubrum* sp (25) and the thermoacidophilic archaeon *Sulfolobus islandicus* (7). Results documenting recombination in *F. acidarmanus* suggest that mosaic genome structure may be a common feature of natural archaeal populations.

The restriction subunit of the type I R/M system described above exemplifies the form of sequence heterogeneity that can exist in the fer1(env) population. There are three distinguishable sequence types for the restriction subunit, exhibiting at least five recombination events across the 2,985-bp gene. The fer1(env) population also contains a type I specificity subunit gene with several allelic variants that show evidence of recombination among them. Recombination among individuals with distinct sequence types may provide the capacity to generate a large number of novel, hybrid restriction enzymes. Overall, R/M system dynamics may factor significantly in natural population structure and evolution (26–28) as heterogeneity in R/M activity (diversity in molecular recognition or the complete presence/absence of such systems) may lead to "recombination isolation." Ultimately, the absence of recombination as a cohesive mechanism hindering genetic divergence among population members may result in an altered evolutionary trajectory (e.g., speciation or extinction).

Population Structure. An important question is the role of selection in shaping genetic variation in natural populations. It has been postulated that genetic homogenization can be achieved when an adaptive genotype providing a selective advantage sweeps through a population (29, 30). Cohesion can also be maintained by high levels of genetic exchange between population members (31). As suggested here for the fer1(env) population and recently for local populations of *Ferroplasma* type II (9) and *S. islandicus* (7), recombination prevents periodic selection from purging total diversity. Because gene variants are not linked to a single genome type, selection removes only the subset of mosaic genome types with less adaptive gene variants. The removal of sequence variation by incomplete selective sweeps may explain the homogeneous genomic regions within the fer1(env) population.

It has been suggested that genetic diversity can be maintained within a clade of very close relatives by one of two mechanisms: (*i*) the clade contains only a single population, and selection for adaptive genotypes is too weak to purge diversity (23, 32), or (*ii*) the clade contains multiple ecologically distinct populations (ecotypes) (33). In this latter model, localized chromosomal regions of sequence identity are explained by rare "globally adaptive" mutations, which are beneficial in the context of any ecotype to which they are introduced (29). When these mutations are transferred by recombination between ecotypes, selection may result in homogenization of the recombined region across ecotypes, while leaving the rest of the genome heterogeneous.

We cannot rule out the possibility that some sequence diversity in the fer1(env) population is due to the partitioning of specific mosaic genome types into different niches within the biofilm. In support of the multiple-ecotype model, it is clear that recombinant genotypes have been amplified throughout the clade and that heterogeneous loci fall into discrete clusters. However, as our data indicates that nearly a third of the genes have been involved in recent global selection-like events, it is more likely that the fer1(env) organisms share a single niche, are subject to very similar selective pressures, and thus represent a single population. Furthermore, high rates of recombination effectively blurs the distinction between ecotypes (29). Ultimately, definitive discrimination between either model will require additional genomic monitoring of these populations over space and time.

Evidence for Selection. The model for the mosaic population structure requires selection to explain the presence of genes with only one sequence type. To seek evidence for selection we modeled synonymous vs. nonsynonymous nucleotide polymorphism patterns in the fer1(env) population, both between population variants and against the respective fer1 ortholog. Pairwise dN/dS ratios (ratios of nonsynonymous to synonymous nucleotide substitutions) among partial and complete orthologous sequences were determined (34). Approximately seventy-nine percent of the 3,745 pairwise comparisons among variable sequence segments exhibited a dN/dS ratio <0.15, indicating that the majority of the genome is under strong purifying selection. No functional bias in distribution of dN/dS values could be detected (data not shown).

Only six cases were identified with dN/dS values >1.0, suggesting that they are under positive selection. Four of these genes encode small hypothetical proteins, one is a conserved hypothetical protein, and one is a predicted siroheme synthase [siroheme, an iron-containing porphinoid, serves as a cofactor for sulfite and nitrite reductases (35)]. In many cases, elevated dN/dS values were associated with regions of genes, rather than entire genes. More sensitive methods for identifying positive selection (e.g., on a codon-by-codon basis) will require reconstruction of complete environmental variant gene sequences and the inclusion of sequences from more distantly related species. Whereas it is expected that many genes would display low dN/dS ratios (e.g., housekeeping genes), these results suggest that stringent functional constraints [including codon usage bias, the selection for preferred synonymous codons (36)] are characteristics of adaptation to this environment. These results are consistent with analyses revealing strong functional constraints in thermophilic species as revealed by global dN/dS analyses among closely related organisms (37).

To seek evidence for the basis of selection, we examined the physical and functional distribution of homogeneous and heterogeneous genes within the fer1(env) data set. Clusters of genes with only one sequence type may be moved in blocks, possibly by recombination events (Figs. 4 and 5). No significant bias in functional gene category representation was detected among homogeneous and heterogeneous regions (SI Fig. 9). Whereas we might predict that genes with a single sequence type may confer metabolic properties that have been important in recent adaptation, greater sequence depth will be required for definitive analysis.

Population Structure in the Context of Environmental and Geologic History. To understand population dynamics it would be helpful to able to correlate environmental perturbations with their genomic and evolutionary consequences. Geologic and historic records can provide information about the sequence of geochemical changes at the site. Using magnetism recorded in hematite gossans (products of pyrite oxidation), Alpers *et al.*[††] determined that ore bodies at Iron Mountain were first exposed by erosion ≈700,000–1,000,000 years ago. When mining began ≈150 years ago, the physical fracturing of the system resulted in increased porosity and permeability within the mountain thus accelerating pyrite oxidation. As solution chemistry and temperature are strongly controlled by the dissolution of pyrite, this process resulted in a decrease in solution pH to values <1.0 and increased temperatures to ≈40°C in the contemporary underground mine. Seasonal rainfall-driven fluctuations in temperature and ionic strength occur within the mine and have been shown to be factors in species selection (38). These factors may also be important in strain genotype selection.

Our comparative analyses have revealed genomic heterogeneity that may be understood in the context of these environmental changes. First, we determined that the fer1(env) population consists of mosaic genomes in which a small number of sequence types predominate. One possible scenario is that subsurface populations that were previously evolving in isolation from one another were brought together and began to recombine when different regions of the mountain were physically interconnected by mining. Recombination-driven homogenization of distinct but closely related populations at their geographic boundaries may be a common process in many ecosystems. Perhaps a limited number of mosaic genome type rose in abundance relative to a diverse mosaic pool in a relatively recent clonal expansion (39), possibly induced by environmental changes caused by the onset of mining. Under this scenario, genes with only a single sequence type may have derived from any of the ancestral strains.

As recombination rates are inferred to be very fast relative to nucleotide polymorphism fixation rates (otherwise we would see many closely related variant types), rapid recombinatorial mixing of variants into the mosaic genome pool may maintain population cohesion. However, mechanisms such as phage and IS dynamics contribute to strain-level variation over short time scales via the rapid introduction and distribution of novel genes. Ultimately, accumulated variation may diminish the frequency of genetic recombination, promoting divergence and possibly speciation (31).

A second finding is that almost all genes are under strong stabilizing selection. Environmental sequence types that are nearly identical (≥99.9% similar) to the isolate type are found at most loci despite >4 years difference in sample collection time. The exact preservation of a large amount of DNA sequence over 4 years may provide one constraint on rates of nucleotide polymorphism fixation. Such information is needed to quantify rates of genome evolution. At present, it is only possible to speculate about the relationship between strain divergence and site history. One possibility is that all of the observed nucleotide-level heterogeneity arose *in situ* over 700,000–1 million years. If this possibility could be confirmed, the relative rates for the accumulation of the different forms of genome change (Table 1) could be converted to absolute rates. Support for this correlation may be obtained by future genomic monitoring of these populations over space and time and through deductions at other sites for which dates of onset of colonization can be determined. This approach may provide a route to the long-term objective of mapping genome change onto environmental change so as to discover how these processes are interconnected.

Materials and Methods

***F. acidarmanus* fer1 Genome Sequencing and Assembly.** *F. acidarmanus* fer1 was isolated from enrichment cultures inoculated with mine water and sediment samples collected in July 1997 from the five-way site within the Richmond Mine (11). Details of the fer1 genome assembly and annotation are presented in *SI Materials and Methods*.

fer1(env) Genome Assembly and Comparative Analyses. We compared the isolate fer1 genome with genomic information obtained by shotgun sequencing small insert libraries constructed from DNA extracted directly from an environmental biofilm. The biofilm was sampled in 2002 from the five-way site of the Richmond Mine at the same location from which the fer1 isolate was collected. Tyson *et al.* (9) used 76 Mb of environmental sequence data to reconstruct large

[††] Alpers, C. N., Nordstrom, D. K., Verosub, K. L., Helm, C. M. (1999) Geological Society of America Cordilleran Section Centennial Meeting, Abstracts with Programs, Vol. 31, No. 6, p. A-33 (abstr).

genome scaffolds from five species, one of which was closely related to *F. acidarmanus*. The partial composite genome reported for this fer1(env) population was previously referred to simply as *Ferroplasma* type I (9). In this work, an additional 27 Mb of community small insert library sequence not reported in ref. 9 was coassembled with the initial 76 Mb of data. The total community genome data set used in the present study consisted of 141,312 small-insert library reads (average trimmed length 737 bp) for a total of $\approx$103 Mb of environmental sequence data. Assembly was performed by using PHRAP (minmatch 50; penalty -15; minscore 40) (40) and ATLAS (k-mer size 32 bp, MaxMismatch 10%, band size 7; minmatch 12, minscore 20, penalty -2, forcelevel 10) (41). Newly generated genomic contigs were binned by alignment with existing contigs and scaffolds described in Tyson *et al*. (9).

Scaffolds and contigs were initially assigned to fer1(env) based solely on assembly and subsequent analyses including % G+C, read depth, and profiling of nucleotide frequencies using TETRA (42). Subsequently, community genome contigs and scaffolds assigned to fer1(env) were verified by comparison with the *F. acidarmanus* fer1 genome by using BLAST (43) and MUMmer 3.0 software package's NUCmer (44). Approximately 14,000 environmentally derived sequencing reads constituted the genomically sampled fer1(env) population. In total, 172 scaffolds belonging to fer1(env) were used in the comparative analyses totaling $>$1.79 Mb of assembled sequence. Genome circular diagrams were generated from .embl format genome files exported from ARTEMIS (45) by using "circular diagram.pl" (Pathogen Group, The Wellcome Trust Sanger Institute). Synteny dot plots were generated by alignment of the fer1 genome sequence with all fer1(env) composite contigs and scaffolds by using NUCmer and MUMMERPLOT (44). Orthologous sequences between fer1 and the composite fer1(env) sequences were identified by using the reciprocal smallest distance method described by Wall *et al*. (46). To determine genome-wide sequence variation in the fer1(env) population, all fer1(env) environmental sequencing reads were aligned to fer1 genes by using BLASTN, and sequence identity was analyzed on a gene-by-gene basis. Multiple alignment outputted blast results were also generated, and mate-pair consistency for all reads localized to a given fer1 gene were verified on a gene-by-gene basis to verify that environmental reads aligned to fer1 genes in the correct genomic context.

Analysis of Selection in the fer1(env) Population. To screen for evidence of selection within the fer1(env) population and against the orthologous fer1 sequences, the set of predicted fer1 orfs was queried against all reads from the community genome sequence by using BLASTN with a strict E value of $e-75$. To verify correct assignment of environmental reads to the fer1(env) population, reads comprising independently assembled fer1(env) contigs were extracted and checked for consistency against all blast matches. All alignments were preformed in CLUSTALW. dN/dS ratios for pairwise comparisons between fer1 orfs and orthologous fer1(env) reads were determined by using the method of Nei and Gojobori (34) in SNAP (47). Correct assignment of environmental sequences to the fer1(env) genome and the accuracy of alignments were confirmed for all pairwise comparisons resulting in dN/dS values $>$0.9 by reconstructing variant alleles and estimating dN/dS values in DnaSP (48).

We thank Ted W. Arman (owner, Iron Mountain Mines), Richard Sugarek, Don Dodds, and Rudy Carver for mine access and on-site assistance; Val Burland and other members of the Blattner laboratory at the University of Wisconsin, Madison for contributions to the initial isolate genome sequencing effort; Fred Cohan for critical reading of the manuscript; and three anonymous reviewers. Research was supported by the Department of Energy (DOE) Microbial Genome Program (J.F.B.), the National Science Foundation (NSF) Biocomplexity Program (J.F.B.), the National Aeronautics and Space Administration Astrobiology Institute (J.F.B.), and a NSF Postdoctoral Research Fellowship in Microbial Biology (to E.E.A.). Sequencing was performed under the auspices of the DOE's Office of Science, Biological and Environmental Research Program.

1. Bhattacharyya A, Stilwagen S, Ivanova, N., D'Souza M, Bernal A, Lykidis A, Kapatral V, Anderson I, Larsen N, Los T, *et al*. (2002) *Proc Natl Acad Sci USA* 99:12403–12408.
2. Nelson KE, Fouts DE, Mongodin EF, Ravel J, DeBoy RT, Kolonay JF, Rasko DA, Angiuoli SV, Gill SR, Paulsen IT, *et al*. (2004) *Nucleic Acids Res* 32:2386–2395.
3. Eppinger M, Baar C, Raddatz G, Huson DH, Schuster SC (2004) *Nat Rev Microbiol* 2:872–885.
4. Holden MT, Titball RW, Peacock SJ, Cerdeno-Tarraga AM, Atkins T, Crossman LC, Pitt T, Churcher C, Mungall K, Bentley SD, *et al*. (2004) *Proc Natl Acad Sci USA* 101:9786–9791.
5. Deng W, Burland V, Plunkett G, III, Boutin A, Mayhew GF, Liss P, Perna NT, Rose DJ, Mau B, Zhou S, *et al*. (2002) *J Bacteriol* 184:4601–4611.
6. Hanage WP, Fraser C, Spratt BG (2005) *BMC Biol* 3:6.
7. Whitaker RJ, Grogan DW, Taylor JW (2005) *Mol Biol Evol* 22:2354–2361.
8. Ochman H, Santos SR (2005) *Infect Genet Evol* 5:103–108.
9. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) *Nature* 428:37–43.
10. Allen EE, Banfield JF (2005) *Nat Rev Microbiol* 3:489–498.
11. Edwards KJ, Bond PL, Gihring TM, Banfield JF (2000) *Science* 287:1796–1799.
12. Dopson M, Baker-Austin C, Hind A, Bowman JP, Bond PL (2004) *Appl Environ Microbiol* 70:2079–2088.
13. Golyshina OV, Pivovarova TA, Karavaiko GI, Kondrateva TF, Moore ER, Abraham WR, Lunsdorf H, Timmis KN, Yakimov MM, Golyshin PN (2000) *Int J Syst Evol Microbiol* 50:997–1006.
14. She Q, Brugger K, Chen L (2002) *Res Microbiol* 153:325–332.
15. Schneider D, Lenski RE (2004) *Res Microbiol* 155:319–327.
16. Ciampi MS, Schmid MB, Roth JR (1982) *Proc Natl Acad Sci USA* 79:5016–5020.
17. Prentki P, Teter B, Chandler M, Galas DJ (1986) *J Mol Biol* 191:383–393.
18. Nevers P, Saedler H (1977) *Nature* 268:109–115.
19. Schneider D, Duperchy E, Coursange E, Lenski RE, Blot M (2000) *Genetics* 156:477–488.
20. Martusewitsch E, Sensen CW, Schleper C (2000) *J Bacteriol* 182:2574–2581.
21. Wick LM, Weihong Q, Lacher DW, Whittam TS (2005) *J Bacteriol* 187:1783–1791.
22. Ohnishi M, Kurokawa K, Hayashi T (2001) *Trends Microbiol* 9:481–485.
23. Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, Polz MF (2005) *Science* 307:1311–1313.
24. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF, Chisholm SW (2006) *Science* 311:1768–1770.
25. Papke RT, Koenig JE, Rodriguez-Valera F, Doolittle WF (2004) *Science* 306:1928–1929.
26. Milkman R (1997) *Genetics* 146:745–750.
27. Milkman R, Raleigh EA, McKane M, Cryderman D, Bilodeau P, McWeeny K (1999) *Genetics* 153:539–554.
28. Jeltsch A (2003) *Gene* 317:13–16.
29. Majewski J, Cohan FM (1999) *Genetics* 152:1459–1474.
30. Cohan FM (2002) *Genetica* 116:359–370.
31. Lawrence JG (2002) *Theor Popul Biol* 61:449–460.
32. Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF (2004) *Nature* 430:551–554.
33. Cohan FM (2002) *Annu Rev Microbiol* 56:457–487.
34. Nei M, Gojobori T (1986) *Mol Biol Evol* 3:418–426.
35. Murphy MJ, Siegel LM, Tove SR, Kamin H (1974) *Proc Natl Acad Sci USA* 71:612–616.
36. Lynn DJ, Singer GAC, Hickey DA (2002) *Nucleic Acids Res* 30:4272–4277.
37. Friedman R, Drake JW, Hughes AL (2004) *Genetics* 167:1507–1512.
38. Edwards KJ, Gihring TM, Banfield JF (1999) *Appl Environ Microbiol* 65:3627–3632.
39. Smith NH, Dale J, Inwald J, Palmer S, Gordon SV, Hewinson RG, Smith JM (2003) *Proc Natl Acad Sci USA* 100:15271–15275.
40. Ewing B, Green P (1998) *Genome Res* 8:186–194.
41. Havlak P, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Weinstock GM, Gibbs RA (2004) *Genome Res* 14:721–732.
42. Teeling H, Waldman J, Lombardot T, Bauer M, Glöckner FO (2004) *BMC Bioinformatics* 5:163.
43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) *J Mol Biol* 215:403–410.
44. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) *Genome Biol* 5:R12.
45. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream, M-A, Barrell B (2000) *Bioinformatics* 16:944–945.
46. Wall DP, Fraser HB, Hirsch AE (2003) *Bioinformatics* 19:1710–1711.
47. Korber B (2000) in *Computational Analysis of HIV Molecular Signatures*, eds Rodrigo AG, Learn GH (Kluwer, Dordrecht, The Netherlands), pp 55–72.
48. Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) *Bioinformatics* 19:2496–2497.