Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities

Vincent J. Denef[a], Linda H. Kalnejais[a,1], Ryan S. Mueller[a], Paul Wilmes[a], Brett J. Baker[a], Brian C. Thomas[a], Nathan C. VerBerkmoes[b], Robert L. Hettich[b], and Jillian F. Banfield[a,2]

[a]University of California, Berkeley, CA 94720; and [b]Oak Ridge National Laboratory, Oak Ridge, TN 37831

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2006.
Contributed by Jillian F. Banfield, November 13, 2009 (sent for review July 21, 2009)

Bacterial species concepts are controversial. More widely accepted is the need to understand how differences in gene content and sequence lead to ecological divergence. To address this relationship in ecosystem context, we investigated links between genotype and ecology of two genotypic groups of *Leptospirillum* group II bacteria in comprehensively characterized, natural acidophilic biofilm communities. These groups share 99.7% 16S rRNA gene sequence identity and 95% average amino acid identity between their orthologs. One genotypic group predominates during early colonization, and the other group typically proliferates in later successional stages, forming distinct patches tens to hundreds of micrometers in diameter. Among early colonizing populations, we observed dominance of five genotypes that differed from each other by the extent of recombination with the late colonizing type. Our analyses suggest that the specific recombinant variant within the early colonizing group is selected for by environmental parameters such as temperature, consistent with recombination as a mechanism for ecological fine tuning. Evolutionary signatures, and strain-resolved expression patterns measured via mass spectrometry–based proteomics, indicate increased cobalamin biosynthesis, (de)methylation, and glycine cleavage in the late colonizer. This may suggest environmental changes within the biofilm during development, accompanied by redirection of compatible solutes from osmoprotectants toward metabolism. Across 27 communities, comparative proteogenomic analyses show that differential regulation of shared genes and expression of a small subset of the ~15% of genes unique to each genotype are involved in niche partitioning. In summary, the results show how subtle genetic variations can lead to distinct ecological strategies.

geomicrobiology | genome evolution | niche partitioning | community genomics | community proteomics

Closely related microbes may play distinct roles in natural environments, but the level and form of genomic variation required for functional differentiation remains unclear (1). Organisms that differ by 3–10% in their 16S rRNA gene sequences have been shown to have distinct ecological distributions (2, 3). There have been attempts to explain such patterns based on differences in gene content. For example, specific gene content has been correlated with specialization to the gut environment in *Bacteroides* isolates (4), differences between planktonic and host-associated populations of *Cenarcheaum symbiosum* (5), different temperatures in microbial mats in *Synechococcus* ecotypes (6), and with low and a high light in *Prochlorococcus* ecotypes (7).

Microdiversity (<1% divergence in 16S rRNA gene sequences) has been observed in many microorganisms from multiple systems, from surface ocean community members, including SAR11 (8), to species belonging to the *Firmicutes* and *Bacteroidetes* lineages in the human distal gut (9). Resource partitioning between co-occurring populations of *Vibrio splendidus* has been demonstrated (10). This might be explained by the unexpectedly large genomic heterogeneity observed between isolates from these distinct populations (11), although no specific links have been made. At even finer levels, community genomic data reveal extensive genomic variation within natural populations (12). However, it is unknown whether this level of differentiation results in altered ecological behavior.

To date, most studies have focused on ecotype-specific gene content and have rarely considered the roles of sequence variation or regulation of shared genes in adaptation. Gene expression differences are likely the first manifestations of organismal divergence because regulatory elements probably evolve faster than the genes they regulate (13). Microbial strain–level differences in gene expression have been observed in laboratory experiments (14), and directly in the environment as well (15). Although limited data are available, such differences can result in ecological differentiation (16). Gene expression divergence has been extensively studied for its role in ecotype and species differentiation in complex Eukaryotes (17). Although much gene expression variation can be attributed to neutral drift correlated with genetic distance (18), adaptive variation to environmental factors has been shown to be important as well (19).

Our understanding of microbial evolution and community functioning may benefit from studies that directly connect the ecology of closely related microbes with their gene content, gene sequence, and protein abundance levels. Here, we explore these links by studying two genotypic groups belonging to *Leptospirillum* group II that differ by 0.3% in their 16S rRNA gene sequences and that inhabit the same ecosystem. *Leptospirillum* group II bacteria play a crucial role in pyrite oxidation, the chemical reaction responsible for the environmental problem of acid mine drainage (AMD) (20). They generate energy from iron oxidation, fix carbon (21, 22), and are the pioneers that condition the environment for further biofilm community development (23). A series of recombinants of two ancestral sequence types of *Leptospirillum* group II (24) and their fine-scale variants (12) dominate the microbial communities in the Richmond Mine (Iron Mountain, CA) (Fig. 1). Types II, III, IV, and V are composed of blocks of the type I and type VI genomes, which were assembled from community genomic datasets from biofilms growing in the Richmond Mine at the five-way and UBA locations, respectively (Fig. 2*A*). Types II–VI comprise the UBA genotypic group, and type I is the only representative thus far of the five-way CG genotypic group. The existence of this series provides the opportunity to study the link between genotype and ecological behavior in environmental con-

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

The authors declare no conflict of interest.

[1]Present address: University of New Hampshire, Durham, NH 03824.

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

This article contains supporting information online at www.pnas.org/cgi/content/full/0907041107/DCSupplemental.

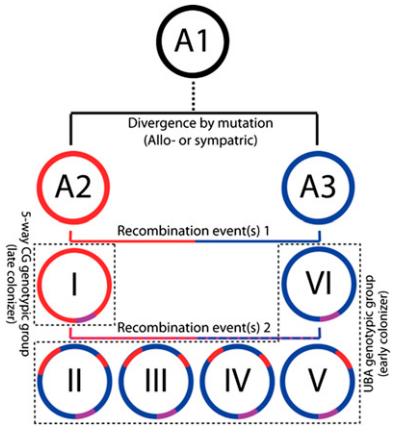

Fig. 1. Evolutionary relationships between the *Leptospirillum* group II genotypic groups. Genotypic groups are defined based on similarity to the UBA (type VI) and five-way CG variants (type I) for which genome sequences are available (21, 24, 44). A1, common ancestor; A2, ancestor of type I; A3, ancestor of type VI. Types I and VI resulted from recombination events between A2 and A3, whereas types II–V were most likely the result from recombination events between types I and VI.

text. In the current study, we analyzed 45 natural biofilms collected over a 4-year period to elucidate links between different forms of genomic variation, activity levels, and ecological divergence.

Results

Distribution of Genomic Types. We used two independent sets of probes for the fluorescent in situ hybridization (FISH) survey, one targeting Bacteria, Archaea, and *Leptospirillum* group III and the other targeting the *Leptospirillum* genus, *L.* group II UBA type, and *L.* group II five-way CG type. By combining these data sets, we showed that *Leptospirillum* group II dominated more than 45 biofilms that grow at the air–AMD solution interface (Fig. 2). The *Leptospirillum* group II FISH probes effectively discriminated UBA and five-way CG genotypic groups (Fig. S1). Although the UBA genotypic group predominated, both occurred over the range of observed temperatures (36–46°C), pH (0.7–1.2), and metal concentrations (Fe = 110–330 mM, Cu = 0.45–4.11 mM, As = 0.24–2.00 mM) (Fig. 2 and Tables S1 and S2). In each sample, the fraction of all cells hybridized to the *Leptospirillum* group III probe approximated the fraction of all cells hybridized to the *Leptospirillum* probe but not to either of the *Leptospirillum* group II type–specific probes, indicating all *Leptospirillum* group II cells were accounted for (Table S2).

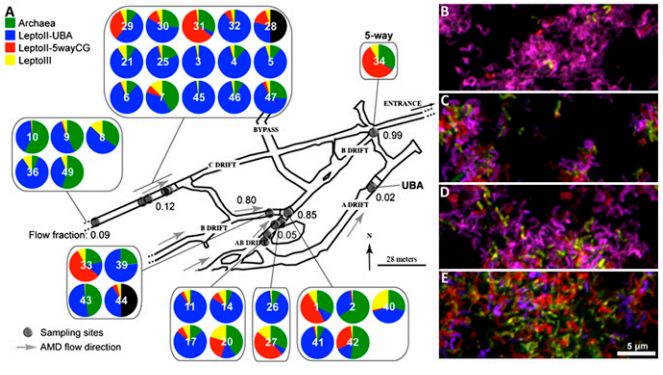

Fig. 2. Community composition of biofilms in the Richmond Mine. (*A*) Layout of Richmond mine tunnel system with indication of the sampling locations (gray dots). Each pie diagram represents FISH data for that sample, numbered as in Table S2 and listed chronologically (left to right, up to down) for each sampling location. Numbers next to the sampling locations indicate estimates for the fraction of the total flow at the mine outlet that passes at that point. Black segments indicate archaeal data were not available for these samples. Examples of FISH imaging for samples 11, 14, 17, and 20 are presented in (*B*), (*C*), (*D*), and (*E*), respectively (purple, *Leptospirillum* group II UBA; yellow, *Leptospirillum* group II five-way; red, other *Leptospirilla*).

We evaluated which environmental parameters best predict community composition, including the relative abundance of recombinant *Leptospirillum* group II genotypes. The BIOENV analysis, a statistical method to identify correlations between environmental factors and community composition, considered two sample subsets: the maximum number of samples (30 samples) and seven environmental parameters, and only the 22 samples for which 15 environmental parameters were available (Table 1). In addition, we considered either all organisms or only *Leptospirillum* group II (as fractions of all cells). Biofilm maturity (determined based on its thickness), more than geochemical conditions, correlated positively with the relative abundance of the five-way CG type (Table 1 and Fig. 3*A*). The power of the environmental parameters to predict which *Leptospirillum* group II genotypic group dominated increased slightly when either time (maximum samples) or conductivity of the AMD solution (maximum parameters) was included with developmental stage (Table 1); exclusion of the developmental stage significantly weakened the correlations, reducing the maximum correlation with a single factor from 0.44 to 0.26 (conductivity).

We extended the analysis by taking into account the proteomics-based, genome-wide typing of the *Leptospirillum* group II populations (Table S2). The five recombinant variants that constituted the UBA genotypic group were most strongly correlated with temperature, with recombinants carrying the largest fractions of five-way CG type genes occurring at higher temperatures (Table 1 and Fig. 3*B*). The inclusion of a flow dynamics parameter slightly improved the correlation. When considering the sample subset with 15 environmental parameters, inclusion of the Zn concentration increased the correlation with the biological data. It has to be noted that the high number of considered variables compared to the number of samples reduces the power of the analysis. However, for all but one analysis, random permutations of the biological data between samples (rows in the biological data table) resulted in lower maximum correlation scores in more than 95% of cases (Table 1). Another trend, consistent with higher diversity in more mature biofilms based on FISH, was the increase in number of *Leptospirillum* group II genotypes as biofilms matured (Tables S1 and S2).

Comparative Genomic and Proteomic Analysis of the *Leptospirillum* Group II Types. Our analysis focused on the 2,625 UBA type and the 2,588 five-way CG type proteins encoded by the composite genomes and excluded proteins only present in minor strain variant regions. Of these, 2,204 are orthologs (~84% of the gene complement of each organism). Overall, 1,595 *Leptospirillum* group II proteins (53% of the pangenome) were identified by proteomics in at least two samples, 1,473 orthologs (67%), and 122 of the proteins specific to either the UBA or five-way CG genome type (15% of all type-specific proteins).

Community proteomics data allowed us to evaluate the relative importance of proteins unique to either genotypic group. Based on the average normalized spectral abundance factor (NSAF), which provides a semiquantitative measure of each protein's relative abundance, the genotypic group–specific proteins contributed less than 1.5% of the total protein abundance and were identified by proteomics in considerably fewer samples than orthologs (Figs. S2 and S3*A*). Notable exceptions are proteins from two UBA type–specific regions that mostly encode hypothetical proteins/proteins of unknown function and proteins typical of mobile elements. Both regions contained multiple metal ion transporters, several of

Table 1. Correlation of biological data with environmental parameters as determined using the BIOENV procedure

Data set	Parameters	Best match	*R*	*P* value
Not considering recombinants				
All organisms, max samples (30)	Source, FDP*, time†, DS‡, flow§, pH, T	**DS**, time	0.46 (0.41)	1.0E-03 (2.0E-03)
Leptospirillum II, max samples (30)		**DS**, time	0.52 (0.44)	<1.0E-03 (2.0E-03)
All organisms, max parameters (22)	Source, FDP, time, DS, flow, pH, T,	**DS**, cond, FDP	0.42 (0.35)	2.1E-02 (6.5E-02)
Leptospirillum II, max parameters (22)	cond¶, Fe^{II}, Fe^{T}, Fe^{II}:Fe^{III}, Cu, As, Zn, Ca	**DS**, cond	0.47 (0.42)	1.3E-02 (3.4E-02)
Considering recombinants				
All organisms, max samples (23)	Source, FDP Time, DS, Flow, pH, T	**T**	0.48	1.0E-03
Leptospirillum II, max samples (23)		**T**, FDP, time	0.43 (0.37)	1.0E-03 (5.0E-03)
All organisms, max parameters (15)	Source, FDP, Time, DS, Flow, pH, T,	**T**, FDP, Zn	0.73 (0.55)	<1.0E-03 (1.7E-02)
Leptospirillum II, max parameters (15)	cond, Fe^{II}, Fe^{T}, Fe^{II}:Fe^{III}, Cu, As, Zn, Ca	**FDP**, Zn, T	0.70 (0.48)	<1.0E-03 (2.9E-02)

Correlation of biological data with environmental parameters was determined using the BIOENV procedure (45). The best combination of parameters is presented with its corresponding correlation factor and *P* value (based on a random permutation test). The parameter in boldface type is the strongest single correlating factor (correlation coefficient between parentheses). Biological data used for this analyses were the arcsin (square root) transformed fractions of the total community as determined by FISH for archaea, *L.* group III, *L.* group II five-way CG, and *L.* group II UBA and by PIGT for the five variants constituting the UBA genotypic group.

*Fluid dynamics parameter; the higher the more rapid flow and/or the narrower the flow path.

†Time of sampling.

‡Developmental stage of the biofilm, as defined by in situ observation of the biofilm thickness and scaled based on microscopic analysis of biofilm cross-sections (23).

§Daily discharge rate at the mine entrance.

¶Conductivity of the AMD solution.

which were identified (UBAL2_80620056,91,93,97), although only in a limited number of samples. Another set of unique genes that encode proteins that were consistently identified by proteomics is located at the center of a large five-way CG type-specific prophage region. Apart from proteins of unknown function, these include TonB-, TolQ-, and ExbD-like proteins (CGL2_1127 7177,178,176), which transduce proton motive force from the cytoplasmic membrane to outer membrane transporters. Although no genes encoding transport proteins were adjacent to this protein set, TonB-dependent transporters are generally involved in translocating large molecules such as iron siderophores and vitamin B12 but can also mediate phage infection.

Proteins that were identical between the two genotypic groups were less commonly identified than divergent proteins (Fig. S3*A*; Pearson's χ^2 test, $P < 0.05$). When examining this set of orthologs in more detail, it became clear that their reduced identification rate was mostly due to the low identification frequency (<25% of proteins were identified) of the subset of proteins located in integrated plasmid/phage regions (Fig. S3*B*).

Overall, more highly conserved proteins (>97.5% amino acid identity) contributed a disproportionally large fraction of the total cellular protein pool (Fig. S3*A*). Some functional categories were more dominated by highly conserved proteins than others (Fig. S3*A*). The proportion and abundance of proteins in different functional categories with high sequence divergence between the genotypic groups is interesting, given their potential roles in niche partitioning. Divergent proteins (<90% identity) that were frequently identified (more than one half of all samples for core proteins and more than one third of all samples for type-specific proteins) are listed in Table S3. In addition, the distribution among functional categories of proteins with low levels of sequence divergence may be informative in terms of shared traits required for the adaptation of both organisms to the AMD environment.

More than 60% of the total protein pool was composed of proteins from the "Translation," "Defense mechanism," "Post-translational modification, turnover and chaperones," "Energy," and "Function unknown" categories (Fig. S3*A*). Highly conserved proteins dominated the protein pool related to "Defense mechanisms" and "Post-translational modification, turnover and chaperones." These include thioredoxins, chaperones, and virus defense proteins (CRISPR-associated proteins in a region encoding proteins identical in both genotypic groups). Many restriction modification system genes unique to the five-way CG or UBA type were assigned to the "Defense mechanisms" category, but they were rarely identified by proteomics. "Translation" and "Energy" functional categories were also dominated by highly conserved proteins, including abundant ribosomal proteins and cytochromes. Most of the protein pool of the "Replication, recombination and repair" category originated from highly conserved proteins such as histone-like DNA-binding proteins, RecA and DNA replication machinery (Fig. S3*A*). Also included in this category were many transposases unique to one of the two genotypic groups, but few were identified by proteomics.

As compared with the other functional categories, greater than average numbers of identified proteins in the "None" and "Unknown function" categories were highly divergent between the two genotypic groups (<95% amino acid identity categories of

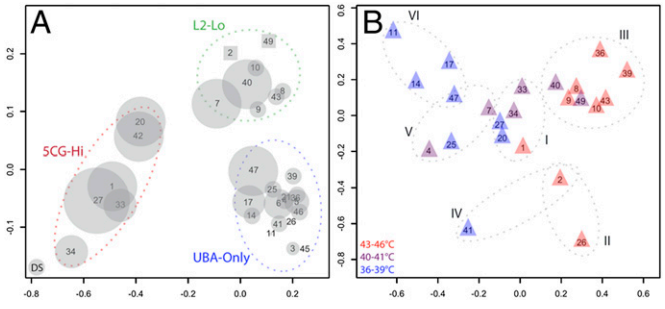

Fig. 3. Nonmetric multidimensional scaling (NMDS) of the biological data. (*A*) *Leptospirillum* group II data (fraction of whole community) for the maximum number of samples (Table 1). Final stress from NMDS, 2.8. Circle diameter is scaled based on the developmental stage (DS) of the biofilm. Samples 2 and 49 are represented as cubes, as they have an atypical biofilm morphology that does not fit the standard succession model. Groups: 5CG-Hi, samples with high levels of *Leptospirillum* group II five-way CG type; L2-Lo, samples with relatively low levels of *Leptospirillum* group II; UBA-Only, samples for which the *Leptospirillum* group II population only contained the UBA type. (*B*) All data including the quantitative estimates for the recombinants determined by PIGT (Table S2), when including maximum number of samples (Table 1). Final stress from NMDS, 14.8. Triangle color indicates the temperature of the AMD solution at the time of sampling. Groups I–VI indicate the dominant *Leptospirillum* group II genotype present in these samples. In both ordinations, sample groupings are intended as a guide for interpretation and are not statistically supported.

Fig. S3*A*). The "Membrane, cell wall and outer membrane biosynthesis," "Carbohydrate biosynthesis and transport," and "Cofactor biosynthesis and transport" categories also showed higher relative levels of sequence divergence. Notably, considering their direct role in environmental sensing, the "Signal transduction" category included a relatively high proportion of divergent and unique proteins (Fig. S3*A*). Although more than half of the so-called Signal transduction protein pool was contributed by 100% identical proteins, this pattern originated mostly from one protein (UBAL2_86920026).

Proteins with lower dN/dS values (i.e., under stronger purifying selection) were, on average, more abundant and identified in more samples (Fig. 4 and Fig. S4). Only two proteins had a dN/dS > 1, suggesting positive selection, one of which was identified by proteomics (UBAL2_82410389, a putative SAM-dependent methyltransferase). The average abundance and identification rate decreased with increasing dN/dS. Outliers of this trend include 17 proteins with average NSAF values that were more than five times the average for proteins with the same dN/dS. Another 14 proteins were found in significantly more samples than average for their respective dN/dS (Fig. 4, Fig. S4, and Table S3). These 31 proteins were predominantly proteins of unknown function, but also included proteins involved in key processes such as iron oxidation (cytochrome 572), compatible solute biosynthesis (trehalose-6-phosphate phosphatase) and nutrient uptake (PstS).

Strain-Resolved Comparative Proteomics. The best set of biofilm communities available to compare protein abundance patterns between coexisting genotypic groups originated from the C +10 m location, collected at different times. These samples were dominated by only two genotypes, type I and/or type V. Clustering of the six strain-resolved type V and four strain-resolved type I proteomic sets using a filtered set of 271 proteins (*Materials and Methods*) revealed a stronger clustering by sequence type than by biofilm sample (Fig. 5). Only the strain-resolved dataset of the type V population in sample 29 failed to cluster according to its sequence type. The clustering by genotype persisted when only proteins identified in all samples were included as well as when an additional normalization by dataset (which sets the sum of squares of the values for each protein to 1) was performed as part of the clustering procedure. This confirms that the clusters are robust and highlights the distinctiveness of these populations. Cofactor biosynthesis and motility-related proteins were overrepresented within the subset of genes that were more highly expressed in the type I strain, whereas proteins involved in energy production and

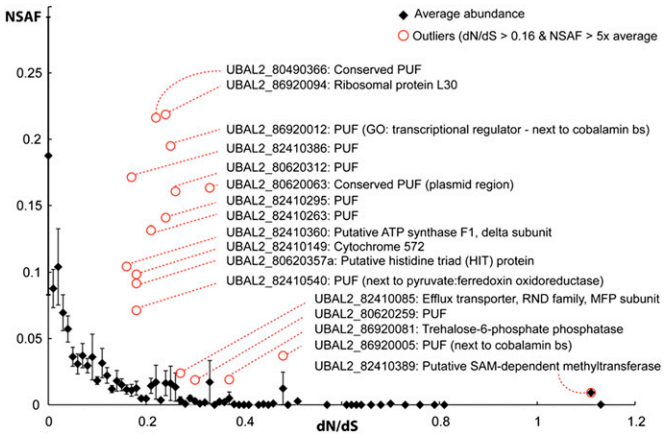

Fig. 4. Relationship between dN/dS and protein abundance. Average (and standard error) NSAF for all proteins with a dN/dS = x (black) as well as outliers (red, >5× the average NSAF for all proteins with its respective dN/dS) are presented.

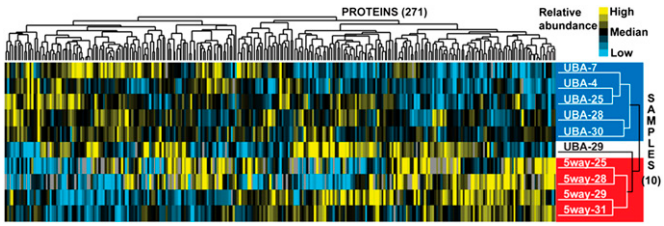

Fig. 5. Strain-resolved proteomic analysis. Average linkage hierarchical clustering of 10 strain-resolved datasets (horizontal) and proteins (vertical) based on the Kendall Tau distance matrix of the proteomics data (S-NSAFs). Blue and red shadings group strain-resolved protein abundance data sets that were used for SAM analysis (Fig. S5).

conversion were overrepresented within the group of more abundant proteins in the type V strain (Fisher's exact test, $P < 0.1$).

We determined which proteins were significantly more or less abundant in the late colonizer as compared to the early colonizer using the Significance Analysis for Microarrays (SAM) package. Based on the hierarchical clustering, the two groups compared were as follows: (*i*) strain-resolved data for the UBA genotype (type V) from samples 4, 7, 25, 28, and 30; and (*ii*) strain-resolved data for the five-way CG genotype (type I) from samples 25, 28, 29, and 31. Analysis was performed at a false-discovery rate <5% (Table S4), and proteins with significant differences in abundance with at least a 1.5-fold change are listed in Fig. S5.

Integrated Analysis. Differentially expressed proteins, proteins with an elevated dN/dS and an anomalously high frequency of identification or abundance, and expressed divergent proteins were integrated into one list and ordered by presumed cellular function (Table S3). Additional information on whether these proteins were located in recombinant regions was added as well. This list included proteins that constitute a coherent signal for an increased role in the late colonizer of (de)methylation reactions, cobalamin biosynthesis, and the glycine cleavage complex (Table 2).

FISH on Cryosectioned Biofilms. Three-dimensional reconstruction of CLSM z-stacks of a moderately thick (~100 μm) (23) biofilm growing into an AMD stream at the C +10 m location (samples 4–7) demonstrated an exclusionary and patchy distribution of five-way CG colonies within UBA-type dominated biofilms (Fig. 6).

Discussion

Several studies have examined the relationship between environmental conditions and microbial community structure (25, 26). However, investigations that consider microdiversity are rare (10, 27). Here, we showed that two *Leptospirillum* group II genotypic groups that differ by a mere 0.3% at the 16S rRNA locus displayed a different distribution in the same ecosystem. We found genotypic as well as protein expression differences that could have contributed to ecological divergence, advancing our understanding of how molecular evolution and ecological differentiation of these bacteria interrelate. The results highlight the potential importance to community functioning of what have previously been considered to be ecological redundancies (e.g., when organisms <1% divergent at the 16S rRNA gene are considered as a single group in studies correlating environmental conditions to community structure).

The main ecological difference between the two *Leptospirillum* group II genotypic groups in the Richmond Mine is a preference of the UBA genotypic group for early developmental stages and of the five-way CG genotypic group (represented by type I) for later developmental stages. The weak correlations between geochemical parameters and ecological distribution indicate that both genotypic groups can exist over the range of conditions prevailing in the system. This is associated with high sequence

Table 2. Proteins potentially underlying ecological divergence between the two genotypic groups

UBAL2*	dN/dS†	SAM‡	Rec§	EDP¶	%id‖	Ab**	Cnt††	Function	Notes
80620018	—	2.04/1.32	—	—	92.3	0.036	27	Uroporphyrin-III C-methyltransferase/ synthase (CobA)	Cobalamin biosynthesis
82410481	—	1.51/1.69	—	—	92.3	0.008	25	Putative cobalamin biosynthesis protein (CbiG)	Cobalamin biosynthesis
85240113	—	—	—	Yes	88.8	0.007	17	Putative cobalbumin biosynthesis enzyme (CobU)	Cobalamin biosynthesis
85240123	—	—	—	Yes	87.2	0.004	17	Cobyrinic acid a,c-diamide synthase (CbiA)	Cobalamin biosynthesis
86920005	0.37	—	—	Yes	85.8	0.019	20	Protein of unknown function	Cob neighbor
86920012	0.25	1.49/1.29	—	—	93.3	0.195	27	Protein of unknown function	Cbi neighbor/potential transcriptional regulator
79310208	—	2.15/1.69	—	—	99	0.117	27	Adenosylhomocysteinase	SAM biosynthesis
82410132	—	—	II, III, IV	Yes	81.4	0.003	13	Probable radical SAM family protein	Catalyzes methylations
82410389	1.11	—	—	—	91.4	0.009	23	Putative SAM-dependent methyltransferase	Methyl donor is SAM
85240195	—	1.74/1.29	—	—	97.4	0.107	27	Putative SAM-dependent methyltransferase	Methyl donor is SAM
85240110	—	—	—	Yes	86.4	0.003	15	Conserved protein of unknown function	Cob and thiopurine S-methyltransferase neighbor
85240112	—	3.7/1.82	—	—	98	0.057	27	Putative thiopurine S-methyltransferase	Methyl donor is SAM
82410229	—	2.08/1.73	II	—	97.4	0.03	27	Glycine dehydrogenase (decarboxylating) subunit 2	glycine -> CO_2 + 5, 10-MethyleneTHF + NH_4^+
82410231	—	1.58/1.68	II	—	93.3	0.059	27	Glycine cleavage system H protein	
82410232	—	—	II	Yes	88.9	0.028	27	Glycine cleavage system T protein	
82410616	—	−2.87/0.57	—	—	98.3	0.076	27	Glycine hydroxymethyltransferase	5,10-methyleneTHF+ glycine + H_2O = THF + L-serine.

Selection of proteins from Table S3 involved in cobalamin biosynthesis, (de)methylation reactions, and one carbon metabolism.
*Locus tag for *Leptospirillum* group II UBA (UBAL2_).
†Outlying values presented only.
‡Score (d) / fold change determined by SAM for proteins significantly more abundant at FDR < 5% (d > 0 indicates protein more abundant in five-way CG).
§Located in recombinant region of recombinant types X.
¶Expressed divergent protein, <90% amino acid identity and identified in >50% of samples.
‖Amino acid identity between UBA and five-way CG orthologs.
**Average protein abundance (NSAF).
††Number of samples in which the protein was identified.

conservation between abundant orthologous proteins involved in functions crucial to success in this extreme environment (e.g., molecular chaperones, peroxidoredoxins, rubryerythrin). The abundance of highly conserved energy production proteins, specifically particular cytochromes, and proteins presumably involved in carbon fixation (modified reverse TCA cycle) (22) reflects their common role in the community as primary producers.

Recombination between the early and late colonizer genotypes occurred recently, based on sequence conservation in shared regions across all genotypes and the record of the transfer in the CRISPR locus (28). We can consider allo- or sympatric divergence before these recombination events. The rise to high frequency of the recombinants could suggest divergence from a common ancestor to the ancestral UBA and five-way CG types after spatial isolation, followed by dispersal, mixing, and recombination (Fig. 1). A similar reasoning was proposed to explain recombination patterns within and between two closely related archaeal populations that coexist in the Richmond Mine (21, 29) and between two subpopulations of *Leptospirillum* group II five-way CG (12). The distinctive complement of transposases (and other mobile elements) in the UBA and five-way CG composite genome sequences also supports this scenario. If indeed the parental types diverged allopatrically, the current ecological partitioning of the two genotypic groups is not necessarily reflective of the context in which parental genome divergence occurred. Instead, adaptations may have arisen in two different environments that shared some features with the early and late biofilm developmental stage. Nevertheless, we cannot exclude the possibility that the observed patterns are the consequence of sympatric divergence. For this to be plausible, selective pressures exerted by different biofilm developmental stages should have been high enough to cause sufficient ecological separation and avoid homogenization due to recombination. The current proliferation of recombinant types could then be explained by recent environmental changes that have opened new niches.

The distribution of variants formed by recent recombination between the early and late colonizer genomes as a function of environmental parameters suggests ecological fine tuning. The much larger number of genotypes (five vs. one) comprised primarily of type VI rather than type I blocks indicates that cells in early-stage biofilms are more exposed to environmental gradients than in mature biofilms. The increased strength in correlation between environmental parameters, including temperature, and fine-scale community composition confirms this (Table 1). More pronounced exposure to the environment of the early colonizer is consistent with elevated abundances of proteins involved in arsenic resistance (Table S3), and the presence of additional metal transporters. Although the two genome types that occur at higher temperatures (type II and III; Fig. 1) (24) share a segment of late colonizer type genes that is absent in the other early colonizers, it is unclear which, if any, specific genes in this region affect this distribution. Once established, the biofilm matrix likely creates an environment less subject to variations in the AMD solution, limiting fitness effects of recombination events with type I genotypes as the receptor. On the other hand, the increased number of genotypes co-occurring in mature as compared to early stage biofilms might reflect the presence of multiple microniches.

During development of biofilms, nutrient and oxygen gradients develop (30). Genetic differences that confer a benefit to the late colonizer in this modified environment could include adaptations

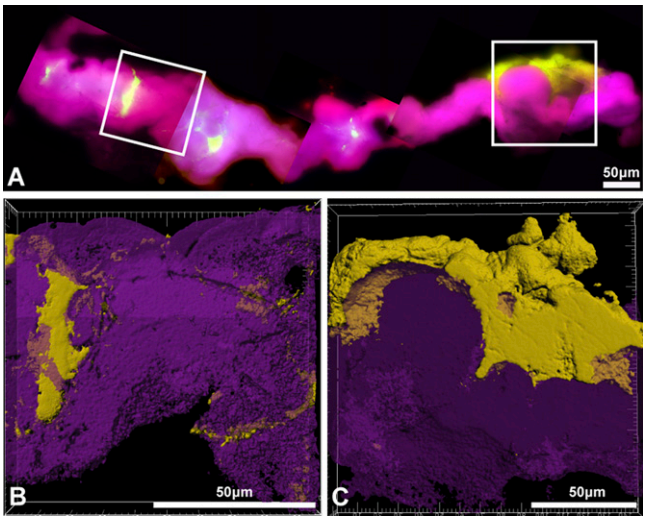

Fig. 6. Fine-scale distribution of the two genotypic groups. FISH on a sectioned biofilm from the C +10 m location (November 2006) to reveal the distribution of the *Leptospirillum* group II UBA (purple) and five-way CG genotypic groups (yellow). (*A*) Montage of epifluorescence micrographs. (*B* and *C*) Confocal microscopy imaging of boxed regions in (*A*). Not visible are archaea and fungi, which are mixed throughout the biofilm in stream biofilms (23). The biofilm is partly refolded on itself.

to low nutrient availability. The high-affinity phosphate transporter (PtsS) that displayed an elevated dN/dS despite relatively high expression, high sequence divergence, and elevated protein abundance in the late colonizer, could reflect an adaptive benefit during phosphate limitation. The higher abundance of motility proteins in the late colonizer could imply movement toward specific resources within the biofilm. However, the physicochemical gradients are not simply vertical, given the patchy colonization patterns observed in the imaged biofilm sections.

Strain-resolved proteomic data allow an initial investigation of the contributions to ecological divergence of proteins differentially expressed between the two genotypic groups. These differences are the result of both neutral drift, the magnitude of which correlates with genetic distance, as well as directed evolution for adaptation to the specific environmental conditions in which they evolved (19). Our approach differed from traditional measurements of within- and between-population expression, which are normally carried out under controlled conditions; rather, we integrated in situ expression data with comparative genomics to reveal variation that may form the adaptive basis of ecological differentiation between the two genotypic groups that coexist in a subset of cases.

Results point to the increased role in the late colonizer of cobalamin biosynthesis and cobalamin-dependent, *S*-adenosyl methionine–dependent methyltransferases, and glycine cleavage (Table 2). The late colonizer might obtain a competitive edge in mature biofilms because of an increased C_1 flux generated by the glycine cleavage complex, a major source of C_1 building blocks (31), and their use in methylations in anabolic pathways. The lower abundance in the late colonizer of the protein that generates serine from a C_1 and glycine is consistent with the use of the glycine cleavage complex for this purpose. The release of ammonium from glycine cleavage also might be important in more nutrient-limited mature biofilms.

No overall increase is apparent in amino acid or peptide uptake and degradation to supply C and N to the cell in mature biofilms. Conversion of serine or threonine to glycine is not supported by our findings either. Instead, our data suggest that the source of glycine is the compatible solute betaine (trimethyl glycine). *Leptospirillum* group II can transport betaine produced by other mature biofilm community members into the cell (ectoine/betaine/proline transporter, UBAL2_82410060), or produce this compound itself by choline degradation or less likely, glycine methylation. The known betaine degradation pathway occurs by sequential demethylation (32) and seems to be incomplete in *Leptospirillum* group II, although certain (de)methylation proteins of unknown function that are more abundant in the late colonizer could be involved. If so, the high sequence divergence of proteins involved in betaine conversion and their increased abundance suggest a higher turnover of compatible solutes in the late colonizer. This could help explain the late colonizer's increased competitiveness in the later stages of biofilm development. Our hypothesis is supported by elevated sequence divergence, dN/dS, and abundance in the early colonizer of the protein catalyzing the final step of trehalose biosynthesis, another compatible solute (UBAL2_86920081 is twice as abundant, but not listed in Tables S3 and S4, as data were available for only 8 of 10 strain-resolved data sets).

The higher fitness of the five-way CG type in later stages of biofilm development is supported by increased abundance of several main ribosomal operon proteins when the two types co-occurred (Table S3). In contrast, a ribosomal protein located outside of the main operons was more abundant in the early colonizer. Discordance in ribosomal protein abundance patterns was also noted when comparing exponential to stationary growth in SAR11 (33). Whether these proteins are truly more abundant in the less active organism or whether posttranslational modifications in the more active organism lower identification of its peptides by MS is yet to be tested.

The ecological divergence between the two genotypic groups can be viewed as an example of *r*- vs. *K*-selection (34). The UBA group proliferates optimally in the absence of competition from other organisms, thanks to adaptations that allow it to rapidly propagate in the AMD environment. Adaptations of the five-way CG genotypic group allow its proliferation in conditions with high inter- and intraspecific competition for resources. Distinct ecological strategies between closely related organisms exemplify how fine-scale variation within ecological functional groups can have significant effects on community structure and functioning.

Conclusion

The availability of genome sequences for closely related microorganisms has at the same time clarified and complicated our view of species delineation. Although the 16S rRNA gene based classification generally corresponds to genomic and ecological differences (35), organisms grouped as one species often display both significant gene content variation (11) as well as resource partitioning (10). Uptake and/or loss of genes is clearly a mechanism used for adaptation to the local environment (36), and relevance of unique gene content to ecological divergence was evident when comparing more distantly related *Leptospirillum* group II and III species in the same system (8% 16S rRNA divergence) (22). However, our current data show that only a small fraction of the gene pool unique to either of the *Leptospirillum* group II genotypic groups was expressed in natural communities. This supports the argument that much of the laterally transferred gene pool found in closely related isolate genomes is of a transient, nonadaptive nature (37–39). Our data emphasize the role of sequence and expression variation of shared genes in ecological divergence. As such, we highlight an interesting parallel to higher organisms, in which evolution of gene expression has been suggested as an important factor in species differentiation (19). By integrating the analysis of genotypic differentiation and its translation to divergent expression of niche-determining genetic loci, we have shown how subtle genomic differences between coexisting bacteria contribute to their ecological differentiation within an ecosystem.

Materials and Methods

Samples. Forty-five natural community samples, all from biofilms growing at the liquid–air interface, were collected from the Richmond Mine, Iron Mountain, CA (40° 40′ 38.42″ N and 122″ 31′ 19.90″ W, elevation of ∼900 m) between March 2002 and August 2007 (Table S1 and Fig. 2). Samples 11–13, 14–16, 17–19, 21–22, and 23–25 were biological replicates taken from biofilms at the same location and time.

Physical and Geochemical Measurements. Temperature, pH, and conductivity (as a measure of total ionic strength) of acid mine drainage solution were measured in situ from each location that biofilm was sampled. Samples for metal analysis were collected, preserved, and analyzed according to USGS protocols (40) as described further in Table S1.

The developmental stage of the biofilm was represented by a number ranging between 0.1 and 5 to reflect the change in biofilm thickness as biofilms mature. This corresponds to a range in thickness of ∼5 to ∼250 μm, as determined in recent imaging experiments (23). Essentially all AMD from the Richmond Mine is piped from the system and this total flow was measured near the mine entrance. The source of the AMD solution at the sampling locations was evaluated based on the mine configuration (41) (Fig. 2; tunnel A, 1; tunnel B, 2; tunnel C, 3). In cases in which two tunnels join upstream of the sampling location, the source of the AMD solution was assigned a number that reflects the contribution of flow from the different tunnels, as indicated in Fig. 2 and Table S1. To differentiate sampling locations based on local flow characteristics, a fluid dynamics parameter was calculated by dividing the local contribution to the total flow by the dimensions of the flow path.

FISH. Two probes were designed specifically targeting the 23S rRNA of *Leptospirillum* group II five-way CG type (FITC-L2CG353: 5′-TcggtcctccgctgcgttT-3′) and *Leptospirillum* group II UBA-type (Cy3-L2UBA353: 5′-Tcggccctccgcc-gcgctT-3′). These two probes were used in combination with a probe targeting *Leptospirillum* groups I–III (Cy5-LF655) (42). Another probe mixture was used to determine the fraction of archaeal (Cy5-ARC915) (43), bacterial (Cy3-EUB338-mix) (43), and *Leptospirillum* group III cells (FITC-LF1252) (42). Hybridizations were performed using standard protocols (43) at a formamide concentration of 35%. Newly designed probes were optimized using isolate cells of *Leptospirillum* group II UBA and five-way CG types (44). Cells were counted after epifluorescence microscopy imaging (Leica DMRX) and are reported as fractions of the total cell count obtained by DAPI staining.

An intermediate developmental stage biofilm at the C +10 m location (November 2006; samples 4–7) was cryoembedded, sectioned, and hybridized as previously described (23) using FITC-L2CG353, Cy3-L2UBA353 probes. Biofilm sections were imaged on a on a standard epifluorescent microscope (Leica DMRX; Leica Microsystems, Bannockburn, IL) and a confocal laser scanning microscope (Zeiss LSM 510 META; Carl Zeiss MicroImaging Inc., Thornwood, NY). Z-stacks were combined, reconstructed and projected using the IMARIS software package (Bitplane AG, Zürich, Switzerland).

Statistical Analysis of FISH Data. To determine which of the measured environmental parameters correlated best to the biological data, we used the BIOENV procedure. This is an R implementation of the multivariate statistical method devised by Clarke and Ainsworth (45) to link community structure to environmental variables (developed by J. Oksanen as part of the VEGAN package). Briefly, dissimilarity matrices using Euclidean distance measures for the environmental parameters and Bray-Curtis dissimilarity measures for the biological data (arcsin of the square root of the transformed relative abundance data determined by FISH or proteomics-inferred genome typing [PIGT]) were calculated. Environmental distance matrices were calculated for each possible combination of factors (from one factor to all factors). Subsequently, Spearman rank correlations were calculated between the biological distance matrix and each environmental distance matrix. Significance levels of correlations were determined by comparing them to the distribution of maximum BIOENV correlations observed in 1,000 permutations obtained by randomizing row order in the biological data table. For visualization of the similarity between samples based on their biological data and an overlay of the most strongly correlated environmental parameter, nonmetric multidimensional scaling (NMDS) was performed in R.

Comparative Genomics. We previously reconstructed and manually curated genomic assemblies for *Leptospirillum* group II UBA and five-way CG from community genomic datasets from samples from the UBA [June 2005 (22, 44)] and five-way locations [March 2002 (12, 21)] in the Richmond Mine. Orthologous genes shared by the UBA and five-way CG *Leptospirillum* group II types were determined earlier (44). Functional categories were assigned through manual curation using the clusters of orthologous groups classification (COG) ontology with addition of "Unknown transport" (Table S5). The category "None" contained predicted proteins without homologs outside of *Leptospirillum* group II (annotated as hypothetical proteins or protein of unknown function when identified by proteomics), whereas the category "Unknown function" contained proteins with no assigned function that have homologs outside of *Leptospirillum* group II (annotated as conserved hypothetical proteins and conserved proteins of unknown function).

Support for positive versus purifying selection was sought based on the relative proportions of synonymous (S) and nonsynonymous substitutions (N) between orthologs. Orthologous proteins were aligned using Clustalw2 (http://www.ebi.ac.uk/Tools/clustalw2/), replaced by their nucleotide sequences, and dN/dS ratios calculated using a modified version of the SNAP perl script by B. Korber based on work by Nei and Gojobori (46).

Proteomics Data Acquisition. All proteomics data derived from datasets generated previously (24). For 27 samples, a whole-cell protein fraction was extracted from 1 g of biofilm and analyzed by nano-2D-LC (strong cation exchange – reversed phase) - MS/MS on a hybrid LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, San Jose, CA). Peptide identification relied on spectral assignments using the Sequest algorithm (47) using the search database described in detail in (24). At least two peptides had to be identified in the same run at Xcorr > 1.8 (+1), 2.5 (+2) and 3.5 (+3) and ΔCN > 0.08 to deem a protein identified. All spectra and DTAselect files are available online at (http://compbio.ornl.gov/biofilm_amd_PIGT). Data from three technical replicates were used for each sample. The total and unique sequence count, total and unique spectral count, and sequence coverage for each protein (cumulative over the three replicates) were used in further analyses. Abundance estimates were calculated based on an adapted version of the normalized spectral count abundance factor (NSAF) (48). In brief, we divided total spectral count for a protein by its amino acid length to normalize for the correlation between protein length and peptide identification probability. This number was divided by the sum of length-normalized spectral counts of all proteins to obtain the NSAF. To obtain estimates for the fraction of the total *Leptospirillum* group II proteome contributed by a specific protein, the denominator was calculated by summing the length-normalized spectral counts for *Leptospirillum* group II proteins only.

Comparative Proteomics. We determined the protein expression distribution across the 22 functional categories in *Leptospirillum* group II using the fraction of proteins in each category identified by proteomics and a measure of protein abundance (NSAF). We limited the analysis to proteins that were identified at least twice in the 27 communities to minimize effects caused by false-positive identification (which we previously showed to occur at a rate of <1%) (44). To identify categories that might be key to ecological divergence, we subdivided the proteins in each functional category into proteins shared between and unique to each *Leptospirillum* group II genotypic group, and also based on the level of amino acid identity between orthologs. The amino acid identity values for subcategory boundaries were chosen so that each subcategory contained an approximately equal number of proteins.

Calculation of Strain-Resolved Normalized Spectral Abundance Factor. To compare protein abundances of co-occurring *Leptospirillum* group II UBA and five-way CG types, we performed strain-resolved proteomic analyses on samples 4, 7, 25, 28, 29, 30, and 31. These samples all originate from the same location (C +10 m) and are composed of different mixtures of type I, similar to the type characterized in the five-way community genomic dataset, and type V, a recombinant variant closely related to the genome type characterized in the UBA community genomic dataset. Strain-resolved NSAF values were determined for the subset of proteins that had at least one unique spectral count for one of both protein variants (characterized in the UBA or five-way community genomic dataset) but were not located in the regions where the recombinant variant genome (type V) was of the five-way CG type. For each protein variant, the "strained" spectral count was calculated by adding the unique spectral count for that variant and the fraction of the nonunique spectral count determined by the ratio of unique spectral counts for the ortholog pair. For each of the two genomic types, the strain-resolved normalized spectral abundance factor (S-NSAF) of each protein was calculated by dividing the length-normalized strain-resolved spectral count by the sum of all length-normalized strain-resolved spectral counts for the respective genome type.

S-NSAF Clustering. The strain-resolved normalized data were clustered using Cluster 3.0 and visualized using Treeview (49). Each S-NSAF value was transformed by taking the arcsin of the square root, which approximates the distribution of proportional data to a normal distribution. To equalize the

effect of each protein's abundance profile on the sample clustering, values for each protein were placed on a single scale. Specifically, the transformed S-NSAF values for each protein were median-centered across all samples, and normalized to set the sum of the squares of the values to 1 (resulting in a common blue (low) to yellow (high) scale in the visualization of each protein with Treeview). Proteins were clustered using self-organizing maps (SOM) using the default parameters of Cluster 3.0 and the Kendall Tau distance matrix metric. To achieve a relatively smooth transitioning between groups of clustered proteins, SOM clusters were used to guide the final hierarchical clustering. We used average linkage clustering based on the Kendall Tau distance matrix of both proteins (horizontal axis in figure) and strain-resolved proteomic data sets (vertical axis in figure). As the primary factor influencing clustering of five-way CG variants in samples 4, 7, and 30 and UBA variants in sample 31 was the small number of protein identifications, these values were excluded from the clustering analysis. Only proteins for which variants were identified in at least 90% of the remaining strained datasets (nine of 10) were used for clustering. These filtering procedures limited the analysis to 271 proteins.

Statistical Analysis of Protein Expression. We determined which of the 271 proteins used for clustering were significantly more abundant in either genotypic group. Groups delineated by hierarchical clustering were compared using the R implementation of the Significance Analysis of Microarrays (SAM) procedure (50), which has been applied to proteomics (51). S-NSAF data were $\log_2$ transformed, and significance analysis was performed at a false discovery rate < 0.05 (at a $\delta = 0.68$).

ACKNOWLEDGMENTS. We thank Mr. T. W. Arman (President, Iron Mountain Mines Inc.) and Mr. R. Sugarek (U.S. Environmental Protection Agency) for site access, and Mr. R. Carver for on-site assistance. We thank Banfield laboratory members for their contributions to sample collection. B. Suttle (Imperial College, UK) and C. Miller (University of California Berkeley) are thanked for assistance with and discussion of statistical analyses. We thank C. Miller for critical reading of the manuscript. P. Abraham, M. Lefsrud, M.B. Shah, and D. Schmoyer (Oak Ridge National Laboratory [ORNL]) for their assistance with proteomic measurements and analysis. D.K. Nordstrom and B. McCleskey (U. S. Geological Survey, Boulder) for advice on field protocols and assistance with the metal analyses. We thank Dr. J.P. Gogarten and Dr. M.F. Polz for critically reviewing our paper before publication. ORNL is managed by University of Tennessee–Battelle LLC for the Department of Energy under contract DOE-AC05-00OR22725. This project was funded by Grant DE-FG02-05ER64134 from the U.S. DOE Genomics: GTL program (Office of Science).

1. Wilmes P, Simmons SL, Denef VJ, Banfield JF (2009) The dynamic genetic repertoire of microbial communities. *FEMS Microbiol Rev* 33:109–132.
2. Carlson CA, et al. (2009) Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J* 3:283–295.
3. Maixner F, et al. (2006) Nitrite concentration influences the population structure of *Nitrospira*-like bacteria. *Environ Microbiol* 8:1487–1495.
4. Xu J, et al. (2007) Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biol* 5:1574–1586.
5. Hallam SJ, et al. (2006) Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc Natl Acad Sci USA* 103:18296–18301.
6. Bhaya D, et al. (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J* 1:703–713.
7. Kettler GC, et al. (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* 3:2515–2528.
8. Acinas SG, et al. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* 430:551–554.
9. Eckburg PB, et al. (2005) Diversity of the human intestinal microbial flora. *Science* 308: 1635–1638.
10. Hunt DE, et al. (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320:1081–1085.
11. Thompson JR, et al. (2005) Genotypic diversity within a natural coastal bacterioplankton population. *Science* 307:1311–1313.
12. Simmons SL, et al. (2008) Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol* 6:1427–1442.
13. Vicente M, Mingorance J (2008) Microbial evolution: The genome, the regulome and beyond. *Environ Microbiol* 10:1663–1667.
14. Konstantinidis KT, et al. (2009) Comparative systems biology across an evolutionary gradient within the *Shewanella* genus. *Proc Natl Acad Sci USA* 106:15909–15914.
15. Wilmes P, et al. (2008) Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J* 2:853–864.
16. Mandel MJ, Wollenberg MS, Stabb EV, Visick KL, Ruby EG (2009) A single regulatory gene is sufficient to alter bacterial host range. *Nature* 458:215–218.
17. Whitehead A, Crawford DL (2006) Variation within and among species in gene expression: Raw material for evolution. *Mol Ecol* 15:1197–1211.
18. Khaitovich P, et al. (2004) A neutral model of transcriptome evolution. *PLoS Biol* 2:682–689.
19. Whitehead A, Crawford DL (2006) Neutral and adaptive variation in gene expression. *Proc Natl Acad Sci USA* 103:5425–5430.
20. Schrenk MO, Edwards KJ, Goodman RM, Hamers RJ, Banfield JF (1998) Distribution of thiobacillus ferrooxidans and leptospirillum ferrooxidans: Implications for generation of acid mine drainage. *Science* 279:1519–1522.
21. Tyson GW, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43.
22. Goltsman DSA, et al. (2009) Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing "Leptospirillum rubarum" (Group II) and "Leptospirillum ferrodiazotrophum" (Group III) bacteria in acid mine drainage biofilms. *Appl Environ Microbiol* 75:4599–4615.
23. Wilmes P, et al. (2009) Natural acidophilic biofilm communities reflect distinct organismal and functional organization. *ISME J* 3:266–270.
24. 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.
25. Zhou J, et al. (2002) Spatial and resource factors influencing high microbial diversity in soil. *Appl Environ Microbiol* 68:326–334.
26. Van der Gucht K, et al. (2007) The power of species sorting: Local factors drive bacterial community composition over a wide range of spatial scales. *Proc Natl Acad Sci USA* 104:20404–20409.
27. Ramette A, Tiedje JM (2007) Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem. *Proc Natl Acad Sci USA* 104: 2761–2766.
28. Tyson GW, Banfield JF (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* 10:200–207.
29. Eppley JM, Tyson GW, Getz WM, Banfield JF (2007) Genetic exchange across a species boundary in the archaeal genus *ferroplasma*. *Genetics* 177:407–416.
30. Stewart PS, Franklin MJ (2008) Physiological heterogeneity in biofilms. *Nat Rev Microbiol* 6:199–210.
31. Pizer LI (1965) Glycine synthesis and metabolism in *Escherichia coli*. *J Bacteriol* 89: 1145–1150.
32. Meskys R, Harris RJ, Casaite V, Basran J, Scrutton NS (2001) Organization of the genes involved in dimethylglycine and sarcosine degradation in *Arthrobacter* spp.: Implications for glycine betaine catabolism. *Eur J Biochem* 268:3390–3398.
33. Sowell SM, et al. (2008) Proteomic analysis of stationary phase in the marine bacterium "Candidatus Pelagibacter ubique". *Appl Environ Microbiol* 74:4091–4100.
34. Andrews JH, Harris RF (1986) R-selection and K-selection and microbial ecology. *Adv Microb Ecol* 9:99–147.
35. Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* 361:1929–1940.
36. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.
37. Kuo C-H, Ochman H (2009) The fate of new bacterial genes. *FEMS Microbiol Rev* 33: 38–43.
38. Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* 3:679–687.
39. Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. *Trends Genet* 25:107–110.
40. McCleskey RB, Nordstrom DK, Naus CA (2004) U.S. Geological Survey Open File Report, 2004-1341 (USGS, Denver, CO).
41. Druschel GK, Baker BJ, Gihring TM, Banfield JF (2004) Acid mine drainage biogeochemistry at Iron Mountain, California. *Geochem Trans* 5:13–32.
42. Bond PL, Banfield JF (2001) 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.
43. Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59:143–169.
44. Lo I, et al. (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* 446:537–541.
45. Clarke KR, Ainsworth M (1993) A method of linking multivariate community structure to environmental variables. *Mar Ecol Prog Ser* 92:205–219.
46. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418–426.
47. 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.
48. Florens L, et al. (2006) Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* 40:303–311.
49. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95:14863–14868.
50. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98:5116–5121.
51. Roxas BA, Li Q (2008) Significance analysis of microarray for relative quantitation of LC/MS data in proteomics. *BMC Bioinformatics* 9:187.