Annu. Rev. Genet. 2004. 38:525–52
doi: 10.1146/annurev.genet.38.072902.091216

First published online as a Review in Advance on July 14, 2004

METAGENOMICS: Genomic Analysis of Microbial Communities

Christian S. Riesenfeld,[1,2] Patrick D. Schloss,[1] and Jo Handelsman[1,2]

Department of Plant Pathology,[1] Microbiology Doctoral Training Program,[2] University of Wisconsin-Madison, Madison, Wisconsin 53706; email: joh@plantpath.wisc.edu

Key Words microbial ecology, environmental genomics, community genomics, culture-independent, and unculturable bacteria

■ **Abstract** Uncultured microorganisms comprise the majority of the planet's biological diversity. Microorganisms represent two of the three domains of life and contain vast diversity that is the product of an estimated 3.8 billion years of evolution. In many environments, as many as 99% of the microorganisms cannot be cultured by standard techniques, and the uncultured fraction includes diverse organisms that are only distantly related to the cultured ones. Therefore, culture-independent methods are essential to understand the genetic diversity, population structure, and ecological roles of the majority of microorganisms. Metagenomics, or the culture-independent genomic analysis of an assemblage of microorganisms, has potential to answer fundamental questions in microbial ecology. This review describes progress toward understanding the biology of uncultured Bacteria, Archaea, and viruses through metagenomic analyses.

CONTENTS

INTRODUCTION

Obtaining bacteria in pure culture is typically the first step in investigating bacterial processes. However, standard culturing techniques account for 1% or less of the bacterial diversity in most environmental samples (2). Although some significant breakthroughs have resulted from recent attempts to culture the as-yet-unculturable bacteria (56, 89, 99, 127), a suite of culture-independent techniques are needed to complement efforts to culture the thousands or millions of unknown species in the environment.

A new era of microbial ecology was initiated when sequencing of ribosomal RNAs and the genes encoding them was introduced to describe uncultured bacteria in the environment. The first approach was to sequence clones from a 5S rRNA cDNA library derived from the symbiotic community within the tubeworm *Riftia pachyptila* (109). Variations of this method generated a set of culture-independent techniques to (*a*) reconstruct phylogenies, (*b*) compare microbial distributions among samples using either nucleotide sequence or restriction fragment length polymorphisms (RFLPs), and (*c*) quantify the relative abundance of each taxonomic group using membrane hybridization or fluorescent in situ hybridization (2, 47, 57, 78–80).

The most startling result of the many microbial diversity studies that have employed 16S rRNA culture-independent methods is the richness of the uncultured microbial world. As of April 1, 2004, GenBank contained 21,466 16S rRNA genes from cultured prokaryotes and 54,655 from uncultured prokaryotes, according to the search terms described by Rappé & Giovannoni (90), and many of those from uncultured organisms affiliate with phyla that contain no cultured members. When Woese (121) originally proposed a 16S rRNA-based phylogeny, 12 bacterial phyla were recognized, each with cultured representatives. Since then, 14 additional phyla with cultured representatives have been identified. In addition, 16S rRNA gene sequence analysis suggests 26 candidate phyla that have no known cultured representatives (90). Therefore, half of the known microbial phyla have no cultured representatives.

Among the phyla that contain cultured members, a few contain many isolates and the rest contain too few to represent the full spectrum of diversity in the phylum. For example, Hugenholtz (53) found that 97% of prokaryotes deposited in the Australian Culture of Microorganisms in 2001 were members of just four phyla: the Proteobacteria (54%), Actinobacteria (23%), Firmicutes (14%), and

Bacteroidetes (6%). Within GenBank, 76% of the 16S rRNA gene sequences of cultured prokaryotes are from these four groups. But other phyla may be more diverse, prevalent, and ecologically consequential in the environment. 16S rRNA gene sequences from the Acidobacterium phylum are among the most abundant in clone libraries obtained from soil and have been found in all soils examined, suggesting that the Acidobacteria play important roles in soil ecosystems. However, of the 684 Acidobacterium 16S rRNA gene sequences in GenBank, only 19 (2.8%) are from cultured isolates, providing an inadequate collection to describe the physiological diversity of the phylum. Other than 16S rRNA gene sequences, little is known about the bacteria within the 22 poorly cultured phyla and 26 candidate phyla. Many terms, such as unculturable, uncultivated, as yet uncultured, and not yet cultured, are used to refer to microorganisms that we know of only through culture-independent means. In this review, we refer to them as uncultured.

Describing the phylogenetic diversity of uncultured microorganisms is only the first step. A greater challenge is to assign ecological roles to them. The uncultured microbiota must play pivotal roles in natural environmental processes and are a large untapped resource for biotechnology applications. Exploiting the rich microbial biodiversity for enzyme and natural product discovery is an active research area that has been reviewed elsewhere (39, 45, 46, 65, 66, 77, 97, 104). This review discusses the application of culture-independent genomics-based approaches to understand the genetic diversity, population structure, and ecology of complex microbial assemblages (26, 93, 94).

METAGENOMICS DEFINED

"Metagenomics" describes the functional and sequence-based analysis of the collective microbial genomes contained in an environmental sample (Figure 1) (45). Other terms have been used to describe the same method, including environmental DNA libraries (110), zoolibraries (55), soil DNA libraries (68), eDNA libraries (13), recombinant environmental libraries (22), whole genome treasures (77), community genome (114), whole genome shotgun sequencing (115), and probably others. In this review, we use metagenomics to describe work that has been presented with all of these names because it is the most commonly used term (15, 27, 35, 59–61, 65, 66, 82, 105, 107, 117, 118), was used for the title of the first international conference on the topic ("Metagenomics 2003" held in Darmstadt, Germany), and is the focus of an upcoming issue of the journal *Environmental Microbiology*. The definition applied here excludes studies that use PCR to amplify gene cassettes (52) or random PCR primers to access genes of interest (17, 32), since these methods do not provide genomic information beyond the genes that are amplified. Many environments have been the focus of metagenomics, including soil, the oral cavity, feces, and aquatic habitats, as well as the hospital metagenome, a term intended to encompass the genetic potential of organisms in hospitals that conribute to public health concerns such as antibiotic resistance and nosocomial infections (20).

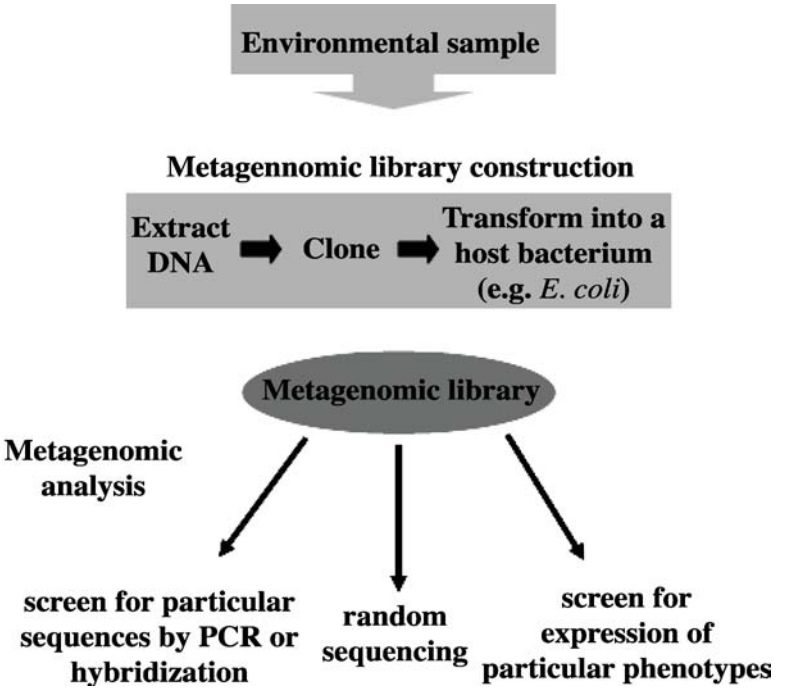

Figure 1 Metagenomics involves constructing a DNA library from an environment's microbial population and then analyzing the functions and sequences in the library.

The concept of cloning DNA directly from an environment was initially suggested by Pace (79) and first implemented by Schmidt et al. (106), who constructed a λ phage library from a seawater sample and screened it for 16S rRNA genes. Advances by the DeLong group in cloning DNA directly from seawater provided the landmark work that launched the field (110). Development of metagenomic analyses of soil was slower than with seawater because of the technical challenges of cloning DNA from the complex matrix of soil, which contains many compounds that bind to DNA or inhibit the enzymatic reactions required for cloning. Significant progress has been made, producing libraries that have substantially advanced understanding the functions in the soil community (96). The past eight years have witnessed an explosion of interest and activity in metagenomics, accompanied by advances in technology that have facilitated studies at a scale that was not feasible when the field began. For example, the seminal paper in 1996 by Stein et al. (110) reported the sequencing and reconstruction of a 40-kb fragment from an uncultured marine archaeon, which was a major undertaking at the time. In 2004, Venter et al. (115) reported their attempt to sequence the entire metagenome of the Sargasso Sea by obtaining over 1 million kb of nonredundant sequence. The advances in sequencing technology have expanded the approaches and questions that can be

considered with metagenomics, providing access to a staggering amount of genomic information. Metagenomic technology has been successful at all scales—it has been used to study single genes (e.g., cellulases, 48), pathways (e.g., antibiotic synthesis, 96), organisms (e.g., Archaea, 110), and communities (e.g., acid mine drainage biofilm, 114). Approaches that involve massive sequencing to capture entire communities will likely become more common with further advances in sequencing technology.

LINKING PHYLOGENY AND FUNCTION WITHIN SPECIES

Phylogenetic Anchors

The first metagenomic studies aimed to link a function with its phylogenetic source, providing information about one species within a community. One of the challenges with this approach is to link a phenotype with the identity of the original host. Three approaches have been taken: Screen a metagenomic library for a phenotype and then attempt to determine the phylogenetic origin of the cloned DNA (Table 1), screen clones for a specific phylogenetic anchor (e.g., 16S rRNA) or gene and then sequence the entire clone and search for genes of interest among the genes flanking the anchor (Table 2), or sequence the entire metagenome and identify interesting genes and phylogenetic anchors in the resulting reconstructed genomes (Table 3).

Function Then Phylogeny

Diverse activities have been discovered by functional analysis of metagenomic libraries. New antibiotics (11–14, 36, 68, 96, 119, 120), hydrolytic and degradative enzymes (21, 48–50, 59, 60, 91, 96, 117), biosynthetic functions (31, 61), antibiotic resistance enzymes (22, 92), and membrane proteins (69) have been identified. The diversity of functionally active clones discovered in metagenomic libraries validates the use of functional screens as one means to characterize the libraries. Antimicrobial screens have revealed new antibiotics such as terragine (119), turbomycin A and B (36), and acyl tyrosines (13), as well as previously described antibiotics such as indirubin (68) and violacein (12). Most of these compounds are structurally based on common cell substituents, such as amino acids, and none requires more than a few genes for its synthesis. The goal of identifying new polyketide, macrolide, and peptide antibiotics (45) may require different methods. Enhancing expression of genes in metagenomic libraries may lead to discovery of a wider array of natural products. This will be accomplished by moving the libraries into alternative hosts, such as *Streptomyces*, which was the basis for discovery of terragine (119). Alternative hosts may enhance gene expression or provide starting materials that *Escherichia coli* does not contain. *E. coli* can be engineered to express a wider range of functions by introducing genes encoding new sigma factors, rare tRNAs, or functions required to synthesize starting materials

TABLE 1 Metagenomics discovery based on functional screens

Environment	Number of clones	Insert size (kb)	Total DNA (Gb)	Activity of interest	Reference
Soil	n.s.	Cosmid	—	Fatty acid enol esters	(11)
Soil	n.s.	Cosmid	—	Pigments	(12)
Soil	700,000	Cosmid	~26	Antimicrobials	(13)
Soil	n.s.	Cosmid	—	Fatty acid enol esters	(14)
Marine	825,000	Plasmid	~4.0	Chitinases	(21)
Feces and soil	4 × 6000–35,000	30–40	~3	Biotin biosynthesis	(31)
Anaerobic digester	15,000	1–12	0.10	Cellulases 4-hydroxybutyrate utilization	(48) (49)
Soil	3 × ~300,000	5–8 kb	5.9	Lipases Antiporter activity	(50) (69)
Soil & river sediment	1 × 80,000 2 × 240,000	3–5 kb	2.2	Dehydratase	(59)
Soil & river sediment	4 × 100,000	3–6 kb	1.8	Alcohol oxidoreductase	(60)
Soil	3 × 320,000–580,000	3–5 kb	6.2	Carbonyl formation	(61)
Soil	1.5×10^7	37 kb	560	Antimicrobial	(68)
Human mouth	450	Plasmid	~0.001	Antibiotic resistance	(27)
Various	n.s.	n.s.	n.s.	Amylase	(91)
Soil	4 × 58,000–650,000	3–4 kb	4.2	Antibiotic resistance Antimicrobials and novel enzymes	(92) (96)
Soil	3648 24,576	27 44.5	1.2	Antimicrobials	(36)
Soil	n.s.	30 kb	—	Novel biocatalysts	(117)
Soil	n.s.	n.s.	—	Antimicrobials	(119)
Geothermal sediment	37,000	5	0.2	Pigments	(120)

TABLE 2 Metagenomics discovery of homologues of targeted genes or gene families

Environment	Number of clones	Insert size (kb)	Total DNA (Gb)	Genes of interest	Reference
Marine	6240	80		16S rRNA Archaea 16S rRNA Photosystem II (*psbA*)	(5) (9) (126)
Marine	7200	40	2.9	Archaea 16S rRNA	(6)
Marine	32,000	10–20 kb	0.48	16S rRNA	(106)
Polychaete symbionts	n.s.	Fosmid	—	16S rRNA	(18)
Marine	5000	50	0.25	Polyketide synthase	(22)
Marine	n.s.	BAC	—	Proteorhodopsin	(25)
Sediment	n.s.	Fosmid	—	Archaea 16S rRNA methyl coenzyme M reductase A	(43)
Tubeworm symbiont	1500	Fosmid	~0.06	Histidine protein kinase	(55)
Soil	3600 24,600	27 45	1.2	16S rRNA	(63, 96)
Marine	6107	35–40		16S rRNA	(64)
Beetle symbionts	n.s.	Fosmid	—	Polyketide synthase	(81)
Beetle and sponge symbionts	n.s.	Cosmid	—	Polyketide synthase	(82)
Soil	25,278	35–40	0.95	16S rRNA	(84)
Soil	56,000	33–44	2.2	Acidobacteria 16S rRNA	(85)
Sponge symbionts	n.s.	Fosmid	—	*RadA*	(100)
Sponge symbionts	n.s.	40	—	Archaea 16S rRNA	(101)
Sponge symbionts	n.s.	40	—	Archaea 16S rRNA	(103)
River biofilm	n.s.	40	—	Hybridization analysis	(107)
Marine	3552	40	0.14	Archaea 16S rRNA Planctomycetes 16S rRNA	(110) (116)

TABLE 3 Metagenomics studies based on random sequencing

Environment	Insert size (kb)	Sequence reads	Base pairs of sequence	Goal	Reference
Feces	LASL[a]	532	~37,000	Random viral clone sequencing	(15)
Marine	LASL[a]	1061	~740,000	Random viral clone sequencing	(16)
Drinking water network biofilms	Plasmid and Cosmid	2496	2×10^6	Random clone sequencing	(105)
Acid mine drainage	3.2	103,462	76.2×10^6	Reconstruct genome of microbial community	(114)
Marine	2.6	2×10^6	1.63×10^9	"Pilot study" of large-scale whole community sequencing	(115)

[a]LASL: Linker amplified shotgun library

for antibiotic biosynthesis that are deficient in *E. coli*. Alternatively, sequences that carry conserved regions of genes associated with antibiotic biosynthesis, such as the polyketide synthases and peptide synthetases, may be identified by sequenced-based screens that do not require heterologous gene expression. This approach successfully identified clones carrying a novel hybrid polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of a beetle (81).

Novel enzymes have been revealed in metagenomic libraries by screening clones directly for activity (49, 50, 96). Pigments have been identified by visual inspection (12, 36, 68). These methods require handling individual clones, usually in an array format. Because the frequency of active clones is low, high-throughput methods are essential for efficient screening. Selection for the ability to grow on hydroxyl-butyrate as the sole carbon and nitrogen source provided a powerful selection for clones carrying new degradative enzymes (49), and selection for antibiotic resistance identified new antibiotic resistance determinants from soil (22, 92) and from oral flora (27).

Linking new functions with the organisms from which they were cloned will facilitate ecological inferences and may lead to culturing strategies for uncultured species. Several approaches have been used to identify the phylogenetic origin of functionally active clones. First, to determine which phylogenetic groups are represented in a metagenomic library, 16S rRNA gene libraries have been constructed using DNA from the metagenomic library as the template for PCR (43, 63, 68, 96, 117). Comparing the phylogenetic distribution of 16S rRNA sequences in the environmental sample and the metagenomic library can reveal biases in library construction (63). The nucleotide sequence of the genes that flank the

region of functional interest can provide the basis for inferences about phylogeny that are supported by similarity of the flanking genes to genes of known function in GenBank (31, 48, 59–61, 81). Conclusions from these analyses must be treated cautiously because horizontal gene transfer and the lack of functional homologues in the database may confound the results, leading to matches in flanking DNA originating in two different phyla (48). However, the presence of many cloned genes on a clone that all show similarity to genes from related organisms can bolster phylogenetic claims. Analyses of G + C content and codon usage within the region of interest and the flanking gene sequences may suggest the phylogenetic origin of the cloned DNA, although this approach has not been widely successful to date (19).

Phylogeny Then Function

Concerted effort devoted to finding clones carrying phylogenetic anchors from the least known taxons has produced impressive collections of clones derived from some of these groups. Sequence analysis of the DNA that flanks phylogenetically informative genes has provided the first glimpses into the genetic potential of taxa that contain no cultured members (Table 2). rRNA genes are the most widely used anchors (6, 9, 43, 63, 64, 84, 85, 102, 110, 116), and *radA/recA* homologues (100) have been informative as well. Other genes that contain phylogenetic information but have not been used in metagenomic analyses include DNA gyrase (125), chaperonin-60 (38, 51, 108), RNA polymerase β-subunit (24, 73), ATPase β-subunit (67), elongation factor TU (67), heat shock protein 70 (41), σ^{70}-type sigma factor (42), and tRNA synthetases (122, 123). The phylogenetic anchor approach has been a rich source of information that can be used to develop hypotheses about the function and physiology of uncultured members of microbial communities that were previously known only by their 16S rRNA gene signature. Use of more phylogenetic anchors will increase the frequency of functionally active clones that also contain anchors.

A consistent result of most studies that initially screen for a phylogenetic anchor is the identification of a large number of ORFs that are either hypothetical or have no known function. Initially, this information does little to describe the niche that the organism fills in the environment, but the sequence provides G + C content, codon usage, promoter sites, and other characteristics that may be helpful in achieving expression of genes from phyla with no cultured representatives. These sequences also enrich the databases, offering insight subsequently when the function is determined for a homologue of the same gene family.

Acidobacterium Phylogeny and Function

Genomic information about uncultured bacteria may hasten development of media for their culture (5, 18, 63, 85, 116). This will be a needed boost to the laborious work that has led to significant recent advances in culturing technology (56, 89, 99, 127). The Acidobacterium phylum has been largely recalcitrant to culturing

and has attracted attention because of its abundance and wide distribution on Earth (63, 85). This phylum has been divided into eight groups (4, 54). Janssen et al. (56, 99) have made tremendous advances in culturing members of the phylum, although all cultured members are from three of the eight groups.

Metagenomics has provided the first information beyond 16S rRNA gene sequences about the uncultured Acidobacterium subgroups based on partial and full-length sequencing of six fosmid clones (~35 kb each; 85) and 12 BAC clones (25 kb; 63) from uncultured soil Acidobacterium members. Two of the clones are from members of Acidobacterium Group III, five are from Group V, and 11 are from Group VI. These clones provide a preliminary indication that the Acidobacterium phylum contains substantial genetic diversity. Among the Group V clones, for example, the G + C content of one is 56% and the other four range from 62% to 68%, while the one cultured member of this group, *A. capsulatum*, has a G + C content of 60%. Some of the metagenomic clones that affiliate with the Acidobacterium phylum have been fully sequenced, revealing many genes with homology to housekeeping genes involved in DNA repair, transport, cell division, translation, and purine biosynthesis. Other gene sequences include those with homology to genes encoding cyclic β 1′-2′ glucan synthetase, polyhydroxybutyrate depolymerase, *Bacteroides fragilis* aerotolerance functions, and an operon distantly related to the lincomycin biosynthesis pathway of *Streptomyces lincolnensis* (85), which provide hints about ecological roles of the Acidobacteria.

Ten of the 22 ORFs on one Acidobacterium fosmid clone showed homology to genes from members of the *Rhizobiales* within the α-Proteobacteria phylum (85). Comparison of the fosmid sequence to the gene sequences from *Rhodopseudomonas palustris* and *Bradyrhizobium japonicum*, Quaiser et al. (85) revealed a colinear 10-kb region containing eight ORFs, which had homology to genes encoding a penicillin-binding protein, zinc metalloprotease, hydroxybutyrate depolymerase, and a highly conserved two-component histidine kinase response regulator. A phylogenetic analysis of the response regulator indicated that one gene affiliated within the *Rhizobiales* and the second gene affiliated with other groups in the α-Proteobacteria. This 10-kb region was contiguous to an Acidobacterium-type *rrn* operon, suggesting that this region might be the product of horizontal gene transfer (85).

Archaeal Phylogeny and Function

Just as culture-independent methods recalibrated thinking about the dominant organisms in soil, indicating that the Acidobacteria were far more abundant than had been established by culturing, the discovery of 16S rRNA gene sequences that affiliate with the Archaea in diverse terrestrial and marine environments on Earth has significantly altered the microbiologist's image of Archaea. Like the Acidobacteria, Archaea in the Crenarchaeota phylum have been refractory to culturing, making it challenging to elucidate their role in the environment. There has been significant interest in applying metagenomics to learn more about the

members of the Archaea in soil (84) and as planktonic organisms in seawater (6, 9, 43, 64, 110). Six clones produced in fosmids, cosmids, or BACs, which contain 16S rRNA genes that affiliate with the Archaea, have been sequenced.

A particularly fruitful application of metagenomics to Archaea has been the study of the symbiotic community of an *Axinella* sp. sea sponge (100, 101, 103). In a culture-independent 16S rRNA gene survey, Preston et al. (83) found that 65% of the symbiotic community associated with the sponge was represented by a single archaeal 16S rRNA gene sequence. They proposed the name *Cenarchaeum symbiosum* for the uncultured archaeal symbiont. Schleper et al. (101, 103) then constructed fosmid libraries of the prokaryotic community and identified 15 unique clones that harbored 16S rRNA genes.

Analysis of the 15 clones revealed genetic variation within the *C. symbiosum* population and provided insight into the role of the population within the sponge (101). After identifying two 16S rRNA sequence variants that differed by two point mutations over a 590-bp region, they sequenced one fosmid clone from each variant. The 16S genes of the two clones were 99.2% identical while the 28-kb colinear region that they shared had 87.8% overall DNA identity, and 91.6% similarity in ORF amino acid sequence. Of the 17 ORFs in the 28-kb region shared by the two variants, eight had no known function, and the others had functions related to heme and menaquinone biosynthesis, glycolysis, DNA replication and repair, protein folding, and DNA methylation. The DNA polymerase found in this region was expressed and characterized in an *E. coli* host (103). The complete genome sequence of *C. symbiosum* will contribute to understanding its biology and symbiotic relationship with its sponge host. There is precedent for this in the symbiosis between the aphid *Baizongia pistacea* and the bacterium *Buchnera aphidicola*. *B. aphidicola* is found in pure culture as an endosymbiont of its host. The complete genome sequence for the bacterium revealed a complex biochemical symbiosis between the partners in which each partner had lost biochemical functions that the other conducted for both of them (112). The *B. aphidicola* system illustrates the power of genomics to elucidate the biology of uncultured microorganisms.

Proteorhodopsin Function and Phylogeny

Discovery of the rhodopsin-like photoreceptors in marine Bacteria exemplifies the type of biological surprise that can be revealed through metagenomic analysis. Previously, rhodopsins had been found only in Archaea, not in members of the domain Bacteria. Béjà et al. (5) sequenced a 130-kb fragment that contained the 16S rRNA operon of an uncultured γ-Proteobacterium (the SAR 86 group) and discovered a bacteriorhodopsin, which indicated a novel taxon of marine phototroph. The bacteriorhodopsins couple light-energy harvesting with carbon-cycling in the ocean through nonchlorophyll-based pathways, and the new homologue was expressed in *E. coli* and shown to bind retinal and form an active, light-driven, proton pump. Subsequent studies showed that many marine Proteobacteria harbor "proteorhodopsins," that are optimized for various light wavelengths at different

ocean depths (7, 8, 25, 70, 98). This line of research was successful at showing that bacteria that harbor proteorhodopsin variants are widespread, and recent work by Venter et al. (115) using a shotgun sequencing approach revealed that the class of proteorhodopsins previously observed is a small subset of the total proteorhodopsin diversity.

LINKING PHYLOGENY AND FUNCTION IN MICROBIAL COMMUNITIES: METAGENOME RECONSTRUCTION

As sequencing technology has improved, it has become feasible to sequence the entire metagenome of an environmental sample. Most environments contain communities far too complex for it to be possible to sequence a complete metagenome, and even the simple communities contain microheterogeneity that makes most genome reconstructions simplified versions of reality. However, it is useful to refer to a metagenome, just as it is useful to refer to the human genome, although it is widely recognized that the true human genome is far more complex and variable than the published genomic sequence, which is based on a few of the 6×10^9 members of the species. Reconstruction was initially pursued for viral communities in the ocean and human feces (15, 16) and has since been attempted in an acid mine drainage (AMD) biofilm (114) and the Sargasso Sea (115). The AMD biofilm community was ideal for complete metagenome sequencing because 16S rRNA gene sequencing indicated that there were three bacterial and three archaeal species in the biofilm. Marine communities contain far greater species richness, on the order of 100 to 200 species per ml of water (23), making the sequencing and assembly effort considerably more difficult. Further out on the continuum of biological complexity is soil, with an estimated species richness on the order of 4000 species per gram of soil (23, 113). Sequencing the soil metagenome requires faster and less expensive sequencing technology than currently available. Meanwhile, soil metagenomics continues to focus on targeted biological questions, thereby elucidating a slice of the community's function by genomic analysis.

Metagenomic Analysis of Bacteriophage

Microbial communities are dominated by bacteria, and bacterial populations are dominated by bacteriophage. Bacteriophages influence the diversity and population structure of microbial communities (124). Like their bacterial counterparts, most bacteriophages have never been studied in the laboratory because they represent staggering diversity (95), and many of their hosts have not been cultured. Complete viral genome sequences are also necessary for viral phylogeny studies since there is no equivalent to the 16S rRNA gene for virus phylogeny. Using metagenomic analyses, two recent studies examined phage diversity of the human gut and marine environments.

The microbial community within the human gut is complex, consisting of more than 400 species (15). To investigate the phage population in the human gut, Breitbart et al. (15) conducted a metagenomic analysis of the viral community in human feces. They constructed a library containing random fragments of viral DNA from a preparation of virus particles isolated from a 500-g human fecal sample, end-sequenced 532 clones, and found that 59% of them did not contain significant similarity to previously reported sequences. The viral community contains approximately 1200 genotypes, which probably outnumber the bacterial species in the human intestine. The viral community may affect community structure by infecting and lysing particular members of the bacterial community and enhancing its diversity by mediating genetic exchange between bacteria.

In another study, Breitbart et al. (16) described a metagenomic analysis of marine phage collected at two locations. In total, almost 2000 viral sequences were obtained. The results suggest that the phage populations differed between the two marine locations and between the marine and fecal samples. For instance, T7-like podophages comprised over 30% of the marine phage types (16) and less than 6% of the fecal phage types (15). The predominance of gram-positive bacteria in the gut and gram-negative bacteria in seawater is at least partly responsible for the substantial difference in viral communities.

Metagenome of the Microbial Community in Acid Mine Drainage

Acid mine drainage results from bacterial iron oxidation, which leads to acidification due to dissolution of pyrite in abandoned mines (29). The microbial biofilm growing in the AMD in the Richmond mine at Iron Mountain, California, has a pH of 0.83, temperature of 43°C, and high concentrations of Fe, Zn, Cu, and As (114). Sequences of the 5′ and 3′ ends of 384 16S rRNA genes obtained from the biofilm revealed members of *Leptospirillum* groups II and III, *Sulfobacillus* sp., *Ferroplasma* sp., "A-plasma," and "G-plasma." There are also protists containing *Rickettsiales*-type endosymbiotic bacteria (3).

METAGENOME RECONSTRUCTION *Ferroplasma acidarmanus* fer1, an archeon in the mine has been grown in pure culture, and its genome has been sequenced, although attempts to grow other members of the community in culture have been unsuccessful. To sequence the genomes of the uncultured Bacteria and Archaea in the biofilm, Tyson et al. (114) extracted DNA directly from the biofilm, constructed a small insert library (average insert size of 3.2 kb), and obtained 76.2 Mbp of sequence data from 103,462 reads. They partitioned the community into populations and identified five sequence "bins": high G + C scaffolds with 3x and 10x coverage, low G + C scaffolds with 3x and 10x coverage, and short scaffolds with poor coverage. Each of the four bins with greater than 3x coverage contained a single 16S rRNA gene fragment, enabling them to assign a phylogenetic identity to each bin with confidence. Based on this analysis, they obtained near-complete genome

sequences of *Leptospirillum* group II and *Ferroplasma* type II and partial genome sequences for *Leptospirillum* group III, *Ferroplasma* type I, and G-plasma.

BIOGEOCHEMISTRY With a metagenome sequence, Tyson et al. (114) set out to determine the ecological role of each of the five prokaryotes in the acid mine drainage. Genes that would enable most of the organisms to fix carbon via the reductive acetyl coenzyme A pathway were identified within each of the genome sequences. Based on the many genes in the *Ferroplasma* type I and II genomes that have significant similarity to sugar and amino acid transporters, they predicted that these *Ferroplasma* spp. preferred a heterotrophic lifestyle. The only N_2 fixation genes that were identified in the metagenome belonged to the genome of the *Leptospirillum* group III population. Because of its specialized role in the environment and relatively low abundance (10%), *Leptospirillum* group III was proposed as a keystone species. They tested and supported this hypothesis by isolating *Leptospirillum* group III in pure culture from a N_2-based enrichment (J.F. Banfield, personal communication). Other genes that are potentially responsible for microaerophilic survival, biofilm formation, acid tolerance, and metal resistance were observed.

Many of the hypotheses suggested in the AMD study will be evaluated by constructing microarrays and developing new techniques for isolating the uncultured prokaryotes (114). The simplicity of the community and the differences in G + C content of its members facilitated the powerful genomic reconstruction in the AMD biofilm. The genomics, coupled with keen insight into the chemistry of the environment, produced inferences and hypotheses that will lead to future studies to culture each of the community's members and unravel the complex interactions that produce this extreme environment.

Metagenome of the Microbial Community in the Sargasso Sea

Venter et al. (115) conducted a massive sequencing project focused on the microbiota of the nutrient-limited Sargasso Sea, an intensively studied marine environment. Two million random sequences yielded over 1.6 billion base pairs of sequence information, including approximately 1.0 billion base pairs of nonredundant sequence.

BIOGEOCHEMISTRY Although the interpretation of this huge dataset has barely begun, new insights into the biogeochemistry of marine ecosystems have emerged. First, contrary to the entrenched dogma that oceanic nitrification is mediated by bacteria, the authors identified an archaeal scaffold that contains an ammonium monooxygenase. Second, among the sequences derived from the phosphorus-limited Sargasso Sea are genes thought to be involved in uptake of phosphorus in various forms; these include polyphosphates, pyrophosphates, phosphonates, and other inorganic phosphorus. These results augment a relatively thin literature about the phosphorus cycle. Compared with the five other major elements (C, N, O, H, S), relatively little is known about how phosphorus enters biological systems and changes oxidation state. Therefore, examining a phosphorus-limited

environment is a likely place to search for new mechanisms of phosphorus acquisition. The Sargasso Sea study provides the basis for functional studies to determine how these phosphorus acquisition genes are deployed to promote survival in a phosphorus-deficient community.

GENOME ORGANIZATION AND RECONSTRUCTION Overlapping sequences were assembled into scaffolds and then sorted into tentative "organism bins" based on three criteria: oligonucleotide frequencies, read depth, and similarity to previously sequenced genomes. Due to the immense microbial diversity of the Sargasso Sea, relatively few complete microbial genomes were assembled, despite the large amount of sequence data accumulated. When "reconstructing" genomes from the environment, care must be taken not to generate chimeric genomes that do not exist in nature. Venter et al. (115) set a good standard for the field by making the trace files of individual sequences available so that others can check the validity of the assembled scaffolds or reanalyze the data as sequence assembly algorithms improve.

In comparison to culture-independent methods that are based on individual genes (usually 16S rRNA), large-scale sequencing of environmental genomes provides insights into microbial diversity at much higher resolution. Unlike traditional genome sequencing projects, which start with a homogeneous clonal population, environmental samples are likely to contain multiple strains of any given species (or phylotype). Heterogeneity makes the accurate assembly of discrete genomes difficult, but the sequences offer unprecedented opportunities to understand evolutionary events within natural microbial populations.

IMPACT ON PUBLIC DATABASES The vast amount of data from the Sargasso Sea study contribute to metagenomics and microbial ecology, providing the largest genomic dataset for any community on Earth. But the data have also skewed genomic analysis: As of April 1, 2004, 5% of GenBank was from the Sargasso Sea scaffold collection. A BLAST analysis of one sequence read from their collection against GenBank will often identify 50 similar DNA fragments of no known function that are all from the Sargasso Sea, making annotation laborious. It might be useful for users of GenBank to have the option to exclude or include environmental DNA sequences from their searches, just as users of the Human EST database can select sublibraries to search. It is critical that users of the databases are aware that finding matches to sequences from the Sargasso Sea is more likely to be due to the abundance of sequences from this study than to ecological similarities.

CHALLENGES WITH METAGENOMIC ANALYSIS

Phylogenetic Anchors

The ideal phylogenetic anchor would be equally represented in all species. The 16S rRNA genes do not meet this standard because microorganisms differ in the number of *rrn* operons they carry in their genomes, with a range of 1 to 15 (58). If the number

of *rrn* operons is positively correlated with growth rate, as has been postulated (58), then slow-growing, difficult to culture bacteria would be poorly represented in 16S rRNA libraries generated by PCR, and their 16S rRNA genes would occur less often in metagenomic clones than the 16S rRNA genes of their rapidly growing counterparts. Once genomes have been reconstructed, one *rrn* operon per genome is sufficient to determine the phylogenetic affiliation of the source of the genes in the genome, but in the absence of metagenome reconstruction for an entire environment, phylogenetic anchors that are found at frequent intervals in genomes are essential. When it is not possible to identify a 16S rRNA gene, another anchor is needed.

An alternative to finding phylogenetic anchors on a DNA fragment encoding a function of interest is to find fragments of the genome that are linked to the one of interest and search for phylogenetic anchors on them. This method increases the effective size of the contiguous piece of DNA that is being analyzed without requiring an increase in the size of the inserts in the library. To find fragments that are linked in a simple community, a library with high redundancy is needed. The clones can be blotted on a membrane and then probed with the clone of interest to identify those that have an end overlapping with it (110) or clones can be screened by PCR for overlapping regions (84). Neither of these methods has been reported as successful, but this is an area of rapid development and functional methods will likely be established in the future.

Size of Metagenomes

Constructing metagenomic libraries from environmental samples is conceptually simple but technically challenging. If seawater contains 200 species per ml (23), then the metagenome would contain 1 Gbp of unique DNA. To obtain greater than single sequencing coverage, the size of a metagenomic library would need to be many times the size of the metagenome. Because members of a community are not equally represented, it is likely that a metagenomic library of minimum coverage would only represent the genomes of the most abundant species. To obtain substantial representation of rare members ($<1\%$) of the community, the library would likely need to contain 100- to 1000-fold coverage of the metagenome. A library of 500 Gbp might be required to capture the species richness in 1 ml of seawater. Cloning the metagenome of soil, with a species richness 20-fold higher than seawater (23), would be a considerably more daunting prospect (10,000 Gbp). These examples illustrate several challenges in constructing and interpreting information from metagenomic libraries: (*a*) a large amount of DNA must be isolated and cloned from a small sample, (*b*) many clones and sequences must be processed to provide meaningful data, and (*c*) lognormal-type population distributions make it difficult to represent the minor species from a sample. Each of these challenges is being addressed, and several studies evaluating methods for library construction have been published (10, 33, 34, 37, 44).

Size of Inserts

Strategies for library construction vary depending on the intended study of the resulting library. Libraries containing large DNA fragments are constructed in lambda phage, cosmid, fosmid, and BAC vectors. Most of the reported large-insert metagenomic libraries contain fewer than 100,000 clones (Tables 1 and 2) and are several orders of magnitude too small to capture the entire microbial diversity present in the complex communities they represent. Although increasing the library size is a worthy goal, existing libraries have provided useful insights into the microbial ecology of several ecosystems in the absence of complete metagenome coverage.

Small-insert libraries have a significant advantage over large insert libaries because to obtain small inserts the microorganisms can be lysed by harsh methods that would shear DNA too much to obtain large inserts. Bead beating, for example, extracts DNA from diverse Bacteria and Archaea, providing a good representation of the community. Small-insert libraries are not useful for capturing complex pathways requiring many genes, but they provide an appropriate resource for discovery of new metabolic functions encoded by single genes and for reconstructing metagenomes (114, 115). The past reconstruction studies, however, did not take advantage of the harsh lysis methods that are accommodated by small-insert library construction, instead using chemical lysis, which does not access DNA from as diverse a group of organisms. Although both reconstruction studies reported sequences from Archaea, which are notoriously difficult to lyse, the Sargasso Sea study did not report members of the phyla containing only uncultured organisms, so the effect of the DNA extraction on the diversity of DNA in the metagenomic libraries method remains unclear.

The two metagenome reconstruction studies illustrate the difficulty in representing all of the members of a community in the library. For example, five species were identified within the AMD metagenome and the two archaeal *Ferroplasma* and *Leptospirillum* spp. were represented similarly in the metagenomic libraries, although fluorescent in situ hybridization (FISH) showed that the *Leptospirillum* spp. represented 85% of the community (114). The discrepancy between FISH and genome-sequencing data suggests that there may be a cloning bias, which is advantageous for this application because it led to high representation in the library of minor members of the community. In the Sargasso Sea study (115), the genomes of five species were represented with more than 3x coverage in the libraries, but the seawater contains an estimated 1800 species, suggesting that at least five of the species are far more abundant than others, illustrating the limitation of metagenomics in providing access to rare community members.

As with any genomics study, gene product toxicity is a concern in metagenomic analyses. However, a high-copy vector was used to construct the AMD metagenomic library (114), and the small number of gaps suggests little impact on the composition of the library. High-copy vectors are useful because it is easier to obtain sufficient DNA for further analysis, and for this reason, several plasmids that

have copy number that can be modulated from 1 to 50 per cell have been developed (44). These plasmids can be maintained at low copy during clone isolation to avoid toxicity and then can be amplified for screening or plasmid isolation.

Identifying Sequences of Interest in Large Metagenomic Libraries

The library size required to obtain sufficient coverage of the metagenome of even the simplest community presents a significant challenge for screening. Brute-force sequencing (114, 115) has provided tremendous insight into the libraries and the communities from which they were derived, but the information gleaned from sequencing is limited by the annotation of genes in the existing databases and the available sequencing capacity. In the Sargasso Sea study, only 35% of the genes identified had significant sequence similarity to genes in the public databases. Functional screening has the potential to identify interesting genes that would not be recognizable based on their sequences, but sequence-based screening can identify sequences that would not be expressed in the host species carrying the library. A combination of sequence-based methods and functional screening is critical to advancing the field because neither can define the full diversity of gene function in the libraries. High-throughput methods are needed to identify clones carrying functionally active genes, phylogenetic anchors, and novel genes.

FUNCTIONAL SCREENING Advances in screening for active clones will increase the knowledge mined from metagenomic libraries. Key approaches will include new selections that facilitate identification of active clones from among millions of clones. Another productive approach is to construct reporter fusions that respond to expression of the genes of interest. If clones expressing the reporter can be identified rapidly by selection for antibiotic resistance or fluorescence-activated cell sorting, then libraries of sufficient size to represent the diversity of a natural environment could be screened. Implementation of such screens will provide comprehensive functional information that will complement complete sequencing of the metagenome.

SEQUENCE-BASED SCREENING Screening libraries for genes of interest using primers or probes based on conserved sequences identifies homologues of known genes. This has proved effective to identify phylogenetic anchors and genes encoding enzymes with highly conserved domains (81). A challenge associated with screening libraries for clones carrying phylogenetic anchors is detecting the anchor on the cloned DNA without detecting the homologue in the chromosome of the host cell. This can be circumvented by using a vector with high or inducible copy number (44). When the cloned gene is in 50-fold excess to the chromosomal copy, the signal is sufficient to detect. Alternatively, "terminator PCR" can be used to block amplification of the host cell's homologue. Terminator PCR was used to identify clones carrying 16S rRNA genes, and terminator oligonucleotides specific for the *E. coli* 16S rRNA gene prevented detection of the host cell genes (63, 96). Finally, libraries can be screened with taxon-specific oligonucleotide probes and

PCR primers for 16S rRNA genes of interest that will not detect the *E. coli* genes (116).

New approaches are directed toward identifying sequences that are unique to uncultured microorganisms or those specific to a particular environment. These methods involve profiling clones with microarrays that identify previously unknown genes in environmental samples (107), subtractive hybridization to eliminate all sequences that hybridize with another environment, or subtractive hybridization to identify differentially expressed genes (35), and genomic sequence tags (28). These methods will enhance the efficiency of screening and aid in identifying minor components in communities and genes that define community uniqueness.

INTEGRATING METAGENOMICS AND COMMUNITY ECOLOGY

Metagenomics is a powerful approach for exploring the ecology of complex microbial communities. Its power will be realized when it is integrated with classical ecological approaches and efforts to culture previously unculturable microorganisms, which will likely be facilitated by clues about the physiology of the uncultured microorganisms derived from metagenomic analysis. Microscopy and stable isotope analysis are two approaches that will be particularly informative when linked to metagenomics.

Microscopy

Metagenomics complements direct observation of microorganisms in situ with FISH (1), RING-FISH (128), and FISH-MAR (62). Fluorescent in situ hybridization (FISH) is commonly used in microbial ecology studies to visualize microorganisms that contain rRNA that hybridizes with a fluorescently labeled probe, most often directed toward the 16S rRNA gene. Until recently, FISH has been limited to detection of highly expressed genes, such as the 16S rRNA genes, because it detects abundant RNA and not single-copy genes. To overcome this limitation, Zwirglmaier et al. (128) described an adaptation of FISH called RING-FISH (*r*ecognition of *in*dividual *g*enes-FISH) that facilitates visualization of plasmid or chromosomal genes in situ by increasing the sensitivity of detection. Another variation on FISH is to link it with microautoradiography to identify taxons that utilize particular substrates in natural microbial communities (40, 62, 76). Further augmentation of metagenomics with microscopy that can monitor genes, gene expression, and environmental conditions on a microscale will provide new insights into the workings of microbial communities.

Stable Isotopes

The use of stable isotopes to understand cycling of elements in microbial communities presents a singular opportunity in metagenomics. Members of communities fed substrates labeled with stable carbon or nitrogen isotopes incorporate the

TABLE 4 Bacterial phyla represented in metagenomic libraries

Bacterial phyla with cultured members[a]	Candidate bacterial phyla (no cultured representatives)
Acidobacteria* (63, 85, 96)	ABY1
Actinobacteria (63, 115)	BD1-5 group
Aquificae	BRC1
Bacteroidetes (63, 96, 111, 115)	Guaymas1
Caldithrix	Marine Group A
Chlamydiae	NC10
Chlorobi (115)	NKB19 (63)
Chloroflexi* (115)	OP1
Chrysiogenes	OP3
Coprothermobacter	OP5
Cyanobacteria (111, 115)	OP8
Deferribacteres	OP9
Deinococcus-Thermus (115)	OP10 (63)
Desulfurobacterium	OP11*
Dictyoglomus	OS-K
Fibrobacteres (111)	SBR1093
Firmicutes (96, 115)	SC3
Fusobacteria (115)	SC4
Gemmatimonadetes* (63)	Termite Group 1
Nitrospira (114)	TM6
Planctomycetes* (63)	TM7
Proteobacteria (63, 96, 111, 115)	VadinBE97
Spirochaetes (115)	WS2
Synergistes	WS3 (63)
Thermodesufobacteria	WS5
Thermotogae	WS6
Verrucomicrobia* (63, 111)	

[a]The phylum names reviewed by Rappé & Giovannoni (90) are listed alphabetically. Highlighted text and references indicate bacterial phyla for which at least one clone within a metagenomic library has been reported; phylogenetic anchors (usually 16S rRNA) provide evidence of these phyla within metagenomic libraries.

*Indicates phyla that Hugenholtz et al. (54) or Rappé & Giovannoni (90) noted as particularly abundant in environmental samples but underrepresented in culture.

isotopes into their DNA. Building metagenomic libraries from such communities can be used either to determine which community members are metabolically active (using a labeled substrate such as glucose, which can be metabolized by many organisms) or which utilize the labeled substrate (using a specialized substrate, such as an environmental pollutant). The DNA from the fraction of the community that used the labeled substrate will contain DNA with the stable isotope, making it separable from unlabeled DNA by density centrifugation. Metagenomic libraries and 16S rRNA gene libraries can be constructed with the labeled DNA to enrich for genes associated with the active species (71, 72, 74, 75, 86–88). Innovative approaches such as these will advance metagenomic analysis to the next level, linking phylogeny and function more precisely.

CONCLUDING REMARKS

Constructing metagenomic libraries captures the phylogenetic and genetic diversity in environmental samples. Clones derived from one third of the 52 bacterial phyla have been reported, including representatives of several candidate phyla and most of the phyla that are particularly abundant in environmental samples but underrepresented by cultured isolates (Table 4). The assessment of phylogenetic diversity in metagenomic libraries underestimates the biodiversity because a small proportion of clones contain phylogenetic anchors, and the phylogenetic representation of very few libraries has been characterized. The genetic potential of the libraries has only begun to be tapped. The small molecules and enzymes already discovered indicate the potential of metagenomics to mine the environment for fundamental knowledge and products for biotechnology. But effective mining will require high-throughput functional screens and selections and rapid methods for identifying sequences of interest. The advances in sequencing technology make it possible to accumulate vast amounts of DNA sequence, which has proved a powerful source of discovery, but more directed methods will lead to larger collections of genes of a particular type or function. Future advances in understanding the differences among communities or environments will be derived from "comparative metagenomics" in which libraries prepared from different sites or at different times can be compared. Our knowledge of 1% of the microbial world through culturing may be predictive of much that we will find among the uncultured organisms, but abundant surprises await us in this unknown world.

The *Annual Review of Genetics* is online at http://genet.annualreviews.org

LITERATURE CITED

1. Amann RI, Krumholz L, Stahl DA. 1990. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental-studies in microbiology. *J. Bacteriol.* 172:762–70
2. Amann RI, Ludwig W, Schleifer KH. 1995. Phylogenetic identification and *in*

situ detection of individual microbial cells without cultivation. *Microbiol. Rev.* 59:143–69

3. Baker BJ, Hugenholtz P, Dawson SC, Banfield JF. 2003. Extremely acidophilic protists from acid mine drainage host Rickettsiales-lineage endosymbionts that have intervening sequences in their 16S rRNA genes. *Appl. Environ. Microbiol.* 69:5512–18
4. Barns SM, Takala SL, Kuske CR. 1999. Wide distribution and diversity of members of the bacterial kingdom Acidobacterium in the environment. *Appl. Environ. Microbiol.* 65:1731–37
5. Béjà O, Aravind L, Koonin EV, Suzuki MT, Hadd A, et al. 2000. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* 289:1902–6
6. Béjà O, Koonin EV, Aravind L, Taylor LT, Seitz H, et al. 2002. Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl. Environ. Microbiol.* 68:335–45
7. Béjà O, Spudich EN, Spudich JL, Leclerc M, DeLong EF. 2001. Proteorhodopsin phototrophy in the ocean. *Nature* 411:786–89
8. Béjà O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, et al. 2002. Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* 415: 630–33
9. Béjà O, Suzuki MT, Koonin EV, Aravind L, Hadd A, et al. 2000. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* 2:516–29
10. Berry AE, Chiocchini C, Selby T, Sosio M, Wellington EMH. 2003. Isolation of high molecular weight DNA from soil for cloning into BAC vectors. *FEMS Microbiol. Lett.* 223:15–20
11. Brady SF, Chao CJ, Clardy J. 2002. New natural product families from an environmental DNA (eDNA) gene cluster. *J. Am. Chem. Soc.* 124:9968–69
12. Brady SF, Chao CJ, Handelsman J, Clardy J. 2001. Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA. *Org. Lett.* 3:1981–84
13. Brady SF, Clardy J. 2000. Long-chain N-acyl amino acid antibiotics isolated from heterologously expressed environmental DNA. *J. Am. Chem. Soc.* 122:12903–4
14. Brady SF, Clardy J. 2003. Synthesis of long-chain fatty acid enol esters isolated from an environmental DNA clone. *Org. Lett.* 5:121–24
15. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. 2003. Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* 185:6220–23
16. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. 2002. Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. USA* 99:14250–55
17. Brzostowicz PC, Walters DM, Thomas SM, Nagarajan V, Rouviere PE. 2003. mRNA differential display in a microbial enrichment culture: Simultaneous identification of three cyclohexanone monooxygenases from three species. *Appl. Environ. Microbiol.* 69:334–42
18. Campbell BJ, Stein JL, Cary SC. 2003. Evidence of chemolithoautotrophy in the bacterial community associated with *Alvinella pompejana*, a hydrothermal vent polychaete. *Appl. Environ. Microbiol.* 69: 5070–78
19. Coenye T, Vandamme P. 2003. Extracting phylogenetic information from whole-genome sequencing projects: the lactic acid bacteria as a test case. *Microbiology* 149:3507–17
20. Coque TM, Oliver A, Perez-Diaz JC, Baquero F, Canton R. 2002. Genes encoding TEM-4, SHV-2, and CTX-M-10 extended-spectrum beta-lactamases are carried by multiple *Klebsiella pneumoniae* clones in a single hospital (Madrid, 1989 to 2000). *Antimicrob. Agents. Chemother.* 46:500–10

21. Cottrell MT, Moore JA, Kirchman DL. 1999. Chitinases from uncultured marine microorganisms. *Appl. Environ. Microbiol.* 65:2553–57
22. Courtois S, Cappellano CM, Ball M, Francou FX, Normand P, et al. 2003. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl. Environ. Microbiol.* 69:49–55
23. Curtis TP, Sloan WT, Scannell JW. 2002. Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci. USA* 99: 10494–99
24. Dahllof I, Baillie H, Kjelleberg S. 2000. *rpo* B-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. *Appl. Environ. Microbiol.* 66:3376–80
25. de la Torre JR, Christianson LM, Béjà O, Suzuki MT, Karl DM, et al. 2003. Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proc. Natl. Acad. Sci. USA* 100:12830–35
26. DeLong EF. 2002. Microbial population genomics and ecology. *Curr. Opin. Microbiol.* 5:520–24
27. Diaz-Torres ML, McNab R, Spratt DA, Villedieu A, Hunt N, et al. 2003. Novel tetracycline resistance determinant from the oral metagenome. *Antimicrob. Agents. Chemother.* 47:1430–32
28. Dunn JJ, McCorkle SR, Praissman LA, Hind G, Van Der Lelie D, et al. 2002. Genomic signature tags (GSTs): a system for profiling genomic DNA. *Genome Res.* 12:1756–65
29. Edwards KJ, Bond PL, Druschel GK, McGuire MM, Hamers RJ, et al. 2000. Geochemical and biological aspects of sulfide mineral dissolution: lessons from Iron Mountain, California. *Chem. Geol.* 169:383–97
30. Eisen JA. 1995. The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *J. Mol. Evol.* 41:1105–23
31. Entcheva P, Liebl W, Johann A, Hartsch T, Streit WR. 2001. Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. *Appl. Environ. Microbiol.* 67:89–99
32. Eschenfeldt WH, Stols L, Rosenbaum H, Khambatta ZS, Quaite-Randall E, et al. 2001. DNA from uncultured organisms as a source of 2,5-diketo-D-gluconic acid reductases. *Appl. Environ. Microbiol.* 67:4206–14
33. Fiandt M. 2000. Construction of an environmental genomic DNA library from soil using the EpiFOS fosmid library production kit. *Epicentre Forum* 7:6
34. Gabor EM, de Vries EJ, Janssen DB. 2003. Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. *FEMS Microbiol. Ecol.* 44:153–63
35. Galbraith EA, Antonopoulos DA, White BA. 2004. Suppressive subtractive hybridization as a tool for identifying genetic diversity in an environmental metagenome: the rumen as a model. *Environ. Microbiol.* In press
36. Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, et al. 2002. Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Appl. Environ. Microbiol.* 68:4301–6
37. Gillespie DE, Rondon MR, Handelsman J. 2004. Metagenomic libraries from uncultured microorganisms. In *Molecular Microbial Ecology*. Bios Sci. Publ. In press
38. Goh SH, Facklam RR, Chang M, Hill JE, Tyrrell GJ, et al. 2000. Identification of *Enterococcus* species and phenotypically similar *Lactococcus* and *Vagococcus* species by reverse checkerboard hybridization to chaperonin 60 gene sequences. *J. Clin. Microbiol.* 38:3953–59
39. Gray KA, Richardson TH, Robertson DE, Swanson PE, Subramanian MV. 2003.

Soil-based gene discovery: a new technology to accelerate and broaden biocatalytic applications. *Adv. Appl. Microbiol.* 52:1–27

40. Gray ND, Howarth R, Pickup RW, Jones JG, Head IM. 2000. Use of combined microautoradiography and fluorescence in situ hybridization to determine carbon metabolism in mixed natural communities of uncultured bacteria from the genus *Achromatium*. *Appl. Environ. Microbiol.* 66:4518–22
41. Griffiths E, Gupta RS. 2001. The use of signature sequences in different proteins to determine the relative branching order of bacterial divisions: evidence that Fibrobacter diverged at a similar time to Chlamydia and the Cytophaga-Flavobacterium-Bacteroides division. *Microbiology* 147:2611–22
42. Gruber TM, Bryant DA. 1998. Characterization of the group 1 and group 2 sigma factors of the green sulfur bacterium *Chlorobium tepidum* and the green non-sulfur bacterium *Chloroflexus aurantiacus*. *Arch. Microbiol.* 170:285–96
43. Hallam SJ, Girguis PR, Preston CM, Richardson PM, DeLong EF. 2003. Identification of methyl coenzyme M reductase A (*mcr* A) genes associated with methane-oxidizing archaea. *Appl. Environ. Microbiol.* 69:5483–91
44. Handelsman J, Liles MR, Mann DA, Riesenfeld CS, Goodman RM. 2003. Cloning the metagenome: culture-independent access to the diversity and functions of the uncultivated microbial world. In *Functional Microbial Genomics*, ed. B Wren, N Dorrell, pp. 241–55. New York: Academic
45. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5:R245–49
46. Handelsman J, Wackett LP. 2002. Ecology and industrial microbiology: microbial diversity—sustaining the Earth and industry. *Curr. Opin. Microbiol.* 5:237–39
47. Head IM, Saunders JR, Pickup RW. 1998. Microbial evolution, diversity, and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms. *Microb. Ecol.* 35:1–21
48. Healy FG, Ray RM, Aldrich HC, Wilkie AC, Ingram LO, et al. 1995. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. *Appl. Microbiol. Biotechnol.* 43:667–74
49. Henne A, Daniel R, Schmitz RA, Gottschalk G. 1999. Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Appl. Environ. Microbiol.* 65:3901–7
50. Henne A, Schmitz RA, Bomeke M, Gottschalk G, Daniel R. 2000. Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. *Appl. Environ. Microbiol.* 66:3113–16
51. Hill JE, Seipp RP, Betts M, Hawkins L, Van Kessel AG, et al. 2002. Extensive profiling of a complex microbial community by high-throughput sequencing. *Appl. Environ. Microbiol.* 68:3055–66
52. Holmes AJ, Gillings MR, Nield BS, Mabbutt BC, Nevalainen KM, et al. 2003. The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environ. Microbiol.* 5:383–94
53. Hugenholtz P. 2002. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* 3:REVIEWS0003.1–8
54. Hugenholtz P, Goebel BM, Pace NR. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* 180:4765–74
55. Hughes DS, Felbeck H, Stein JL. 1997. A histidine protein kinase homolog from the endosymbiont of the hydrothermal vent

tubeworm *Riftia pachyptila*. *Appl. Environ. Microbiol.* 63:3494–98
56. Janssen PH, Yates PS, Grinton BE, Taylor PM, Sait M. 2002. Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions *Acidobacteria*, *Actinobacteria Proteobacteria*, and *Verrucomicrobia*. *Appl. Environ. Microbiol.* 68:2391–96
57. Kent AD, Triplett EW. 2002. Microbial communities and their interactions in soil and rhizosphere ecosystems. *Annu. Rev. Microbiol.* 56:211–36
58. Klappenbach JA, Dunbar JM, Schmidt TM. 2000. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* 66:1328–33
59. Knietsch A, Bowien S, Whited G, Gottschalk G, Daniel R. 2003. Identification and characterization of coenzyme B-12-dependent glycerol dehydratase- and diol dehydratase-encoding genes from metagenomic DNA libraries derived from enrichment cultures. *Appl. Environ. Microbiol.* 69:3048–60
60. Knietsch A, Waschkowitz T, Bowien S, Henne A, Daniel R. 2003. Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on *Escherichia coli*. *Appl. Environ. Microbiol.* 69:1408–16
61. Knietsch A, Waschkowitz T, Bowien S, Henne A, Daniel R. 2003. Metagenomes of complex microbial consortia derived from different soils as sources for novel genes conferring formation of carbonyls from short-chain polyols on *Escherichia coli*. *J. Mol. Microbiol. Biotechnol.* 5:46–56
62. Kong YH, Beer M, Seviour RJ, Lindrea KC, Rees GN. 2001. Structure and functional analysis of the microbial community in an aerobic: anaerobic sequencing batch reactor (SBR) with no phosphorus removal. *Syst. Appl. Microbiol.* 24:597–609
63. Liles MR, Manske BF, Bintrim SB, Handelsman J, Goodman RM. 2003. A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl. Environ. Microbiol.* 69:2684–91
64. Lopez-Garcia P, Brochier C, Moreira D, Rodriguez-Valera F. 2004. Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. *Environ. Microbiol.* 6:19–34
65. Lorenz P, Liebeton K, Niehaus F, Eck J. 2002. Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. *Curr. Opin. Biotechnol.* 13:572–77
66. Lorenz P, Schleper C. 2002. Metagenome—a challenging source of enzyme discovery. *J. Mol. Catal. B* 19:13–19
67. Ludwig W, Neumaier J, Klugbauer N, Brockmann E, Roller C, et al. 1993. Phylogenetic relationships of Bacteria based on comparative sequence analysis of elongation factor Tu and ATP-synthase beta-subunit genes. *Antonie van Leeuwenhoek* 64:285–305
68. MacNeil IA, Tiong CL, Minor C, August PR, Grossman TH, et al. 2001. Expression and isolation of antimicrobial small molecules from soil DNA libraries. *J. Mol. Microbiol. Biotechnol.* 3:301–8
69. Majernik A, Gottschalk G, Daniel R. 2001. Screening of environmental DNA libraries for the presence of genes conferring $Na^+(Li^+)/H^+$ antiporter activity on *Escherichia coli*: characterization of the recovered genes and the corresponding gene products. *J. Bacteriol.* 183:6645–53
70. Man D, Wang W, Sabehi G, Aravind L, Post AF, et al. 2003. Diversification and spectral tuning in marine proteorhodopsins. *EMBO J.* 22:1725–31
71. Manefield M, Whiteley AS, Griffiths RI, Bailey MJ. 2002. RNA stable isotope probing, a novel means of linking

microbial community function to phylogeny. *Appl. Environ. Microbiol.* 68: 5367–73

72. Manefield M, Whiteley AS, Ostle N, Ineson P, Bailey MJ. 2002. Technical considerations for RNA-based stable isotope probing: an approach to associating microbial diversity with microbial community function. *Rapid Commun. Mass Spectrom.* 16:2179–83
73. Mollet C, Drancourt M, Raoult D. 1997. *rpo* B sequence analysis as a novel basis for bacterial identification. *Mol. Microbiol.* 26:1005–11
74. Morris SA, Radajewski S, Willison TW, Murrell JC. 2002. Identification of the functionally active methanotroph population in a peat soil microcosm by stable-isotope probing. *Appl. Environ. Microbiol.* 68:1446–53
75. Murrell JC, Radajewski S. 2000. Cultivation-independent techniques for studying methanotroph ecology. *Res. Microbiol.* 151:807–14
76. Nielsen JL, Christensen D, Kloppenborg M, Nielsen PH. 2003. Quantification of cell-specific substrate uptake by probe-defined bacteria under in situ conditions by microautoradiography and fluorescence in situ hybridization. *Environ. Microbiol.* 5:202–11
77. Oh HJ, Cho KW, Jung IS, Kim WH, Hur BK, et al. 2003. Expanding functional spaces of enzymes by utilizing whole genome treasure for library construction. *J. Mol. Catal. B* 26:241–50
78. Pace NR. 1997. A molecular view of microbial diversity and the biosphere. *Science* 276:734–40
79. Pace NR, Stahl DA, Lane DJ, Olsen GJ. 1985. Analyzing natural microbial populations by rRNA sequences. *ASM News* 51:4–12
80. Pace NR, Stahl DA, Lane DJ, Olsen GJ. 1986. The analysis of natural populations by ribosomal RNA sequences. *Adv. Microbial Ecol.* 9:1–55
81. Piel J. 2002. A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of *Paederus* beetles. *Proc. Natl. Acad. Sci. USA* 99:14002–7
82. Piel J, Hui D, Fusetani N, Matsunaga S. 2004. Targeting modular polyketide synthases with iteratively acting acyltransferases from metagenomes of uncultured bacterial consortia. *Environ. Microbiol.* In press
83. Preston CM, Wu KY, Molinski TF, DeLong EF. 1996. A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen. nov., sp. nov. *Proc. Natl. Acad. Sci. USA* 93:6241–46
84. Quaiser A, Ochsenreiter T, Klenk HP, Kletzin A, Treusch AH, et al. 2002. First insight into the genome of an uncultivated crenarchaeote from soil. *Environ. Microbiol.* 4:603–11
85. Quaiser A, Ochsenreiter T, Lanz C, Schuster SC, Treusch AH, et al. 2003. Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. *Mol. Microbiol.* 50:563–75
86. Radajewski S, Ineson P, Parekh NR, Murrell JC. 2000. Stable-isotope probing as a tool in microbial ecology. *Nature* 403:646–49
87. Radajewski S, Murrell JC. 2002. Stable isotope probing for detection of methanotrophs after enrichment with $^{13}CH_4$. *Methods Mol. Biol.* 179:149–57
88. Radajewski S, Webster G, Reay DS, Morris SA, Ineson P, et al. 2002. Identification of active methylotroph populations in an acidic forest soil by stable-isotope probing. *Microbiology* 148:2331–42
89. Rappe MS, Connon SA, Vergin KL, Giovannoni SJ. 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* 418:630–33
90. Rappe MS, Giovannoni SJ. 2003. The uncultured microbial majority. *Annu. Rev. Microbiol.* 57:369–94
91. Richardson TH, Tan X, Frey G, Callen W, Cabell M, et al. 2002. A novel,

high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable alpha-amylase. *J. Biol. Chem.* 277:26501–7

92. Riesenfeld CS, Goodman RM, Handelsman J. 2004. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ. Microbiol.* In press
93. Rodriguez-Valera F. 2002. Approaches to prokaryotic biodiversity: a population genetics perspective. *Environ. Microbiol.* 4:628–33
94. Rodriguez-Valera F. 2004. Environmental genomics, the big picture? *FEMS Microbiol. Lett.* 231:153–58
95. Rohwer F. 2003. Global phage diversity. *Cell* 113:141
96. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, et al. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* 66:2541–47
97. Rondon MR, Goodman RM, Handelsman J. 1999. The Earth's bounty: assessing and accessing soil microbial diversity. *Trends Biotechnol.* 17:403–9
98. Sabehi G, Massana R, Bielawski JP, Rosenberg M, Delong EF, et al. 2003. Novel proteorhodopsin variants from the Mediterranean and Red Seas. *Environ. Microbiol.* 5:842–49
99. Sait M, Hugenholtz P, Janssen PH. 2002. Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. *Environ. Microbiol.* 4:654–66
100. Sandler SJ, Hugenholtz P, Schleper C, DeLong EF, Pace NR, et al. 1999. Diversity of *rad* A genes from cultured and uncultured archaea: comparative analysis of putative RadA proteins and their use as a phylogenetic marker. *J. Bacteriol.* 181:907–15
101. Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, et al. 1998. Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum*. *J. Bacteriol.* 180: 5003–9
102. Schleper C, Holben W, Klenk HP. 1997. Recovery of Crenarchaeotal ribosomal DNA sequences from freshwater-lake sediments. *Appl. Environ. Microbiol.* 63:321–23
103. Schleper C, Swanson RV, Mathur EJ, DeLong EF. 1997. Characterization of a DNA polymerase from the uncultivated psychrophilic archaeon *Cenarchaeum symbiosum*. *J. Bacteriol.* 179:7803–11
104. Schloss PD, Handelsman J. 2003. Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol.* 14:303–10
105. Schmeisser C, Stockigt C, Raasch C, Wingender J, Timmis KN, et al. 2003. Metagenome survey of biofilms in drinking-water networks. *Appl. Environ. Microbiol.* 69:7298–309
106. Schmidt TM, DeLong EF, Pace NR. 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* 173:4371–78
107. Sebat JL, Colwell FS, Crawford RL. 2003. Metagenomic profiling: microarray analysis of an environmental genomic library. *Appl. Environ. Microbiol.* 69:4927–34
108. Sigler PB, Xu Z, Rye HS, Burston SG, Fenton WA, et al. 1998. Structure and function in GroEL-mediated protein folding. *Annu. Rev. Biochem.* 67:581–608
109. Stahl DA, Lane DJ, Olsen GJ, Pace NR. 1984. Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science* 224:409–11
110. Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF. 1996. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment front a planktonic marine archaeon. *J. Bacteriol.* 178:591–99
111. Suzuki MT, Béjà O, Taylor LT, Delong EF. 2001. Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal

marine bacterioplankton. *Environ. Microbiol.* 3:323–31
112. Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, et al. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. *Science* 296:2376–79
113. Torsvik V, Goksoyr J, Daae FL. 1990. High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.* 56:782–87
114. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43
115. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74
116. Vergin KL, Urbach E, Stein JL, DeLong EF, Lanoil BD, et al. 1998. Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order Planctomycetales. *Appl. Environ. Microbiol.* 64:3075–78
117. Voget S, Leggewie C, Uesbeck A, Raasch C, Jaeger KE, et al. 2003. Prospecting for novel biocatalysts in a soil metagenome. *Appl. Environ. Microbiol.* 69:6235–42
118. Wackett LP. 2001. Soil DNA and the microbial metagenome. An annotated selection of World Wide Web sites relevant to the topics in Environmental Microbiology Web alert. *Environ. Microbiol.* 3:352–53
119. Wang GY, Graziani E, Waters B, Pan W, Li X, et al. 2000. Novel natural products from soil DNA libraries in a streptomycete host. *Org. Lett.* 2:2401–4
120. Wilkinson DE, Jeanicke T, Cowan DA. 2002. Efficient molecular cloning of environmental DNA from geothermal sediments. *Biotechnol. Lett.* 24:155–61
121. Woese CR. 1987. Bacterial evolution. *Microbiol. Rev.* 51:221–71
122. Woese CR, Olsen GJ, Ibba M, Soll D. 2000. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* 64:202–36
123. Wolf YI, Aravind L, Grishin NV, Koonin EV. 1999. Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* 9:689–710
124. Wommack KE, Colwell RR. 2000. Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* 64:69–114
125. Yamamoto S, Bouvet PJ, Harayama S. 1999. Phylogenetic structures of the genus Acinetobacter based on *gyr* B sequences: comparison with the grouping by DNA-DNA hybridization. *Int. J. Syst. Bacteriol.* 49:87–95
126. Zeidner G, Preston CM, Delong EF, Massana R, Post AF, et al. 2003. Molecular diversity among marine picophytoplankton as revealed by *psb* A analyses. *Environ. Microbiol.* 5:212–16
127. Zengler K, Toledo G, Rappe M, Elkins J, Mathur EJ, et al. 2002. Cultivating the uncultured. *Proc. Natl. Acad. Sci. USA* 99:15681–86
128. Zwirglmaier K, Ludwig W, Schleifer KH. 2004. Recognition of individual genes in a single bacterial cell by fluorescence in situ hybridization—RING-FISH. *Mol. Microbiol.* 51:89–96

Annual Review of Genetics
Volume 38, 2004

CONTENTS