Genomics for environmental microbiology

Adam M Deutschbauer[1,2,*], Dylan Chivian[1,*] and Adam P Arkin[1,2,3]

The utilization of natural microbial diversity in biotechnology is hindered by our inability to culture the vast majority of microorganisms and the observation that laboratory engineered bacteria rarely function in the wild. It is now clear that an understanding of the community structure, function and evolution of bacteria in their natural environments is required to meet the promise of microbial biotechnology. To meet these new challenges, microbiologists are applying the tools of genomics and related high-throughput technologies to both cultured microbes and environmental samples. This work will lead to new views on ecosystems and biological function together with the biotechnology enabled by this science.

Addresses
[1] Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail stop 977-152, Berkeley CA 94720, USA
[2] Howard Hughes Medical Institute, Berkeley CA, USA
[3] University of California at Berkeley, Department of Bioengineering, Berkeley CA, USA
[*] These authors contributed equally to this work.

Corresponding author: Arkin, Adam P (aparkin@lbl.gov)

Current Opinion in Biotechnology 2006, **17**:229–235

Edited by David A Stahl and Michael Wagner

Available online 2nd May 2006

0958-1669/$ – see front matter

DOI 10.1016/j.copbio.2006.04.003

Introduction

Although bacteria and their viruses represent only part of the vast interconnected web of life that make up the global ecosystem, in numbers they make up the majority [1]. The variety of environments in which they live, the strategies they use to survive and grow, and the substrates they transform in that service lead to a wealth of forms and functions, the extent of which we are only beginning to understand. The growth and death of subpopulations of microbes in response to environmental change and their invasion into new niches can lead to large changes in the balance of a local ecosystem and can lead to interference with human operation — with effects ranging from the corrosion of oil lines to increasing the prevalence of and the introduction of new pathogenic strains. Beneficial aspects of microbial populations include their contribution to important geochemical cycles, their ability to buffer environmental change through bioremediation, and the possibility that they can provide a wealth of new functions for energy conversion, catalysis and natural product synthesis.

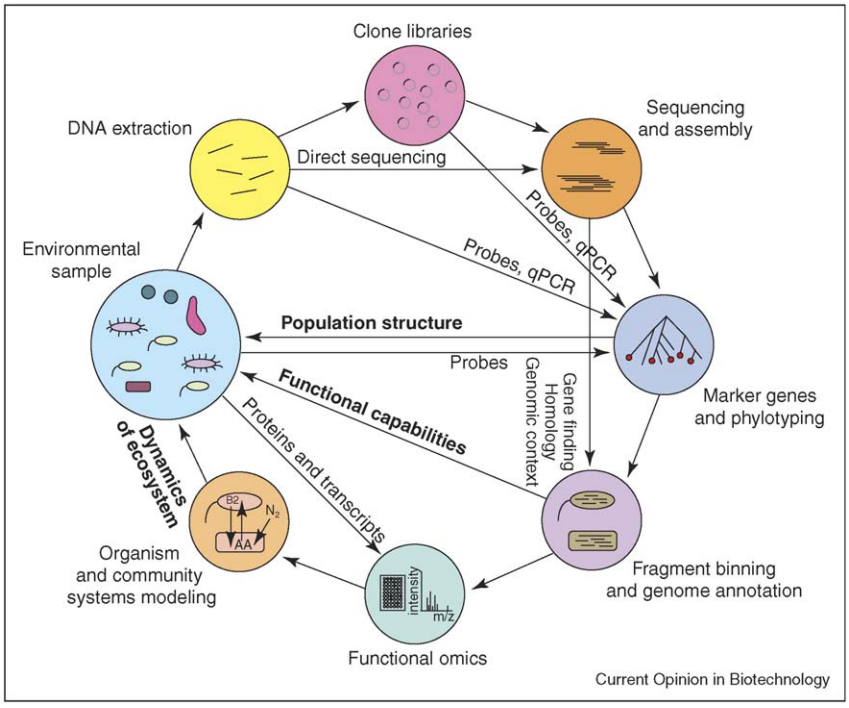
High-throughput sequencing and advances in DNA cloning and amplification technology, coupled with genomic tools, are enabling holistic views into the composition and dynamics of predominantly unculturable microbial communities (Figure 1). This emerging field, termed 'metagenomics' [2], offers new discoveries into the capabilities of microbes that allow them to collaborate and compete to survive in a wide range of environments. Genomic investigations into the diversity of environmental bacteria are leading to insights into ecological dynamics, the evolution of new forms of biological systems, and the discovery of new functions that might be exploited for biotechnological and biomedical purposes. Here we outline the broad questions in environmental microbiology that can be approached using genomic and derivative techniques. The unification of microbial ecology, environment, and gene function that is enabled by this science is discussed as are the associated experimental and bioinformatics challenges that must be overcome to reach this goal.

What is there?

Efforts to understand the biological composition of environments and the nature of engendered ecologies and their place in regional and global geochemistry are considerably aided by the identification of the constituent organisms. Modern approaches, made possible by genomic technologies, provide a much broader ability to access this diversity than traditional microscopic and culturing techniques. Current estimates indicate that less than 1% of microbial species are amenable to growth in isolation under standard laboratory conditions [3]. Instead, sequencing and other techniques for identifying DNA from environmental samples can yield a far more complete picture of the organisms involved in a community and, ultimately, the placement of those organisms into their ecological roles.

Identification and classification of both well-known and novel organisms is greatly aided by phylogenetic marker genes, which result from the commonality of certain tasks such as transcription and translation. Although conserved protein-coding genes can be used to identify bacterial lineages [4], the most commonly used phylogenetic markers are genes for RNA subunits of the ribosome, most frequently the small subunit (16S rRNA gene); these have been used in numerous studies to determine the presence and relative abundance of taxonomic groups within environmental samples. These studies have either sequenced rRNA genes directly [5], used PCR amplification to scan rRNA genes in large-insert clones built from environmental samples or employed sequencing followed

Figure 1



The application of genomics and derivative technologies yields insight into ecosystems. The use of genomics, functional genomics, proteomic and systems modeling approaches allows for the analysis of community population structure, functional capabilities and dynamics. The process typically begins with sequencing of DNA extracted from an environmental sample, either after cloning the DNA into a library or by affixing to beads and direct sequencing. After the sequence is assembled, the computational identification of marker genes allows for the identification and phylogenetic classification of the members of the community and enables the design of probes for subsequent population structure experiments. The assignment of sequence fragments into groups that correspond to a single type of organism (a process called 'binning') is facilitated by identification of marker genes within the fragments, as well as by other characteristics such as G+C content bias and codon usage preferences. Computational genome annotation, consisting of the prediction of genes and assignment of function using characterized homologs and genomic context, allows for the description of the functional capabilities of the community. Knowledge of the genes present also enables functional genomic and proteomic techniques, applied to extracts of protein and RNA transcripts from the sample. These latter studies inform systems modeling, which can be used to interpret and predict the dynamics of the ecosystem and to guide future studies. qPCR, quantitative polymerase chain reaction.

by computational identification of the 16S rRNA gene. One key benefit has been the identification of novel species, clades and divisions, which guides future research into a more balanced understanding of the tree of life. Additionally, these studies have revealed that the diversity of different communities can vary dramatically from just a few species to thousands [6•], often to a much greater extent than had been expected, and possess members that had not been previously identified. For example, one study using the 16S rRNA gene discovered a new clade that is one of the most abundant members of bacterioplankton communities in the ocean [7], but which had escaped detection by other approaches. Subsequent culturing and genomic sequencing of one of the members of this clade, *Pelagibacter ubique*, revealed a streamlined genome with fewer genes than any free-living bacteria to date [8], adding to its interest and highlighting the value of initial survey studies in discovering important organisms for further investigation. Similar studies have led to the isolation and characterization of other important marine bacteria [9,10] and archaea [11].

Temporal and spatial studies of microbial population structure beyond initial surveys are also of great importance, and will facilitate comparative analyses of community composition that will yield insight into the relationship of the ecology with the conditions that favor one population structure over another (see also Update). In light of this, we expect less labor-intensive approaches than sequencing that capture the presence of organisms in environmental samples, such as those based on the hybridization of probes to the 16S rRNA gene, to prove useful as researchers attempt to more rapidly characterize populations, potentially even in the field. One such approach takes advantage of fluorescence *in situ* hybridization to ribosomal genes within the sample [12], whereas others make use of microarray technology [13–17]. In the former approach, identification of only a small number of types is

possible in a given experiment, owing to the uniform application of the probes and the limited number of fluorescent dyes that can be utilized at once. In the latter approach, the separation of the probes allows for the identification of far more types. Additionally, probes of varying taxonomic specificity can be used, allowing for rapid classification, and perhaps quantification, of the organisms present within a sample. However, one drawback of techniques based on probes is that such investigations are limited to the identification of known groups and will fail to capture the presence of truly novel organisms. These approaches might be best applied once a better understanding of the organisms expected to be present is achieved by an initial survey to allow for the synthesis of custom probes. Nevertheless, it is the expectation of the authors that as such spatial and temporal population structure studies are one of the essential directions for the field, the rapidity and relative inexpensiveness of probe-based approaches will lead to their frequent use in future studies.

What is it doing?

In contrast to phylogenetic marker-based studies that survey the microbes present in an environment, DNA sequencing of environmental samples addresses the functional capabilities of the constituent organisms through analysis of the community gene complement. Recent studies utilizing 'shotgun' environmental sequencing reflect the challenges associated with these studies and the conclusions that can be drawn from them. The primary challenge to 'piece together' fragmentary sequences to determine the genetic content of each species in a community is greatly affected by the complexity of the sample, the comprehensiveness of the sequencing, and the length of the fragments themselves. For example, the shotgun approach taken in studies of the Sargasso Sea [18•], soil and whale carcasses [6•] yielded hundreds to thousands of unique species, but reads that were possible to group together were primarily from those species that already had sequenced genomes. Additionally, although these undertakings were huge in scope, ultimately the sequencing was not comprehensive over all regions of the genomes of each species, and could not be comprehensive owing to the large amount of strain variation. However, the nature of microbial genomes (gene-rich with small genes) permitted the quantitative assessment of gene repertoires in each of the sampled environments. These 'environmental gene tags' could be used to distinguish environments using differences in the inferred metabolic activities and functional roles of each microbial community [6•]. In contrast to the challenges presented by more complex communities, the relatively low species complexity of an acid mine drainage (AMD) biofilm lent itself more readily to the assembly of two nearly complete genomes and the partial assembly of three additional genomes [19•]. This achievement allowed for the assignment of the roles played by the members of the community, such as the appearance that only one of the members possesses the capacity for essential nitrogen fixing, and highlights the potential of metagenomic studies to reveal the keystone organisms within a given ecology and the interactions between community members.

Given the cost and complications of genome assembly associated with environmental shotgun sequencing, it is often practical and informative to sequence large-insert fosmid or BAC (bacterial artificial chromosome) clones to meet research goals. The ability to retrieve archived DNA in the form of a large-insert library greatly aids phylogenetic identification of the clones, assembly of metagenome DNA sequence, linking of genes with organisms, and the linking of genes and organisms to environmental function. For example, Hallam *et al.* [20] sequenced a marine sediment-derived fosmid library enriched for archaeal DNA to demonstrate that an uncultured archaea contains the gene complement required to oxidize methane. Similarly, DeLong *et al.* sequenced fosmid clones derived from marine microbial communities isolated from multiple depths at a single site in the Pacific Ocean to link microbial diversity to oceanic parameters like nutrients, salinity, temperature and the availability of light [21•].

Despite the insight into microbial functional capacity provided by environmental DNA sequences, these data in isolation are typically not sufficient to determine gene function. This is highlighted by the large number of both conserved and non-conserved genes with unknown function in individual bacterial genomes and metagenomes. The discrepancy between our ever-increasing sequencing capacity and our inability to systematically determine gene function is exasperated by the realization that a comprehensive understanding of microbial life requires the elucidation of complex interactions and dynamics between genes, organisms and their environment. It is clear that omics level technologies derived from primary sequence information are necessary to make the transition from gene and genome catalogues to functional significance.

Microarray-based gene expression profiling provides a quantitative assessment of transcript abundance and can be used to predict gene function based on the hypothesis that functionally related genes are more likely to be transcriptionally coregulated. In natural microbial communities, microarray technology can be applied both as a tool to monitor critical gene activities across a diverse spectrum of genomes [22,23] or to access the transcriptome of single microbial strains in a complex community. Regardless of the nature of the study, substantial challenges (e.g. efficient RNA extraction, detection of signal above background noise for complex samples, and cross-hybridization) need to be overcome before environmental

gene expression studies approach the reproducibility of similar laboratory-based analyses. Finally, compared with large-scale sequencing, gene expression studies are more amenable to time-course studies. The significance of this is that the dynamics of a complex microbial community in a changing environment can be tracked with a single, portable experimental tool.

In addition to gene expression, environmental proteomics are enabled by the availability of near-complete microbial metagenomes. Through alignment of mass-spectrometry-generated peptide signatures to the assembled AMD biofilm metagenome [19•], high-confidence detection was achieved for ~2000 proteins including ~50% of the predicted proteins from the high-abundance *Leptospirillum* group II strain [24••]. Although DNA sequence illustrates the metabolic and functional potential of an organism, the detection of expressed proteins in a community provides critical insight into the important cellular activities at temporal and spatial environmental resolution. In the AMD proteomics study, many genes with a role in oxidative stress and protein folding were highly expressed, potentially reflecting the challenge in maintaining cellular integrity in a harsh environment [24••].

The classic approach to assess gene function is to identify which genes are required for fitness in a given condition through gene disruption. One attractive mutagenesis technology that can be employed directly in the environment is the 'tagging' of individual mutants in an approach analogous to bacterial signature tagged mutagenesis [25] and the parallel phenotypic analysis of the yeast deletion collection [26]. In these strategies, each tag is a unique DNA sequence that serves to mark a single mutant strain. The presence of common PCR priming sites surrounding the unique tags enables the amplification of all tags in a complex pool of mutants in a single reaction. The relative abundance of each mutant can then be assessed by hybridization of the tags to a microarray containing the tag complements. In this manner, all pooled mutants that did not survive an experimental selection can be identified in parallel. For environmental studies, the main advantage of tagged mutagenesis is that the tag signals can potentially be PCR-amplified from the environmental 'noise' and quantified using a microarray without the need for culturing the pooled mutants after addition to the environment. Such experiments would identify genes required for survival in a natural environment. Groh *et al.* [27] applied the signature tagged mutagenesis approach in the metal-reducing bacteria *Shewanella oneidensis* and *Desulfovibrio desulfuricans*. Pools of 60 tagged mutants were analysed for survival in an artificial anaerobic sediment environment using a custom microarray. Simulation of the natural environment will identify genes required for fitness under more natural conditions when studies cannot be performed in the field. As more environmental microbes are cultured [11], often enabled by the blueprint of the genome sequence [28], tagged mutagenesis will become increasingly applicable.

What is novel and useful for biotechnology?

The biotechnology applications derived from microbial diversity range from the isolation of genes encoding novel functionality for industrial or biomedical applications [29] to the cleanup of environmental pollutants using engineered microbes [30]. Both companies and academic groups have constructed libraries of environmental DNA from diverse sources such as soil [31] and seawater [32] to identify genes with particular characteristics, such as those conferring antibiotic resistance [33] or encoding specific enzymatic activities [34], or more generally to gain a better understanding of the variety and range of a protein family of interest [32]. The challenges associated with this approach include potential problems with expressing heterologous DNA in a surrogate host (typically *Escherichia coli*), insufficient homology to identify clones using PCR, and the laborious task of screening through thousands of clones for rare 'hits'.

Methods to rapidly screen or select clones of interest from the thousands in a standard metagenome library are required to bring this technology to the average research laboratory lacking high-throughput infrastructure. One promising development is the substrate-induced gene expression screening (SIGEX) technique [35•]. Taking advantage of the observation that most catabolic genes are induced by their substrates, a groundwater metagenome library was cloned in an operon-trap vector driving expression of the gene encoding green fluorescent protein (*gfp*). Upon induction by a hydrocarbon substrate, *gfp*-expressing clones (presumably containing catabolic genes involved in the degradation of the hydrocarbon) were identified and separated from non-induced clones using fluorescence-activated cell sorting.

The use of individual microbes for complex environmental tasks such as bioremediation of contaminated and polluted sites represents a great challenge for environmental biotechnology on several levels. Foremost, there is substantial discord between the laboratory conditions where the organism is manipulated and the *in situ* environment that is targeted by the microbe. Consequently, it is not surprising that genetically modified bacteria rarely function in a natural environment [30]. How do we cope with the laboratory-environment discrepancy and how do we design laboratory experiments that adequately represent natural conditions? One solution is to take a global, systems biology approach by examining the numerous stress responses, regulatory systems, and genes critical for the desired biological activity such as bioremediation. The key to this approach will be the integration of gene expression, proteomics, physiological, mutant phenotype, and metabolic data into working cellular models that can accurately predict the response of the organism to a given

environment [36]. Meeting these goals of microbial systems biology will additionally require the development of computational resources and infrastructure [37,38] that link services such as data storage and integration into coherent, testable models. Finally, the functionality of an environmentally introduced, engineered microbe(s) will be aided by the culture-independent technologies described previously to determine the impact of the endogenous microbial population, track the activity and progress of the engineered microbe over time, and to understand the ecological impact of the human intrusion.

Perspectives and future directions

Genomic-based analyses in environmental microbiology are in their infancy. Meeting the challenges associated with applying experimental techniques in the environment, analysing complex data, and meshing biogeochemical cycles with the relevant microorganism will go a long way towards realizing the biotechnological potential of natural microbial diversity. Currently, because of the complexity and cost of technologies necessary for environmental genomics, these projects are often accomplished through the formation and collaborative effort of large research teams (e.g. the US Department of Energy Genomics:GTL programme [39]) such that cost and expertise are distributed. However, we anticipate that technological innovations will lead to a severe cost reduction in DNA sequencing and other related technologies and make large-scale metagenomics more accessible to the individual researcher. Similar to the development of standard laboratory genomics, the availability of the necessary tools to a larger number of researchers will spur future discoveries in environmental genomics.

As environmental microbiology data accumulates we can begin to move from 'What is there' and 'What is it doing' towards higher order questions regarding the generation and maintenance of genetic diversity and the impact of environmental change on microbial evolution. These questions will be aided by analyses into the prevalence and function of viruses, transposable DNA elements, plasmids, and horizontally transferred genes within and across communities. It is becoming increasing clear, both from comparative studies of whole microbes [40–44] and from studies of sequences obtained in environmental samples [45], that the horizontal transfer of genes plays a large role in the spread of functional abilities within communities and in enabling the adaptation of organisms to changing niches [46]. It appears that, in addition to the measures taken by bacteria and archaea to confer fitness upon their brethren, phage might also provide a means for the transfer of useful genetic elements between microbes [47,48], and perhaps even contribute to the evolution of novel functions. Therefore, metagenomic studies of viruses [49] are an important, and perhaps essential, complement to genomic studies of microbial communities. We look forward to the day when our grasp of evolutionary mechanisms allows us to fully understand the origins of microbial diversity in the complex and dynamic communities that serve as the foundation for life on earth.

Update

Johnson *et al.* [50] have reported the distribution of six *Prochlorococcus* ecotypes across the Atlantic Ocean. The abundance of individual ecotypes was significantly correlated with temperature, thus providing a link between an environmental gradient and population structure.

Acknowledgements

We thank Terry Hazen and David Stahl for comments on the manuscript. Research in the laboratory of APA is supported by the Howard Hughes Medical Institute. This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Programme:GTL through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the US Department of Energy.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Whitman WB, Coleman DC, Wiebe WJ: **Prokaryotes: the unseen majority**. *Proc Natl Acad Sci USA* 1998, **95**:6578-6583.

2. Handelsman J: **Metagenomics: application of genomics to uncultured microorganisms**. *Microbiol Mol Biol Rev* 2004, **68**:669-685.

3. Kaeberlein T, Lewis K, Epstein SS: **Isolating 'uncultivable' microorganisms in pure culture in a simulated natural environment**. *Science* 2002, **296**:1127-1129.

4. Santos SR, Ochman H: **Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins**. *Environ Microbiol* 2004, **6**:754-759.

5. Stahl DA, Lane DJ, Olsen GJ, Pace NR: **Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences**. *Appl Environ Microbiol* 1985, **49**:1379-1384.

6. • Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC *et al.*: **Comparative metagenomics of microbial communities**. *Science* 2005, **308**:554-557.
Comparative study of metagenomics sequence from different environments reveals different diversity profiles and gene complements.

7. Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ: **SAR11 clade dominates ocean surface bacterioplankton communities**. *Nature* 2002, **420**:806-810.

8. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M *et al.*: **Genome streamlining in a cosmopolitan oceanic bacterium**. *Science* 2005, **309**:1242-1245.

9. Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V, Duprat S, Galperin MY, Koonin EV, Le Gall F *et al.*: **Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome**. *Proc Natl Acad Sci USA* 2003, **100**:10020-10025.

10. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR *et al.*: **Genome**

divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* 2003, **424**:1042-1047.

11. Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA: **Isolation of an autotrophic ammonia-oxidizing marine archaeon**. *Nature* 2005, **437**:543-546.

12. DeLong EF, Wickham GS, Pace NR: **Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells**. *Science* 1989, **243**:1360-1363.

13. Desantis TZ, Stone CE, Murray SR, Moberg JP, Andersen GL: **Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray**. *FEMS Microbiol Lett* 2005, **245**:271-278.

14. El Fantroussi S, Urakawa H, Bernhard AE, Kelly JJ, Noble PA, Smidt H, Yershov GM, Stahl DA: **Direct profiling of environmental microbial populations by thermal dissociation analysis of native rRNAs hybridized to oligonucleotide microarrays**. *Appl Environ Microbiol* 2003, **69**:2377-2382.

15. Palmer C, Bik EM, Eisen MB, Eckburg PB, Sana TR, Wolber PK, Relman DA, Brown PO: **Rapid quantitative profiling of complex microbial populations**. *Nucleic Acids Res* 2006, **34**:e5.

16. Peplies J, Lau SC, Pernthaler J, Amann R, Glockner FO: **Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton**. *Environ Microbiol* 2004, **6**:638-645.

17. Sebat JL, Colwell FS, Crawford RL: **Metagenomic profiling: microarray analysis of an environmental genomic library**. *Appl Environ Microbiol* 2003, **69**:4927-4934.

18. • Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al.*: **Environmental genome shotgun sequencing of the Sargasso Sea**. *Science* 2004, **304**:66-74.
The shotgun sequencing of a Sargasso Sea microbial community introduces many of the technical and bioinformatics challenges of analysing metagenome sequence data.

19. • Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment**. *Nature* 2004, **428**:37-43.
Reports the elucidation of the structure of a simple community using metagenomic sequencing.

20. Hallam SJ, Putnam N, Preston CM, Detter JC, Rokhsar D, Richardson PM, DeLong EF: **Reverse methanogenesis: testing the hypothesis with environmental genomics**. *Science* 2004, **305**:1457-1462.

21. • DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR *et al.*: **Community genomics among stratified microbial assemblages in the ocean's interior**. *Science* 2006, **311**:496-503.
Comparative population and functional studies from samples taken from different depths in the water column suggest depth-variable community trends in carbon and energy metabolism, attachment and motility, gene mobility, and host–viral interactions.

22. Dennis P, Edwards EA, Liss SN, Fulthorpe R: **Monitoring gene expression in mixed microbial communities by using DNA microarrays**. *Appl Environ Microbiol* 2003, **69**:769-778.

23. Rhee SK, Liu X, Wu L, Chong SC, Wan X, Zhou J: **Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays**. *Appl Environ Microbiol* 2004, **70**:4303-4317.

24. •• Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake RC II, Shah M, Hettich RL, Banfield JF: **Community proteomics of a natural microbial biofilm**. *Science* 2005, **308**:1915-1920.
First proteomic study of a natural microbial community, a biofilm growing in an extreme acid mine drainage. Mass-spectrometry-based techniques were used to validate the expression of ~2000 proteins in the community and to provide a link between microbial cellular activity and the geochemistry of the environment. In addition, the expression and cellular localization of hundreds of novel proteins were verified.

25. Hensel M, Shea JE, Gleeson C, Jones MD, Dalton E, Holden DW: **Simultaneous identification of bacterial virulence genes by negative selection**. *Science* 1995, **269**:400-403.

26. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B *et al.*: **Functional profiling of the *Saccharomyces cerevisiae* genome**. *Nature* 2002, **418**:387-391.

27. Groh JL, Luo Q, Ballard JD, Krumholz LR: **A method adapting microarray technology for signature-tagged mutagenesis of *Desulfovibrio desulfuricans* G20 and *Shewanella oneidensis* MR-1 in anaerobic sediment survival experiments**. *Appl Environ Microbiol* 2005, **71**:7064-7074.

28. Tyson GW, Lo I, Baker BJ, Allen EE, Hugenholtz P, Banfield JF: **Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community**. *Appl Environ Microbiol* 2005, **71**:6319-6324.

29. Lorenz P, Eck J: **Metagenomics and industrial applications**. *Nat Rev Microbiol* 2005, **3**:510-516.

30. Cases I, de Lorenzo V: **Genetically modified organisms for the environment: stories of success and failure and what we have learned from them**. *Int Microbiol* 2005, **8**:213-222.

31. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C *et al.*: **Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms**. *Appl Environ Microbiol* 2000, **66**:2541-2547.

32. Sabehi G, Massana R, Bielawski JP, Rosenberg M, Delong EF, Beja O: **Novel Proteorhodopsin variants from the Mediterranean and Red Seas**. *Environ Microbiol* 2003, **5**:842-849.

33. Riesenfeld CS, Goodman RM, Handelsman J: **Uncultured soil bacteria are a reservoir of new antibiotic resistance genes**. *Environ Microbiol* 2004, **6**:981-989.

34. Robertson DE, Chaplin JA, DeSantis G, Podar M, Madden M, Chi E, Richardson T, Milan A, Miller M, Weiner DP *et al.*: **Exploring nitrilase sequence space for enantioselective catalysis**. *Appl Environ Microbiol* 2004, **70**:2429-2436.

35. • Uchiyama T, Abe T, Ikemura T, Watanabe K: **Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes**. *Nat Biotechnol* 2005, **23**:88-93.
This paper describes a novel approach to rapidly screen metagenome libraries for clones containing catabolic genes.

36. Fredrickson JK, Romine MF: **Genome-assisted analysis of dissimilatory metal-reducing bacteria**. *Curr Opin Biotechnol* 2005, **16**:269-274.

37. Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL, Arkin AP: **The MicrobesOnline Web site for comparative genomics**. *Genome Res* 2005, **15**:1015-1022.

38. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I *et al.*: **The integrated microbial genomes (IMG) system**. *Nucleic Acids Res* 2006, **34**:D344-D348.

39. Genomics DOE. **GTL Roadmap: systems biology for energy and environment**. US Department of Energy; 2005. http://doegenomestolife.org/roadmap/index.shtml.

40. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA: **The net of life: reconstructing the microbial phylogenetic network**. *Genome Res* 2005, **15**:954-959.

41. Lerat E, Daubin V, Ochman H, Moran NA: **Evolutionary origins of genomic repertoires in bacteria**. *PLoS Biol* 2005, **3**:e130.

42. Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF: **Do orthologous gene phylogenies really support tree-thinking?** *BMC Evol Biol* 2005, **5**:33.

43. Perna NT, Plunkett G III, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA *et al.*: **Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7**. *Nature* 2001, **409**:529-533.

44. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS *et al.*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'**. *Proc Natl Acad Sci USA* 2005, **102**:13950-13955.

45. Nesbo CL, Boucher Y, Dlutek M, Doolittle WF: **Lateral gene transfer and phylogenetic assignment of environmental fosmid clones**. *Environ Microbiol* 2005, **7**:2011-2026.

46. Pal C, Papp B, Lercher MJ: **Adaptive evolution of bacterial metabolic networks by horizontal gene transfer**. *Nat Genet* 2005, **37**:1372-1375.

47. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW: **Photosynthesis genes in marine viruses yield proteins during host infection**. *Nature* 2005, **438**:86-89.

48. Daubin V, Ochman H: **Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli***. *Genome Res* 2004, **14**:1036-1042.

49. Edwards RA, Rohwer F: **Viral metagenomics**. *Nat Rev Microbiol* 2005, **3**:504-510.

50. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW: **Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients**. *Science* 2006, **311**:1737-1740.