METAGENOMICS: Genomic Analysis of Microbial Communities

Christian S. Riesenfeld,1,2 Patrick D. Schloss,1 and Jo Handelsman1,2

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 .................................................... 526
METAGENOMICS DEFINED ........................................ 527
LINKING PHYLOGENY AND FUNCTION WITHIN SPECIES .......... 529
Phylogenetic Anchors ............................................. 529
Function Then Phylogeny ......................................... 529
Phylogeny Then Function ....................................... 533
Acidobacterium Phylogeny and Function .......................... 533
Archaeal Phylogeny and Function ................................. 534
Proteorhodopsin Function and Phylogeny ......................... 535
LINKING PHYLOGENY AND FUNCTION IN MICROBIAL COMMUNITIES: METAGENOME RECONSTRUCTION .......... 536
Metagenomic Analysis of Bacteriophage ........................ 536
Metagenome of the Microbial Community in Acid Mine Drainage 537
Metagenome of the Microbial Community in the Sargasso Sea ... 538
CHALLENGES WITH METAGENOMIC ANALYSIS ................... 539
Phylogenetic Anchors ............................................. 539

0066-4197/04/1215-0525$14.00
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 contribute to public health concerns such as antibiotic resistance and nosocomial infections (20).
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 \( \lambda \) 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

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

<table>
<thead>
<tr>
<th>Environment</th>
<th>Number of clones</th>
<th>Insert size (kb)</th>
<th>Total DNA (Gb)</th>
<th>Genes of interest</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marine</td>
<td>6240</td>
<td>80</td>
<td></td>
<td>16S rRNA</td>
<td>(5)</td>
</tr>
<tr>
<td>Marine</td>
<td>7200</td>
<td>40</td>
<td>2.9</td>
<td>Archaea 16S rRNA</td>
<td>(6)</td>
</tr>
<tr>
<td>Marine</td>
<td>32,000</td>
<td>10–20 kb</td>
<td>0.48</td>
<td>Archaea 16S rRNA, Photosystem II (psbA)</td>
<td>(9)</td>
</tr>
<tr>
<td>Polychaete symbionts</td>
<td>n.s.</td>
<td>Fosmid</td>
<td>—</td>
<td>16S rRNA</td>
<td>(126)</td>
</tr>
<tr>
<td>Marine</td>
<td>5000</td>
<td>50</td>
<td>0.25</td>
<td>Polyketide synthase</td>
<td></td>
</tr>
<tr>
<td>Marine</td>
<td>n.s.</td>
<td>BAC</td>
<td>—</td>
<td>Proteorhodopsin</td>
<td>(25)</td>
</tr>
<tr>
<td>Sediment</td>
<td>n.s.</td>
<td>Fosmid</td>
<td>—</td>
<td>Archaea 16S rRNA methyl coenzyme M reductase A</td>
<td>(43)</td>
</tr>
<tr>
<td>Marine</td>
<td>6107</td>
<td>35–40</td>
<td>16S rRNA</td>
<td>Histidine protein kinase</td>
<td></td>
</tr>
<tr>
<td>Soil</td>
<td>3600</td>
<td>27</td>
<td>1.2</td>
<td>16S rRNA</td>
<td>(63, 96)</td>
</tr>
<tr>
<td>Soil</td>
<td>24,600</td>
<td>45</td>
<td></td>
<td>Acidobacteria 16S rRNA</td>
<td></td>
</tr>
<tr>
<td>Marine</td>
<td>6107</td>
<td>35–40</td>
<td>16S rRNA</td>
<td>RadA</td>
<td>(100)</td>
</tr>
<tr>
<td>Beetle symbionts</td>
<td>n.s.</td>
<td>Fosmid</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beetle symbionts</td>
<td>n.s.</td>
<td>Cosmid</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beetle and sponge</td>
<td>n.s.</td>
<td>Fosmid</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soil</td>
<td>25,278</td>
<td>35–40</td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soil</td>
<td>56,000</td>
<td>33–44</td>
<td>2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sponge symbionts</td>
<td>n.s.</td>
<td>Fosmid</td>
<td>—</td>
<td></td>
<td>(100)</td>
</tr>
<tr>
<td>Sponge symbionts</td>
<td>n.s.</td>
<td>40</td>
<td>—</td>
<td></td>
<td>(101)</td>
</tr>
<tr>
<td>Sponge symbionts</td>
<td>n.s.</td>
<td>40</td>
<td>—</td>
<td></td>
<td>(103)</td>
</tr>
<tr>
<td>River biofilm</td>
<td>n.s.</td>
<td>40</td>
<td>—</td>
<td></td>
<td>(107)</td>
</tr>
<tr>
<td>Marine</td>
<td>3552</td>
<td>40</td>
<td>0.14</td>
<td></td>
<td>(110)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(116)</td>
</tr>
</tbody>
</table>
TABLE 3  Metagenomics studies based on random sequencing

<table>
<thead>
<tr>
<th>Environment</th>
<th>Insert size (kb)</th>
<th>Sequence reads</th>
<th>Base pairs of sequence</th>
<th>Goal</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feces LASL</td>
<td>532</td>
<td>∼37,000</td>
<td></td>
<td>Random viral clone sequencing</td>
<td>(15)</td>
</tr>
<tr>
<td>Marine LASL</td>
<td>1061</td>
<td>∼740,000</td>
<td></td>
<td>Random viral clone sequencing</td>
<td>(16)</td>
</tr>
<tr>
<td>Drinking water</td>
<td>2496</td>
<td>2 × 10^6</td>
<td></td>
<td>Random clone sequencing</td>
<td>(105)</td>
</tr>
<tr>
<td>biofilms Plasmid</td>
<td>103,462</td>
<td>76.2 × 10^6</td>
<td></td>
<td>Reconstruct genome of microbial community</td>
<td>(114)</td>
</tr>
<tr>
<td>Acid mine drainage</td>
<td>3.2</td>
<td>2 × 10^6</td>
<td>1.63 × 10^9</td>
<td>“Pilot study” of large-scale whole community sequencing</td>
<td>(115)</td>
</tr>
<tr>
<td>Marine</td>
<td>2.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

aLASL: 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 sequence-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 \textit{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 \(\beta\)-subunit (24, 73), ATPase \(\beta\)-subunit (67), elongation factor TU (67), heat shock protein 70 (41), \(\sigma^{70}\)-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 Acidobacteriaceae-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. Béja 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 $6 \times 10^9$ 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 of 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 N₂ 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* in pure culture from a N₂-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 monoxygenase. 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 rRNA 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 libraries 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 (recognition of individual genes-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

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

\(^a\)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


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


111. Suzuki MT, Bejá O, Taylor LT, DeLong EF. 2001. Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal...


CONTENTS

MOBILE GROUP II INTRONS, Alan M. Lambowitz and Steven Zimmerly 1
THE GENETICS OF MAIZE EVOLUTION, John Doebley 37
GENETIC CONTROL OF RETROVIRUS SUSCEPTIBILITY IN MAMMALIAN CELLS, Stephen P. Goff 61
LIGHT SIGNAL TRANSDUCTION IN HIGHER PLANTS, Meng Chen, Joanne Chory, and Christian Fankhauser 87
CHLAMYDOMONAS REINHARDTII IN THE LANDSCAPE OF PIGMENTS, Arthur R. Grossman, Martin Lohr, and Chung Soon Im 119
THE GENETICS OF GEOCHEMISTRY, Laura R. Croal, Jeffrey A. Gralnick, Davin Malasarn, and Dianne K. Newman 175
CLOSING MITOSIS: THE Functions OF THE CDC14 PHOSPHATASE AND ITS REGULATION, Frank Stegmeier and Angelika Amon 203
RECOMBINATION PROTEINS IN YEAST, Berit Olsen Krogh and Lorraine S. Symington 233
DEVELOPMENTAL GENE AMPLIFICATION AND ORIGIN REGULATION, John Tower 273
THE FUNCTION OF NUCLEAR ARCHITECTURE: A GENETIC APPROACH, Angela Taddei, Florence Hediger, Frank R. Neumann, and Susan M. Gasser 305
GENETIC MODELS IN PATHOGENESIS, Elizabeth Pradel and Jonathan J. Ewbank 347
MELANOCYTES AND THE MICROPHTHALMIA TRANSCRIPTION FACTOR NETWORK, Eirikur Steingrimsson, Neal G. Copeland, and Nancy A. Jenkins 365
EPIGENETIC REGULATION OF CELLULAR MEMORY BY THE POLYCOMB AND TRITHORAX GROUP PROTEINS, Leonie Ringrose and Renato Paro 413
REPAIR AND GENETIC CONSEQUENCES OF ENDOGENOUS DNA BASE DAMAGE IN MAMMALIAN CELLS, Deborah E. Barnes and Tomas Lindahl 445
MITOCHONDRIA OF PROTISTS, Michael W. Gray, B. Franz Lang, and Gertraud Burger 477
CONTENTS

METAGENOMICS: GENOMIC ANALYSIS OF MICROBIAL COMMUNITIES,
Christian S. Riesenfeld, Patrick D. Schloss, and Jo Handelsman 525

GENOMIC IMPRINTING AND KINSHIP: HOW GOOD IS THE EVIDENCE?,
David Haig 553

MECHANISMS OF PATTERN FORMATION IN PLANT EMBRYOGENESIS,
Viola Willemsen and Ben Scheres 587

DUPlication AND Divergence: THE EVOLUTION OF NEW Genes
And Old Ideas, John S. Taylor and Jeroen Raes 615

Genetic Analyses from Ancient DNA, Svante Pääbo,
Hendrik Poinar, David Serre, Viviane Jaenicke-Despres, Juliane Hebler,
Nadin Rohland, Melanie Kuch, Johannes Krause, Linda Vigilant,
and Michael Hofreiter 645

Prion Genetics: New Rules for a New Kind of Gene,
Reed B. Wickner, Herman K. Edskes, Eric D. Ross, Michael M. Pierce,
Ulrich Baxa, Andreas Brachmann, and Frank Shewmaker 681

Proteolysis as a Regulatory Mechanism, Michael Ehrmann and
Tim Clausen 709

Mechanisms of MAP Kinase SignalinG Specificity in
Saccharomyces Cerevisiae, Monica A. Schwartz
and Hiten D. Madhani 725

rRNA Transcription in Escherichia Coli, Brian J. Paul, Wilma Ross,
Tamas Gaal, and Richard L. Gourse 749

Comparative Genomic Structure of Prokaryotes,
Stephen D. Bentley and Julian Parkhill 771

Species Specificity in Pollen-Pistil Interactions,
Robert Swanson, Anna F. Edlund, and Daphne Preuss 793

Integration of Adeno-Associated Virus (AAV) and
Recombinant AAV Vectors, Douglas M. McCarty,
Samuel M. Young Jr., and Richard J. Samulski 819

Indexes
Subject Index 847

Errata
An online log of corrections to Annual Review of Genetics chapters
may be found at http://genet.annualreviews.org/errata.shtml