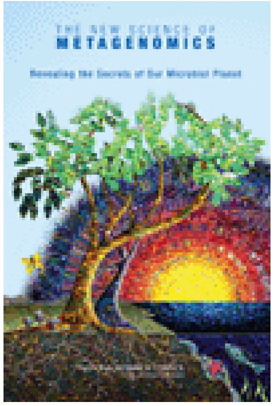


The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet

Committee on Metagenomics: Challenges and Functional Applications, National Research Council

ISBN: 0-309-10677-X, 170 pages, 6 x 9, (2007)

THE NEW SCIENCE OF METAGENOMICS

Revealing the Secrets of Our Microbial Planet

Committee on Metagenomics: Challenges and Functional Applications

Board on Life Sciences
Division on Earth and Life Studies

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Contract MCB—0544539 between the National Academy of Sciences and the National Science Foundation (NSF), Contract N01-OD-4-2139 between the National Academy of Sciences and the Department of Health and Human Services, National Institutes of Health (NIH), and Contract DE-AT01-05ER64072 between the National Academy of Sciences and the Department of Energy (DOE). The content of this publication does not necessarily reflect the views or policies of NIH, NSF, or DOE, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

International Standard Book Number-13: 978-0-309-10676-4
International Standard Book Number-10: 0-309-10676-1

Cover: Design by Francesca Moghari; artwork by Nicolle Rager Fuller (*www.sayo-art.com*).

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, NW, Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. A. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON METAGENOMICS: CHALLENGES AND FUNCTIONAL APPLICATIONS

JO HANDELSMAN *(Cochair)*, University of Wisconsin, Madison
JAMES TIEDJE *(Cochair)*, Michigan State University, East Lansing
LISA ALVAREZ-COHEN, University of California, Berkeley
MICHAEL ASHBURNER, University of Cambridge, United Kingdom
ISAAC K. O. CANN, University of Illinois, Urbana-Champaign
EDWARD F. DeLONG, Massachusetts Institute of Technology, Cambridge
W. FORD DOOLITTLE, Dalhousie University, Halifax, Nova Scotia, Canada
CLAIRE M. FRASER-LIGGETT, University of Maryland School of Medicine, Baltimore
ADAM GODZIK, Burnham Institute for Medical Research, La Jolla, CA
JEFFREY I. GORDON, Washington University School of Medicine, St. Louis, MO
MARGARET RILEY, University of Massachusetts, Amherst
MOLLY B. SCHMID, Keck Graduate Institute, Claremont, CA

Staff

ANN H. REID, Study Director
FRANCES E. SHARPLES, Director, Board on Life Sciences
ANNE F. JURKOWSKI, Senior Program Assistant
MERC FOX, Program Assistant
NORMAN GROSSBLATT, Senior Editor

BOARD ON LIFE SCIENCES

KEITH YAMAMOTO (*Chair*), University of California, San Francisco
ANN M. ARVIN, Stanford University School of Medicine, Stanford, CA
JEFFREY L. BENNETZEN, University of Georgia, Athens
RUTH BERKELMAN, Emory University, Atlanta, GA
DEBORAH BLUM, University of Wisconsin, Madison
R. ALTA CHARO, University of Wisconsin, Madison
JEFFREY L. DANGL, University of North Carolina, Chapel Hill
PAUL R. EHRLICH, Stanford University, Stanford, CA
MARK D. FITZSIMMONS, John D. and Catherine T. MacArthur Foundation, Chicago, IL
JO HANDELSMAN, University of Wisconsin, Madison
ED HARLOW, Harvard Medical School, Boston, MA
KENNETH H. KELLER, University of Minnesota, Minneapolis
RANDALL MURCH, Virginia Polytechnic Institute and State University, Alexandria
GREGORY A. PETSKO, Brandeis University, Waltham, MA
MURIEL E. POSTON, Skidmore College, Saratoga Springs, NY
JAMES REICHMAN, University of California, Santa Barbara
MARC T. TESSIER-LAVIGNE, Genentech, Inc., South San Francisco, CA
JAMES TIEDJE, Michigan State University, East Lansing
TERRY L. YATES, University of New Mexico, Albuquerque

Staff

FRANCES E. SHARPLES, Director
KERRY A. BRENNER, Senior Program Officer
ANN H. REID, Senior Program Officer
MARILEE K. SHELTON-DAVENPORT, Senior Program Officer
EVONNE P. Y. TANG, Senior Program Officer
ROBERT T. YUAN, Senior Program Officer
ADAM P. FAGEN, Program Officer
ANNA FARRAR, Financial Associate
ANNE F. JURKOWSKI, Senior Program Assistant
TOVA JACOBOVITS, Senior Program Assistant
MERC FOX, Program Assistant

Acknowledgments

This report has been reviewed in draft form by persons chosen for their diverse perspectives and technical expertise in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of the independent review is to provide candid and critical comments that will assist the institution in making the published report as sound as possible and to ensure that the report meets institutional standards of objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following for their review of the report:

Gary Anderson, Lawrence Berkeley National Laboratory, Berkeley, CA
Jeffrey Dangl, University of North Carolina, Chapel Hill
Julian E. Davies, University of British Columbia, Vancouver, BC, Canada
Jed Fuhrman, University of Southern California, Los Angeles
Dennis Mangan, University of Southern California School of Dentistry, Los Angeles
Victor Markowitz, Lawrence Berkeley National Laboratory, Berkeley, CA
Randall Murch, Virginia Polytechnic Institute and State University, Blacksburg
Norman R. Pace (NAS), University of Colorado, Boulder
David Relman, Stanford University, Stanford, CA
Edward Rubin, Lawrence Berkeley National Laboratory, Berkeley, CA
George Weinstock, Baylor College of Medicine, Houston, TX

Although the reviewers listed above have provided constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of the report was overseen by **John Wooley,** University of California, San Diego. Appointed by the National Research Council, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of the report rests entirely with the author committee and the institution.

The committee benefited from briefings provided by several speakers. At its second meeting, on May 2, 2006, the committee was briefed by: **Michael Gray** (by telephone), Professor and Department Head, Canada Research Chair in Genomics and Genome Evolution, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada; **Mitchell Sogin,** Senior Scientist and Director of the Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, The Woods Hole Biological Laboratory, Woods Hole, MA; and **Robert Edwards,** San Diego State University and Burnham Institute, San Diego, CA. At its third meeting, on July 27, 2006, the committee was briefed by: **David J. Lipman,** Director, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Rockville, MD; **Rolf Apweiler,** Head of Sequence Database Group, European Bioinformatics Institute, Cambridge, UK; **Victor Markowitz,** Head of Lawrence Berkeley National Lab's Biological Data Management and Technology Center, Berkeley, CA; **Paul Gilna,** Executive Director, CAMERA, San Diego, CA; and **Amaranth Gupta,** Associate Research Scientist, Director Advanced Query Processing Lab, San Diego Supercomputer Center, University of California, San Diego.

The committee extends heartfelt thanks to Ann Reid who served as Study Director for this report. The product reflects both Ann's attention to our charge and her ability to provoke us into addressing it thoroughly. Her outstanding editing contributed greatly to the clarity and logic of the report. We also thank Anne Jurkowski for her dedication to this report and its authors. Throughout the process, the committee relied on Anne's administrative prowess and her willingness to do whatever was necessary to get the report done or the committee on track. Anne's aesthetic intuition and visual acuity shaped the report as well as its derivative materials.

We thank Dr. Patrick Schloss for his assistance in building the metagenomics bibliography and Dr. Luke Moe, Snow Brook Peterson, and Dr. Ainslie Little for helpful discussion and Christina Matta for assuring historical accuracy.

Contents

APPENDIXES

Summary

THE DAWNING OF A NEW MICROBIAL AGE

Microbes run the world. It's that simple. Although we cannot usually see them, microbes are essential for every part of human life—indeed all life on Earth. Every process in the biosphere is touched by the seemingly endless capacity of microbes to transform the world around them. It is microbes that convert the key elements of life—carbon, nitrogen, oxygen, and sulfur—into forms accessible to all other living things. For example, although plants tend to get credit for photosynthesis, it is in fact microbes that contribute most of the photosynthetic capacity to the planet. All plants and animals have closely associated microbial communities that make necessary nutrients, metals, and vitamins available to their hosts. The billions of benign microbes that live in the human gut help us to digest food, break down toxins, and fight off disease-causing microbes. We also depend on microbes to clean up pollutants in the environment, such as oil and chemical spills. All these activities are carried out by complex microbial communities—intricate, balanced, and integrated entities that adapt swiftly and flexibly to environmental change. Some of the communities, like those in soil, may contain thousands of interdependent kinds of microbes. Microbial communities not only are key players in maintaining environmental stability and the health of individual plants and animals, they can also live in extreme environments, at temperatures, pressures, and pH levels in which no other organisms can survive. Microbes have developed countless strategies for survival, their genomes contain the directions for countless biochemical transformations, and their communities have adapted through countless individual generations and billions of years of environmental change. In addition to their essential activities throughout the biosphere, microbes have been the source of numerous technologies that have improved the human condition. They are used commercially to produce most of the antibiotics and many other drugs in clinical use, to remediate pollutants in soil and water, to enhance crop productivity, to produce biofuels, to ferment many human foods, and to provide unique signatures that form the basis of microbial detection in disease diagnosis and forensic analysis.

Historically, the study of microbes has predominantly focused on single species in pure laboratory culture, and so understanding of microbial communities lags behind understanding of their individual members. Only recently have the tools become available to study microbes in the complex communities where they actually live and thus to begin to understand what they are capable of and how they work. Traditional microbiological approaches have already shown how useful microbes can be; the new approach of metagenomics will greatly extend scientists' ability to discover and benefit from microbial capabilities.

The opportunity that stands before microbiologists today is akin to a reinvention of the microscope in the expanse of research questions it opens to investigation. Metagenomics provides a new way of examining the microbial world that not only will transform modern microbiology but has the potential to revolutionize understanding of the entire living world. In metagenomics, the power of genomic analysis is applied to entire communities of microbes, bypassing the need to isolate and culture individual bacterial community members. The new approach and its attendant technologies will bring to light the myriad capabilities of microbial communities that drive the planet's energy and nutrient cycles, maintain the health of its inhabitants, and shape the evolution of life. Metagenomics will generate knowledge of microbial interactions so that they can be harnessed to improve human health, food security, and energy production.

Metagenomics combines the power of genomics, bioinformatics, and systems biology. Operationally, it is novel in that it involves study of the genomes of many organisms simultaneously. It provides new access to the microbial world; the vast majority of microbes cannot be grown in the laboratory and therefore cannot be studied with the classical methods of microbiology. Although community ecology is not new to microbiology, the ability to bring to bear the power of genomics in the study of communities initiates an unparalled opportunity.

WHAT IS METAGENOMICS?

Like genomics, metagenomics is both a set of *research techniques*, comprising many related approaches and methods, and a *research field*. In Greek, *meta* means "transcendent." In its approach and methods, metagenomics overcomes the twin problems of the unculturability and genomic diversity of most microbes, the biggest roadblocks to advancement in clinical and environmental microbiology. *Meta* in the first sense means that this new science seeks to understand biology at the aggregate level, transcending the individual organism to focus on the genes in the community and how genes might influence each other's activities in serving collective functions. In the second sense, *meta* also recognizes the need to develop computational methods that maximize understanding of the genetic composition and activities of communities so complex that they can only be sampled, never completely characterized.

Metagenomics, still a very new science, has already produced a wealth of knowledge about the uncultured microbial world because of its radically new ways of doing microbiology. All metagenomics studies take the same first step: DNA is extracted directly from all the microbes living in a particular environment. The mixed sample of DNA can then be analyzed directly, or cloned into a form maintainable in laboratory bacteria, creating a library that contains the genomes of all the microbes found in that environment (see Box S-1). The library can then be studied in several ways, based primarily either on analyzing the nucleotide sequence of the cloned DNA or on determining what the cloned genes can do when they are expressed as proteins. It is important to recognize that the library is not organized into neat volumes, each containing the genome of one community member. Instead,

BOX S-1
Clones and Libraries

The word *clone* can have several different meanings in biology. In the context of this report, the word is used to describe a process whereby fragments of DNA isolated from a microbial community are inserted—or *cloned*—into circular pieces of DNA called plasmids. Laboratory bacteria can be manipulated to take up all the plasmids; when the bacteria subsequently divide, they replicate the plasmid along with their genomic DNA. When a large collection of plasmids containing all the DNA fragments from a given community is cloned into a bacterial culture, the resultant collection of bacteria is called a *library*—a living repository of all of the DNA from a microbial community.

it consists of millions of clones, each holding a random fragment of DNA. A metagenomics library is like thousands of jigsaw puzzles jumbled into a single box—putting the puzzles together again is one of this new science's great challenges. The metagenomics approach is now possible because of the availability of inexpensive, high-throughput DNA sequencing and the advanced computing capabilities needed to make sense of the millions of random sequences contained in the libraries.

Sequence-based metagenomics captures a massive amount of information on the microbial community under study. A study of the metagenome of the microbial inhabitants of the Sargasso Sea, for example, generated sequences of about a million genes and revealed whole classes of genes that were more diverse than could ever have been anticipated on the basis of studies of cultured organisms. At the other end of the spectrum, studies of a simple microbial community that lives in the extremely acidic water draining from metal mines demonstrated the potential of metagenomics to dissect detailed interactions among microbial-community members.

Metagenomics, however, is more than just large-scale sequencing. In *function-based metagenomics*, millions of random DNA fragments in a library are translated into proteins by bacteria that grow in the laboratory. Clones producing "foreign" proteins are then screened for various capabilities, such as vitamin production or antibiotic resistance. This enables researchers to access the tremendous genetic diversity in a microbial community without knowing anything about the underlying gene sequence, the structure of the desired protein, or the microbe of origin. New antibiotics and resistance mechanisms have already been discovered using function-based metagenomics.

STAGING THE FUTURE OF METAGENOMICS

The landscape of metagenomics is as expansive as microbiology itself. Microbial communities live virtually everywhere, and we are largely ignorant of their inhabitants and ecology; so there are literally millions of potential metagenomics projects. Each project would generate massive amounts of DNA sequence and functional data. To understand the potential of this new field and to determine how best to stage its development and encourage its success, several US government agencies—the National Science Foundation, five institutes of the National Institutes of Health, and the Department of Energy—asked the National Research Council to undertake an 18-month study of the emerging field of metagenomics. The Committee on Metagenomics: Challenges and Functional Applications was charged with describing the current state of the field and identifying obstacles that current researchers are facing. The committee was also asked to recommend the most promising directions for future metagenomics research and possible mechanisms for addressing infrastructure needs and improving communication and collaboration among groups studying different microbial communities. The committee met four times in 2006, including two short workshops: one on the implications of the massive amount of data generated by metagenomics and one on the questions of how and whether the nonbacterial members of environmental communities could be included in metagenomics studies (see Statement of Task, Appendix A).

Until recently, the complex microbial communities inhabiting nearly every environment and organism on Earth have essentially been invisible. With metagenomics, the astonishing genetic and metabolic diversity of the microbial world will be increasingly revealed. The practical applications of knowledge of these previously unseen realms of nature will be only part of the result. It is likely that as new biological strategies are brought to light, fundamental biological concepts will be affected. Basic ideas that organize biologists' understanding of the living world may need refinement in the face of greater understanding of how microbial communities function. New concepts of genomes, species, evolution, and ecosystem robustness will have effects beyond the specific field of microbiology. The questions that must be asked are "deep" ones, but answers will in all cases inform and guide the work of putting increased knowledge of microbial communities to practical use.

MAJOR ACADEMIC, GOVERNMENTAL, AND COMMERCIAL STAKEHOLDERS

There are many potentially beneficial collaborations among various academic disciplines in metagenomics projects, including atmospheric, ocean, soil, and water studies; geology; medicine; veterinary science; agricultural science; environmental; and bioengineering. It is, however, perhaps the field of biology that will be most affected by increasing knowledge of microbes. Virtually all biologists—whether they work on evolution, development, ecology, or cancer and whether they study yeasts, plants, corals, insects, birds, or mammals—will find that greater understanding of microbial communities has something to contribute to their research.

Because the applications are so broad, the government stakeholders in metagenomics are numerous. Metagenomic study of microbial communities has the potential to contribute to the missions of many government agencies. Fortunately, there is already a mechanism for 12 US government agencies with interests in microbiology to share information about their activities. The Microbe Project is an interagency working group formed in August 2000. The mission of the Microbe Project is "to maximize the opportunities offered by genome-enabled microbial science to benefit science and society, through coordinated interagency efforts to promote research, infrastructure development, education and outreach." The committee hopes that this existing mechanism will prove useful in ensuring that the development of the field of metagenomics occurs in the context of continuing communication and coordination among the interested government agencies. Besides the United States, metagenomics projects are also under way in the European Community, Canada, China, Brazil, Singapore, South Korea, and Japan, and including these and other international groups in planning for the field of metagenomics would be worthwhile.

DIFFICULTIES FACING CURRENT RESEARCHERS

The sequence-based metagenomics approach has already been applied to many environments, including the ocean, many soils, coral reefs, whale carcasses, thermal vents, and hot springs. The microbial communities associated with different organisms—including humans, termites, aphids, and worms—have been studied. Function-based metagenomics has been used to identify novel antibiotics and proteins involved in antibiotic resistance, vitamin production, and pollutant degradation. Much has been learned from the early efforts, and it is starting to become clear which steps in the process commonly present difficulties and obstacles.

The starting material for a metagenomics study is a mixture of DNA from a community of cells that may include bacterial, archaeal, eukaryotic, and viral species at different levels of diversity and abundance. In some projects, sample collection may be confounded because too little DNA is present or because compounds are present that interfere with DNA extraction. Contaminating DNA from a microbial community's host or from eukaryotic members of a community needs to be excluded from current metagenomic analyses because the amount of DNA they contain overwhelms both sequencing capacity and computational analysis. The quality and completeness of data obtained from metagenomic analysis of any community will be only as good as the procedures used for the extraction of DNA from an environmental sample.

Determining how best to sample a microbial community for metagenomics is also fraught with challenges. Change in habitats over time is one of the most interesting aspects of communities, and their responses to changing conditions are central to understanding community structure, function, and robustness. Similarly, understanding the role of host-associated microbial communities in host development and health requires not only sampling from the same host over time, but also understanding host-to-host variation. But habitat and host variability exacerbate the sampling conundrum. Over time, as biological and computational methods become more efficient, we will be able to draw more robust conclusions from more complex communities in more variable habitats. No matter the power of the methods now or in the future, it is essential to consider sampling issues and limitations at the beginning and throughout any metagenomic study of a complex community, and the sampling scheme must inform the interpretation of results.

Extracting maximal information from metagenomic libraries will continue to be challenging, primarily because of the massive size and complexity of the datasets. Determining the complete genome of any individual community member from pooled sequence data is extremely difficult and currently achievable only for very simple communities. The problem is exacerbated by the uneven abundance of members of microbial communities, which leads to sampling the most abundant organisms over and over and often missing the rare ones entirely. New technologies that allow much greater depth of sequencing or that remove redundant DNA would make it possible to detect important members that may be rare. Finally, improvements in bioinformatics tools, culturing techniques, and physical separation methods—with the generation of complete genome sequences for model microbes—will all make it easier to interpret the metagenome sequence data and in some cases to assemble whole genomes from metagenomic sequence data.

Function-driven metagenomics has already unearthed many proteins that would not have been recognized by their sequences alone. The potential for discovery is staggering but would greatly benefit from the development of new techniques and host organisms to allow genes from a wide variety of microbes to be expressed in the laboratory.

RECOMMENDATIONS

The opportunity afforded by metagenomics to study microbial communities in their natural state represents an endless frontier. Given the intense competition for science funding, some priority-setting is necessary to ensure that the most possible value is gained from early metagenomics investments. The diversity of habitats on Earth, the complexity of microbial communities, and the myriad functions governed by microbes suggest that highly productive metagenomics research will be possible in decentralized, *small-project settings*. However, no individual researcher is likely to have the capability and resources to achieve a comprehensive characterization of a complex microbial community. Therefore, there is also a substantial need for *medium-sized, collaborative projects* that involve multiple investigators. Both mechanisms of funding are tested and proven effective in advancing new fields of science. The mixture of single- and multi-investigator projects maximizes the diversity of scientific approaches, assures that many avenues of research are pursued simultaneously, presents an opportunity to study many habitats, and engages a broad community, thereby utilizing the creativity of many investigators. All these benefits are essential for the advancement of the field.

Metagenomics, however, differs from much of the science that precedes it in its complexity, multidisciplinarity, and in the magnitude of its unknowns. Its very nature departs from each of the fields—microbiology, ecology, and genomics—that fuse to form this new science. Consequently, metagenomics presents a number of conceptual and technical obstacles that limit the productivity of all metagenomics researchers. The committee believes that the needs of the metagenomics field are not entirely met by current funding mechanisms. Encouraged by the example of the human and other model organism genome projects, the committee believes that the best way to spur these advances is through a multi-scale approach. The committee recommends the establishment of a Global Metagenomics Initiative that includes a small number of *large-scale, comprehensive projects* that use metagenomics to understand model microbial communities, a larger number of middle-sized projects, and many small projects.

The committee believes that the field of metagenomics would be greatly advanced by the establishment of a few large, internationally coordinated projects with the goal of characterizing in great detail a small number of carefully chosen microbial communities. These large-scale model metagenomics projects would enable collaboration and coordination that are difficult to achieve in smaller projects. Large-scale projects could unite scientists of multiple disciplines around the study of a particular sample, habitat, function, or analytical challenge—an approach that is more likely to illuminate themes and advance technical approaches than would a disparate group of small projects by researchers with different goals and nonuniform methods. These large-scale projects would also serve as incubators for the development of novel technologies, analytical techniques, and community databases and would equip smaller-scale projects with the knowledge to design efficient sampling schemes, make informed choices about habitats to study, and identify fruitful strategies for identifying specific functions. Moreover, large projects would furnish the basis for developing a new conceptual framework for microbial ecology, as well as a new community of young scientists, that will guide the design of predictive models about community behavior.

Because the study of microbial communities has the potential to contribute to the missions of so many government agencies, it is likely that each will support a portfolio of small-scale metagenomics projects relevant to its particular mission. However, the metagenomics research community, which will include scientists working on a broad array of habitats and funded by many agencies, should be encouraged to work together to disseminate advances, agree on common standards, and develop guidelines on best practices in metagenomics that would be of use to all the funding agencies interested in supporting metagenomics research. This should include attention to bringing sample collection into alignment with international agreements and local values.

Information from metagenomics studies will be exploited fully only if appropriate data management and analysis methods are in place. Furthermore, metadata—information on the sampling method, sample treatment and data about the sampled habitat—are essential for the analysis of metagenomics sequence data. If metagenomics data are to be used to their fullest advantage, a metadata infrastructure is an urgent need. No metadata standard will be appropriate to all habitat types, but there should be close collaboration and coordination among the communities of scientists developing metadata standards.

In the genomic-sequencing community, many of the major species being studied have special community genomics databases, for example, *FlyBase* for the fruitfly *Drosophila,*[1] and *TAIR* for the model plant *Arabidopsis.*[2] This model—community databases organized to accommodate metagenomics data from particular environments or organisms—appears to be a promising approach to providing convenient access to the data of metagenomics projects.

One major challenge faced by metagenomics databases in contrast with "conventional" genomics databases will be the demand for community input into the annotation process. Annotation is the process of assigning functional, positional, and species-of-origin information to the genes in a database. In conventional genomics, primary responsibility for annotating data falls on the authors, and annotations are not often updated. In metagenomics projects, annotations will change as additional data (or metadata) are collected by other groups and an annotation database must be able to accept and integrate individual and large-scale (computational) annotations of metagenomic data continually. The need for dynamic and flexible annotation may make it essential that community metagenomics databases be provided sufficient resources to support ongoing, professional curation.

The analysis of genomics data is absolutely dependent on computer software. In general, grants for metagenomics projects will require an even higher percentage of funds for bioinformatic and statistical support than have genomics projects or than may be typical for other kinds of biological research. It is common for software developed for a particular project gradually to find widespread use in the community. Providing a mechanism whereby analytical tools that have proved their value to the community can be brought up to robust, engineered, documented form would be very

[1] *http://www.flybase.org/.*

[2] *http://www.arabidopsis.org/.*

worthwhile. This is a pipeline that is poorly supported by traditional grant-funding mechanisms.

The rise of genomics has been characterized by both technological and scientific innovations and by novel practices in data dissemination. In the early 1980s the scientific community in Europe and the United States established community archives for nucleic acid sequence data. These data immediately became accessible in a form suitable for computer analysis and were freely available, without impediment to all researchers, whether in academe or in industry. It is no exaggeration to state that without these publicly accessible databanks, the success of the Human Genome Project and similar genome projects would not have been possible. It is vital that the metagenomics community continue to adhere to the practice of publicly depositing, in a timely manner, all relevant data.

It should also be remembered that the more is known about microbes, the greater value metagenomics data will have. Thus, it is extremely important that basic microbiology research not be neglected, but instead be strengthened and deepened. Active communication between metagenomics researchers and members of other subdisciplines of microbiology and their representatives in funding agencies will help to guide the various fields in complementary directions.

TRAINING AND PUBLIC OUTREACH

Metagenomics presents some specific challenges for training experts and some global opportunities for educating the public about microbiology. The interdisciplinary nature of the science of metagenomics necessitates deployment of new training programs to encourage scientists to broaden their skills beyond those learned in their own disciplines. Graduate programs, intensive courses, fellowship programs, and sabbatical support are all mechanisms that can be used to develop investigators with the necessary configuration of skills and knowledge. Metagenomics also offers an opportunity to integrate public communication into graduate training. Each metagenomics project should design ways of teaching graduate students the principles of effective public outreach and then provide opportunities for them to use their new skills.

The dazzling power and opportunity of metagenomics as well as the "Big Science" nature of the large-sized projects in the Global Metagenomics Initiative will attract public interest in microbiology. The sense of delving into a truly unknown world, the potential for deriving human benefit from microbes, and the sheer power of microbes to influence just about every earthly function provide an irresistible draw for the public. Therefore, both large and small projects can be used as catalysts for teaching microbiology. Each large project should have a budget for developing materials that explain its scientific basis and implications in accessible and interesting ways. All metagenomics scientists should be encouraged to teach about their science in their local communities. In turn, these outreach efforts would provide a training ground for a new generation of scientists who are skilled in communicating science to the public.

1

Why *Meta*genomics?

Microbes run the world. It's that simple. Although we can't usually see them, microbes are essential for every part of human life—indeed all life on Earth. Every process in the biosphere is touched by the seemingly endless capacity of microbes to transform the world around them. The chemical cycles that convert the key elements of life—carbon, nitrogen, oxygen, and sulfur—into biologically accessible forms are largely directed by and dependent on microbes. All plants and animals have closely associated microbial communities that make necessary nutrients, metals, and vitamins available to their hosts. Through fermentation and other natural processes, microbes create or add value to many foods that are staples of the human diet. We depend on microbes to remediate toxins in the environment—both the ones that are produced naturally and the ones that are the byproducts of human activities, such as oil and chemical spills. The microbes associated with the human body in the intestine and mouth enable us to extract energy from food that we could not digest without them and protect us against disease-causing agents.

These functions are conducted within complex communities—intricate, balanced, and integrated entities that adapt swiftly and flexibly to environmental change. But historically, the study of microbes has focused on single species in pure culture, so understanding of these complex communities lags behind understanding of their individual members. We know enough, however, to confirm that microbes, as communities, are key players in maintaining environmental stability.

By making microbes visible, the invention of microscopes in the late 18th century made us aware of their existence. The development of laboratory cultivation methods in the middle 1800s taught us how a few microbes make their livings as individuals, and the molecular biology and genomics revolutions of the last half of the 20th century united this physiological knowledge with a thorough understanding of its underlying genetic basis. Thus, almost all knowledge about microbes is largely "laboratory knowledge," attained in the unusual and unnatural circumstances of growing them optimally in artificial media in pure culture without ecological context. The science of metagenomics, only a few years old, will make it possible to investigate microbes in their natural environments, the complex communities in which they normally live. It will bring about a transformation in biology, medicine, ecology, and biotechnology that may be as profound as that initiated by the invention of the microscope.

WHAT IS METAGENOMICS?

Like genomics itself, metagenomics is both a set of *research techniques*, comprising many related approaches and methods, and a *research field*. In Greek, *meta* means "transcendent." In its approaches and methods, metagenomics circumvents the unculturability and genomic diversity of most microbes, the biggest roadblocks to advances in clinical and environmental microbiology. *Meta* in the first context recognizes the need to develop computational methods that maximize understanding of the genetic composition and activities of communities so complex that they can only be sampled, never completely characterized. In the second sense, that of a research field, *meta* means that this new science seeks to understand biology at the aggregate level, transcending the individual organism to focus on the genes in the community and how genes might influence each other's activities in serving collective functions. Individual organisms remain the units of community activities, of course, and we anticipate that metagenomics will complement and stimulate research on individuals and their genomes. In the next decades, we expect that the top-down approach of metagenomics, the bottom-up approach of classical microbiology, and organism-level genomics will merge. We will understand communities, and the collection of communities that forms the biosphere, as a nested system of systems of which humans are a part and on which human survival depends. In some situations, it will be possible to apply the new understanding to problems of urgency and importance.

Metagenomics in either sense will probably never be circumscribed tightly by a definition, and it would be undesirable to attempt to so limit it now, but the term includes cultivation-independent genome-level characterization of communities or their members, high-throughput gene-level studies of communities with methods borrowed from genomics, and other "omics" studies (see Box 1-1), which are aimed at understanding transorganismal

BOX 1-1
The Other "Omics" Sciences

The term *genome* was first proposed by Hans Winkler, a professor of botany at the University of Hamburg, Germany, in 1920 (Winstead 2007). It was coined to describe the total hereditary material contained in an organism long before it was known that genetic information is encoded by DNA. Today *genome* is used to describe all the DNA present in a haploid set of chromosomes in eukaryotes, in a single chromosome in bacteria, or all the DNA or RNA in viruses. The suffix *ome* is derived from the Greek for "all" or "every." In the past several years, many related neologistic *omes* have come into use to describe related fields of study that encompass other aspects of large-scale biology. Some of them are:

- The *proteome*, the total set of proteins in an organism, tissue, or cell type; **proteomics** is the associated field of study.
- The *transcriptome*, the total set of RNAs found in an organism, tissue, or cell type.
- The *metabolome*, the entire complement of metabolites that are generated in an organism, tissue, or cell type.
- The *interactome*, the entire set of molecular interactions in an organism.

The list of "omes" and "omics" is growing longer as scientists develop new tools and approaches for carrying out large-scale studies of biological systems.

behaviors and the biosphere at the genomic level. Although in its current early implementation (and for the purposes of this report) metagenomics focuses on non-eukaryotic microbes (see Box 1-2), there is no doubt that its concepts and methods will ultimately transform all biology. In just this way has genomics, a science developed to aid the advancement of biomedicine and the understanding of our own species, transformed the science of all organisms and the application of that science in epidemiology, clinical microbiology, virology, agriculture, forestry, fisheries, biotechnology, microbial forensics, and many other fields.

In conceptualizing metagenomics, we might simply modify Leroy Hood's definition of *systems biology* as "the science of discovering, modeling, understanding and ultimately managing at the molecular level the dynamic relationships between the molecules that define living organisms" (Hood 2006). We need only replace the last word, *organisms,* with the phrase "*communities and the biosphere.*"

BOX 1-2
A Note on Terminology

What is a microbe? In practice, the term *microbe* is used to describe living things invisible to the human eye, that is, generally less than about 0.2 mm. The terms *microbe*, *microorganism, bacteria*, *germ,* and even *bug* are often used interchangeably by nonscientists to describe these small organisms. Microbiologists have specific names for the various microbes, which include Bacteria, Archaea and some members of the Eukarya. The first two groups (domains), although unlike in many ways, share a type of cellular organization known as prokaryotic. They lack membrane-enclosed organelles, such as mitochondria, chloroplasts and, most notably, a nucleus. The genomes of Bacteria and Archaea typically contain little non-coding DNA and range in size from 0.5 to 10 million base pairs. By contrast, members of life's third domain, Eukarya, which comprises animals, plants, fungi, algae, and protozoa have larger genomes with substantially more non-coding DNA. Some eukaryotes are also too small to be seen individually except under a microscope and thus have been traditionally studied by microbiologists. Included among these small eukaryotes are many fungi, such as baker's yeast and the human pathogen *Candida*, and many of the algae and protozoa (harmless paramecia, for instance, and the malaria parasite *Plasmodium*). Viruses, although arguably not alive, in that they can replicate only inside cells and have no metabolism or cell structure of their own, are also encompassed in the science of microbiology. In this report, we address primarily metagenomics projects that focus on Bacteria, Archaea and viruses. Because of their larger genomes, microbial eukaryotes have received less attention, a situation which should be remedied as sequencing becomes less expensive and bioinformatic methods become more powerful.

WHAT MICROBES CAN DO: FOUR EXAMPLES

We start with examples. There are countless ways in which microbes influence daily life. Earth is a biological entity as much as it is a physical one, and most of the vital biology, on which all life depends, is *micro*biology (see Box 1-2). But because microbes are individually invisible, we (even microbiologists) need to be reminded of our debt to them. Here are four of the thousands of reasons.

Microbes Modulate and Maintain the Atmosphere

Carbon is the most abundant chemical element in all living things, including humans (excluding the hydrogen and oxygen in the water, which makes up the bulk of our weight). Carbon dioxide (CO_2) in the atmosphere is the most abundant source of carbon on Earth, but in this form it is inaccessible to animals and most bacteria. Plants and some bacteria "fix" carbon through photosynthesis, a light-driven conversion of CO_2 to sugars that generates the oxygen that fuels all aerobic forms of life. Although plants tend to get most of the credit, bacteria are responsible for about half of the photosynthesis on Earth (Pedros-Alio 2006).

Ocean microbes, collectively present at billions of cells per liter, grow at rates of about one doubling per day in surface waters and are consumed at about the same rate (Whitman et al. 1998). The organisms that carry out photosynthesis turn over rapidly in the ocean as well, on the average about once per week. Net primary productivity in the global ocean is estimated to fix 45-50 billion tons of CO_2 per year (Falkowski et al. 1998). Chemical transformations mediated by marine microbes play a critical role in global biogeochemical cycles (see Figure 1-1). The collective metabolism of marine microbial communities has global effects on fluxes of energy and matter in the sea, on the composition of Earth's atmosphere, and on global

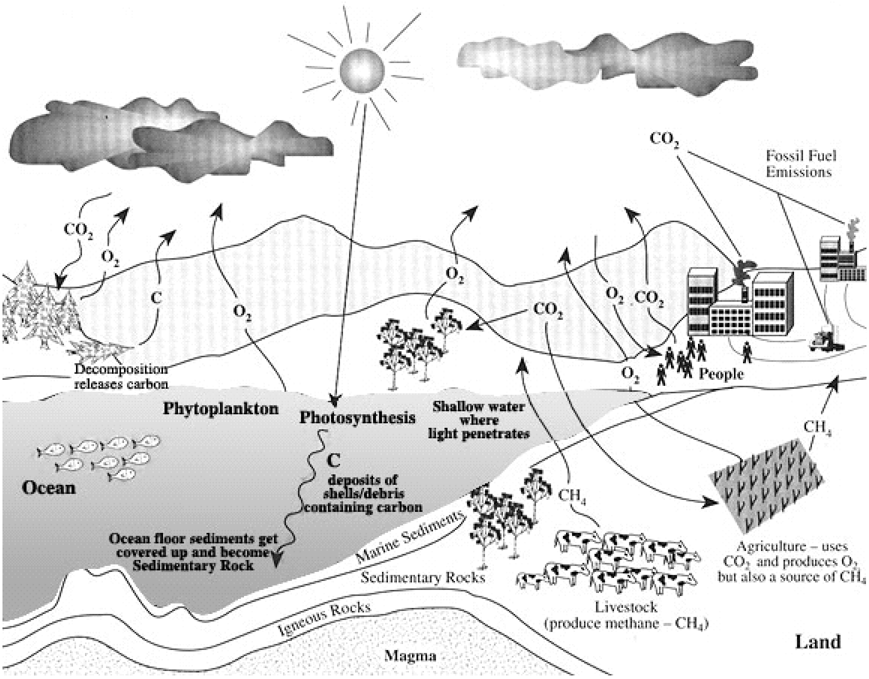

CO_2
O_2
C
O_2
Decomposition releases carbon
O_2
CO_2
O_2 CO_2
CO_2
Fossil Fuel Emissions
People
O_2
Phytoplankton
Photosynthesis
Shallow water where light penetrates
CH_4
C
deposits of shells/debris containing carbon
Ocean
CH_4
Ocean floor sediments get covered up and become Sedimentary Rock
Marine Sediments
Sedimentary Rocks
Agriculture – uses CO_2 and produces O_2 but also a source of CH_4
Livestock (produce methane – CH_4)
Igneous Rocks
Magma
Land

FIGURE 1-1 The global carbon cycle. SOURCE: *http://www.bigelow.org/foodweb/carbon_cycle.jpg.*

climate. In essence, the combined activities of microbial communities affect the chemistry of the entire ocean and maintain the habitability of the entire planet. Hidden within the population dynamics of these complex communities are fundamental lessons of environmental response and sensing, species and community interactions, gene regulation, and genomic plasticity and evolution. Microbes are the stewards of Earth's biosphere and are Nature's biosensors par excellence.

Perhaps most obviously today, the living oceans play a critical role in the global carbon cycle (Falkowski et al. 1998). The coupling of the upper ocean and the atmosphere results in higher concentrations of dissolved CO_2 in surface seawater than in the rest of the ocean. Much of the elevated carbon input can move through the action of the ocean's "biological pump," which depends on microbial communities in the surface water that transform inorganic CO_2 into organic carbon. The organic carbon can either be respired and recycled back to the upper ocean-atmosphere system or sink out of the surface water and be sequestered in the deep ocean. Complex microbial community interactions help to regulate the proportion of recycled versus sequestered carbon. The structure of the phytoplankton community, the rates at which phytoplankton are attacked and destroyed by viruses, and the capacity of other microbes to turn organic carbon back into CO_2 all influence the fate of carbon, and the ability of the ocean to act as a source of, or a sink for, CO_2. CO_2 is a very important greenhouse gas, so photosynthetic bacteria serve the planet in two ways: they convert carbon into biologically accessible forms and they remove CO_2 from the atmosphere, thereby mitigating some of the anthropogenic release of CO_2 and other greenhouse gases.

Microbes Keep Us Healthy

It should come as no surprise that in the microbe-dominated biosphere, close relationships between microbes and animals are an ancient theme. Humans are no exception. The numbers are staggering. The microbes that reside on the surface of the human body alone outnumber human cells by about a factor of 10. The genomes of members of our indigenous microbial communities (the human metagenome) contain thousands of times more genes than the human genome (Gill et al. 2006). Microbial communities also inhabit the human mouth, skin, and respiratory and female reproductive tracts. The compositions of these communities change over time and, for some body sites, like the oral cavity, there is already evidence that certain community compositions are associated with periodontal disease. Understanding how microbial community structure affects health and disease may contribute to better diagnosis, prevention, and treatment of disease. The vast majority of these microbial partners live in the intestine, where a diverse community of microbes, 10 to 100 trillion in number, perform functions that humans have not had to evolve, including the extraction of calories from otherwise indigestible components of our diet and the synthesis of essential vitamins and amino acids. The complex communities of microbes that dwell in the human gut shape key aspects of postnatal life, such as the development of the immune system, and influence important aspects of adult physiology, including energy balance. Gut microbes serve their host by functioning as a key interface with the environment; for example, they defend us from encroachment by pathogens that cause infectious diarrhea, and they detoxify potentially harmful chemicals that we ingest (intentionally or unintentionally). In light of the crisis in management of infectious pathogens due to emergence of antibiotic resistance, we would be well served to understand the role of microbial communities in protecting us from infectious agents. Our microbes are master physiological chemists: identifying the chemical entities that they have learned to manufacture and characterizing the functions of human genes and gene products that they manipulate should lead to valuable additions to our 21st-century medicine cabinet (pharmacopeia).

Microbes Support Plant Growth and Suppress Plant Disease

The microbial communities on and around plants play a central role in the health and productivity of crops. The most complex of these communities reside in the soil, which is a composite of mineral and organic materials teeming with bacteria and archaea. Some functions of these microbes are well known. Some bacteria fix atmospheric nitrogen, converting it from dinitrogen gas—a form unusable by plants and animals—to ammonia, which is readily used. Other soil microbes recycle nutrients from decaying plants and animals, and others convert elements, such as iron and manganese, to forms that can be used for plant nutrition. Soil microbial communities determine whether plants will become infected by pathogens. A lingering mystery is the “suppressive soil” phenomenon (Mazzola 2004). In some soils, plants stay healthy even when pathogens are present at high density; when the soil is sterilized, the disease suppression disappears, suggesting a biological basis of the phenomenon. However, in only very few cases has a single microbe isolated from a soil been able to duplicate the suppression. After decades of wrestling with the enigma of suppressive soils, plant pathologists have concluded that in many cases a complex community is responsible for the suppressive activity, which is hugely beneficial to agriculture. No organism has been found to provide the same effect in isolation, because the community members modify each other’s behavior.

Microbes Clean Up Fuel Leaks

There are hundreds of thousands of underground storage tanks in this country, most of which are used for storing gasoline. In fact, almost every corner gasoline station in the United States uses three or more of these tanks to dispense regular, premium, and super-premium versions of gasoline. The sad truth about these underground tanks is that the vast majority of them are already leaking or will leak and send gasoline into the subsurface, where it has the potential to contaminate the groundwater. Given the ubiquity and magnitude of the gasoline leaks and the fact that 50% of the US population relies on groundwater as a drinking-water source, one must wonder how it is that we are not all drinking water contaminated with gasoline!

The answer is that we are being protected by the omnipresent and vastly adaptable subsurface microbial community (Mazzola 2004). As gasoline is released into the subsurface, relatively dormant members of the microbial community are triggered to become active and biodegrade the gasoline constituents. Gasoline is composed of thousands of organic chemicals and a variety of microbes containing complementary metabolic systems are required to degrade them all. Furthermore, because there is too little of any single electron acceptor in the subsurface to react with all the electron donors of gasoline, different bacteria with different respiratory capabilities are required to complete the gasoline remediation. For example, when oxygen is depleted in the groundwater in the vicinity of a gasoline spill, bacteria that can respire nitrate take over, followed by bacteria that respire iron, manganese, sulfate, and, eventually, CO_2. This complicated community of microbes works together in a self-organized pattern triggered by the movement of the leaking gasoline until the contaminants have been transformed into harmless CO_2 and water. The microbial community then becomes dormant again, awaiting the next influx of substrate (either natural or anthropogenic) to return to activity.

INVISIBLE COMMUNITIES: GLOBAL IMPACT

Modulating the atmosphere, keeping humans and plants healthy, and cleaning up leaking gasoline are just a few examples of the many things that microbial communities can do. The combined activities of microbial communities shape the face of the biosphere on a global scale. The power of these communities lies hidden in the metabolic versatility of their component species that, acting together, regulate the vast majority of matter and energy transformations on Earth. In a loose analogy, the entire biosphere can be imagined as a sort of “superorganism.” Its many systems for the recycling of carbon, oxygen, nitrogen, and phosphorus can be compared with the organs of the human body working in unison to facilitate circulation, nutrient acquisition, respiration, waste processing, and so forth. Unquestionably, humans depend on these global geochemical cycles, and microbes are vital players in the cycles' operation and stability. Microbes can "eat" rocks, "breathe" metals, transform the inorganic to the organic, and crack the toughest of chemical compounds. They achieve these amazing feats in a sort of microbial "bucket brigade"—each microbe performs its own task, and its end product becomes the starting fuel for its neighbor. For complex transformations, no microbe can do it alone—it takes a community. For example, no microbial species is capable of completely oxidizing ammonia to nitrate, but teams of microbes do it efficiently. One microbial group oxidizes ammonia to nitrite, and its waste becomes the fuel for another species that transforms nitrite to nitrate, completing the "bucket brigade." Virtually all elemental cycles—including the generation, consumption and flux of greenhouse gases (or, as noted above, the remediation of spilled gasoline)—involve similar sorts of microbial collaborations that are tightly regulated and coupled through microbial community interactions. So the bucket brigades are themselves interconnected laterally—an interwoven web of chains. In this way, microbial communities play essential roles in the transformations of energy and matter, producing the air we breathe and shaping the biosphere and climate that we enjoy on Earth today.

Larger organisms play key roles, too, of course: about half of all carbon is fixed and half of all oxygen produced by trees, grasses, and other macroscopic plant life. But these larger organisms also depend on microbes; for example, plants depend on the nitrogen fixation carried out by symbiotic microbes in the roots of legumes and other plants that form symbiotic associations. Humans might survive in a world lacking other macroscopic life forms, but without microbes all higher plants and animals, including humans, would die. Not only can many individual systems—for example, the human gut or such processes as the bioremediation of toxic hydrocarbons—be seen to be the tasks of complex and dynamic microbial communities, but these communities are themselves constituents of even larger systems, predominantly microbial, that collectively make up the biggest and most complex functioning system we know: the biosphere. Whatever the causes, extent, and consequences of the global climate change now upon us, the biosphere's response to the changes—and human survival—will depend on its microbes and their activities.

We live in a time of unprecedented and dramatic global change, in which the effects of human activities challenge the ability of natural ecosystems to buffer them. The industrial revolution marked the beginning of rapid environmental transformation. For example, until the early 20th century, all nitrogen entering the biosphere was produced from atmospheric nitrogen by microbes, providing the organic nitrogen required for new plant growth. In the early 1900s, the Haber-Bosch process was invented to perform the same job to produce vast amounts of nitrogenous plant fertilizer from atmospheric nitrogen; this industry now produces more organic nitrogen than all biological processes combined (Socolow 1999). Another obvious and dramatic change in the global environment is the enormous amount of CO_2 released by the burning of fossil fuels, previously stored as relatively inert reservoirs deep in Earth. Present concentrations of atmospheric CO_2 are higher than they have been in 420,000 years and, given current trajectories, will continue to rise dramatically (Petit 1999).

Understanding the dynamic role of microbial communities in this rapidly changing environment is a critical and currently unmet challenge. How resilient are microbial communities in the face of such rapid global change? Can microbial communities, versatile as they are, help to buffer and mediate key elemental cycles now undergoing rapid shifts? Can changes in microbial communities serve as sensors and early-alarm systems of environmental perturbation? To what extent can we "manage" microbial communities to modulate the effects of human activities on natural elemental cycles sensibly and deliberately? Never before have such questions had such urgency.

UNDERSTANDING MICROBIAL COMMUNITIES

Given that the microbial collective profoundly influences geochemical and greenhouse-gas cycles, as well as climate and environmental change, it is relevant to ask how well we understand microbial communities. In the past, it was difficult to study microbes in their own environments; microbiologists studied individual species one by one in the laboratory. It now appears that many microbes function in nature as multicellular, often multi-species, entities, sometimes even physically connected (as in biofilms) and often metabolically connected.

The Limits of Pure Culture

Even into the 19th century, some scientists believed that microbes were generated spontaneously from nonliving matter or from other organisms. Establishing that such tiny entities were organisms that belonged to definable, fixed species was difficult. Fixity of species was especially important in theories of disease causation; fixed species were essential if a single bacterial species was to be held responsible for a single infectious disease. Agriculturalists and botanists had long suspected that some sort of unseen organisms were associated with plant disease; in 1726, for example, the association farmers saw between barberry rust and wheat rust led the Connecticut colonial legislature to ban the bushes (Campbell et al. 1999). Over a century later, the German botanist Anton de Bary demonstrated the correlation between the life cycle of *Phytophthora infestans* and the disease cycle of late blight of potato. In a series of experiments conducted in the late 1850s and early 1860s, he built on the previous work of J. Speerschneider and Marie-Anne Libert and established that *P. infestans* was indeed the cause of the disease (Matta 2007).

Demonstrating that microorganisms were not spontaneously generated and had distinct species was fundamental to bacteriology as well. Robert Koch published his description of the life cycle of *Bacillus anthracis* (the cause of anthrax) in 1876 and then published a series of papers in which he established an experimental method for confirming the specific causes of various infectious diseases. In an 1884 paper on tuberculosis, he outlined his four "postulates" for proof of microbial causation: an organism must be found in all cases of the disease but not in healthy hosts, the organism must be isolated from the host and grown in pure culture, reintroduction of the organism from such cultures must cause disease in healthy hosts, and the organism must again be isolatable from such infected hosts (Munch 2003). That rigorous approach, particularly the emphasis on pure cultures (a culture that contains organisms of only one type) set the standards for microbiology as a whole. By the middle of the 20th century, even with "environmental microbes" (the vast majority of harmless and beneficial bacteria, archaea and microbial eukaryotes), pure cultures became a gold standard for experimentation and the basis of almost all recent knowledge of medical bacteriology, biochemistry, and molecular biology.

In the pure-culture paradigm, the presence of multiple species in the same culture medium means "contamination," and species whose growth requires metabolic products of other species are impossible to detect, study, or even name. Not surprisingly, microbes that grow well as single cells suspended in a liquid medium and that can easily form discrete colonies on Petri plates became the model for much of modern biology. Indeed, many microbiologists came to view the "planktonic" state as the natural condition of microbes—complex communities and slimy biofilms being somehow an aberration and unworthy of serious scientific attention. On the contrary, it is now becoming clear that many microbes live in communities whose members interact and communicate in complex ways. Microbial communities often interact through the medium (water or soil) in which they grow, exchanging nutrients, biochemical products, and chemical signals without direct cell-to-cell contact. Some grow on surfaces (on suspended particles, on the walls of pipes, on teeth) where they are in physical contact with others of their own kind and with other species. Biofilms, which are aggregates of microbial cells embedded in an extracellular polysaccharide matrix, exhibit a great diversity of complex structures. The composition of such communities is far from accidental. Many microbes have evolved to grow together in surface communities and many of their collective activities, whether vital to the biosphere or detrimental to human health, reflect the physical structure and division of labor within the communities.

The study of microbes in culture will continue to be important, but it falls short of telling us about environmental processes, biofilms, microbial bucket brigades of energy and matter flux, and the future trajectory of biogeochemical cycles. Understanding microbial communities will require that the traditional techniques of pure culture be supplemented with new approaches.

The Genomics Promise

One approach that has contributed greatly to understanding all organisms is genomics—learning about the evolution and capabilities of organisms by deciphering the sequence of their DNA. Genomics has also greatly advanced microbiology, but, like pure culture, traditional genomics is limited in its ability to elucidate the dynamics of microbial communities.

The precipitous decline in the cost of gene sequencing, spurred in part by the Human Genome Project, has made it possible to generate genomic sequences for a great variety of organisms. The first microbial genome sequenced, that of the pathogen *Haemophilus influenzae*, was published in 1995 (Fleischmann et al. 1995). Microbial genome sequences have since appeared at an exponentially increasing rate: the genome sequences of 399 bacteria, 29 archaea, and almost 30 eukaryotic microbes are publicly available at the time of this writing. Pathogenic bacteria and eukaryotes—such as the causative agents of plague, anthrax, tuberculosis, Lyme disease, candidiasis, malaria, and sleeping sickness—have received much attention. But many nonpathogenic archaea and bacteria have also been sequenced, including such beneficial organisms as several species of *Prochlorococcus* and *Synechococcus*, major producers of oxygen in the ocean; *Dehalococcoides ethenogenes*, effective in the bioremediation of soils contaminated with chlorinated hydrocarbons*; Lactobacillus acidophilus,* used in making yogurt; *Bradyrhizobium japonicum*, a nitrogen-fixing symbiont of soybeans; and *Saccharomyces cerevisiae* (baker's yeast).[1]

When attention turned to sequencing the genomes of microbes, the preference for working in pure culture was reinforced. No one knew how difficult it might be to sequence an entire genome, but it was obvious that assembly (using a computer to put the sequenced fragments together in complete genomes) would be vastly more complicated if the pieces belonged to several different organisms (see Box 1-3). Until recently, all microbial genome sequences were determined from pure cultures. But in the last few years, more than a dozen microbes that can be physically separated

[1] See *http://www.ncbi.nlm.nih.gov/Genomes/* for more information.

BOX 1-3
Blueprints for the Living World: Genes, Genomes, and Genomic Sequences

Genes are made of DNA, and the exact sequence of the four canonical DNA bases (designated A, T, C, and G) in any gene specifies the product (usually a protein) that it encodes. In bacteria and archaea, genes are about 1,000 base pairs long. These microbes have 500-10,000 genes, usually arrayed on a single circular DNA molecule (a *chromosome*), some 600,000-12 million base pairs long (there is some space between genes for regulatory signals). Eukaryotic microbes typically have more and longer genes and multiple chromosomes. Together, all the genes in a microbe's chromosome or chromosomes and any in accessory genetic elements, such as plasmids, make up its *genome*.

For complete genome sequencing, the whole genome shotgun approach has proved effective. All the DNA from a pure culture is fragmented randomly into pieces of one to a few thousand base pairs. Fragments totaling some 6-10 times the genome's length are sequenced so that overlaps between them can be used to establish the order of the fragments in the intact genome and verify the accuracy of the sequencing. This step, called *assembly*, is computationally intensive. So is the next step, *annotation*, which is the prediction of gene boundaries, regulatory regions, and the properties and function of the proteins (or sometimes RNAs) that the genes encode. Annotation usually involves finding a similar gene sequence for which a function has already been determined in another organism, although at present typically one-third of the genes in any newly sequenced microbe will not have any obvious similarity to genes with known or proposed functions. Finally, the data are released to a public data repository, such as GenBank, maintained by the National Center for Biotechnology Information (National Library of Medicine) in Bethesda, Maryland.

from other major sources of DNA or that greatly predominate where they are found in nature have also been sequenced. *Treponema pallidum* and *Mycobacterium leprae* (which cause syphilis and leprosy, respectively) are among the former, and two species predominant among acid-mine drainage site biofilms (*Ferroplasma acidarmanus* and a species of *Leptospirillum*) are examples of the latter. Sequencing such physically purified or environmentally concentrated (and thus naturally "pure") microbes crosses the boundary between genomics and metagenomics as far as methods are concerned.

Soon, there will be thousands of sequenced microbial genomes. If all microbial species were culturable and if such species were easily defined and limited in number (even a number in the tens of thousands), the ultimate goal of microbial genomics might be to determine all these genome sequences once the per-genome cost fell far enough. Then the *meta* in *meta*genomics might parallel its use in meta-analysis and mean bringing together individual databases in search of a common set of truths about nature. But not all species are culturable, few are easily defined at the genomic level, and indeed the number of different genomes in nature turns out to be uncountably large. We discuss these problems in turn.

WHY GENOMICS IS NOT ENOUGH

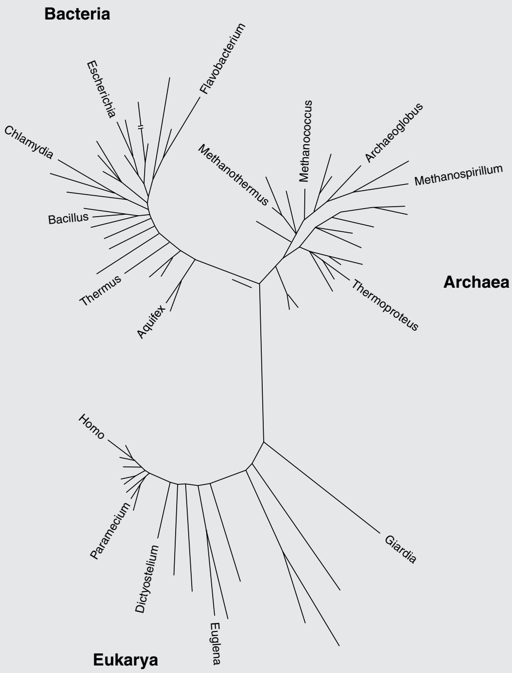
Most Microbes Cannot Be Cultured

In 1985, Staley and Konopka reviewed data on scientists' ability to bring microbes from the environment into laboratory cultivation. The "great plate-count anomaly" they identified was this: the vast majority of microbial cells that can be seen in a microscope and shown to be living with various staining procedures cannot be induced to produce colonies on Petri plates or cultures in test tubes. It is estimated that only 0.1-1.0% of the living bacteria present in soils can be cultured under standard conditions; the culturable fraction of bacteria from aquatic environments is ten to a thousand times lower still. The application of genomics-inspired moderate- to high-throughput nutrient screening methods and nontraditional approaches to monitoring growth responses will no doubt bring many recalcitrant organisms into culture. Indeed, two recent successes are the cultivation (and genome sequencing) of *Pelagibacter ubique*, a bacterium representative of one of the most common microbial phylogenetic groups found in the open ocean, and the isolation of several acidobacteria, the most abundant organisms in soil (Sait et al. 2002; Field et al. 1997; Martinez and Rodriguez-Valera 2000; Brown and Fuhrman 2005; Rappe et al. 2002). Both successes depended on the nontraditional molecular (rRNA-based) method discussed below for monitoring growth. But the fraction of organisms cultivatable in isolation will likely always be low, and for most the reason will be that, for growth, it takes a community. Culturing always favors the recovery of organisms that are best able to thrive under laboratory conditions (colloquially "lab weeds"), not necessarily the dominant or most influential organisms in the environment.

Given the evidence that many microbes resist being cultured, *culture-independent* methods for identifying and enumerating microbes in the environment have come to play a larger and larger role over the last several decades. Predominant among them is ribosomal RNA (rRNA) *phylotyping*, a powerful technique—indeed, an independent research paradigm—developed by Pace and his colleagues (Pace 1997). This method is based on the enormous database of rRNA gene sequences (more than 200,000) that have been collected for the purpose of reconstructing the universal Tree of Life (see Box 1-4). By determining the sequence of an organism's rRNA genes,

BOX 1-4
Ribosomal RNA and the Tree of Life

Ribosomal RNAs (rRNAs) are essential structural and functional components of ribosomes, the cellular factories on which proteins are made according to the information encoded in DNA. Information from DNA is transmitted to the ribosomes through an intermediate, "messenger RNA." All organisms have rRNAs similar enough to each other that they can be recognized as the "same molecule" but different enough that the differences are a good measure of evolutionary distance. The same is true of the genes encoding the rRNAs, on which the phylotyping method is based. Thus, two closely related organisms (for example, the benign *Escherichia coli* laboratory strain K12 and its sometimes lethal diarrhea-producing relative, strain O157:H7) will have almost identical rRNA gene sequences, whereas two remotely related species (such as *E. coli K12* and an archaean, such as *Picrophilus torridus*) will have very different sequences. With enough sequences and suitably sophisticated computational tools, relationships between organisms measured by the differences in the sequences in their rRNA genes can be converted to a tree-like picture of their evolutionary histories. The widely accepted rRNA-based three-domain Tree of Life, which we owe to the pioneering work of Carl Woese and the heroic efforts of his many colleagues and students, is shown below (adapted from Pace [1997]; SOURCE: Hazen 2005).



one can position it on the appropriate branch of the Tree of Life and infer that its biology and ecology are likely to be similar to those of its closest relatives, the nearest branches on the tree. An organism does not have to be culturable to determine its phylotype. The *polymerase chain reaction (PCR)* allows rRNA (or other) genes to be detected and copied directly from environmental samples, then cloned and sequenced. If the environmental sample contains many types of organisms, there will be many different rRNA sequences, the diversity of which will be a measure of the complexity of the community and which, in the context of the Tree, will tell us "who is there." Phylotyping has revolutionized the field of microbial ecology, and hundreds of environments—from dry Antarctic valleys to deep-sea hydrothermal vents ("black smokers") to sewage-treatment plants and methane-producing reactors—have been studied in this way. Very often, new lineages whose rRNA gene sequences are little like anything that has been cultured are discovered. Indeed, the majority of the 50-plus major divisions of Bacteria that have been delineated through their rRNA genes do not yet have any cultured representatives. Community rRNA sequencing and phylogenetic analysis, in itself, is not considered metagenomics (because it focuses on only one gene, not entire genomes), but it can be a useful preliminary step in a metagenomics project because it provides a phylogenetic assessment of the diversity of a community.

Microbial Diversity and Variation Have No Limits

When genetic information from macroscopic organisms (animals or plants) is organized into phylogenetic trees to examine how they are related to one another, one can assume that all the individuals of a given species have virtually identical genomes. For example, the genomes of humans differ from one another by only 0.1%. In contrast, microbial phylotyping coupled with genome sequencing has shown that even if culturability ceased to be a problem, diversity will always be a challenge; indeed, it is a greater challenge than might have been imagined. Hundreds of thousands or even millions would be too low an estimate of the number of genomes that would have to be sequenced in any kind of whole-genome-based metagenomics program. This is due in part to the large numbers of species of microbes in most environments. But also it reflects genomic diversity *within* what scientists had been calling species. Almost all phylotyping surveys of almost all environments yield not a single phylotype for each likely microbial species contributor to community dynamics, but dozens or hundreds of very close but unquestionably nonidentical phylotypes that form *microdiverse clusters* (see Figure 1-2). In addition to differing slightly in the sequences of the marker genes used in phylotyping, these organisms—supposedly members

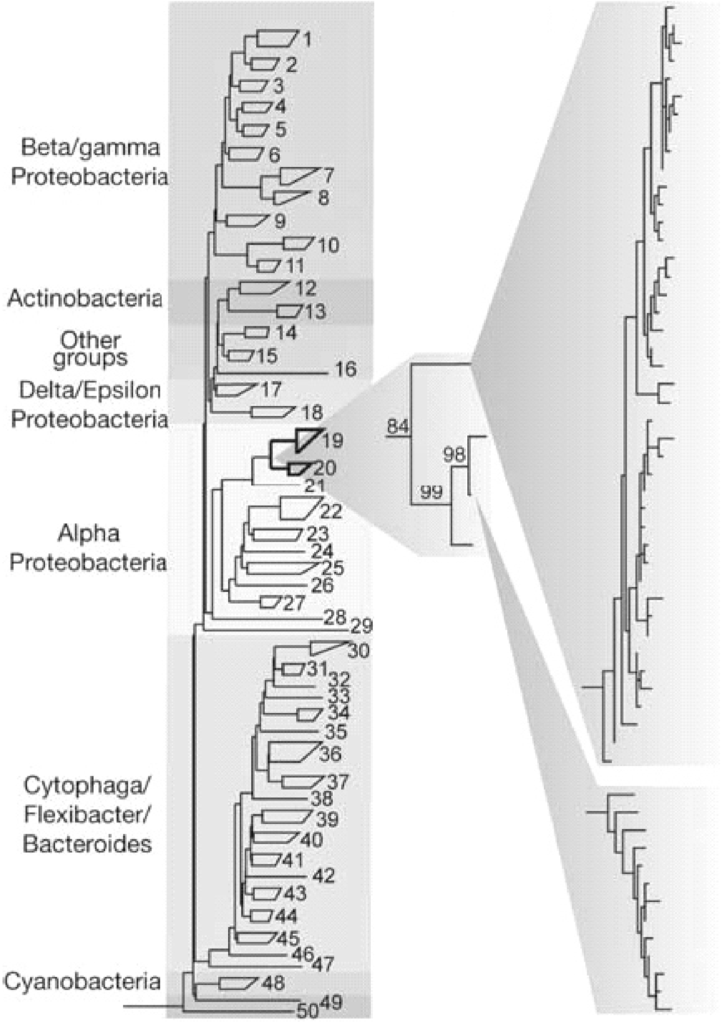


FIGURE 1-2 Microdiversity of environmental 16S rRNA sequences. PCR-amplified 16S rRNA gene sequences from an environmental DNA sample, showing a pattern of clustering often interpreted to be indicative of species divisions. Reprinted by permission from Macmillan Publishers Ltd: *Nature* 430:551, copyright 2004.

of the same species—differ substantially (by up to 30%) in the genes that their genomes contain.

In a recent survey of the diversity and genome sizes (gene contents) of strains of the environmental bacterium *Vibrio splendidus*, Polz and coworkers documented astonishing diversity (up to 25% difference in apparent gene content) in a small area (a single site off a beach in Massachusetts) (Thompson et al. 2005). They were forced to conclude that "this group consists of at least a thousand distinct genotypes, each occurring at extremely low environmental concentrations (on average less than one cell per milliliter)." All this means that no single collection of genes can be said to be "the *V. splendidus* genome" or "the *E. coli* genome" or indeed the genome of almost any designated bacterial or archaeal species, and no amount of complete genome sequencing will be enough to map the genomic diversity of the microbial world. Perhaps the biggest challenge faced by microbial ecology as a science today is to understand the ecological significance of such phylotypic microdiversity and genomic variability, and this challenge cannot be met with a traditional genomic approach.

METAGENOMICS OFFERS A WAY FORWARD

The pure culture paradigm has not only limited what microbiologists have studied; it has also limited how they think about microbes. Microbes have been studied as sovereign entities and examined only for their responses to the simple chemicals that can be added to their media. We know little about their behavior as partners in the strategic alliances that are metabolic consortia, such as the consortia that decontaminate drinking water or that make up the complex structured biofilms that keep dental hygienists busy. The invisible members of a microbial community can differ vastly in their biochemical activities and interactions, not only between species but also within species. Phylotyping gives some reliable information about "Who is there?" but because of within-species genomic diversity, only imperfect guesses as to "What are they doing?" Metagenomic methods, which will be discussed later, go a long way toward answering the second question. In the end, it may be possible to view ecosystems themselves as biological units with their own genetic repertoires and to sidestep consideration of individual species. Then, both "Who is there?" and "What are they doing?" could be replaced with "What is being done by the community?"

Such understanding can be achieved only with methods that go beyond the pure-culture and single-whole-genome approaches that have dominated microbial genomics. We must move directly to the genes, to defining environments by the potential and realized biochemical and geochemical activities of the genes that are there, and the complex patterns of interactions within and between cells that regulate their responses to changes in their

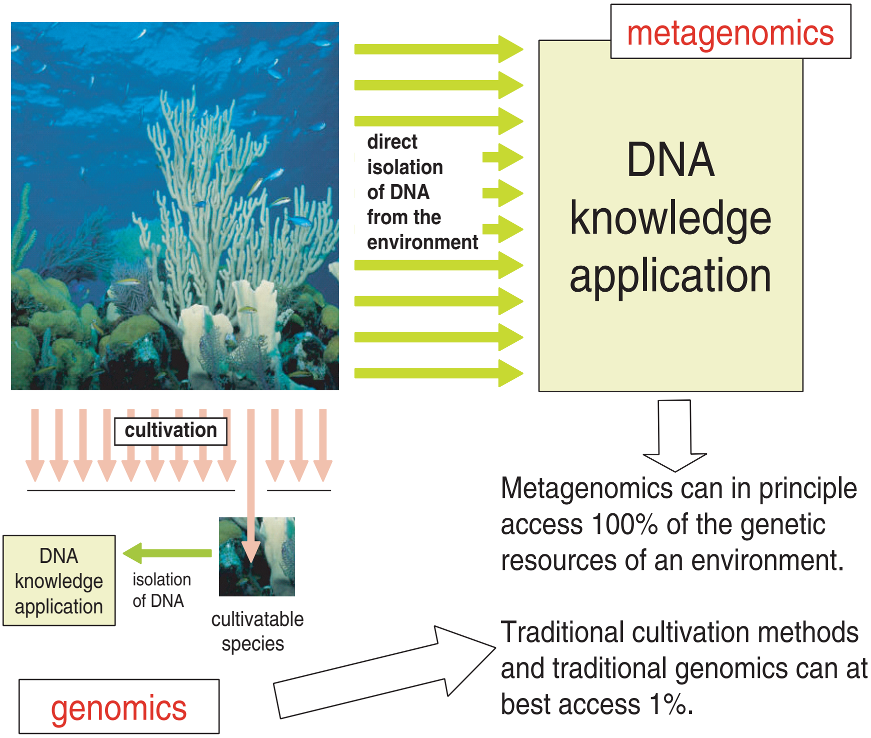

metagenomics
DNA knowledge application
direct isolation of DNA from the environment
cultivation
DNA knowledge application
isolation of DNA
cultivatable species
genomics
Metagenomics can in principle access 100% of the genetic resources of an environment.
Traditional cultivation methods and traditional genomics can at best access 1%.

FIGURE 1-3 How metagenomics differs from microbial genomics. Image provided by W. Ford Doolittle.

physical and biological surroundings. We must do this while recognizing that—except in restricted environments and specialized consortia with limited numbers of genetically homogeneous constituents—we will be dealing with enormous amounts of data that will represent an incomplete sampling of the genetic diversity present. In short, we must adopt the methods of metagenomics (see Figure 1-3).

Pioneering steps in this direction, which illustrate the character and range of such methods, are described later in this report; but in metagenomics, necessity not only is the mother of invention but will be the grandmother of a paradigm shift. It will refocus us one level higher in the biological hierarchy (molecules, cells, organisms, species, populations, *communities*, the biosphere). It will shift the emphasis from individuals to interactions, from parts to processes—a change that would be timely and highly desirable even if it were not also technologically necessary.

Not coincidentally, this shift will parallel the new focus of organismal genomics on interactions between cellular components and how they are coordinated within the complex systems called organisms. This new focus is called *systems biology*. Metagenomics will be the systems biology of the biosphere.

Metagenomics provides a means for studying microbial communities on their own "turf." Complex ecological interactions—including lateral gene transfer, phage-host dynamics, and metabolic complementation—can now be studied with the lens of metagenomics. Community composition, function, and dynamics can now be measured and modeled in the environment with universal microbial-community genomic approaches. These approaches have the potential to provide new insights into the environmentally relevant microbial communities and activities that control matter and energy flux on Earth. With such information in hand, it will become possible to interpret the interplay between natural cycles and human activities that together shape the future of the planet.

METAGENOMICS CAN CONTRIBUTE TO ADVANCES IN MANY FIELDS

Metagenomics offers a means of solving practical problems facing humanity. Cracking the secrets of some of Earth's countless microbial communities will reveal ways to meet myriad challenges in biomedicine, agriculture, and environmental stewardship. These are among the most important potential contributions:

- **Earth Sciences:** the development of genome-based microbial ecosystem models to describe and predict global environmental processes, change, and sustainability.
- **Life Sciences:** the advancement of new theory and predictive capabilities in community-based microbial biology, ecology, and evolution.
- **Biomedical Sciences:** the description, on a global scale, of the role of the human microbiome (the collective genome of our symbionts) in health and disease in individuals and populations, and the development of novel diagnostic and treatment strategies based on this knowledge.
- **Bioenergy:** the development of microbial systems and processes for new bioenergy resources that will be more economical and environmentally sustainable and less vulnerable to disruption by world politics.
- **Bioremediation:** the development of tools for monitoring environmental damage at all levels (from climate change to leaking gas-storage tanks) and microbe-based (green) methods for restoring healthy ecosystems.

- **Biotechnology:** the identification and exploitation of the remarkably versatile and diverse biosynthetic capacities of microbial communities to generate beneficial industrial, food, and health products.
- **Agriculture:** the development of more effective and comprehensive methods for early detection of threats to food production (crop and animal diseases) and food safety (monitoring and early detection of dangerous microbial contaminants) and the development of management practices that maximize the beneficial attributes of microbial communities in and around domestic plants and animals.
- **Biodefense and Microbial Forensics:** the development of more effective vaccines and therapeutics against potential bioterror agents, the deployment of genomic biosensors to monitor microbial ecosystems for known and potential pathogens, and the ability to precisely identify and characterize microbes that have played a role in war, terrorism, and crime events, thus contributing to discovering the source of the microbes and the party responsible for their use.

2

A New Light on Biology

At the dawn of the 21st century, scientific understanding of microbes is uneven—sophisticated in some ways, primitive in others. Decades of genetic, molecular, and biochemical dissection of microbial life have revealed the detailed structure and inner workings of several bacteria and archaea. Although there is much more to learn even about model organisms, such as *E. coli,* many individual pathways for nutrient cycling, gene regulation, and reproduction are understood at a satisfying level of precision. But these processes in the majority of microbes remain unknown and knowledge of the evolution and ecology of microbial communities lags far behind cellular microbiology. Basic ideas that organize biologists' understanding of the living world may need refinement in the face of greater understanding of community function. New concepts of genomes, species, evolution, and ecosystem robustness will have effects beyond the specific field of microbiology. The questions that must be asked are "deep" ones, but answers will in all cases inform and guide the work of putting increased knowledge of microbial communities to practical use. This chapter focuses on some of the more fundamental questions raised by the study of microbial communities that can be addressed through metagenomics.

WHAT IS A GENOME?

When the first microbial genome sequence was completed in 1995, informed opinion was that a few dozen more genomes, chosen to be appropriately diverse, would exhaust the range of variability in how genes could be assembled to make microbes. But as the number of fully sequenced genomes approaches 500, there seems to be no end to the ways in which genes can be arranged—on linear chromosomes or circular, on one or many, tightly compacted or (in many eukaryotic microbes) separated by "junk" DNA 10 times their length. The number of genes in the genome of a free-living bacterium ranges from 500 to 10,000 or more; the largest bacterial genomes are more than twice the size of the smallest eukaryotic genomes. In contrast, the genomes of many parasitic or symbiotic microbes are highly reduced, with not nearly enough genes to support them independently of their hosts.

Even within a single clonal culture established from a single cell, there will probably be multiple forms of the genome. Many bacteria, especially pathogens, have elaborate mechanisms for rearranging their genes. The mechanisms serve as mutational switches, ensuring that as the microbe's environment changes due to shifts in chemical, physical, or biological conditions, there will be variants in the population that can flourish. For example, no matter what defenses a host's immune system mounts against the pathogen, there will be some resistant variants in the pathogen's population. Variability is also achieved by exchange between genomes: recombination (similar to the genetic exchange that occurs in sexual reproduction) constantly reshuffles the variants (alleles) of genes in the population, generating new adaptive combinations. Plasmids, small and often self-transmissible packets of genes that encode environmentally relevant functions, are rife.

It is, however, the pervasiveness of lateral gene transfer *between* species that most profoundly challenges the notion that a single bacterial species has a single genome. Several natural processes—transport by viruses, bacterial "mating," and the direct uptake of DNA from the environment—carry genetic information from one species to another. These processes are regulated and evolutionarily preserved; they are turned on when they are most likely to result in gene transfer, and genes that must function together are often transferred together, forming genomic "islands" (pathogenicity islands, symbiosis islands, or biodegradation islands). Genomic plasticity is an evolved strategy. No single sequence can be said to be *the* genome sequence of the bacterial species *Escherichia coli.* And the variations are decidedly not like the trivial differences that account for much of the 0.1% sequence variation among humans. When genomes of multiple strains of the same species (like *E. coli* K12, O157:H7, and another dozen now available) are compared, they differ up to 25% in genome size and in the number and kinds of genes they carry. Indeed, the genes that are shared by all sequenced *E. coli* strains amount to less than 40% of the genes present in the species as a whole (see Figure 2-1). Microbial genomicists have started to think in terms of microbial "species genomes" or *pangenomes*, which comprise a core of genes shared by all strains of a species and a library,

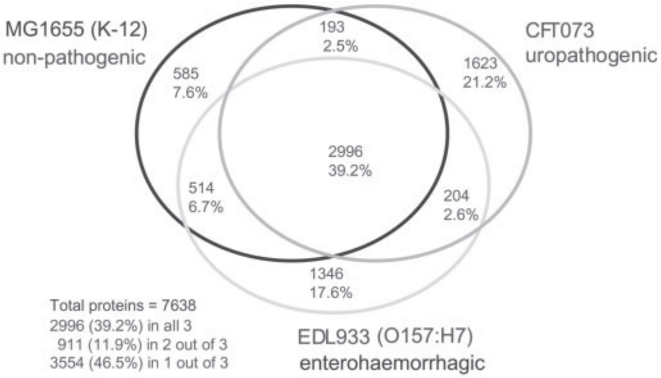


FIGURE 2-1 A Venn diagram showing strain-specific and shared genes for the genomes of three *E. coli* strains. SOURCE: Welch et al. (2002) PNAS 99: 17020-24. Copyright 2002 National Academy of Sciences, U.S.A.

perhaps much larger, of auxiliary genes that are in some members of the species but not all (Fraser-Liggett 2005).

Probing the extent of genomic diversity is an enormous task best carried out by metagenomic approaches. With suitable experimental and computational methods, environmental gene sequences can be binned (statistically grouped) into provisional pangenomes on the basis of compositional characteristics and site of recovery. As more data accumulate, the definition of what constitutes a microbial genome will be better informed and underlying principles governing genomic plasticity in microbes may emerge. If having a more flexible and dynamic genome structure is a fundamental life-strategy difference between bacteria and archaea, on the one hand, and eukaryotes, on the other, what are its advantages and limits? Can understanding the phenomenon help to explain the emergence of multicellular organisms that have more fixed genomes?

WHAT IS A SPECIES?

In many eukaryotes (especially animals and higher plants), a species contains individuals that can breed and produce fertile offspring together. There is no equivalent definition for bacteria and archaea, because they usually reproduce by binary fission, which does not require sexual compatibility. Moreover, as discussed above, their sexual lives are not limited by relatedness: bacteria and archaea transfer DNA to organisms that are distantly related, even in different phyla, thereby providing no indication of an orderly classification. Traditional bacterial classification has (partly for that reason) been based on cell appearance, motility, and physiology. The field of bacteriology developed around these classification methods, and most of the names used for common bacteria today are relics of that system. For many purposes, such traditional classification remains useful. But its connection to genomic information is problematic: as discussed above, the very nature of the microbial genome—a fluid entity subject to invasion by large segments of alien DNA—is the problem.

Although molecular methods, in particular the use of rRNA gene sequences (see Box 1-4, page 26), have transformed bacterial classification in the field and in the clinical laboratory, they have not provided entirely satisfying or conclusive answers. Bacteria or archaea that carry similar or even identical rRNA genes can have deeply diverging genomic structure and content because of horizontal gene transfer. Conventions that enable a standard for assigning species names have been established, such as the rule that no examples of the same species should vary in their 16S rRNA gene sequence by more than 3% (Gevers et al. 2005). Those conventions are convenient and informative, but they are controversial, conceptually ungrounded, and thus somewhat arbitrary. Some "species" defined by the above convention contain highly similar members, for example, and others exhibit remarkable differences in gene content and important features of phenotype (see Box 2-1).

Such concerns are not purely academic. What does it mean, for example, for regulatory standards to require that products be free of particular species (for example, *E. coli* in food products) if the definition of *species* is uncertain? Medical diagnosis of an infectious disease usually requires determining the species of the pathogen. How closely related to a particular pathogen does an organism need to be to be considered that pathogen? Registering products that contain live microbes, such as those used to control pests and pathogens of crops, requires naming the organisms in the product. How can entities that cannot be clearly defined be dealt with in patent applications? What if the biocontrol agent is in the same species as a human pathogen but does not appear to be pathogenic? On what basis can the organisms be deemed sufficiently safe (see Box 2-1)? How can microbial evidence be used effectively in a court proceeding or policy decision if the microbes cannot be fully and precisely identified and linked to a source with confidence and certainty?

Metagenomics will steer microbiology closer to a more realistic, flexible, and predictive classification scheme and a more rational (if possibly more pluralistic) species concept. The power of such an approach is that it will be predicated on a far more extensive dataset than the one that has informed past attempts at classification and will make use of new computational strategies to cluster and split groups of organisms in ways not predictable with today's limited information. The definition of species is

BOX 2-1
What's in a Name?

Names carry important legal and regulatory implications. A glaring example is the "*Bacillus cereus* group," which contains *B. cereus*, *B. thuringiensis*, and *B. anthracis* (Priest et al. 2004). The first species in the group contains strains that induce food poisoning, strains that prevent plant disease, and others that produce unusual antibiotics. Some *B. cereus* strains can perform more than one of these activities. *B. thuringiensis* is the most widely used bioinsecticide in the world; it produces crystal proteins that are toxic to certain insect pests. Some *B. thuringiensis* strains also produce the toxins associated with food poisoning in humans caused by *B. cereus*, but this was not recognized when *B. thuringiensis* was first registered for use in 1961. *B. anthracis* is the causal agent of anthrax, a disease deadly to both cattle and people. Modern phylogeny and genomics indicate that the three species probably make up a single species with a few dramatic phenotypic differences due to a small number of genes. If we call them all by the same name, how will regulatory agencies, the courts, and the public respond to the idea of spraying trees and crops with an organism that has the same name as the anthrax pathogen? If the similarities of the species had been recognized earlier, might the production of the human enterotoxins by *B. thuringiensis* have been recognized sooner and prevented registration of this bacterium? What we call bacteria does make a difference.

less important than intelligent and flexible application of species concepts so that estimates of species richness or organisms' names imply a similar degree of relatedness across groups and can be of genuine utility in the development of ecological theory and environmental applications.

WHAT IS THE ROLE OF MICROBES IN MAINTAINING THE HEALTH OF THEIR HOSTS?

Most multicellular organisms have a closely associated microbial community that provides a variety of functions, from digestion to defense against pathogens. All plants and animals, including humans, can be considered superorganisms composed of many species—animal, bacterial, archaeal, and viral. Historically, the study of physiology has not focused on these host-associated microbial communities; metagenomics offers an opportunity to understand their physiological role and evolutionary significance.

Using the human as an example, the human "metagenome" might be considered an amalgamation of the genes contained in the *Homo sapiens* genome and in the microbial communities that colonize the body inside and out. The organisms within these communities are collectively known as the human "microbiome." The metagenome of these communities encodes physiological traits that humans have not had to evolve, including the ability to harvest nutrients and energy from food that would otherwise be lost because we lack the necessary digestive enzymes (see Figure 2-2). Recent studies suggest that the gut microbiome may play a role in obesity (Turnbaugh et al. 2006). Without understanding the inhabitants of the human microbiome and the mutualistic human-microbial interactions that it supports, our portrait of human biology will remain incomplete.

Metagenomics will enable us to address a number of fundamental questions about ourselves. Is there an identifiable core microbiome shared by all humans? How is each individual's microbiome selected? What is the role of host genotype? Should differences in each individual's microbiome be viewed, with the immune and nervous systems, as features of our biology that are profoundly affected by individual environmental exposures? How is the human microbiome evolving (within and between individuals) over different time scales as a function of changing diets, lifestyle, and biosphere? What are the functional correlates of diversity in the membership of a microbiome, and how does this diversity affect the robustness of a community and the host's ability to respond to various physiological or pathophysiological states? How redundant or how modular are the contributions of individual microbial constituents to community function and to host biology? How should such constituents be defined given that mutualists, like pathogens, do not have a single genomic structure but rather have pangenomes with various degrees of openness to acquisition of genes from other microbes? How can this knowledge be used to manipulate microbial communities to optimize their performance in a person or in a population? Most obviously, how does the microbiome affect health, and vice versa? When we know more, previously unrecognized microbial involvement with disease states will be uncovered. Many host physiological states with primary genetic or biochemical causation will affect the microbiome in ways that may aid in diagnosis. Of course, these questions do not apply only to humans—study of host-associated microbial communities will contribute to understanding of the physiology of all organisms.

HOW DIVERSE IS LIFE?

"How many?" is a fundamentally human question. How many people are there on Earth? How many grains of sand on the beach? How many planets in the universe? Defining the dimensions and limits of an entity is often the first quest of scientific discovery. But as suggested above, the question "How many species of bacteria are there on Earth?" is far from simple. Metagenomics is likely to contribute to a more flexible and useful

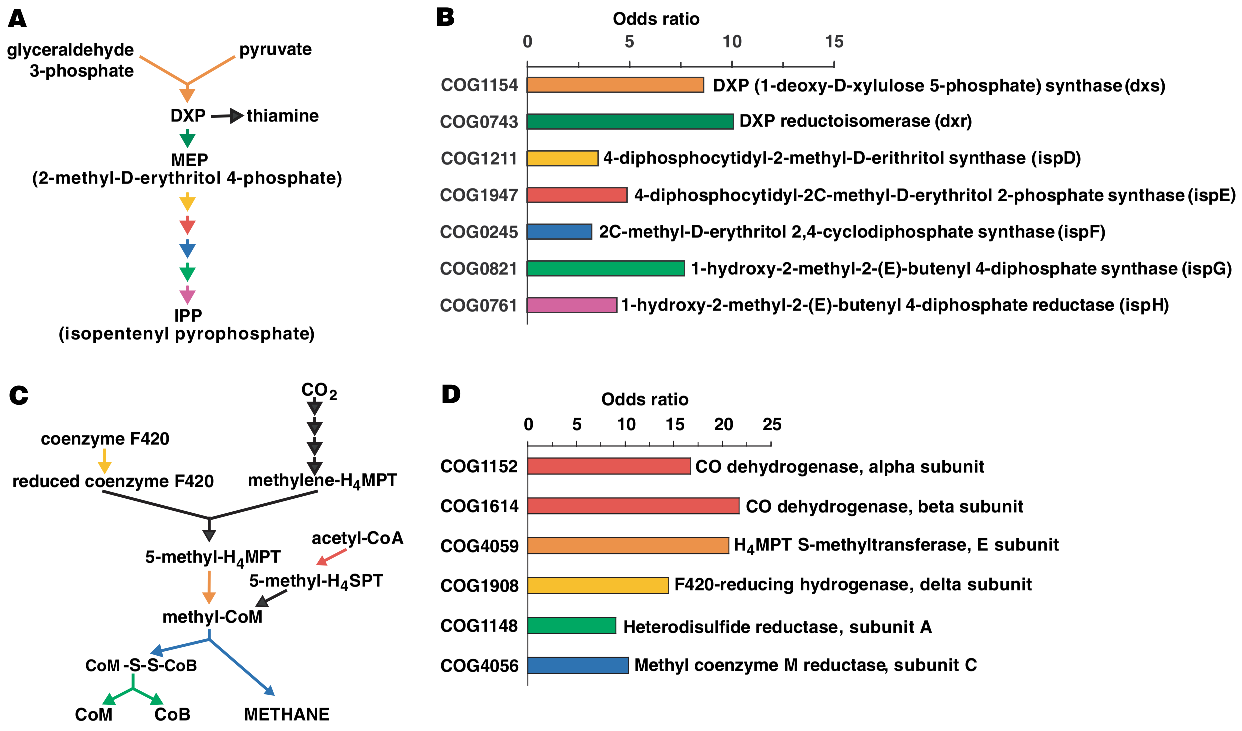


FIGURE 2-2 Some of the metabolism on our own distal gut (isoprenoid synthesis and methane formation) that is the responsibility of genes encoded in the genomes of our microbiota. Results of the metagenomic survey of Gill et al. “Odds ratios” indicate extent to which genes for the indicated biochemical reaction are over-represented in the gut microbiota. SOURCE: Gill et al. (2006) *Science* 312: 1355-9. Reprinted with permission from AAAS.

definition of microbial species, and no matter how microbial "species" are defined, metagenomics will aid in describing the extent of microbial diversity. In some cases, it may be important to know how many different species—however defined—are present. For other purposes, it may be that what is important is the overall genetic content of an environment, not the number of species it contains. The degree to which genetic diversity and species composition affect the capabilities and stability of a microbial community is another fundamental conceptual question to which metagenomics can contribute.

Soil, for example, is estimated to contain a few hundred species per gram on the basis of culturing, a few thousand per gram on the basis of 16S rRNA gene sequencing and mathematical modeling, and a few million per gram on the basis of DNA-DNA reassociation kinetics and kinetic modeling (Schloss and Handelsman 2006). Although with current tools and knowledge the number of "species" in soil cannot be counted with any confidence, with molecular methods soil's complexity can be compared with that of other habitats at the gene level.

Molecular and, in particular, high-throughput metagenomic methods that sample all classes of gene, not just phylogenetic markers like 16S rRNA, will guide many aspects of basic and applied microbiology (see Figure 2-3). They will inform the design of experiments that are directed toward capturing or describing diversity. For example, knowing the extent of diversity in a particular habitat will aid in estimating sample sizes required to draw robust conclusions, and knowing the diversity in a biological grouping may determine the choice of habitat for particular types of study. Searches for antibiotics might focus on environments that contain a high diversity of Actinobacteria, the phylum that has yielded the most antibiotic-producing cultured organisms. Metagenomics will provide information about microbial diversity that is intrinsically linked to information about functional attributes of members of microbial communities and that will aid researchers in making strategic choices.

At a more conceptual level, metagenomics will enable us to begin to explore the reasons for the observed genetic diversity. Do communities with extensive *genetic* diversity also have more *functional* diversity? Do they respond differently to environmental change? Does genetic diversity correlate with environmental stability or resource availability, or is it a matter of chance and history? Such questions will be addressable through metagenomic approaches.

HOW DO MICROBIAL COMMUNITIES WORK?

Generally speaking, biological community interactions are as important to evolutionary and ecological processes as is the surrounding physical and

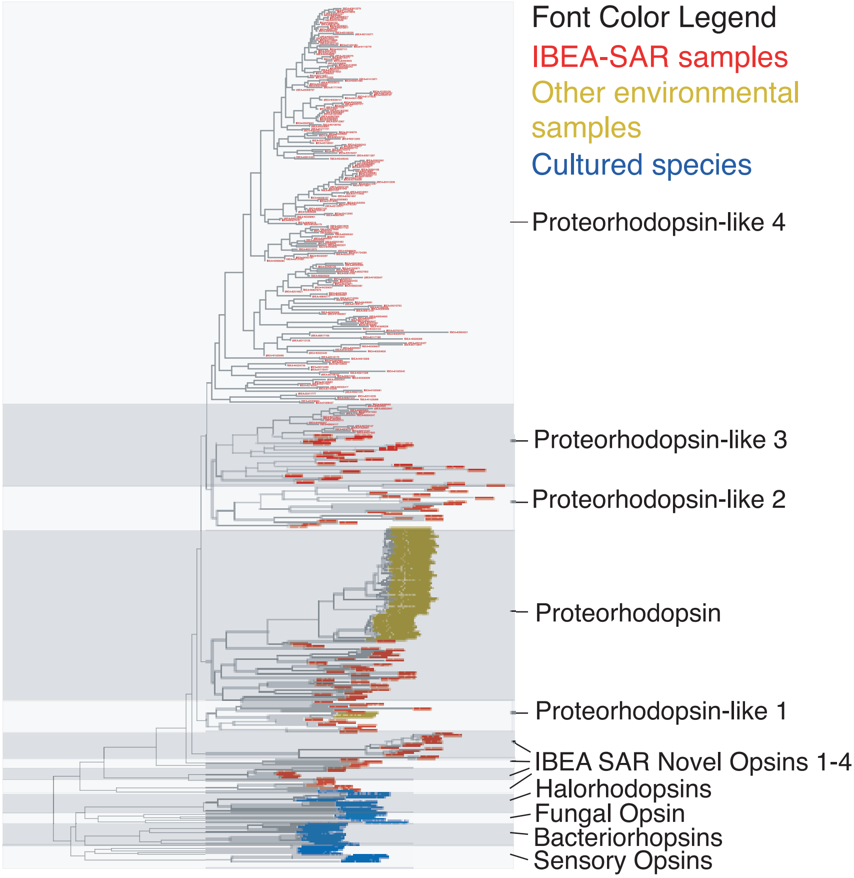


FIGURE 2-3 Diversity of proteorhodopsin sequences in the Sargasso Sea. Proteorhodopsin, a light-driven proton pump, is encoded in many bacterial genomes. Proteorhodopsin genes show a distribution characteristic of lateral gene transfer. As with many other genes, environmental surveys reveal vast and hitherto unexpected numbers and variations of gene sequences. The Sargasso Sea project revealed not only many new gene sequences but whole new classes of sequences. Only those indicated in blue were previously known from cultured organisms. SOURCE: Venter et al. (2004) *Science* 304: 66-74. Reprinted with permission from AAAS.

chemical environment, and community interactions can shape the properties of the surrounding environment as much as the environment shapes the community. A good example is the influence of microbial communities on the oxidation-reduction potential in their surrounding environment (producing anoxic conditions in sediments, for instance), which in turn shapes the spatial organization of the associated communities. Similarly, community interactions mold biological properties. Lateral gene transfer, cell-cell communication, metabolic complementarity, trophic interactions, interspecies competition and predation, and biogeochemical cycling are all results of community processes.

Macroscopic plant and animal communities have been studied for a long time, but parallel description of natural microbial communities has been more challenging. The typical approach has been to dissect microbial communities into their various components, often by isolating individual microbial strains in pure culture. Even if they were readily cultivated, it is impossible with standard cultivation methods to characterize the hundreds and thousands of microbial strains that make up any single community. Furthermore, the vast intraspecies diversity and variability typically seen in natural microbial populations is usually not examined with most culture-based approaches, which focus on clonal populations. And microbial interactions that in part define community structure and functional relevance (competition, predation, lateral gene exchange, metabolic complementation and syntrophy, and allelopathy) are not readily modeled in most laboratory settings.

Metagenomics promises a new view of microbial-community genomic structure, functional properties, and potential interactions. By mitigating many analytical constraints and using high-resolution community-wide genomic information, we can describe the composition, function, and emergent properties of integrated microbial communities more accurately. The effects of microbial community activities span enormous ranges of time and space, from nanoscale molecular interactions to global-scale biogeochemical cycles. Metagenomic data provide a foundational microbial-community database from which the properties and dynamics of biological organization at a variety of levels (genes, genomes, proteins, metabolic pathways, cells, organisms, populations, and communities) can be inferred. For example, metagenomic datasets provide information about the structure, type, and organization of genes in individual genomes, about within-population allelic variability, and about the patterns of organismal and gene occurrence. Reading metagenomic information may make it possible to infer emergent properties and dynamics of interacting genomes and the relationship of the interactions to the functionality of natural microbial ecosystems.

As metagenomic techniques begin to be applied in natural settings, a number of fundamental questions about microbial communities can be better addressed. For example, there is evidence that microbial communities may self-assemble in nonrandom ways (Crump and Hobbie 2005; Fuhrman et al. 2006). What are the genetic and physiological drivers? If genetic instructions in part encode species interactions and community assembly, what are the "assembly rules"? Do founder effects influence the nature of spatially structured microbial ecosystems? What are the differences in community organization and interspecies interactions in biofilms vs planktonic communities? Is substantial functional redundancy built into all microbial communities? Does intrapopulation allelic variation have functional significance? All these questions—whose answers have important consequences for understanding evolutionary, ecological, and environmental processes—can potentially be addressed by metagenomics.

HOW DO MICROBIAL COMMUNITIES REACT TO CHANGE?

Robustness is defined as resistance to and recovery from change. It has implications for fundamental understanding of communities and for their management to bring about beneficial outcomes. Many communities maintain their structures across space and time despite continually changing biological and physical conditions, whereas others are more easily destabilized by external disturbance, introduction of new members, or internal processes. Little is known, however, about the basis of robustness or vulnerability to change. Communities are dynamic assemblages governed by dependence and antagonisms among the members, so robustness is likely to depend, in part, on interactions among the members. But this is surmised, not evidence-based; in few communities are the factors that influence robustness known (see Box 2-2).

Community robustness is critical for stability of natural ecosystems. When communities are vulnerable to *change*, vagaries in weather, seismic activity, or human activity can lead to collapse of a community or the ecosystem in which it resides. The practical implications of community vulnerability are enormous. Managed communities that perform services for humans, such as those in soil or sewage sludge, need to be predictable and steady in their behavior (see Box 2-3). Agricultural productivity depends on the soil community's protecting plants from disease, transforming minerals, and decomposing organic matter. Similarly, human health is shaped by the robustness of the microbial shield that prevents pathogen invasion of the skin, mouth, and gut.

Metagenomics will add tremendously to our currently limited understanding of robustness by providing large datasets that will facilitate the identification of functional traits or groups of traits that are correlated with robustness or vulnerability to change. Comparing metagenomic data-

BOX 2-2
Robustness and the Gut Community

The human gut illustrates some key applications of the principle of community robustness. In some situations, robustness of the gut community is desirable. When a person takes antibiotics that alter the gut community, robustness is depended on to return the community to its original structure and function. The inverse of robustness is vulnerability to invasion, and the success of gut pathogens depends on their invasive ability; this highlights another implication of robustness and suggests processes that could be managed better if we understood it.

Sometimes, invasion of the gut community is desirable, and robustness interferes with the desired outcome. For example, probiotics, (treatments containing live organisms, such as lactobacilli and bifidobacteria) might be more effective if they survived and colonized the gut. Most healthy gut communities are highly resistant to invasion, providing "colonization resistance" that maintains gut community integrity. Little is known about what makes a gut community resistant to or able to recover from invasion, so there is little rational basis for the design or choice of successful invaders for probiotics. Moreover, there are no predictive models to explain why some people's gut communities are more robust than others.

bases for many communities, both robust and vulnerable, over time and space, and applying analytical mathematical tools that can extract patterns will reveal how membership, community structure, specific functions, and functional redundancy and complexity influence robustness. Such analyses might distinguish the characteristics that are associated with all robust communities from those that are specialized or unique to certain habitats or functional units.

HOW DO MICROBES EVOLVE?

Microbial genome variation and the practical and theoretical problems that gene exchange poses for defining species suggest that microbial evolution differs in tempo and mode from the evolution of animals and plants. Current understanding of evolution in general is built on eukaryotes, so a more broadly synthetic evolutionary theory is needed to reconstruct the history of microbial life, to model microbial ecology, and to integrate microbial with eukaryotic evolutionary theory.

The microbial evolutionary model that dominated until recently emphasized *clonality* and *periodic selection*. In this model, bacteria are primarily asexual beings. Their populations comprise clones—descendants of a single progenitor cell. Adaptation and divergence are the consequences of favorable mutations in clonal populations that confer advantage on the genomes

BOX 2-3
Community Robustness: The Case of Sludge

The removal of phosphorus from wastewater by microbes by a process known as enhanced biological phosphorus removal (EBPR) depends upon the stability and robustness of the microbial community responsible for phosphorus accumulation (Levantesi et al. 2002; Garcia Martin et al. 2006). A single organism, *Candidatus Accumulibacter phosphatis*, supplies all the required biochemical functions to remove phosphorus in many systems. However, although *A. phosphatis* can be enriched to high numbers in laboratory scale bioreactors, the organisms remain recalcitrant to growth in pure culture, and this suggests a role for additional community members in their maintenance.

Although EBPR is generally stable and was first used in full-scale wastewater treatment facilities over thirty years ago, these facilities must continue to maintain backup chemical phosphorus-removal systems to respond to periodic crashes of the biological systems. The cause of crashes is not well understood, but they are hypothesized to result from particular biological and environmental perturbations that destabilize the phosphorus-accumulating microbial community. In laboratory-scale reactors that mimic the wastewater treatment plant cycling, small perturbations in pH and the type of carbon supplied can stimulate the growth of competitors of the phosphorus accumulators and result in less efficient or completely abolished phosphorus removal. In addition, homogeneity of the population of *A. phosphatis* may leave the community vulnerable to infection by bacteriophage. Greater understanding of the interactions sustaining the EBPR microbial community will lead to more reliable phosphorus-removal systems.

in which they occur and on the cells that harbor them. Episodes of selection of favored mutants periodically purge populations of genetic and genomic diversity and maintain the cohesiveness (genome-to-genome and cell-to-cell similarity) that allows us to recognize and define species.

However, discoveries of the last decade indicate that gene transfer between similar but nonidentical genomes is, at least in some bacteria, more often the cause of genetic diversity than are new mutations in clones. Indeed, recombination may well be the principal generator of evolutionary novelty in such groups and has parallels to the role of sex in the evolution of animal species. But in other respects there are important differences between microbes and animals: the boundaries of cross-species homologous recombination may be much less distinct, and lateral gene transfer, almost by definition a transgressor of species boundaries, clearly is an important cause of divergence and adaptation in bacteria.

Debate will continue to rage over the frequency and evolutionary importance of such cross-species transfer. Metagenomics, by focusing on genes in an environmental rather than an organismal context, will recast the terms of the debate, as it will of the question "What is a species?" Understanding the genetic and ecological processes that determine the structure and function of microbial metagenomes cannot but lead to new ways of describing patterns in Nature and could lead to the emergence of new theories integrating microevoutionary and macroevolutionary principles.

WHAT ECOLOGICAL AND EVOLUTIONARY ROLES DO VIRUSES PLAY?

Viruses are important not only as pathogens, but also as agents of lateral gene transfer and catalysts that generate tremendous genetic variation in their specific hosts. Viral activity also has important consequences for turnover of the elements, for example, in carbon cycling in aquatic systems. It has only recently been recognized that virus particle numbers are enormous, often exceeding those of co-occurring cellular life. For example, seawater contains 10 times more bacteriophage than cellular microbes. Estimates suggest the biosphere harbors perhaps as many as 10^{31} viral particles (Edwards and Rohwer 2005). Given these vast numbers, the influence of viruses on biodiversity and evolutionary catalysis, and their role in biogeochemical cycling, there is considerable interest in characterizing naturally occurring virus populations. Metagenomics has recently provided an important avenue for exploring these ubiquitous and biologically important entities.

Of special interest is the recent evidence that viruses infecting marine cyanobacteria carry genes involved in photosynthesis (Lindell et al. 2004). Presumably that prolongs the lives of infected hosts (and thus increases virus yields), but another effect is to serve as a genetic bridge between different host species, coupling their evolution, at least as far as such genes are concerned.

Viruses present several unique and interesting opportunities and challenges for metagenomic analyses. Their numbers are large, their genomes are small, and their diversity is impressive. Viruses typically evolve rapidly, so gene sequence conservation is typically much less than that in cellular organisms. Practically speaking, although their numbers are great, their biomass is small, and cloning of viral genes has sometimes been problematic because of modified nucleotides and the cellular toxicity of some of their genes. Metagenomic methods, especially newer sequencing technologies that do not require cloning, may mitigate some of these problems.

3

From Genomics to Metagenomics: First Steps

Genomics as a discipline is at most three decades old. The notion that it might be possible to sequence the genome of our own species began to be discussed in the early 1980s and was seriously considered at federally sponsored workshops in 1984 and 1985; pilot projects began in 1986 (Lambright 2002). The completion of the Human Genome Project (HGP) in 2000 has not only greatly accelerated biomedical science, it has also transformed it. Many questions first asked at the level of individual molecules and genes have better and more complete answers at the level of genomes and systems. And this is true not only for humans: it is now nearly unthinkable to launch a major comprehensive initiative in the biology of any species without sequencing its genome. We have genome sequences for many species of fungi; for nematodes, fruit flies and zebrafish (all highly useful models for human biology); for *Arabidopsis* (a model plant) and rice; for dog, cow, chimpanzee, and many more eukaryotes; and for almost five hundred bacteria and archaeans.

Furthermore, the many associated "omic" sciences (transcriptomics, proteomics, structural genomics, and metabolomics), all using similar high-throughput systems approaches, have revolutionized understanding of what genes do and how they work together (see Box 1-1, page 14). Genomic scientists have returned to hypothesis-testing, making predictions about the behavior of biological systems that can be tested through the acquisition of genome-level data by comparative sequencing and the application of the new "omic" methods.

The field of metagenomics would not be possible without the technological advances and bioinformatics tools that grew out of the HGP.

Descriptions of some of the earliest metagenomics projects illustrating different approaches to characterizing microbial communities are presented later in this chapter.

SEQUENCING IS JUST ONE KIND OF METAGENOMICS

As with genomics, many early metagenomics projects concentrated on gathering enough sequence information to characterize complete genomes. For metagenomics projects, the assembly of complete genomes from samples that are not pure cultures requires the physical recovery of organism-specific clones from environmental-DNA libraries or the computational recovery from environmental-DNA sequence databases of overlapping target-organism-specific sequences (“contigs”). For environments of low complexity, such as the acid mine drainage described below, it is possible to assemble several genomes simultaneously from an environmental sequence database by using various sophisticated “binning” methods (see Box 3-1). Other early metagenomics efforts, including the ones that first applied the term *metagenome*, used the term to describe a resource (all the genes in a particular community) to be mined for specific genes by assessing biochemical functions performed by large-insert clones in suitable hosts (Rondon et al. 2000). This kind of project is now called functional metagenomics, but that term is also sometimes taken to have a meaning analogous to functional genomics. In functional genomics, the goal is to determine not just the sequence of the genome but each gene’s function in the organism in which it is found. The metagenomic analogue would assess functions of the genes found in a community (or a sampling thereof) rather than in an individual species.

Many other “omics” techniques can be borrowed across disciplines. DNA microarrays, when bearing multiple rRNA (or other phylogenetic marker) gene sequences as probes, can be used to track variations in population structure and thus (indirectly) in community function over time and space. Microarrays based on selected genes (and gene variants) involved in processes of particular interest can be used to assess a community’s ability to perform a collective function (such as biodegradation of contaminants) and monitor changes in it over relevant periods (for example, during bioremediation). Community transcriptomics and metabolomics are still subdisciplines in their infancy because of the lability of mRNA and the complexity of communities, but metaproteomics (separation and identification through mass-spectrometric methods of many of the proteins in an environmental sample) is surprisingly well-advanced. And in communities where several genomes are known, it is beginning to be possible to develop community-interaction maps. Meta-omic monitoring of microbial communities as they function and change with time—for instance, genetic,

BOX 3-1
Organizing Metagenomic Sequence Data

Clustering: An approach to data analysis in which a large dataset is divided into distinct subsets based on some specific measure. In analyzing DNA or protein sequences, clustering is used to identify groups of sequences that share an evolutionary origin (families) but can also identify larger sets, such as genomes (see *binning*). Genome annotations can be viewed as form of clustering, where individual genes are assigned to well-characterized (or at least previously known) gene families. In metagenomics, direct clustering of DNA sequences is likely to remain a primary annotation method, as most of these sequences will not be easily assigned to any known gene family. In direct clustering, the nucleotide (or predicted protein) sequence itself is the basis of the grouping of sequences.

Binning: A clustering method that uses composition and/or other characteristics of DNA contigs (overlapping individual reads) to divide them into groups (clusters) that belong to specific genomes or groups of genomes. Examples of characteristics that can be used for binning are GC content and codon use. In metagenomic projects in which genome assembly is a goal, this is used as a preliminary step.

Gene annotation: A process of classifying predicted genes into known and well-characterized gene families. In metagenomics, where a substantial percentage of sequences cannot be easily classified, annotations often remain at the preliminary stage of clustering the sequences into groups (families) that are otherwise uncharacterized.

Gene prediction: A process of analyzing genomic DNA sequences to predict which encode biological functions, such as coding for proteins, structural and regulatory RNA, and other regulatory elements. Gene prediction is important for determining the functional repertoire of a microbial community and for comparing the capabilities of different communities.

population, and metabolic processes that affect methane generation in the permafrost as it experiences global warming—is in principle not different from monitoring such changes in a culture of saccharomycetes as it adapts to a new substrate, in a fruit fly embryo as it develops, or in a human tumor as it progresses. Structural genomics—the systematic expression and structural characterization of the products of all the uncharacterized genes in a genome—will also be a boon; so far, this approach has been applied in the organismal context, but all the highly expressed but unidentified genes in a community metagenome would be an ideal target.

New concepts and methods will be developed for metagenomics that will expand the general genomic repertoire. Metagenomics captures microdiversity, or variation among strains of the same species, thereby producing a more nuanced view of microbes. For instance, a comparison of sequenced genomes of *Prochlorococcus* in the Sargasso Sea shotgun-clone database facilitates identification of "genomic islands" that are highly variable within these genomes; in contrast, genomics on a pure culture would typically generate the sequence of only one of the variants, and the subtlety of population variation would be lost (Coleman et al. 2006). The use of such enormous databases to identify regions and mechanisms of variation within genomes or individual genes is a novel contribution of the metagenomic approach. So is DNA-based stable-isotope probing, in which specific incorporation of substrate containing a stable-isotope (such as C_{13}) by cells in a community that can use it allows specific separation (by density) and identification (by sequencing or with microarrays) of their genes (Dumont and Murrell 2005).

PIONEERING PROJECTS IN METAGENOMICS

We illustrate below, through discussion of a few pioneering achievements and projects now under way, what the metagenomic research paradigm embraces and how its practitioners have begun to combine data collection and hypothesis-testing in sophisticated ways. Five types of projects are discussed: a simple community analyzed in depth, a large-scale sequencing survey in an environmental setting, a functional genomic project, a project focused on a microbial community living in a host, and a project focused on viruses.

The Acid Mine Drainage Project

Microbes in collaboration with humans have wreaked havoc on some geologic sites. One example is the production of extremely acidic outflows from metal mines around the world. The acid is produced by oxidation of sulfide minerals that are exposed to air as a result of mining activity. The acidic solutions that form as a result of mining activities are referred to as acid mine drainage (AMD) (see Figure 3-1). The microbial communities that drive the acidification have formed the basis of some remarkable metagenomic analyses designed to explore the distribution and diversity of metabolic pathways involved in AMD (for example, nitrogen fixation, sulfur oxidation, and iron oxidation), to understand the mechanisms by which the microbes tolerate the extremely acidic environment, and to evaluate how the tolerance mechanisms affect the geochemistry of the environment (Allen and Banfield 2005; Tyson and Banfield 2005; Ram et al. 2005; Tyson et al. 2004).

The AMD project has been paradigm-setting in part because the community exhibits just the right level of complexity—only five major players

FIGURE 3-1 An acid mine drainage site. From such a location, metagenomics studies have allowed assembly sequences of a consortium of genome sequences and fostered pioneering studies of gene exchange and expression. Photo provided by Jill Banfield.

(three bacterial and two archaeal species) reproducibly form a dense biofilm at the sites under study—and in part because it has been studied in great depth. Shotgun sequencing of community DNA enabled the nearly complete assembly of two genomes and partial recovery of three others. The challenge of simultaneously assembling multiple genomes was met by several binning procedures that allow provisional assignment of contigs (sequences that have been generated by computer assembly of overlapping individual DNA fragment reads) to different genomes on the basis of such overall characteristics as base composition and frequency of recovery (see Box 3-1).

Bioinformatic analyses have shown how individual community members might collectively interact biochemically, and the sequences themselves have provided evidence of more long-term genetic interaction at the level of recombination and lateral gene transfer. Nitrogen fixation could be assigned (because of infrequent recovery of relevant genes) to a minor "keystone" species, and metagenomic information guided the later cultivation of this species in pure form; one benefit of metagenomics will be that it will allow the cultivation of more currently "uncultivatable" organisms. Metaproteomic analyses of the same biofilms have now been performed. Many (about half) of the proteins predicted from the genomes of the dominant organism could be found, and their relative abundances say much about how the consortium functions bioenergetically. Proteins involved in coping with protein refolding and oxidative stress are highly expressed, and this reflects how difficult it is to live in acid mine drainage. Many abundant proteins appear to be novel (hitherto unknown) and peculiar to this harsh environment; these proteins will be key targets for a structural genomics approach.

The AMD project moved quickly and relatively easily, partly because of the very low complexity of the microbial assemblage studied. However, most microbial assemblages in nature are not nearly so simple, and this AMD biofilm assemblage represents a rare exception rather than the rule. Therefore, many of the challenges and opportunities of microbial-community genomics cannot be fully appreciated from this single, exceptionally simple example.

Further exploration of diverse microbial communities now demonstrates that shotgun sequencing alone cannot easily be used to complete whole microbial genomes, even in communities that are only moderately complex. Newer methods and approaches now being developed (for example, single-cell genome amplification) are likely to be necessary to dissect and compare the moderate- to high-complexity microbial communities common in natural settings. And completing whole genomes will often not be the goal of metagenomics projects: as the next examples will show, much can be learned about communities without identification of their individual members.

The Sargasso Sea Metagenomic Survey and Community Profiling

The world's oceans harbor vast microbial populations that in part regulate the flux of energy, matter, and greenhouse gases in the sea. The biological properties of these globally distributed microbial communities are still only poorly described. In one approach to this problem, one of the largest metagenomic sequencing endeavors conducted to date employed a shotgun sequencing survey of microbial assemblages from the Sargasso Sea—an ocean environment thought to be relatively low in diversity (Venter et al. 2004). The project began by collecting microbial cells and viruses in different size fractions and extracting DNA from them. The single survey reported 1,214,207 identified putative protein-encoding genes, which represented almost 10 times more protein sequences than were present in all curated protein databases at the time (see Figure 3-2) (Coleman et al. 2006). The Sargasso Sea dataset is remarkable not only with regard to its new information, but also because of the sheer volume of data in it. The Sargasso Sea study was one among several recent studies that heralds a sea change in environmental microbiology efforts, and underscores the significant challenges and opportunities now associated with archiving, integrating, and analyzing massive metagenomic sequence datasets. It is becoming clear that metagenomic DNA sequences soon will outnumber all other types of DNA-sequence data combined.

The gene complement of the assembled Sargasso Sea microbial plankton included 1412 individual ribosomal RNA genes—a useful metric for taxonomic calibration. Counts of proteins useful as taxonomic markers were used to estimate species richness. The proteins indicated that there were about 1800 species (as usually defined) in the sample—which was in fair agreement with the total unique rRNA counts. The types of microbes encountered were generally consistent with those known to be prevalent in the ocean (with some exceptions, discussed below), on the basis of cultivation-independent surveys conducted in the sea over the last decade.

The native microbial Sargasso Sea sequence assemblies demonstrated that assembling large, accurate DNA contigs and scaffolds from shotgun-sequence datasets of complex mixed microbial populations is still a difficult problem. In the Sargasso Sea dataset for example, relatively few large contigs, and no whole genomes, could be assembled from the native microbial population. In retrospect, that is perhaps not too surprising in that it is widely recognized that native microbial populations harbor vast amounts of sequence variation. Inherent intrapopulation genetic complexity, combined with variable species richness and evenness, still poses extreme challenges for standard shotgun sequencing and assembly approaches developed for single microbial strains. One lesson learned from the Sargasso Sea analysis was that complementary approaches, including large-insert

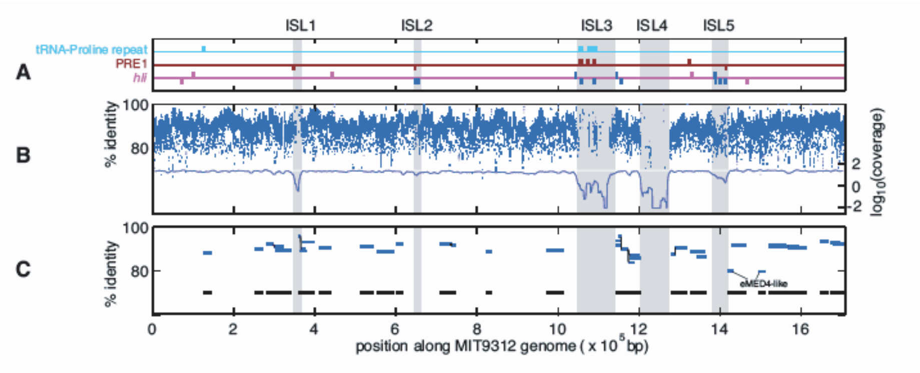


FIGURE 3-2 Genomic islands in the *Prochlorococcus* MIT9312 genome (ISL 1-5) compared to metagenomic data from the Sargasso Sea and the Pacific Ocean. A. PRE1, 48 bp repetitive element; hli, high light inducible genes. B. Blue dashes represent nucleotide identity of individual *Prochlorococcus* DNA fragments in the Sargasso Sea metagenomic dataset; the blue line represents average coverage; C. Blue lines represent individual *Prochlorococcus* large genome fragments (36 kbp) in a Pacific Ocean metagenomic dataset. Black lines represent total coverage of all the fragments across the *Prochlorococcus* MIT9312 genome. SOURCE: Coleman et al. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311: 1768-70. Reprinted with permission from AAAS.

DNA sequencing and single-cell genomic approaches, will be useful and required for thorough characterization of all but the simplest of microbial communities.

Another serious issue with the initial Sargasso metagenomics effort was the microbial contamination that appears to have compromised a large portion of the largest sample, severely limiting its utility for ecological interpretations, and biasing it with non-indigenous microbial genes (DeLong and Karl 2005; Mahenthiralingam et al. 2006). This unfortunate result clearly indicates that metagenomics studies require much more integrated efforts, as opposed to simplistic cloning and sequencing of random environmental samples. Careful sampling and verification procedures, coordination with field experts, and independent sample validation are prerequisites for large-scale metagenomic sequencing efforts. Deep knowledge of the environment sampled, experience with the types and distributions of the indigenous microbes, and independent sample validation methods, will facilitate scientifically rigorous metagenomic studies. Skills from a wide variety of disciplines including environmental science, microbiology, molecular biology, bioinformatics, mathematics, biochemistry, physiology, and ecology, are all necessary to properly gather and interpret metagenomics datasets.

Despite indigenous genomic complexities, potential sampling problems, and analytical challenges, the Sargasso Sea dataset has already proved a useful resource. Among previously discovered genes and proteins, including such photoproteins as proteorhodopsins, the Sargasso Sea dataset revealed new variations on a theme. The dataset's value is evidenced in the large number of papers that have mined its taxonomic information, analyzing new gene sequences or even synthetically producing and characterizing never-before-studied genes and gene products. Detailed studies of whole-genome genomic variability, structural organization, and evolution in taxa that were well represented, including *Prochlorococcus,* have been greatly advanced by these new data. The theoretical and practical discoveries and applications already arising from this single dataset provide ample evidence of the value of large-scale, whole-microbial-community genomic analyses.

In lieu of full genome assemblies, comparative analyses of the Sargasso Sea and other datasets have demonstrated the utility of individual gene comparisons within and between samples. A new approach in microbial ecology, comparative community genomics, is emerging from such studies. A recent study, for example, compared, on a gene-by-gene basis, similarities and differences between community gene-sequence datasets from the Sargasso Sea, a sea-floor whale carcass, and the acid mine drainage community (Tringe et al. 2005). By taking a "gene-centric" approach, as opposed to an assembly-driven "genome-centric" approach, it was possible to compare the patterns of occurrence of specific gene categories and assemble "community profiles" of functional-gene content. The overrepresented specific categories of "environmental gene tags" (EGTs) in different samples (for instance, a disproportionate representation of photosynthetic genes and rhodopsins in the Sargasso Sea sample) verified the utility of the approach for inferring metabolic features associated with specific microbial communities. Judicious sampling can greatly facilitate such comparisons by allowing comparison of communities along well-validated environmental gradients. Another recent study in the Pacific Ocean compared microbial communities along a depth gradient, from surface waters to 4 km deep. Genomic characteristics that covaried with the environment were evident and suggested depth-specific functional and evolutionary themes in microbial communities and genomes (DeLong et al. 2006). This recent work highlights the future promise and utility of comparative community genomic studies.

The Soil-Resistome Project

The discovery of antibiotics transformed medicine in the middle of the 20th century by providing effective treatments for infections that were previously untreatable and often fatal. After decades of use, the power of many antibiotics has diminished because populations of many human pathogens have evolved resistance to them and left us with no retaliatory weapons for some of the most virulent infectious agents. Although many of the genes that enable pathogens to resist antibiotics have been identified, little is known about where they originate in nature. Most of the antibiotics in clinical use were discovered in soil bacteria, so it seems likely that resistance genes also arose there. That the vast majority—perhaps as many of 99.9%—of the microbes in soil are not readily culturable, invites the use of metagenomics to assess the suite of antibiotic-resistance genes, or the "resistome," in soil (D'Costa et al. 2006). The soil-resistome project takes a functional metagenomics approach: fragments of DNA are cloned from soil, and the clones are screened for expression of antibiotic resistance. This differs from the metagenomic studies discussed so far in that the genes are recognized by their activity—antibiotic resistance—rather than by their sequence, and this provides the opportunity to detect genes that might be unrelated to any known resistance genes.

The soil-resistome project has led to isolation of new groups of antibiotic-resistance genes. The strategy has been to clone metagenomic DNA from soil in temperate sites with natural vegetation, mixed grassland in Wisconsin, and a boreal forest in central Alaska. One of the advantages of studying antibiotic resistance is that it provides a selectable phenotype; only the clones of interest will grow in the presence of the antibiotic, so it is possible to screen libraries that contain millions of clones. This is impossible with most screens because they usually require addressing each clone individually and recording some feature of interest. In the soil-resistome project, the libraries are cultured in the presence of each antibiotic of interest, clones that grow are retested, and clones that are confirmed are saved for further study.

The resistome study revealed aminoglycoside resistance genes that encode a group of enzymes called acetyltransferases that are more closely related to each other than any previously described members of the family (Riesenfeld et al. 2004) and genes that encode resistance to β-lactam antibiotics (penicillin-like compounds) that were phylogenetically distinct from previously described enzymes. A gene that encodes an acetyltransferase was also discovered in the Alaskan forest soil, and its closest homologues in the sequence database were the genes discovered in the Wisconsin soil. This result raises intriguing questions: Are the resistance genes seen in clinical isolates the most abundant in the environment? Will some of the new genes find their way to clinical settings? If so, by what route? Will tools used in the past to inhibit the activity of resistance genes work in the future if these "wild" resistance genes reach the clinic?

The greatest challenge in metagenomic analysis based on functional screens is gene expression. Success generally depends on expression, in a laboratory strain of *E. coli*, of exotic genes from exotic organisms. The differences in gene-expression mechanisms among species are likely to prevent detection of many genes by this method. Using multiple host species and tweaking the gene expression machinery of *E. coli* are mechanisms for achieving expression of a wider array of genes that deserve further study to enhance the utility of the function-based approach to metagenomics.

The Human-Microbiome Project

Microbes thrive on us: we provide wonderfully rich and varied homes for our 100 trillion microbial (bacterial and archaeal) partners. Considering that we contain perhaps 10 times more microbial than human cells and at least 100 times more microbial than human genes, it is inescapable that we are superorganisms composed of both microbial and human parts. Bacterial communities play an important role in health and disease in a variety of anatomical locations, such as the female reproductive tract, the skin, the oral cavity, and the respiratory tract. Even after completion of the first reference human genome, our view of the "human" genetic landscape is quite incomplete. We know little about how our microbial component has evolved or about the forces that are shaping it as our biosphere, our lifestyles, and our technologies change. What aspects of our microbiome are uniquely "human," or mammalian? Are we undergoing a form of "microevolution" because of changes in our microbial ecology that is affecting our biology and our predispositions to diseases?

Because the human microbiota has not yet been extensively explored, much of what is known of the contributions of organisms' microbial partners has come from comparisons of germ-free animals (reared with no microbes) with their counterparts that have been colonized with defined components of the mouse or human microbiota (Turnbaugh et al. 2006, 2006; Samuel and Gordon 2006). Comparisons of germ-free and colonized animals have shown, for example, that the gut microbiota regulates energy balance, directs myriad biotransformations (including detoxification of carcinogens), modulates the maturation and activity of the innate and adaptive immune systems, and affects the cardiovascular system. On the basis of these and other observations, the gut microbiota has been invoked as a factor that determines susceptibility to diseases ranging from obesity and diabetes to gastrointestinal and other malignancies, atopic disorders (such as asthma), infectious diarrhea, and various immunopathologic states, including inflammatory bowel diseases.

Initial results of 16S rRNA gene-based enumerations of the microbial communities of a small number of humans have revealed remarkable diversity in a number of habitats, including the gut (Eckburg et al. 2005; Ley et al. 2006). That raises the question of whether there is a core microbiome associated with all humans and, if a shell of diversity surrounds such a core, what it contributes to the differences between individual physiologic properties.

The first truly metagenomic survey of a component of the human microbiota appeared in 2006 (Gill et al. 2006). It involved sequencing the microbial communities harvested from the colons of two healthy adults. Analysis of 78 million base pairs of unique DNA sequence disclosed that, compared with the human genome and previously sequenced microbial genomes, the gut metagenome is enriched in genes involved in the breakdown and fermentation of otherwise indigestible plant-derived polysaccharides that form an important part of modern diets, the detoxification of xenobiotics consumed intentionally or inadvertently, and the synthesis of essential amino acids and vitamins. These findings emphasize that the human metabolome is actually a composite of human and microbial attributes; they also point to a future in which it may be possible to optimize the nutritional status of the overfed or underfed on the basis of knowledge of their gut microbial ecology, or to predict the bioavailability of orally administered drugs, or to forecast the susceptibilities of individuals or populations to particular types of cancer. Greater knowledge of the microbial communities of the oral cavity, skin, and female reproductive tract will similarly improve our ability to prevent, diagnose, and treat diseases at those sites.

We are also host to countless viruses. A recent survey reported that human feces contain about a billion RNA viruses per gram, representing 42 viral "species" (Zhang et al. 2006). Most were plant viruses (probably originating in food), and the RNA of the most abundant (pepper mild mottle virus) was still infectious to its host (the pepper plant); this suggests a role for humans as agricultural-disease vectors.

What remains to be learned about the human microbiome is enormous. The early projects hint at how rich and productive further study will be. The Human Genome Project had to be thoughtfully staged and coordinated, and any project aimed at understanding the human microbiome will need similar careful planning.

Viral Metagenomics

Despite the challenges, early applications of metagenomics to marine virus populations have already provided considerable insight. Studies of naturally occurring phage in aquatic systems first physically separated viruses from co-occurring microbial cells and then used amplification techniques to generate "shotgun" viral DNA libraries, which were then randomly sequenced. The first look at seawater confirmed its huge diversity of viral assemblages: 65% of all the sequences examined from the first seawater viral libraries were novel and bore no significant similarity to any known genes in the databases. The approach appeared reasonably comprehensive for double-stranded DNA phages; it recovered most major families, including those with bacterial or algal hosts. Similar applications in marine sediments yielded parallel results, but with some interesting differences. Among sediment viral assemblages, even greater novelty was detected: more than 75% of the viral sequences recovered resembled nothing in the databases. The double-stranded viral DNA sequences identified in sediments suggested an important role for temperate phages, for example, viruses that can integrate into their host's genome. A large comparative analysis of seawater viral assemblages collected from diverse locales recently indicated that marine viral species have a global distribution (Angly et al. 2006).

Applications for analyzing RNA-based virus assemblages from seawater have also been developed, and have revealed new groups of RNA-viruses that infect marine planktonic protists and animals. In total, the early viral metagenomic analyses provide a solid starting place for exploring and interpreting the genomic diversity in naturally occurring viruses. They also provide a foundation for the next steps in microbial-community genomics, the integrated analyses of viruses and their cellular hosts collected from the same metagenomics samples and analyzed simultaneously.

4

Designing a Successful Metagenomics Project: Best Practices and Future Needs

As the number and diversity of metagenomics studies have grown, so too has an appreciation of the challenges that these studies present as compared with genome-based analysis of single organisms. Many of the challenges are likely to diminish with the development of new technologies and mathematical tools. Nonetheless, many of the criteria of success in metagenomics studies will remain unchanged by new knowledge or methods. This chapter is devoted to the key steps in developing a metagenomics project, the decision points along the way, and the issues that need to be considered at each step.

PARALLELS WITH TRADITIONAL MICROBIAL GENOME SEQUENCING

Metagenomics-based approaches share many features with traditional genome sequencing of cultured bacteria but also present a number of unprecedented challenges. This chapter describes a number of complementary approaches to the study of microbial communities (see Figure 4-1) which, depending on the goals of a particular project, can be applied individually or together to obtain a new understanding of the numbers and abundance of microbial community members, their metabolic capabilities, and how these parameters change in response to external stimuli. This chapter identifies the advantages and limitations of each approach and explores the research needed to overcome barriers to understanding microbial communities.

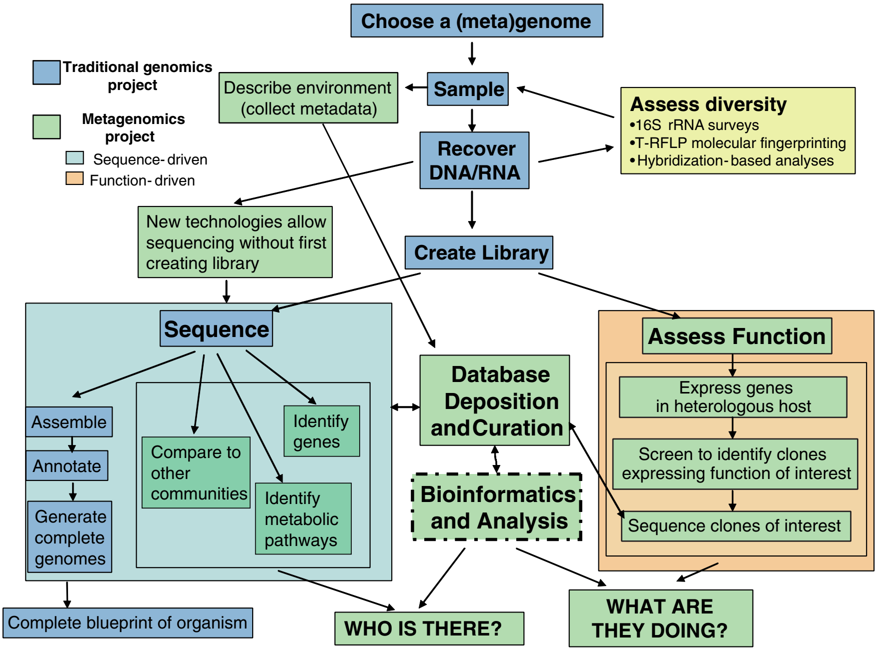

Traditional genomics project
Metagenomics project
Sequence-driven
Function-driven
Choose a (meta)genome
Describe environment (collect metadata)
Sample
Assess diversity
•16S rRNA surveys
•T-RFLP molecular fingerprinting
•Hybridization-based analyses
Recover DNA/RNA
New technologies allow sequencing without first creating library
Create Library
Sequence
Assemble
Annotate
Generate complete genomes
Compare to other communities
Identify genes
Identify metabolic pathways
Database Deposition and Curation
Bioinformatics and Analysis
Assess Function
Express genes in heterologous host
Screen to identify clones expressing function of interest
Sequence clones of interest
Complete blueprint of organism
WHO IS THERE?
WHAT ARE THEY DOING?

FIGURE 4-1 Metagenomics differs from traditional genomic sequencing in many ways. The dark blue boxes show the typical steps in the sequencing of a single organism's genome. Metagenomics requires greater attention to sampling, and assessing the diversity of the sample by various means (yellow box) is necessary to ensure that the sample is representative. Extracting the appropriate nucleic acids from the sample is another step that can be challenging in a metagenomics project. Preparation of a library is often the next step, but new sequencing technology can bypass this step. The DNA from metagenomics samples can then either be sequenced (blue box) or assessed for the functions it encodes (orange box). The sequence can sometimes be assembled into complete genomes of community members, but can also be analysed in other ways (light blue box). Data storage and computational analyses are critical steps in metagenomics projects and must be integrated throughout the project. Overall, a metagenomics project can answer the questions "Who is there?" and "What are they doing?" in addition to assembling genomes.

The development, about 15 years ago, of methods for rapid and efficient sequencing and assembly of large segments of DNA was critical for the revolution in microbial genomics and has led to the completion of more than 460 bacterial and archaeal genome sequences by January 2007.[1] For

[1] *http://www.genomesonline.org.*

most of the projects, the starting material has been DNA extracted from pure cultures of organisms grown in the laboratory or in association with animal or plant cells. Regardless of the organism selected for sequencing, the goal of the projects has been the same: to generate a complete or nearly complete genome sequence that can serve as the substrate for genome annotation and analysis (see Figure 4-2). For metagenomics projects, it will be important to accumulate additional complete genome sequences, especially for currently under-represented taxa. Such sequences should help in the identification of otherwise unidentifiable open reading frames in metagenomic fragments, and facilitate scaffolding of metagenomic data.

Metagenomics projects differ from traditional microbial-sequencing projects in many respects. The starting material is a mixture of DNA from a community of organisms that may include bacterial, archaeal, eukaryotic, and viral species at different levels of diversity and abundance. Most of the organisms will elude attempts at cultivation. In some projects, sample collection may be confounded by the presence of limited amounts of DNA or the presence of contaminating DNA or other compounds that interfere with DNA extraction. These factors make it much more challenging to think about the generation of complete or nearly complete genome sequences from metagenomics projects. Often, generating complete genomes will not

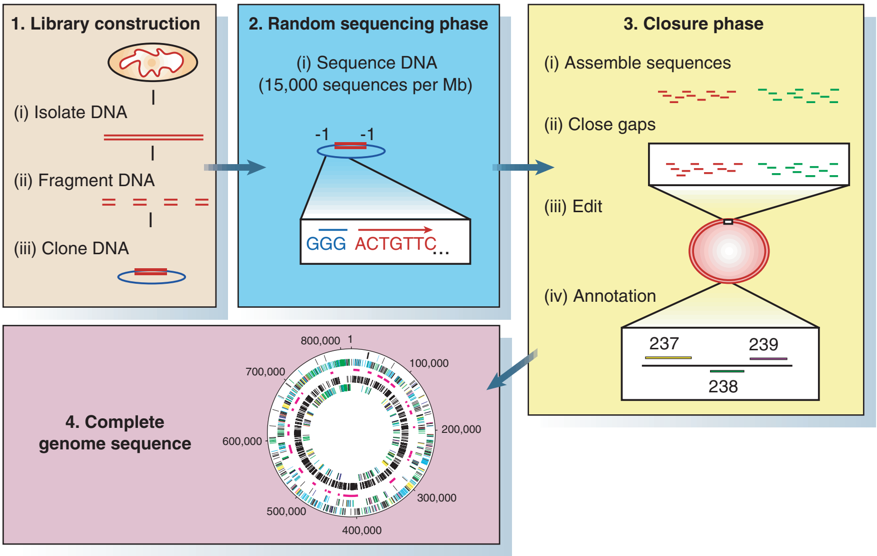


FIGURE 4-2 Steps in a traditional microbial genome project. SOURCE: Fraser et al. Reprinted by permission from Macmillan Publishers Ltd: *Nature* 406:799-803, copyright 2000.

be the focus—not so much because of the difficulty as because the real goal, understanding community composition and function, does not require it. In the study of complex communities, it is often necessary to address the question of how much sequence is enough to understand a community and to carry out comparative analyses of related communities. In many cases ,this information can be obtained by applying various methods based on 16S rRNA sequence that can reveal a tremendous amount of information about microbial diversity and abundance. In other cases, whole-genome shotgun-sequencing data generated by the traditional Sanger sequencing methods, by newer very-high-throughput methods, or by a combination of the two approaches will provide additional information about the gene content of a community and its metabolic potential. Finally, function-driven metagenomic analysis, like the soil resistome project described in Chapter 3, which starts with functional expression of an activity in a surrogate host, followed by sequencing and phylogenetic analysis provides another measure of community potential. Regardless of the methods employed to answer questions about community structure and function, the composition of any microbial community is likely to be profoundly affected by the habitat from which the sample was obtained. Detailed knowledge about the habitat is essential for meaningful biological interpretation of the sequence data.

METAGENOMICS STEP BY STEP

Habitat Selection

The choice of the microbial community to study will be driven by the underlying scientific question being addressed. However, the more information one has about the study habitat—physical, chemical, and ecological—the more insight can be derived from the metagenomic data. Specific hypotheses can be posed and genes sought in genomic data from a well-characterized site. The acid mine drainage study is a case in point. The geochemical conditions that create and maintain that habitat were delineated before the researchers embarked on their metagenomics journey. As a result, the information gleaned in studying the genomes could be placed in a phylogenetic, biochemical, and physiological context. For example, knowledge of the nitrogen budget of the site impelled the researchers to seek nitrogen-fixation genes in the metagenome. When they did not find candidate genes in the dominant members, they examined the minor components of the community and discovered that one of the least abundant members of the community, *Leptospirillum ferrodiazotrophum*, carried the *nif* operon. They then cultured that bacterium by providing N_2 as the only nitrogen source to ensure that only nitrogen-fixing bacteria would grow. The discovery of the keystone species (a community member whose significance to the community is larger than its relative abundance) was made possible by an ecological inference that depended entirely on knowledge of the site.

Exploring habitats that have been well studied by other methods will accelerate progress in metagenomics.

- Well-characterized habitats will leverage the value of metagenomic data.
- Interdisciplinary collaborations with scientists studying the non-microbial aspects of the habitat will inform the analysis.
- Different habitats require different depths of sequencing depending on their complexity and the degree of completeness needed to address the questions being posed. Pilot studies to determine the required depth of sequencing may be necessary.

Sampling Strategy

Sampling is fraught with challenges. Each decision about the type, size, scale, number, and timing of sampling shapes the conclusions and inferences that can be drawn. The labor intensity of producing and analyzing metagenomic libraries aggravates sampling issues that are inherent in all ecological studies. If conclusions about the habitat are to be drawn, the samples must be representative of the habitat. To obtain representative samples, it is critical to know the scale and amplitude of variation in the habitat environment. Soil communities, for example, change on a micrometer scale, following the physical and chemical heterogeneity of the mineral and biological materials that make up the soil. A 1-cm^3 aggregate may contain aerobic and anaerobic regions; clay, silt, and sand particles; plant matter in various stages of decomposition; and a variety of invertebrates, each of which probably has its own associated microbiota. For such a habitat, what is the appropriate sample size? Is it possible to account for the minute microhabitats when 50 g of soil is needed to build a metagenomic library?

Habitat change over time is one of the most interesting aspects of communities. Their responses to changing conditions are central to understanding community structure, function, and robustness. Understanding the role of host-associated microbial communities in host development and health requires not only sampling from the same host over time, but also understanding host-to-host variation. For biodefense and forensics purposes, it may be important to determine whether threat organisms originated in nature or in a laboratory. But the variability of communities creates a sampling conundrum. How many samples are needed to represent the many conditions of a community? How are different types of change accounted for—natural cycles versus catastrophic events, which might be a toothbrushing in the case of oral microbiota and a flood in the case of soil? Even more challenging are the long-term changes, such as global climate change, that both affect and are affected by microbial communities. How much work is needed to differentiate baseline variability from real change?

The answers to most of these questions depend on the complexity of the community, the heterogeneity of the habitat over time and space, and the fineness of the distinctions that need to be made. As biological and computational methods become more efficient, it will be possible to draw more robust conclusions from more complex communities in more variable habitats. No matter the power of the methods now or in the future, it is essential to consider sampling issues and limitations at the beginning and throughout any metagenomics study of a complex community, and the sampling scheme must inform the interpretation of results.

- Sampling strategy should be carefully considered and the variability in the experimental method assessed before sampling (see Table 4-1). If understanding factors that influence change in a community is a research goal, adequate controls should be in place to distinguish baseline variation from real change.
- Pilot projects may be needed to assess diversity, variability, and the appropriateness of different technological approaches (such as targeting of different subgroups or type of sequencing technology) to enable optimization of the project plan.

Macromolecule Recovery

The quality and completeness of data obtained from metagenomic analysis of any community will be only as good as the procedures used for the extraction of DNA from a sample. A currently unanswered question is whether methods like those developed for the direct isolation of DNA from different types of soil are equally effective in recovering DNA from all members of other complex communities. For example, cells from different species differ in their susceptibility to lysis under various conditions, and some members of the same species differ in their susceptibility to lysis in different physiological states. Furthermore, the conditions necessary to lyse the more recalcitrant cells in a community may be sufficiently harsh to cause degradation of DNA from other community members. Another issue that has been tackled recently is the development of methods to distinguish between DNA from viable and dead cells in a given sample—a distinction that may be important in drawing conclusions about the overall metabolic capabilities of a microbial community. Results from a number of studies related to this question suggest that no universal approach is equally effi-

TABLE 4-1 Sampling Considerations in Metagenomic Analyses

Sampling Considerations	Questions
Scale	What is the size of the habitat? What is the size of the sample of the habitat? How representative of the habitat is the sample?
Biological variation	How is biological variation in the site accommodated in the sampling scheme? On what scale is the variation (subsample to subsample, sample to sample, site to site)? How much replication is needed to represent the full variety of properties of the site? How flexible is the community? If very flexible, then what does it mean to take a snapshot in time?
Experimental variability	Where is the experimental variability in process sampling? In extracting DNA? In cloning? In storage of samples? How does the experimental design maximize replication to account for experimental variability?
Reproducibility	If patterns are detected, are they reproducible?
Coordinates of place and time	Is detailed information about the site and time of sampling recorded?
Repository	Can the samples be stored for future analysis? Can they be placed in a central repository?
Singletons	What is the significance of a singleton (a unique sequence or other data point)? If it is never found again, how should its relevance be assessed?

cient for DNA extraction in all environments. These challenges are not insurmountable, and improvements in DNA-extraction methods should be vigorously pursued, but the limitations in DNA extraction methodologies as applied to specific projects should be acknowledged and addressed.

The effect of contaminants on the recovery of DNA or RNA of interest from many different environments presents another technical challenge. Because of the very large differences in genome size between eukaryotic and bacterial cells, even minor contamination of a sample with host (plant, animal, or human) DNA reduces the effective concentration of bacterial DNA available for sequencing and hence increases the cost of generating sufficient useful (nonhost) sequence. It also reduces the chances of recovering low-abundance members of the bacterial community. There appear to be no published reports comparing methods for removing host DNA for bacterial metagenomic analysis. One study used nucleases derived from the host, an insect gut, to digest the host DNA before lysing the bacteria; this was very effective but may not be generally applicable (Guan et al. 2006). One method of building metagenomic libraries from soil involves physically separating the bacteria from the rest of the soil matrix before lysis to minimize contamination with the numerous inhibitors of the enzymes that are used for cloning that are found in soil (Akkermans et al. 1995; Berry et al. 2003). This method and alternatives used in other fields suggest that various filtration, centrifugation, or lysis methods could be adapted to the challenge of separating bacterial cells from eukaryotic cells before library construction. The use of subtractive hybridization (hybridization of the community DNA to immobilized or labeled copies of eukaryotic DNA, from which the unbound bacterial DNA can then be separated), or separation based on GC content (Holben and Harris 1995), will also, in theory, allow enrichment of bacterial DNA at the expense of eukaryotic DNA. Research is needed in more robust nucleic acid extraction procedures that have known effects on recovery of DNA from community members that are more difficult to lyse, that are in different physiological or physical states, or that are rare. Extraction procedures need to be standardized for all habitat types as much as possible to aid comparisons among habitats.

Even after extraction of DNA from the sample, all methods that rely on library construction, including metagenomic approaches, have potential for bias or skewing. Some cloned genes are toxic, and so may be underrepresented in typical clone libraries used for sequencing. Some types of DNA (for example, certain viral DNAs) are chemically modified in such a way that they are difficult to clone. Many of these challenges, however, are reasonably well known from prior molecular biological studies, and can be assessed and addressed by careful analyses that monitor recovery and quantitation issues using parallel independent approaches. In addition, newly evolving sequencing approaches that do not rely on cloning (discussed elsewhere) will alleviate some of the problems associated with cloning bias.

GETTING THE MOST OUT OF METAGENOMICS STUDIES

Obtaining the most information from metagenomics studies will continue to be a challenge primarily because the potentially disparate and incomplete datasets are so large. The approaches to analysis can be divided into three general categories, each of which has advantages and limitations.

16S rRNA-Based Surveys

The first category includes a set of methods based on analysis of 16S rRNA genes, which provide relatively rapid and cost-effective methods for assessing bacterial diversity and abundance. These types of assays are often used as a first step in larger metagenomics projects to evaluate bacterial diversity in potential samples of interest (soil samples from different locations in a defined area, fecal samples from various individuals, etc.) in order to choose the most appropriate samples for more in-depth analysis. These methods can also be used to monitor changes in community composition over time and space without the need to generate other types of sequence data.

One of the simplest ways to assess community structure is based on a method for molecular fingerprinting of microbial communities called terminal-restriction fragment length polymorphism (T-RFLP) analysis. The technique employs polymerase chain reaction (PCR) in which two differentially fluorescently labeled primers are used to amplify a selected region of the 16S rRNA gene from total community DNA. The mixture of dually labeled amplicons is digested with a restriction enzyme (*Msp*I or *Hae*III) releasing the labeled 5′ and 3′ ends—or terminal restriction fragments (T-RFs)—of each individual amplicon. These differentially labeled primer pairs combined with the two restriction enzymes result in six fluorescently labeled T-RFs. T-RFLP profiles can be determined using an automated capillary DNA sequencer and GeneScan® software (Applied Biosystems). T-RFLP profiles reflect differences in the numerical abundance of bacterial populations in the samples (Liu et al. 1997). Changes or differences in microbial community structure can be detected based on the gain or loss of specific fragments from the profiles (Engebretson and Moyer 2003; Forney et al. 2004; Osborn et al. 2000) and statistical clustering analysis of T-RFLP data can identify communities that have similar numerically abundant populations.

Significant insights into species richness, structure, composition, and membership of microbial communities have been gained through analysis of 16S ribosomal RNA (rRNA) gene sequences. PCR amplification with primers that hybridize to highly conserved regions in bacterial or archaeal 16S rRNA genes (or eukaryotic microbial 18S rRNA genes) followed by cloning and sequencing yields an initial description of a microbial community. Powerful computational tools have been developed to assess species richness (FastGroup and DOTUR) in a sample and the similarity between two communities in membership (SONS) or structure (AMOVA, LIBSHUFF, UNIFRAC, and TreeClimber). Analyses with these tools have revealed many challenges still to be resolved. Most communities have many members (that is, they are species-rich) whose abundance is uneven. This presents a sampling issue: how many samples need to be taken to find members of the sparser groups? However, recent estimates based on 16S rRNA sequencing and statistical modeling of soil communities indicate that with decreasing sequencing costs, it is possible to conduct a fairly complete census of soil communities even though these are the most species-rich and uneven in structure of communities studied so far.

The challenges associated with unknown community structure may soon become more manageable.

In addition to sampling challenges, the 16S-based approach to the study of microbial communities has other limitations. First, 16S rRNA sequences provide a phylogenetic framework into which community members can be placed, but that framework does not tell us much about the functional capabilities of the individual members or the entire community under study. In addition, PCR-based studies are inherently biased in that not all rRNA genes amplify equally well with the same "universal" primers. Indeed, in several published metagenomics studies there have been discrepancies between estimates of community diversity derived from PCR-based 16S rRNA gene surveys and those derived from whole-genome shotgun data, although in some studies the estimates are remarkably similar (Liles et al. 2003; Tyson et al. 2004). A third limitation of 16S rRNA gene surveys is that these genes occur in multiple, nonidentical copies in many bacterial and archaeal taxa, which may lead to overrepresentation of some species in 16S rRNA gene libraries; this limitation might be overcome through the use of additional, single-copy phylogenetic markers, such as *recA* or *rpoB*, for initial community surveys. Research is needed for the development of additional genetic markers of community diversity to enhance the phylogenetic and functional resolution of microbial communities.

In parallel with efforts in 16S rRNA sequencing, several groups have been pursuing the development of 16S rRNA-based microarrays (or microarrays based on other phylogenetic marker genes) for high-throughput compositional analysis of microbial communities (Wu et al. 2006). Such phylogenetic oligonucleotide arrays typically carry hundreds to thousands of spots bearing synthesized oligonuceotides as probes matching rRNA gene sequences that are found in databases or are expected to be present in samples. The arrays can be designed to include probes that target bacterial species at different taxonomic levels, from species to phyla. This approach makes it feasible to assess bacterial diversity in large numbers of samples; this facilitates continuous monitoring of microbial communities, and the content of the microarrays can be expanded as new species or phylotypes are revealed in metagenomic studies. One disadvantage of microarray-based approaches to metagenomic analysis is that the information that can be obtained is limited by known bacterial phylogeny represented on the array—that is, the arrays are blind to species that have yet to be discovered. Microarray-based approaches also suffer from a technical challenge that plagues other studies of diverse microbial communities: it may be difficult to distinguish a hybridization signal from a low-abundance community member from background. These caveats aside, microarray-based assays have the potential to provide valuable complementary information in metagenomics projects. For example, the current version (2.0) of the Affymetrix-based PhyloChip targets over 30,000 unique database sequences, totaling almost 9,000 distinct taxonomic groups, with each group represented by a set of 11 or more perfectly matching probes and a corresponding mismatch control probe. The PhyloChip has been successfully used to characterize complex environments such as soil, aquifers, and urban air (Brodie et al. 2007; Desantis et al. 2007). As expected, the PhyloChip detects broader diversity than typical clone library sequence analysis (Brodie et al. 2007; Desantis et al. 2005). Depending on the microbial diversity of the sample, the PhyloChip detects on average twice as many taxa as 16S rRNA gene sequencing (Desantis et al. 2005). The quantitative power of microarray-based assays is somewhat limited to sample comparison at this stage (community dynamics), but the combination of 16S rRNA gene sequencing and arrays is a unique and powerful tool for the characterization of any microbial community because it allows both the discovery of novel phyla and extensive cataloguing of each taxonomic unit present in a given environment.

16S rRNA Phylogenetic and Functional Anchors: A Hybrid Approach

Metagenomic clones can be given a context, or "anchored," by looking for a gene that characterizes the clone or the organism that it came from. The genes most commonly used as anchors are such phylogenetically informative ones as those encoding 16S rRNA or RecA protein. Sequencing all clones that are derived from one phylogenetic group may help to stitch together a picture of the group even in the absence of cultured members. Functional anchors have also been used to collect clones that share a characteristic, in this case an expressed function. The clones expressing the function of interest can be sequenced to search for phylogenetically informative genes to begin to piece together a slice of the community that is related to a particular function.

- Research is needed to develop phylogenetic and functional anchors for use in different microbial communities to advance the process of linking community membership and function.
- New physical methods are needed to enhance the yield of inserts bearing such anchors.
- Novel strategies to obtain gene expression of genes from a wider range of organisms will facilitate this work.

Generation of Large-Scale DNA Sequence

A second important approach to studying microbial communities is based on generating large amounts of DNA sequence using well-described shotgun-sequencing strategies. This is an excellent method for obtaining information on the gene content and functional capabilities of mixed microbial communities. The sequence data, which can potentially provide information on "what are communities doing?" are most informative when coupled with other analyses that help to determine "who is there?" A random shotgun strategy for studying communities can reveal information on community diversity (e.g., bacteriophage and other viruses, eukaryotic species, novel bacteria and Archaea) that is not captured with 16S rRNA gene surveys. Bacteriophage, in particular, are thought to play a critical role in shaping microbial membership and evolution and their abundance and diversity cannot be assessed using the previously described approaches. The fundamental limitation of this approach is the vast number of genes that do not have homologs of known function in the databases.

Assembling Whole Genomes

If the goal of a metagenomics study is to determine the complete genome of some or all of a community's members, many challenges must be overcome. Given that environmental samples contain DNA from many species that are present in different abundance and differ from each other in genome size, the final depth of sequence coverage for each organism at a given level of sequencing will vary. Piecing together all the separately sequenced fragments of a genome is a substantial bioinformatics challenge. In a simple community like the acid mine drainage system described in Chapter 3, there are enough overlapping fragments of the dominant members to assemble their entire genomes. In a more complex community, even the sequence fragments from the dominant members will be sufficiently diluted to preclude assembly, and species of low abundance may be represented by only a few sequences. These differences in sequence coverage can provide information on relative species abundance. It is important to take differential species representation into account in selecting assembly strategies for metagenomic data to avoid classifying sequences from the most abundant species as repeats and throwing them out of assembly algorithms.

In highly diverse microbial communities, even when very large amounts of DNA sequence data are generated (several billion to a trillion base pairs of DNA), it will be difficult to generate assembled genomes, and the less abundant members of any community might be represented only by singleton sequences. New sequencing technologies, now being introduced by a number of companies (see Table 4-2), provide alternative strategies for generating substantially more DNA sequence at a lower cost than current Sanger-based capillary sequencing methods. The new technologies will go a long way toward achieving sequence depth that extends beyond the

TABLE 4-2 New Sequencing Technologies

Applied Biosystems 3730 xl	454 GS FLX Pyrosequencer	Solexa 1G Genome Analyzer	Applied Biosystems 1G SOLiD Analyzer
1-2 Mbp per day/machine	100 Mbp per day/machine	800 Mbp per run/ machine (25 bp) 960 Mbp per run/ machine (30 bp) (assumes 32M features)	1200 Mbp per run/ machine (Frag Library) 2400 Mbp per run/ machine (Mate Pair Library)
Long sequence reads (600-900 bp)	Medium sequence reads (200-300 bp)	Short sequence reads (25-40bp)	Short sequence reads (25-30 bp, 25x2 for mate pair libraries)
Mate pair information[a]	No mate pair information	No mate pair information (promised for future versions)	Mate pair information
Libraries subject to cloning bias	No library cloning bias	No library cloning bias	Libraries may show cloning bias
Can resolve homopolymers	Cannot easily resolve homopolymers	Can resolve homopolymers	Can resolve homopolymers

[a] Mate pairs are two sequencing reads derived from the same clone, or molecule, one from each end. If the length of the clone, or molecule, is known, mate pair information constrains where these sequencing reads can be placed in an assembly.

most abundant members of microbial communities in shotgun sequencing projects, but some applications may still need additional methods (such as normalization, subtraction, or physical separation methods) that will ensure better representation of the lower-abundance community members. Furthermore, the new technologies are still vexed with issues such as shorter read lengths than those that have become routine with Sanger sequencing. The limitations have obvious consequences for assembly, particularly for metagenomics applications in which assembly is already complex and difficult, but considerable effort is also being devoted to figuring out solutions to these technical challenges.

Because the assembled sequence data from metagenomics studies will often be incomplete and it may often be difficult to draw unambiguous conclusions about who is there and what each species is doing metabolically, any additional data that would help with assembly validation and making phylogenetic and functional inferences will be of great utility.

The availability of reference genomes is an example of a kind of data that make assembly easier. In response to the need for such data, the

National Human Genome Research Institute has undertaken a phase 1 human gut microbiome initiative (NIH 2007). The initiative will deliver deep draft genome sequences of 100 cultured bacterial reference species representing each of the divisions known to make up the distal gut microbiota. The strategy adopted for this initiative is to generate 20X sequence coverage of purified DNA with a 454 GS20 pyrosequencer (i.e., 20,000 bp of sequence will be generated for every 1000 bp in the organism's genome) and to combine the resulting data with paired end reads from plasmid whole-genome shotgun subclones from each targeted species (≥5X coverage, produced with a conventional ABI 3730xl capillary machine). This approach will generate hybrid assemblies of each genome ("scaffolds") with nearly complete gene coverage. The completed genome sequences will provide a key reference for metagenomic projects related to the human gut, which will be valuable because relatively few members of the human gut microbiota have been sequenced. In later human metagenomics projects, sequence data generated from microbial-community DNA can be readily aligned with these 100 microbial genome scaffolds to help to validate metagenomics assemblies, answer questions related to phylogeny and metabolism (what species are contributing what genes to the community genome) and assist in the evaluation of gene flow between community members (by providing evidence of lateral gene transfer among community members).

Several kinds of research are needed:

- Although genome assembly is not feasible in complex communities or necessary for answering many questions, it is useful in some cases. Hence, new approaches are needed for such assemblies and for interpretation of consensus genomes or partial genomes of communities.
- Research is needed at both the experimental and computational levels to simplify complexity in complex communities so that patterns are discernable or particular subgroups can be adequately resolved. This includes improvements in bioinformatics tools, sequence binning, normalization, and methods of physical separation, such as flow cytometry and single cell or colony sequencing.
- Metagenomics must be done in concert with improved culture-based science, including improved culturing techniques; generation of complete genome sequences for reference microbes (e.g., the type strains); and the physiological and ecological characterization of these reference organisms.

Gene-Centric Analyses

Because of the current technical limitations and cost associated with generating large amounts of DNA sequence from complex environments and because it will often not be possible to assemble complete or nearly complete genome sequences, it may be necessary in many metagenomics projects to adopt a *gene*-centric rather than a *genome*-centric view of microbial diversity and abundance. One example of a gene-centric approach is the use of environmental gene tags (EGTs), short sequences of DNA that contain fragments of functional genes. Each EGT in a metagenomics dataset may be derived from a different member of a given community, but those genes that are essential for community survival will in theory be represented more frequently, or at least more consistently, than ones that are nonessential or are highly specialized. The collective set of EGTs from a given sample represents a "fingerprint" that can be compared across multiple sites or habitats or over time in the same environment. EGTs that are overrepresented or underrepresented can provide insights into unique metabolic capabilities associated with a particular environment even if it is not possible to assign a given EGT to a particular species. Application of this approach to the metagenomics data from the Sargasso Sea revealed that the community is enriched in genes that encode rhodopsin-like proteins as compared with ocean environments that receive less sunlight (Venter et al. 2004). Newly developed sequencing technologies (see Table 4-2) that do not rely on cloning may be especially useful for identifying EGTs. The new sequencing technologies allow deep sequence coverage and are not subject to potential cloning biases. For identifying tags and quantifying relative gene stoichiometries, they may be particularly useful. To advance the use of gene-centric analysis:

- Non-genome-based methods are needed for the analysis of metagenomic data to identify capabilities that are present in a microbial community and deduce the ecological selection and evolutionary outcomes of the community.
- Improvements in bioinformatics tools, improvements in the ability to deduce function from sequence, and completion of more reference microbial genome sequences are needed.

Hybridization- and Array-Based Analyses

Specific functional gene arrays have also been designed with probes corresponding to genes of interest in an environment (see Figure 4-3) (Wu et al. 2006). They can indicate the diversity of genes performing specific functions at specific sites and assess levels of expression of those genes when community mRNAs (or cDNAs made from them) are the target. Research is needed to:

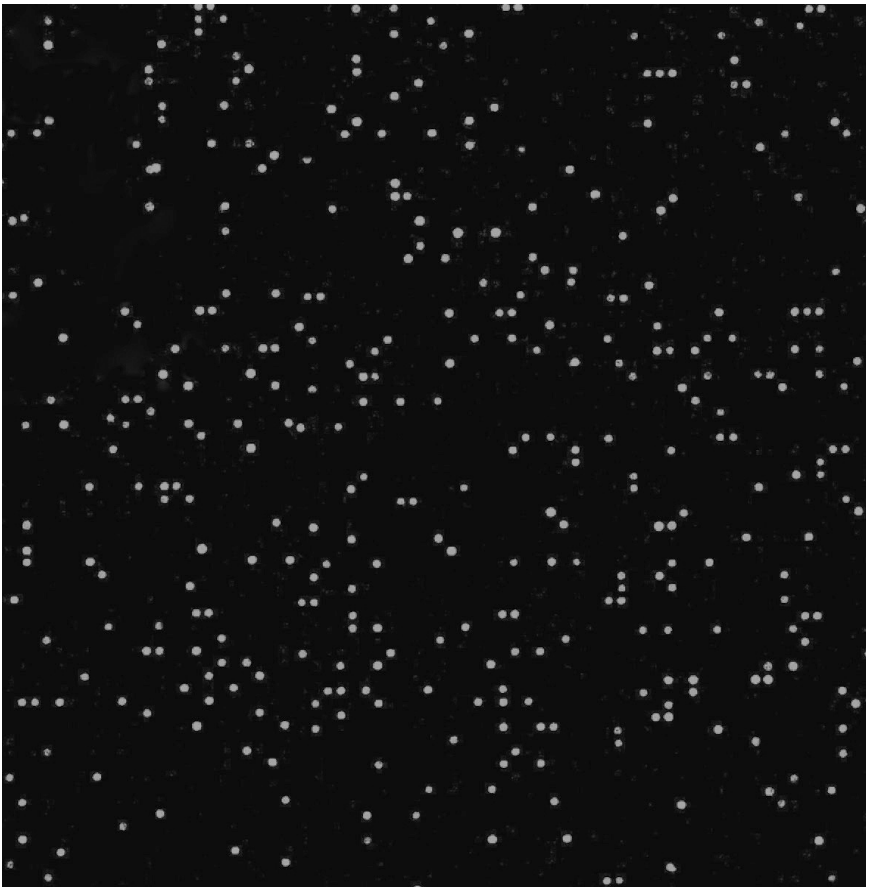


FIGURE 4-3 A functional gene array containing 27,000 probes covering 10,000 functional genes used to monitor microbial community dynamics in an aquifer undergoing uranium bioremediation. Image provided by Jizhong Zhou, University of Oklahoma.

- Improve array approaches for metagenomics applications, including sensitivity, interpreting specificity, speed, cost, and data analysis.
- Improve methods for sensitive and accurate representation and measurement of community mRNA populations.
- Enhance the database of annotated sequences of genes that have important environmental functions, and provide software for easy use in the analysis of metagenomic data and for probe and primer development.

Function-Based Analyses of Microbial Communities

If the ultimate goal of metagenomics is to determine "who is doing what," then sequencing alone is not the answer, because so many genes have unknown functions. Sequencing provides information that is limited by what is in the databases and by the available algorithms for linking sequence to function. Function is inferred when there is statistically significant sequence similarity between genes discovered by metagenomics and those in the databases. Computational tools that can predict secondary structure (how the protein will fold) and recognize a broader array of protein motifs based on the amino acid sequence alone are under development. Sequenced-based studies, however, will always be limited by the completeness of existing data and the accuracy of genome annotation. Furthermore, structural genomics projects that aim to improve the linking of sequence to function face a bottleneck: analysis can be done only on proteins that can be produced in large quantities, purified, and even crystallized. If a newly identified gene has only weak similarity to a gene whose product has been studied biochemically, if a similarity in sequence does not reflect a functional relationship, or if a particular gene can carry out multiple functions in the cell, sequence comparisons may lead to incorrect conclusions about function. Even if annotation and functional assignments were much improved, finding proteins with a defined function may be accomplished more efficiently by taking a functional approach to the metagenomic library.

One way to do this is to screen the metagenomic libraries directly for expressed functions. Function-driven metagenomics has unearthed many proteins that would not have been recognized by their sequences, including those coded for by genes involved in antibiotic biosynthesis, antibiotic resistance, biodegradation of environmental contaminants, and signal-transduction pathways. Finding these genes has increased what is known about the behavior of microbes, has enriched the databases, and has presented opportunities for biotechnology development. The potential for discovery is staggering: there are an estimated 10^{13} (10 trillion) genes in 1 g of soil; because these are derived from at least 10^3 species, there are at least 1 million *different* genes in 1 g of soil (Schloss and Handelsman 2006). Many of the genes will closely resemble genes in other organisms, including cultured and sequenced ones. But some will be novel—unrecognizable by sequence alone—and some will have dramatically new functions. This is one of the potential treasure troves of metagenomics.

Just as staggering as these potential riches are the barriers to discovery of genes by functional screening. The approach is grossly limited by the ability of the organism that is hosting the metagenomic library to express genes from anonymous organisms represented in the library. It is reasonable to imagine that most genes from members of the Enterobacteriaceae will be expressed in *E. coli* and that some, but fewer, genes from diverse organisms, including other γ-Proteobacteria, Firmicutes, Actinobacteria, and Archaea, will also be expressed in *E. coli.* But the variation in gene-expression machinery among microbial groups makes it likely that most genes from the most exotic divisions (those distant from *E. coli*) will not be expressed. Therefore, it is essential to develop techniques that enable *E. coli* to express a greater array of genes (such as providing alternative sigma factors or tRNAs) and to screen libraries in bacteria from other divisions.

- Functional-expression studies would be dramatically advanced by development of vectors and readily culturable host organisms from each phylum of Bacteria and Archaea.
- Research to discern the rules that govern heterologous gene expression will advance this field.
- Methods to expand the repertoire of genes expressed by surrogate hosts, such as *E. coli*, will contribute to function based metagenomics.

ADVANCING THE FIELD

The technical advances described in Table 4-3 need to be coupled with advances in bioinformatics (see Chapter 5) and basic microbiology. Genomic analysis to date has been valuable only because of five decades of comprehensive study of *E. coli* genetics, physiology, and biochemistry; intense study of many other organisms; and 150 years of microbial ecology research. It is imperative that microbiology remain strong and well funded to realize the potential of metagenomics (ASM 2007). This chapter concludes with a discussion of ways, both technological and scientific, in which progress will be most useful for advancing the field of metagenomics.

Sequencing Technology

One of the major challenges for metagenomics studies of complex environments is to capture the extent of bacterial diversity in a population with random-genome shotgun sequencing. As discussed above, without very extensive sequencing coverage of an environmental sample, the less abundant members of low-diversity bacterial communities will probably not be represented in any given dataset. With more complex communities, enormous amounts of DNA-sequence data will be required for assembly of even the most abundant members. Although advances leading to higher throughput and decreased costs for Sanger-based sequencing have occurred in the last 10 years, metagenomics projects will require new, higher-throughput, lower-cost sequencing technologies. For example, the pyrosequencing technology developed for the 454 Life Sciences Genome

TABLE 4-3 Technical Advances Needed in Functional Metagenomics

Current Limitation	Enabling Technical Advances
Not all clones can be expressed in current laboratory hosts; many functions are difficult to screen	Novel gene-expression systems that represent a broader array of organisms Better high-throughput screens
Inadequate number of reference genomes for many habitats. Habitat specific reference genomes would contribute to: • understanding the full metabolic capacity of specific community members and the interactions among community members and metabolic pathways • making it possible to bin sequence data to draw conclusions about microbial communities and environments on the basis of abundance of functional protein groups	Longer reads for 454 or Solexa sequencing Reference genome-sequence data for habitats of interest
Inability to culture organisms from which gene of interest arose	Further refine methods for culturing organisms
Difficulty in associating functions with metadata, such as physical conditions	Further development of methods for analysis of microbial community transcriptomes, proteomes, and metabolomes (which vary with physical conditions more directly than does sequence)
Inadequate information about minor members of communities, which is needed, for example, to identify keystone species	Development of improved methods for isolating single cells by microfluidics or cell sorting and for amplifying DNA and RNA from single cells; development of methods for subtraction and/or normalization of community DNA samples to facilitate the study of rare community members

Sequencer FLX eliminates the need for library construction (in other words, the community DNA can be sequenced directly, without first being cloned into a laboratory host) and can generate more than 100 million base pairs of DNA sequence in a single run. That is equivalent to about 100 runs on the AB 3730xl instrument for approximately 20% of the cost. The Solexa 1G Genome Analyzer also does not require library construction and yields about 1 billion base pairs of sequence per run. Neither of these technologies yet offer the ability to sequence "mate pairs" (see Table 4-2 footnote), which greatly aid the assembly of sequences, but this will be a feature of the Applied Biosystems SOLiD Analyzer, which is projected to have a throughput of about 3 billion base pairs per run when it comes to market in late 2007. Other technologies with similar throughputs are expected from, for example, Helicos, Intelligent-Bio-Systems, and Complete Genomics. In the longer run, a third generation of sequencing technologies, which use single DNA molecule substrates, are expected to reduce the cost even further and increase the throughput of DNA sequencing (Fan, Chee, and Gunderson 2006; Metzker 2005; Shendure et al. 2004).

A feature of the current generation of new sequencing technologies is short or relatively short read lengths (in comparison with Sanger capillary sequencing). Short read-lengths make it more difficult to assemble genomes. However, read-lengths will continue to increase with further development; moreover, the promise of sequencing paired-ends ("mate-pairs") will reduce the disadvantage of short read-lengths dramatically. It is impossible today to predict the advances, and cost, of sequencing technologies; they are changing too fast. It is sufficient to say that the needs of clinical medicine are driving technology development at a rapid pace, and metagenomics sstudies will benefit enormously from the consequent increase in throughput and reduction in cost.

A recent report described the first metagenomic analysis of two samples taken from the Soudan Iron Mine in Minnesota with 454 sequencing technology (Edwards et al. 2006). The analysis revealed interesting differences in metabolic potential between the two sampled environments; but, just as important, it suggested that the 454 sequencing data are remarkably similar to those generated from the same sample with Sanger sequencing, at least in terms of 16S rRNA sequences. Additional studies to validate the utility of the short reads clearly are warranted, but the initial data support the role of new sequencing technologies in future metagenomics studies because they will facilitate deeper sampling of environmental samples than is currently possible. At the same time, it is important that alternative strategies for enrichment of the less abundant members of communities, such as suppressive subtraction hybridization or flow sorting of cells, continue to be developed and implemented.

Gene-Expression Systems

Function-based metagenomics is predicated on expression of genes from anonymous organisms in a surrogate host. The likelihood of gene expression is low in any one host; thus, function-based approaches will be greatly facilitated by the development of vectors that can be maintained in a number of host species.

Aside from *E. coli*, several bacterial hosts are being developed to serve as vehicles for gene cloning in metagenomics. For example, the actinomycetes—which include genera such as *Corynebacterium*, *Myco-*

bacterium, *Nocardia*, and *Streptomyces*—have genomes that are highly GC rich. This group is well known for the production of such natural products as antibiotics, herbicides, and other secondary metabolites. If metagenomics is to be used as a means to discover new antibiotics, it is imperative that efforts to develop gene-expression systems for this group be increased. Streptomycetes are of much promise because they are easy to grow in the laboratory and are useful for the expression of genes from other actinomycetes. The gene-expression machinery of *Streptomyces* is adapted to a genome of high GC content, and their use as hosts for cloning of community DNA may facilitate expression of genes from other high GC content genomes (Wang et al. 2000). Another well-developed bacterial system for heterologous gene expression is *Bacillus subtilis*, which is a better system than *E. coli* for the expression of extracellular proteins (Li et al. 2004). *B. subtilis* is easy to manipulate and grows quickly, although it is known to exhibit plasmid instability, low-level gene expression, and degradation, by its native proteases, of heterologously expressed proteins (Li et al. 2004; Stephenson and Harwood 1998). Attempts have been made to solve the protein-degradation problem by creating protease-deficient *B. subtilis* hosts. The availability of heterologous gene-expression systems for archaea is limited, although there are well-developed gene-manipulation systems for the euryarchaeotes *Halobacterium salinarum* (Peck et al. 2000), *Methanosarcina acetivorans* (Metcalf et al. 1997), and *Methanococcus maripaludis* (Gardner and Whitman 1999) and to some extent for some species of *Sulfolobus* (Worthington et al. 2003), which are representatives of the crenarchaeotes. Development of shuttle vectors for cloning or for heterologous gene expression of the components of large DNA inserts in multiple microbial hosts will accelerate the quest to reap the fruits of metagenomics.

Single-Cell Analyses

A problem that has haunted microbial ecologists since the beginning of the field is the inability to account for minor members of a community. The abundance of species varies so widely that it is unlikely that the least abundant members will be captured in a given metagenomic analysis. Their DNA may well be in the libraries, but the probability of identifying, out of the millions of sequence fragments, the relatively few that came from the same low abundance community member, is low. Another aspect of the natural microbial world revealed from several metagenomics studies is the presence of microheterogeneity among organisms that are now typically grouped as members of the same species. The concept of a microbial pan-genome suggests that species can be defined as a core set of genes that are shared by all members, together with variable genes that differ from one isolate to the next. Thus, a single genome sequence is not sufficient to represent the range of diversity found within a single species. From a functional standpoint, such variability may be critically important in the overall metabolic potential of one community as compared to another. At first it may seem paradoxical to study single cells in order to understand communities, but in fact the function of any given community reflects the contributions of each of its members.

One novel technological development illustrated in Figure 4-4 is the ability to amplify DNA from single cells (or single members of a community) in order to study how they contribute to community function. The method is based on the multiple displacement amplification reaction that uses ϕ29 DNA polymerase and random primers for DNA amplification (Hellani et al. 2004). This technique alone, and with modification, has been applied successfully to the amplification of bacterial sequences from low abundance samples from natural environments with minimal amplifica-

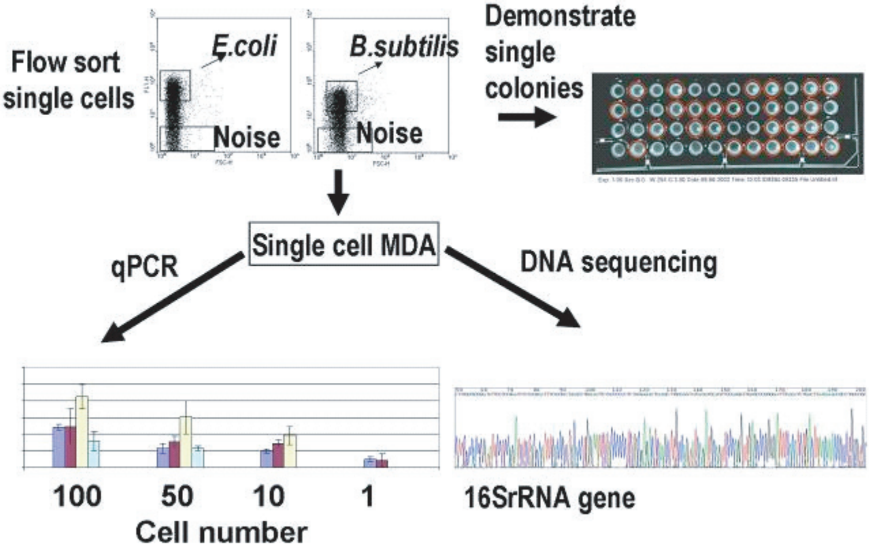


FIGURE 4-4 Molecular Displacement Amplification (MDA) of DNA from single cells. Flow sorted bacterial cells were plated to prove that single colonies could be reliably obtained. Then, having verified that flow sorting was reliable, the DNA from small numbers of cells (1-100) was amplified by MDA and tested in qPCR and DNA sequencing reactions (left panel). SOURCE: Roger S. Lasken, et al., Multiple Displacement Amplification from Single Bacterial Cells. In *Whole Genome Amplification*; Eds: Simon Hughes and Roger S. Lasken, Scion Publishing Ltd, Oxfordshire, UK.

tion bias. The ability to isolate single cells with microfluidics coupled with technology to amplify genomic DNA from single cells will revolutionize the study of unculturable species and the microheterogeneity within species. A number of approaches are being explored to facilitate the capture of information from individual bacteria or minority populations, including fluorescence-activated cell sorting, which can give useful information about size and functional chromophores (Zhang et al. 2006); growing mixed microbial communities in porous microbead columns so that each organism grows as a separate clone (Zengler et al. 2002; Green and Keller 2006); and single cell sequencing by dilution followed by amplification with strand displacement polymerase followed by debranching (Zhang et al. 2006).

The ability to amplify faithfully from single molecules has several applications and advantages relative to shotgun sequencing. Targeted amplification can be done in a way that retains quantitative information or can be normalized to emphasize the rarer species. As a result, information on haplotype or multiple chromosomes per cell can be retained, information about cells (and viruses) bound in pairs or larger aggregates can be captured (yielding data on symbiotic, parasitic, predatory, and other relations), and rare cells can be enriched with a simple single-amplified-cell prescreen (e.g., rRNA) followed by whole genome sequencing. Another rationale for developing methods for capturing and studying single cells in microbial communities reflects the fact that cells in communities are not randomly distributed, but in many cases form highly ordered assemblages of cells whose spatial orientation is essential for proper community function (biofilms on the tooth surface, for example).

Methods for Culturing Uncultured Species

Because the assembly of complete genome sequences is one of the major current limitations in metagenomics research, microbiologists are displaying renewed interest in the art of microbial cultivation. The most notable example is the cultivation of SAR11, a representative strain contained within a ubiquitous and dominant clade of marine heterotrophs, all of which had proved recalcitrant to cultivation for many years. Success in growing this organism was achieved with a combination of high-throughput (microtiter-plate) cultivation techniques using dilute media and rapid and sensitive screening using fluorescent probes specific for the SAR11 cluster (Rappe et al. 2002). Members of other ubiquitous microbial groups poorly represented in culture collections have since been isolated by tweaking of cultivation conditions—even by such simple adjustments as the use of solid vs liquid formulations, reducing nutrient and mineral concentrations, or increasing incubation time (Janssen et al. 2002; Rappe et al. 2002; Sait et al. 2002).

Basic Microbiology

The ultimate goal of metagenomics is to understand the structure and function of microbial communities. It depends on fundamental information about how microbial cells work in isolation and in populations and communities. The vast yield of information from metagenomic analyses conducted thus far has been built on more than a century of intensive study of microbes in pure culture. Recognition of genes based on sequence and making sense of genomic data, functional expression, and phylogenetic analysis depend on more detailed genetic, physiological, and ecological understanding of microbes in the laboratory. The many genes whose functions are not known, even in such well-understood microbes as *E. coli*, indicates the dearth of knowledge. Imperfect understanding of how gene expression machinery differs among species limits the power of function-based metagenomic analysis. And the lack of principles of microbial behavior in communities presents a wall that affects the depth of understanding that can be gleaned from metagenomics studies. Therefore, it is essential for the health and advancement of metagenomics that the study of microbes remains diverse and strong. Basic understanding of genetics, metabolism, gene regulation, cell structure, and responses to the environment needs to advance to aid in the design of metagenomic research strategies and the interpretation of metagenomic data.

Understanding Microbial Habitats and Collecting Metadata

Although seemingly a small component of Figure 4-1, the "Describe environment" box is perhaps the cornerstone of metagenomics studies. Information about the environment is the foundation of all analyses of the genetic and functional data obtained from organisms. Metadata collection, storage, retrieval, and analysis were not required in prior genome-sequencing projects. There were plenty of data in those studies—billions of nucleotides, in fact—but beyond inferring identity, function, and comparison between organisms, little was required of them. In contrast, the goal of metagenomics is to tease out the correlations between species abundance and environment, to link common gene functions to disparate environments, to determine whether the same organisms do the same things in different environments. All these comparisons require sophisticated and fast bioinformatics methods. How to "do it right" is still a fair question and will require creativity and detailed examination by many of the brightest bioinformaticians (discussed further in the next chapter).

Collaboration between microbiologists, geologists, chemists, oceanographers, meteorologists, clinicians, and a host of other scientists will bring rich rewards in the information that can be gleaned from metagenomics datasets. Learning how to coordinate the different kinds of data collected and analyzed by scientists in so many disciplines is a major challenge, both conceptually and bioinformatically.

DOWNSTREAM DEVELOPMENT OF METAGENOMICS

Currently, metagenomics is heavily biased toward sequencing and its associated computational analyses, and pioneering functional analyses. While the current distribution of effort is appropriate for this initial exploratory phase, it will not be sufficient for the next phase of metagenomics, when more value will be desired from the sequence and its metadata. It is important to plan early for the mid- and longer-range development of the field so that both the researchers and agencies plan for and invest in new approaches and capabilities. Many of these downstream uses are difficult to predict in advance but the infrastructure for their encouragement and support can be established. It is important that the field not be slowed by overemphasis on massive sequencing without at least equivalent if not greater advances in other metagenomics approaches.

A 10-year trajectory of possible resource distribution would initially show a shift from emphasis on sequencing toward more computational development and analysis. Later, greater emphasis could be placed on such approaches as proteomics and transcriptomics, more in-depth analyses of metabolic and synthetic pathways (chemical bioinformatics, for example, could focus on detecting the genetic machinery for producing small biomolecules like signaling chemicals), and comprehensive knowledge building. The committee does not intend to deemphasize the importance of adequate sequencing resources, but to point out that the field will need to update its vision, tools, and goals continually, so that resources are appropriately divided between generating sequence and all of the other analytical and experimental approaches that comprise metagenomics.

5

Data Management and Bioinformatics Challenges of Metagenomics

Metagenomics studies are data-rich, rich both in the sheer amount of data and rich in complexity. Biologists now have over two decades of experience in handling and analyzing DNA sequence data, but these are mostly data on reasonably well understood structures—genes and complete genomes. We still do not comprehend the organizing principles of metagenomic data. The expected flood of sequence data from metagenomic studies therefore poses many new challenges that urgently need attention. Information from metagenomics studies will be fully exploited only if appropriate data-management and data-analysis methods are in place.

GENOMIC DATA

The rise of genomics has been characterized by technological and scientific innovations and by novel practices in data dissemination. With the coincident increase in computational power and electronic communication, they were critical for the success of the field of genomics. In the early 1980s the scientific community in Europe and the United States established archives of nucleic acid sequence data. This had several very important consequences. One was that the data were immediately accessible in a form suitable for computer analysis; another was that the data were freely available, without impediment to all researchers, be they in academe or industry. There are three nucleic acid sequence archives, all founded in the 1980s: GenBank, funded by the National Institutes of Health (NIH) through the National Library of Medicine; EMBL-Bank, funded by the European Molecular Biology Laboratory; and the DNA Databank of Japan

(DDBJ), funded by the Ministry of Education, Culture, Sports, Science and Technology of Japan. A formal collaboration between these bodies (the International Nucleotide Sequence Database Collaboration[1]) ensures that the contents of all three are effectively identical at any time. Major achievements of the INSDC have been to make the submission of nucleic acid sequence data to one of the three databases mandatory for publication of any scientific paper that reports new sequence data[2] and to define standards for such submissions. Today, all DNA sequencing done in the public sector is captured in the archives. Their extraordinary growth is shown in Figure 5-1.

It is no exaggeration to state that without the INSDC and the sequences stored in and made available through the collaborating databases, the success of the Human Genome Project and similar genome projects would have been impossible. These databases and the analytical tools whose development was encouraged by the free availability of data allow researchers to access the totality of the world's public DNA and protein sequence data. It is vital for the metagenomics community to continue to adhere to accepted standards with respect to the public deposition of data from community projects[3] and continue to encourage and enable the development of analytical tools and agreed-upon data-management practices (see also Chapter 6).

DNA sequence data submitted to the international archives are *processed* sequences, they are not the "raw" sequences directly from the sequencing machines. In the late 1990s researchers recognized that public access to the raw sequence "traces" would also be of great value. The National Center for Biotechnology Information, the Wellcome Trust, and the European Bioinformatics Institute (EBI) therefore established the Trace Archive[4] for these data. In December 2006, the Trace Archive contained over 1.4 billion traces from over 700 species. Despite the challenges arising from some of the new sequencing methods, timely deposition of raw sequence data to the Trace Archive by the metagenomics community will also be of great long-term community benefit.

The nucleic acid sequence data archives are a primary source of experimentally determined DNA and RNA sequences. Many types of analyses of genomes and individual genes, however, require protein sequences. Although historically these were experimentally determined, the great majority are now computationally predicted from DNA sequence data. This requires

[1] See *www.insdc.org*.

[2] See *http://www.insdc.org/files/documents/open_letter.txt*.

[3] As accepted by the Fort Lauderdale Agreement of 2003: *http://www.wellcome.ac.uk/assets/wtd003207.pdf*.

[4] *http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?; http://trace.ensembl.org/*.

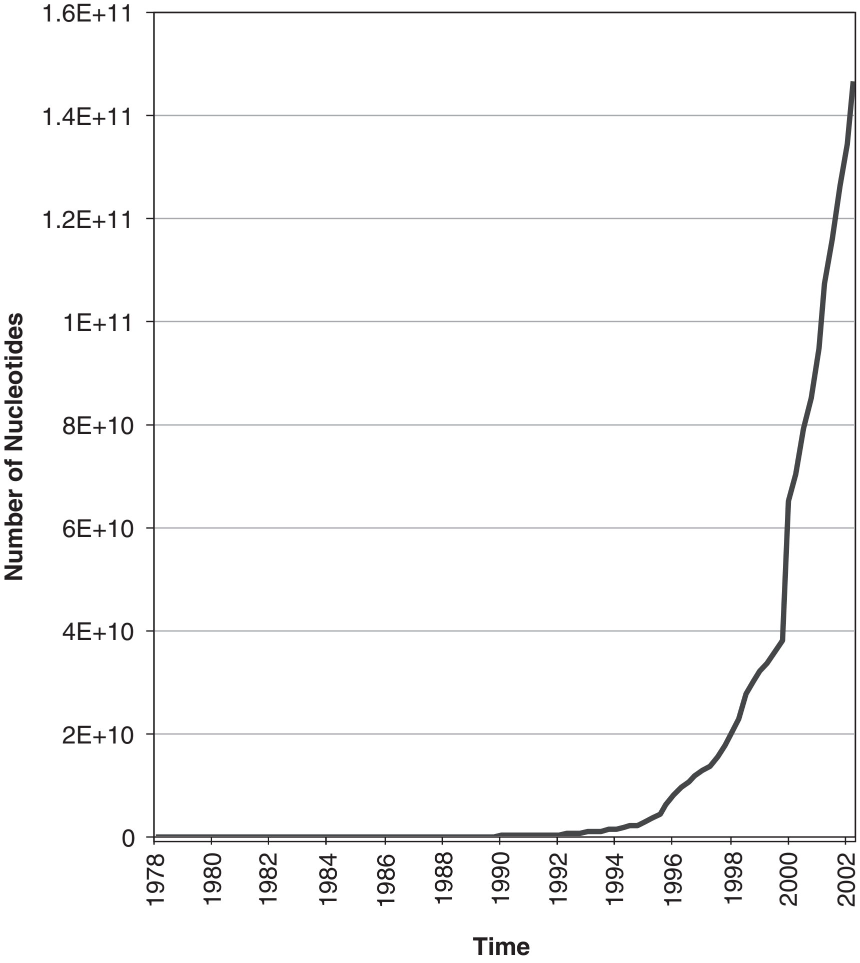


FIGURE 5-1 The growth in the size of the international sequence databases, since their inception in 1982. This graph shows the size of the databases at regular intervals, in numbers of nucleotides. (Data from the European Bioinformatics Institute.)

computational methods to predict which sequences constitute genes, that is, actually code for RNA and proteins. This is far from being a solved problem even for "complete" genomes. It will be even more difficult for the fragmentary sequences that will typically be obtained in metagenomics projects. Two databases, The Protein Information Resource and Swiss-Prot, were established as community resources for protein sequence data, in 1984 and 1986, respectively. They now collaborate closely to produce a common database of protein sequences—UniProt, a product of EBI, the Swiss Bioinformatics Institute, and the National Biomedical Research Foundation at Georgetown University. UniProt is not a primary database, but rather a highly curated database of protein sequences, the vast majority of which are derived computationally from gene models in the nucleic sequence data archive. Not surprisingly, the growth of UniProt has been slower than that of the nucleic acid sequence archive (see Figure 5-2).

METAGENOMIC DATA

In principle, there are no differences between DNA sequence data from "conventional" and metagenomics sequence projects. In both cases, the sequences are simply strings of the four bases A, T, C, and G. In practice, however, metagenomic sequence data require particular infrastructures for management and analysis (see Figure 5-3).

The growth of public DNA sequence data over the last 2 decades has been exponential, with a doubling time of about 14 months. Although predictions must be treated with some caution, early experience with metagenomics projects suggests that they will have a substantially shorter doubling time. Indeed, even the preliminary data from the Global Ocean Survey more than quadrupled the existing (predicted) protein-coding open reading frames (although this increase will be smaller when the GOS data are curated). The predicted growth of metagenomics sequences will result both from the new sequencing technologies now available (see below) and from the fact that metagenomic sequencing is not redundant. Conventional genomic sequencing projects are highly redundant. Multiple coverage of a genome is necessary for assembling a genomic sequence and ensuring accuracy. For example, the human genome was originally sequenced to a "depth," that is, with a redundancy, of 4.5-fold. By contrast, metagenomic sequencing is far less redundant. A consequence is that many more DNA and (predicted) protein data are being generated for the same effort: rather than sequencing the *same* genome 10 times, 10 times as many data are being generated for different sequences. There are other consequences of the relative lack of redundancy of metagenomic sequences.

The major reasons for such high redundancy in conventional genomic sequencing were to achieve "complete" coverage of a genome and to ensure the accuracy of the sequence. For metagenomes, most sequence reads will be unique; instead of several overlapping fragments of the same underlying sequence, there will be on average only one fragment. Therefore, metagenomic sequence data have an intrinsically higher error rate than genomic sequencing data. Moreover, it is now clear that many metagenomics projects will rely on novel sequencing technologies for data collection (see

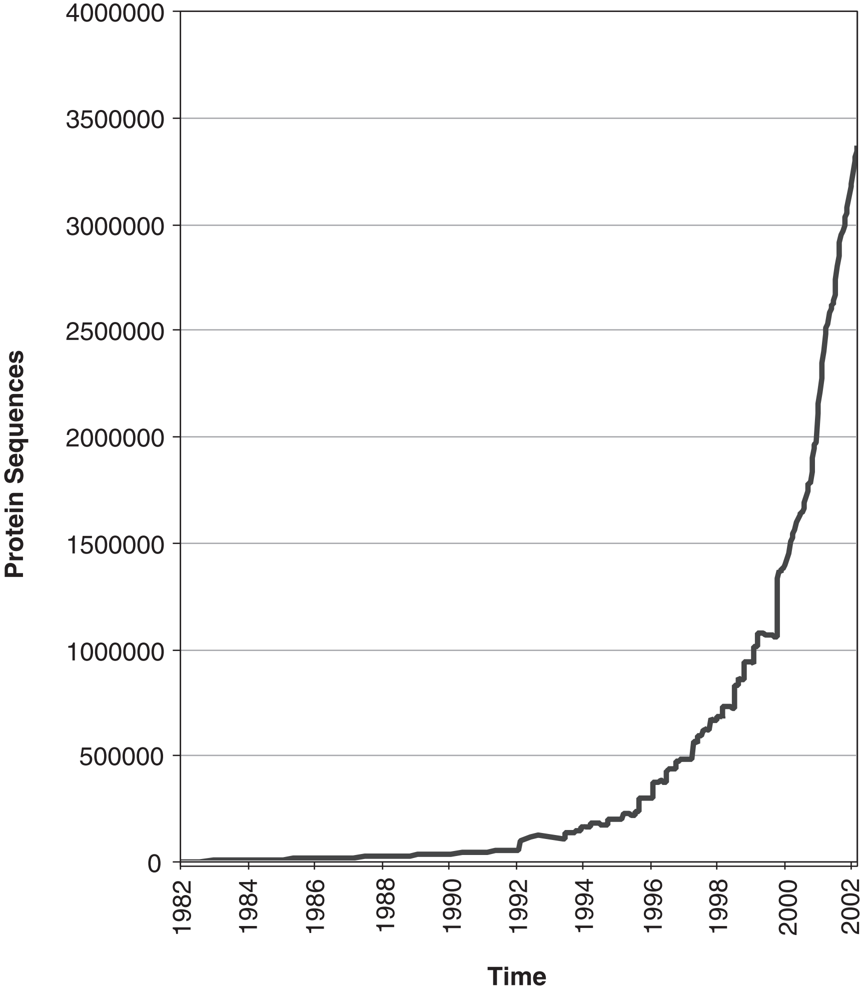


FIGURE 5-2 The growth of known and predicted protein sequences since the founding of Swiss-Prot in 1986. This includes all protein sequences, both those that have been curated and those that have not. (Data from the European Bioinformatics Institute.)

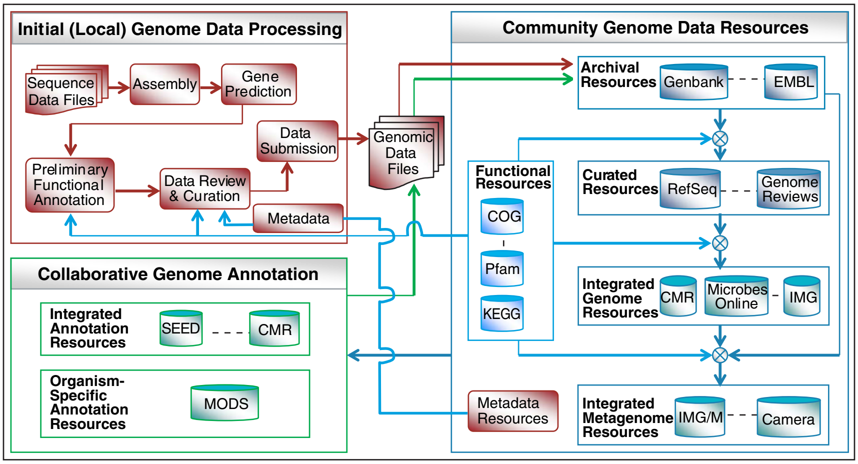


FIGURE 5-3 A typical life cycle of microbial genomic and metagenomic sequence data. This illustrates the pipelines of initial sequence data assembly and annotation, the integration of these data with annotation from other community resources, both integrative and species-specific the deposition of the annotated data into the archives of Genbank, EMBL-Bank and DDBJ; and the derivation of a number of specific integrated community data sets. With thanks to Victor Markowitz, whose original was modified.

Chapter 4). All the new-generation technologies produce sequence read lengths that are short—25-200 bases compared with 800-1000 bases for Sanger capillary sequencing technologies. Moreover, some of the new technologies are rather error-prone. These characteristics of the new technologies and the fact that within any study the sequences may be derived from many different organisms make the assembly of long sequences from the primary sequence data difficult, if not impossible.

THE IMPORTANCE OF METADATA

Metadata are data about data (Gray et al. 2005). They are also about biology. Metadata are the descriptions of sampling sites and habitats that provide the context for sequence information. Metadata are of great importance for metagenomic sequence data for two reasons. First, only by fully describing the samples from which metagenomics sequences have been obtained can one have any possibility of replicating a study. Samples from environmental or biological sources can never be fully replicated, but it is important that samples be sufficiently well described for an independent researcher to have the possibility of resampling. Second, metadata are essential for the analysis of metagenomic sequence data. Metagenomic sequence data that lack an environmental context have no value.

There is an urgent need for the community to agree on what metadata must be included with the submission of any metagenomic sequence data. Appropriate metadata will depend on the type of metagenomics sample. For example, the metadata to be associated with a human gut sample will differ from that to be associated with an ocean sample. Without wishing to determine now what these metadata should be, some examples can be given: [5]

- Detailed, three-dimensional geographic location of the sample, including depth (for water sampling) or height (for land and air samples).
- The general features of the environment of the sample, such as ocean, soil, mine, human, or insect.
- Specific features of the sample site, such as chemical data (pH, salinity, and so on), physical data (temperature, incident light, and so on), time when the sample was taken, and host condition, diet, and habitat.
- Method of sampling, size of sample, and sample preparation.

Some of these data would be collected or recorded when the DNA sample was being collected, and others could be retrieved from other databases, including geospatial databases and weather or ocean-current databases.

The community must address these issues with a sense of urgency. If metagenomic sequence data are to be used to their fullest advantage, a metadata infrastructure, which defines the data that are to be collected and their semantics, is an urgent need. As indicated above, no single metadata standard will be appropriate for all samples. Nevertheless, close collaboration and coordination among the communities developing metadata standards, nationally and internationally, will be important. Much can be learned from standards initiatives in related communities, for example, those of the Microarray Gene Expression Data (MGED) Society.[6] It is relevant that such standards are increasingly adopted and required by the major scientific journals.[7]

[5] See, for example, the database developed by the International Census of Marine Microbes (*http://icomm.mbl.edu/microbis*) and the efforts of the Genomics Standards Consortium (*http://darwin.nox.ac.uk/gsc/gcat*).

[6] See *http://www.mged.org/*.

[7] See *http://www.nature.com/nature/journal/v419/n6905/full/419323a.html*.

DATABASES FOR METAGENOMIC DATA

Absorbing the sequence that will be generated by metagenomics projects will be a challenge for the nucleic acid sequence data archive (GenBank, EMBL-Bank, and DDBJ) (see Box 5-1). But in addition to the archiving challenge, it is clear that the community will require new, secondary databases if the data are to be used effectively. Only a specialized database will be able to offer consistent storage and querying of the rich metadata that metagenomic sequences need. The analysis of metagenomic sequences will require computational programs that are best offered in the context of a specialized database, for example, programs for the clustering of metagenomic sequence reads or for time-series analysis. And metagenomic data must be integrated with data from different projects, such as satellite data on ocean temperatures and time-series data on changes in ocean salinity.

Two large projects have recently been initiated to build an infrastructure for metagenomic sequences and associated metadata. One is the CAMERA project,[8] a joint venture of the University of California, San Diego and the J. Craig Venter Institute in Rockville, MD. CAMERA's objective is to provide cyberinfrastructure tools and resources and bioinformatics expertise to enable the community to use metagenomic data. CAMERA will make raw environmental sequence data and their associated metadata accessible with pre-computed search results and access to high-performance computational resources.

At the Department of Energy's Joint Genome Institute (Walnut Creek, CA), an existing microbial genome database project, Integrated Microbial Genomes, is being extended to cope with metagenomics data in a project called IMG/M.[9] The objective of IMG/M is to integrate conventional microbial genomics data with data from metagenomics projects. In December 2006, it included data from over 680 microbial genomic projects, most of which were aimed at conventional complete genome sequencing.

Several smaller projects around the world have developed various specific data models and interfaces, often for specific metagenomic projects. Examples are the Micro-Mar[10] and MICROBIS[11] databases, in Alicante, Spain, and Woods Hole, MA, respectively.

There are interesting parallels between these projects and the genomic sequencing communities. In the latter, many of the major species being studied have special community genomics databases, for example, the *Saccharomyces* Genome Database for baker's yeast,[12] the Mouse Genome

[8] See *http://camera.calit2.net/*.

[9] See *http://img.jgi.doe.gov/m/doc/about_index.html*.

[10] See *http://egg.umh.es/micromar/index.php*.

[11] See *http://icomm.mbl.edu/microbis/*.

[12] See *http://yeastgenome.org/*.

BOX 5-1
The Metagenomic Data Deluge: Future Data Storage and Access Challenges

From the perspective of sequence data repositories, projected data storage needs for archiving Sanger-based capillary sequence data might not seem overly formidable. Every year disk space gets cheaper, with storage density increasing steadily. Hard drives have experienced a 50-million-fold increase in storage density since their invention. So, is there cause for concern for future metagenomic data storage and retrieval?

Projected future sequence DNA data storage challenges are more complex than simple extrapolation from today's Sanger-based capillary sequence production rates. There are three central reasons why data accumulation is expected to accelerate dramatically, and soon:

1. Technology. New sequencing technologies (see Table 4-2) are poised to increase data throughput and density substantially and pose new platform-specific data storage challenges. Since these platforms also enable individual labs to produce as much sequence data as did large production-scale centers in the past, the data storage and dissemination needs are expected to become even more acute.

The projected throughput of one newly emerging sequencing technology, Solexa, is as much as 10000 Mb per run, compared to 0.07 Mb per run on a Sanger-based capillary machine. Each Solexa run produces 1×10^{12} bytes of image data, which reduces to 1×10^{9} base pairs of raw data per run. Estimates from some sequencing centers suggest that sequence data production and storage needs per annum will approach 10 tera base pairs (Tb) of raw sequence data (1×10^{12}). This estimate does not consider the need for associated metadata (see below), which would increase storage needs by orders of magnitude.

2. Approach. Metagenomic survey approaches can now access vast amounts of biological "sequence space" for study, virtually instantaneously. The days of slower methodical sequencing efforts, one organism at a time, are changing rapidly. Metagenomics sequence datasets will soon dwarf all other sequence databases combined, even in the early stages of development. The metadata required for these data (below) will add to the data storage requirements dramatically.

3. Metadata density and complexity. The magnitude of metadata and associated storage needs for metagenomics datasets are greater than those for straightforward, single organism-based DNA sequencing efforts. Metadata are central and mandatory for metagenomics efforts, because they provide the context for data analyses and interpretation. Metadata are non-homogeneous and add complexity and density to the data storage and dissemination challenge. For example, a single organism's genome requires 1×10^{7} bytes for the raw DNA sequence storage, increasing to 1×10^{10} bytes when sequence annotation is added. By contrast, 1×10^{7} bytes of metagenomic sequence from a single sample with its associated metadata might require 1×10^{12} bytes of storage. Simple data storage projections from DNA alone are deceptive, unless they take these annotation and metadata storage requirements into account.

Database for the laboratory mouse,[13] FlyBase for the fruitfly *Drosophila,*[14] and TAIR for *Arabidopsis.*[15]

These model organism databases are publicly funded (usually by the NIH or the National Science Foundation) and add value to the sequence data deposited in the GenBank, EMBL-Bank, and DDBJ archives. CAMERA, IMG/M, and similar projects promise to be "model organism databases" for metagenomes. Such databases will be essential if data are to be used to the greatest advantage by the scientific and biomedical communities. Different databases will doubtless be required for different needs. Cooperation and collaboration between them, especially in the development of standards for the description of data will be necessary. Not only will it be necessary for databases to include metadata about habitat and sample treatment, it will also be critical to document how the raw data has been processed, filtered, and analyzed. Maintenance and curation of metagenomics databases will greatly add to their value, but are expensive and will require consistent support. Funding for databases requires a different approach than that for research projects: there need to be mechanisms for long-term funding, coupled with community oversight and evaluation. The experiences of the National Center for Ecological Analysis and Synthesis[16] and the National Evolutionary Synthesis Center[17] in providing a community focus for data integration and analyses are examples the metagenomics community might wish to follow.

SOFTWARE

The analysis of genomic data depends on computer software. In general, grants for metagenomics projects will require an even higher percentage of funds for bioinformatic and statistical support than have conventional genomics studies or than may be typical of other kinds of biological research. It is important that appropriate new software be developed, conform to industry standards, and be well documented. The investment in the development of software needs to be timely. If the analytical needs of biologists are still uncertain, a major investment in robust software engineering is premature; but once an analytical technique becomes generally accepted and useful, investment in making the software more user-friendly and reliable is worthwhile. That pipeline is poorly supported by traditional grant-funding mechanisms. Funding agencies should consider a competi-

[13] See *http://www.informatics.jax.org/.*

[14] See *http://flybase.org/.*

[15] See *http://www.arabidopsis.org/.*

[16] *http://www.nceas.ucsb.edu/.*

[17] *http://www.nescent.org/.*

tive funding opportunity providing software engineering support to bring individually developed software programs that have found wide use in the community up to robust, engineered, documented form. Such a program would allow a variety of individual approaches to developing software, with community assessment of the software's value.

ANALYSIS OF METAGENOMIC SEQUENCE DATA

Data from metagenomics projects share features that will require the development of novel computational tools and perhaps a new paradigm for the analysis of DNA data. In genome projects, the organization of the DNA in the organism was well known—a circular chromosome and plasmids in bacteria and multiple chromosomes in eukaryotes. The goal was to recover a sequence of a complete genome of a single organism. In metagenomics, we do not have a clear model of the organization of the DNA in the sample. We do know that each sample contains many different organisms, bacteria, viruses, and small eukaryotes. The different organisms will have different genome structures, linear or circular chromosomes, single or multiple chromosomes, and extrachromosomal elements, and these characteristics will be unknown at the time of data analysis. It is likely that even the most extensive sequencing of a specific sample will provide only partial sampling of the DNA in a given environment; therefore, the data in the sample may have to be used to predict features of the sample, rather than analyzing the features themselves. As an illustration of the complexity of metagenomic sequence analysis we illustrate, in Figure 5-3, an exemplar of a metagenomics project's data "life cycle."

There will be a need to analyze sequences in the context of their metadata, including independent environmental data, such as meteorological and oceanographic. Thus, the analyses that will be done on these sequences will be different from those done on conventional genome sequence data and will involve many questions that will combine data of various types (see Box 5-2).

Because of the unusual features of metagenomic data (fragmentation and high error rate), new computer algorithms and data models will be needed for the clustering of metagenomic sequences (both DNA and predicted peptide sequences) to characterize microbial communities from sequence data and to analyze changes in microbial communities over time. Novel techniques for the visualization of complex data will be needed.

Another major difference between data-management needs for metagenomics projects and those for conventional genomics problems will be the demand for continuing community input into data annotation. In conventional genomics, primary responsibility for annotating data falls on the authors, and this creates inconsistencies in the databases when old

BOX 5-2
Examples of Questions That Illustrate the Utility of Metadata

- Do microbial gene richness and evenness patterns (at some specific sampling density) correlate with other environmental characteristics?
- Which microbial phylotypes or functional guilds co-occur with high statistical probability in different environments?
- Do specific phylotypes track particular geographic or physico-chemical clines (latitudes, isotherms, isopycnals, and so on)?
- Do specific microbial community open reading frames (functionally identified or not) track specific bioenergetic gradients (solar, geothermal, digestive tracts, and so on)?
- What is the percentage of genes with a given role, as a function of some physical feature, such as the average temperature, of the sample sites?
- Do microbial community protein families, amino acid content, or sequence motifs vary systemically as a function of habitat of origin?
- Are specific protein sequence motifs characteristic of specific habitats?

annotations are not updated and thus become inconsistent with new ones. Although there is now a mechanism (called third party annotation[18]) for the community to annotate genomic sequences in GenBank, EMBL-Bank, and DDBJ, the original authors' annotations, even if outdated, remain as primary annotations seen in the database. For instance, annotations added through curation at the appropriate model organism database are only very slowly being incorporated into central databases. In metagenomics projects, where many types of annotations would become possible only after additional data (or metadata) are collected by other groups, an annotation database must be able to accept and integrate both individual and large-scale (computational) annotations of metagenomic data and able to integrate them in a transparent way for their user communities. The need for dynamic and flexible annotation will require ongoing, professional curation—another reason that long-term database funding will be important.

It will be seen that the scientific community will be presented with challenges by the generation of metagenomics data. Many of the challenges will require a high degree of community organization and collaboration. Given the wide array of microbial communities that will be studied—from toxic waste sites to agricultural soil to the human mouth—the interested scientific community will be extremely diverse, and coordination will be

[18] *http://www.ncbi.nlm.nih.gov/Genbank/TPA.html.*

more difficult. No existing body can take the lead to ensure that it occurs. However, the Microbe Project, a US government interagency group, has the appropriate broad membership to facilitate coordination and communication among the interested scientific communities (see Chapter 6).

6

The Institutional Landscape for Metagenomics: New Science, New Challenges

MAJOR STAKEHOLDERS IN METAGENOMICS

The Scientific Community

In new fields, such as metagenomics, the scientific societies are logical foci to build grassroots interest, implement knowledge exchange, and facilitate planning. The societies should organize events to build knowledge and foster communication in their fields and especially among other relevant disciplines. They can provide leadership to build consensus, communicate potential benefits to the public, and facilitate the establishment of leadership groups to foster the coordination and development of metagenomics on a broad scale, both nationally and internationally. One of the successes of the *Arabidopsis* genome project (see below) is that it was organized by the scientific community working in concert with the funding agencies. Metagenomics would be well served by using a similar organizational model.

Funding Agencies

In the United States, 12 federal agencies are members of the Microbe Project, an interagency working group formed in August 2000 under the aegis of the Subcommittee on Biotechnology of the National Science and Technology Council Committee on Science. The mission of the Microbe Project is "to maximize the opportunities offered by genome-enabled microbial science to benefit science and society, through coordinated interagency efforts to promote research, infrastructure development, education and outreach."[1] The 12 members of the Microbe Project are the Department of Agriculture, the Department of Defense, the Department of Energy, the Department of Homeland Security, the Department of the Interior US Geological Survey, the Environmental Protection Agency, the Food and Drug Administration, the National Aeronautics and Space Administration, the National Institutes of Health, the National Institute of Standards and Technology, the National Oceanic and Atmospheric Administration, and the National Science Foundation. These twelve agencies along with the Central Intelligence Agency and the Federal Bureau of Investigation are the federal agencies that because of their missions or responsibilities have benefited from genome-enabled microbiology and would be expected to benefit further from the advances of metagenomics. The Microbe Project's mission makes it ideally suited to convene the necessary working groups to advise on the specifics of the infrastructure needed to enable the science of metagenomics and to articulate a plan that coordinates responsibilities and funding to maximize efficiencies and capture the expected synergies in this new field. Several of the agencies have already funded metagenomics projects, some of which have become models that reveal the promise of the field. Each of the 14 agencies mentioned has its own missions and interests, but much synergy is to be gained by pooling common infrastructure needs, and this is a strong motivator for a well-coordinated effort at the federal level. The Microbe Project should coordinate its work with the scientific societies to involve the scientific community in the development of the field.

Other organizations are and will be interested in funding metagenomics, including foundations with national or international interests, the private sector, and some state agencies. Large projects can have several partners that contribute on large or small scales for targeted components or for general core funding. It is possible to imagine metagenomics projects supported by several countries, funding agencies, and private foundations. Mechanisms will have to be worked out to ensure proper representation and credit while avoiding hindrances of the general goal of work for the public good.

International Coordination

The large-scale nature of metagenomics and the international interest in the field suggest that there will be interest in and value to be derived from international coordination from the beginning. Some metagenomics projects are under way in the European Community, Canada, China, Brazil, Singapore, South Korea, and Japan. Many of the projects have interest in

[1] *http://www.microbeproject.gov/.*

similar but not identical habitats or focal questions. All, however, could benefit from some common infrastructure—most notably metagenomics databases and new analysis tools but also new sampling strategies and data standards, to name the most obvious.

As a first step in addressing how human metagenomics studies might be approached on an international scale, a panel of 75 participants (scientists, physicians, industry representatives, and administrators from funding agencies) from Asia, the Americas, and Europe met in Paris in October 2005 to discuss the feasibility of sequencing the human intestinal metagenome, its importance for human health and industry, possible technical approaches, and possible funding scenarios.[2] The meeting generated a framework for an International Human Gut Metagenome Initiative, including recommendations to generate reference genome sequence data from approximately 1000 gut bacterial species that can be cultured, to develop techniques for sequencing microorganisms that cannot be cultured, and to classify genes of the microbial community based on metagenomic sequencing. Since this meeting in the fall of 2005, a trans-institute NIH committee has been assembled to discuss in more detail its participation in an international human metagenome project. The recent call for proposals under the European Union 7th Framework Programme includes the characterization and variability of the microbial communities in the human body as one of its areas of focus.

International coordination would help to ensure greater efficiency and less duplication of effort, but it should not restrict creativity or the national interests of any country. Besides helping to plan and develop common infrastructure, international coordination would ensure wide communication of ongoing projects and results so that new projects were not undertaken without knowledge of the global landscape. Furthermore, if a few major metagenomics projects are to be undertaken comprehensively and in great depth, they will be more successful if the breadth and resources of the international science and engineering communities are exploited.

The initiation of international coordination is best left to the interested scientific communities—particularly interested scientists and their societies—in communication with national funding agencies. As noted above, the organizational model of the *Arabidopsis* project is useful.

EDUCATION AND TRAINING

Metagenomics will draw on expertise from many disciplines and individuals:

[2] *http://www.international.inra.fr/research/mapping_the_human_intestinal_microbiome.*

- Those with knowledge of microbiology, including microbial genetics, biochemistry, physiology, pathology, systematics, ecology, and evolution.
- Other biologists, including molecular and cellular biologists and those with knowledge of host organisms, such as humans and other mammals, plants, insects, and microbial hosts with important roles in nature or of economic importance.
- Those with knowledge of the environment, including soil and atmosphere scientists, geologists, oceanographers, hydrologists, and ecosystem scientists.
- Computational scientists, including those with knowledge of statistics, computer science, data mining and visualization, database development, modeling, and applied mathematics.
- Those with expertise in scaling information to large ecosystems, and in evaluating the effects of global change and its interface with policy.
- Engineers, physical scientists, and chemists whose skills and insights are potentially field-transforming in their contribution to new methods, chemistry, devices and applications (within and beyond metagenomics), and the understanding of complexity, networks, and system structure.

Metagenomics as defined here is much more than DNA sequences and engages all the "omics" and a broader, microbial-community-based systems biology. To reach the understanding that metagenomics will make possible, new education and training programs will be needed. Experts in a broad array of fields must be integrated into metagenomics projects and provided with appropriate cross-disciplinary knowledge so that their specific expertise can be made the most of and their contributions disseminated to the wider community.

As mentioned in Chapters 4 and 5, metagenomics probably will require proportionally more contributions from computational and bioinformatics scientists than any other field of biology. Hence, it is imperative that this workforce requirement be addressed immediately. It is not easy to identify computational scientists or biologists who have both the interest and the talent in the kind of cross-training that metagenomics projects will require. We recommend establishing several types of training programs to encourage scientists to develop the needed skills. Several mechanisms have been successful in providing cross-discipline training: interdisciplinary training to augment traditional graduate programs, summer courses patterned after the Cold Spring Harbor or Marine Biological Laboratories summer courses, and post-doctoral fellowship programs in which fellows undertake training in new disciplines. Support for faculty to attend metagenomics workshops or to spend sabbaticals in metagenomics research laboratories or facilities would also be beneficial in expanding appropriate training environments.

As described earlier, although metagenomics has similarities to genomics as currently practiced, it also has important differences in the types of data and in questions to be asked, so it is important to recognize that the components and expectations of current genomics training programs will not suffice for metagenomics.

OTHER INSTITUTIONAL ISSUES

Data Release

The rapid release of sequence information has been an important and sometimes contentious issue in genome-sequencing projects. Proponents of rapid release of data cite the relatively long timeframe of sequencing projects and the ability to derive important information even from incomplete data. Opponents of rapid release emphasize the need of those doing the sequencing to have time to analyze and publish the results of their own work before others have publication opportunities. Intellectual-property issues also arise; rapid release of information into the public domain may bar the opportunity to obtain some types of intellectual-property rights.

Data release was a contentious matter in the early days of large-scale sequencing projects. Two meetings, one in Bermuda in February 1996, and one in Fort Lauderdale, FL, in January 2003, grappled with the issues and published recommendations to the community,[3] which were adopted by the major funding agencies, including NIH.[4] At Fort Lauderdale, projects that were funded as community resources were specifically defined: "A 'community resource project' is a research project specifically devised and implemented to create a set of data, reagents or other material whose primary utility will be as a resource for the broad scientific community." Data from such projects should be released immediately for free and unrestricted use by the community. Obligations, however, were imposed on the users of such data, with respect to recognizing the data providers' legitimate interests in publishing and analyzing the data, and in acknowledging the data providers as the source of the data. The Committee's view is that these policies have served the community well and should be explicitly adhered to by metagenomics researchers.

The Fort Lauderdale Agreement recognizes that these policies may not necessarily be appropriate for projects funded by grants to individual investigators, where providing a community resource is not the primary

[3] Bermuda Agreement: *http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml.*

Fort Lauderdale Agreement: *http://www.wellcome.ac.uk/assets/wtd003207.pdf.*

[4] *http://www.genome.gov/page.cfm?pageID=10000910.*

goal. Recognizing that most of the major funding agencies now have data access policies in place,[5] we express the view that even single-investigator projects should be expected to practice release after a specified time period e.g., 6 months.

Intellectual Property

Many companies in several markets tap into the value found in natural resources, and metagenomics constitutes a new way to access natural resources. Advances in DNA and expression technologies provide opportunities to overcome supply issues that in the past limited the value of natural products. The more advanced the technologies become, the more value will be derived and the less destructive sampling of biological materials will become.

The pharmaceutical market is especially large and hence illustrates the potential for intellectual property. Global revenues of this market in 2004 were over US$500 billion; sales in North America, Europe, and Japan made up about 80% of the total. It is estimated that 62% of oncology drugs are derived in some way from natural products (Newman et al. 2003).

Global revenues in industrial, agricultural and healthcare biotechnology in 2004 were $54.6 billion; the United States dominated with 78% of global revenue. Products in this market include enzymes for the textile, detergent, food and feed, and personal-care industries. Many small companies are in the enzyme market, but it is dominated by such large companies as Novozymes and Danisco, which have programs to identify new products by sampling microbes in the environment.

Key patents in metagenomics may affect the ability of researchers to practice some methods of metagenomics. US patents have been issued that claim methods of isolating DNA directly from a mixed population of organisms. These patents may be determined to be infringed by some who are using the metagenomics methods without a license. The uncertainty of the situation poses additional risks to any who seek to commercialize findings arising from metagenomics studies. Patent issues are also associated with bioprospecting (collecting biological material) outside the United States. The Convention on Biological Diversity (see next section) requirement for benefit-sharing poses some threats to the intellectual-property rights of those who wish to commercialize findings from metagenomics studies of samples from outside the United States. Patent issues associated with the convention also may influence full disclosure and information release. New

[5] e.g., *http://www.genome.gov/10506537*; *https://compbio.ornl.gov/microbial/Data_release_policy.htm.*

patent legislation in and outside the United States may require statements of the country of origin about compositions derived from biological sources.

The ownership of genetic resources outside national jurisdictions is uncertain. Collection of samples in areas outside national borders—for example, in deep-sea vents beyond national jurisdictions—is unregulated by international policies. International organizations increasingly recognize the need for such policies (Arico and Salpin 2005).

Metagenomics and the Convention on Biological Diversity

The collection of samples within national borders is now guided by the Convention on Biological Diversity (CBD). Metagenomics studies rely on sample collection, and it will be important that researchers comply with the CBD to prevent charges of "biopiracy."

The CBD is an international treaty that was adopted at the 1992 United Nations Conference on Environment and Development in Rio de Janeiro, Brazil. The three stated goals of the CBD are the conservation of biological diversity, the sustainable use of biological components, and the fair and equitable sharing of benefits arising from use of genetic resources (US Senate Committee on Foreign Relations 1994). The main points of the treaty establish the sovereign rights of states to their own natural resources and make access to the biological resources subject to national rules and legislations. The treaty imposes expectations of access by prior informed consent and of nations' fair and equitable sharing in the benefits of commercial exploitation of their biological resources. Each party nation is expected to establish legislation and policies regarding access and benefit-sharing (ABS). Adequate protection of intellectual property rights is granted, but the treaty also creates expectations that developing nations will be granted access to technology arising from the use of their biological resources, and this can lead to challenges to the rights typically granted by patents. In response to perceived commercial risks and uncertainties, the Biotechnology Industry Organization has created guidelines for bioprospecting.[6] The 1992 Rio "Earth Summit" resulted in over 150 governments signing the CBD, including the United States. More than 187 countries (not including the United States) later ratified the agreement, providing global support and acceptance of the treaty. (In 1994, the US Senate Committee on Foreign Relations approved the treaty, but the full Senate failed to ratify it.)

The UN, which administers the treaty, continues to discuss issues raised by the CBD. A CBD working group on ABS created a December 2005 report that assessed the impact of the CBD policies on the commercial use of biodiversity. The report highlights the importance of the diversity of microbes

[6] *http://bio.org/ip/nternational/200507memo.asp.*

and various pharmaceutical and biotechnology companies' interest in it. The December 2005 CBD report points out that recent metagenomics studies pose "a host of new questions and challenges with regard to access and benefit-sharing, in particular relating to the sovereignty of microbes and the difficulties of ascribing ownership" (Laird and Wynberg 2005). The report confirms that the CBD has changed practices in corporations seeking to exploit biological diversity: "larger or socially responsible companies do not generally consider genetic resources freely available." The difficulties in negotiations between commercial entities and nations are highlighted in the report. Tensions surround the differences between value expectations made by diversity providers and companies' valuations based on commercialization costs. The report also describes regulatory confusion and uncertain policies that hinder the commercial exploitation of biological resources (Laird and Wynberg 2005).

Differences in perspectives regarding intellectual-property protection and rights are often at the center of discussions of ABS. Some nations are increasingly trying to introduce "disclosure of origin" as part of the patent-application process. In the short term, metagenomics has the potential to tap into substantial microbial diversity without venturing abroad. However, as research expands, the unresolved issues raised by the CBD will probably influence metagenomics research. It may be prudent for funding agencies to establish formal sections of proposals in which investigators need to specify how they will comply with the CBD when sampling outside US borders. This would increase awareness of CBD issues and help to protect the agencies against international disputes related to funded research. It should be remembered by all parties that there can be considerable mutual benefit to science and education in well-structured, collaborative, international metagenomics projects.

Biosafety

Metagenomics projects will require the sequencing of DNA arising from unknown organisms with unknown potential for causing human, plant, and animal disease. In that respect, metagenomics projects are much like traditional microbiology, tapping into the unknown microbial diversity in various environments. However, in contrast with traditional microbiology, the end result is DNA clones or DNA sequences, rather than living microbes. Metagenomics projects would thus appear to have fewer biosafety issues than traditional microbiology.

Biological safety levels (BSLs) guidelines for undertaking traditional microbiological and recombinant DNA studies are relatively clear. Because of the potential for cloning genes that might have human health consequences, it would be prudent to undertake metagenomics studies in a BSL2 safety environment whenever pathogenic organisms might be present in significant numbers. There may be some circumstances in which metagenomics studies would be best performed under additional safety standards, such as cloning from an environment that might harbor virulent pathogens, but those circumstances are expected to be uncommon. If the source DNA is considered according to the probability of recovering virulence genes, existing biosafety guidelines appear to be suitable for metagenomics projects.

Outreach

Metagenomics is the kind of accessible and expansive science that can capture the public's imagination. Metagenomics research provides a special opportunity to teach microbiology to the public and train a new generation of scientists to be sophisticated and effective scientific communicators who can bring the thrill of discovery to the public.

7

A Balanced Portfolio: Multi-Scale Projects in the "Global Metagenomics Initiative"

THE VISION

The opportunity afforded by metagenomics to study microbial communities in their natural state represents a vast frontier. Given the intense competition for science funding, some priority-setting is necessary to ensure that the most possible value is gained from early metagenomics investments. The diversity of habitats on Earth, the complexity of microbial communities, and the myriad functions governed by microbes suggest that highly productive metagenomics research will be possible in decentralized, small-project settings. However, no individual researcher is likely to have the capability and resources to achieve a comprehensive characterization of a complex microbial community. Therefore, there is also a substantial need for medium-sized, collaborative projects that involve multiple investigators. Small- and medium-sized projects are familiar to funding agencies and the scientific community in the form of single-investigator grants (National Institutes of Health [NIH] R01s, for example) and interdisciplinary collaborations (National Science Foundation [NSF] and the US Department of Agriculture [USDA] Microbial Observatories; NSF's Long Term Ecological Research [LTER], Frontiers in Integrative Biological Research [FIBR] and National Ecological Observatory Network [NEON] programs; the US Department of Energy's [DOE] GTL program, for example). Both mechanisms of funding are tested and proven effective in advancing new fields of science. The mixture of single- and multi-investigator projects maximizes the diversity of scientific approaches, assures that many avenues of research are pursued simultaneously, presents an opportunity to study many habitats, and engages a broad community, thereby utilizing the creativity of many investigators. All these benefits are essential for the advancement of the field.

Metagenomics, however, differs from much of the science that precedes it in its complexity, multidisciplinarity, and in the magnitude of its unknowns. Its very nature departs from each of the fields—microbiology, ecology, and genomics—that fuse to form this new science. Consequently, metagenomics presents a number of conceptual and technical obstacles that limit the productivity of all metagenomics researchers (detailed in Chapters 4 and 5). The committee believes that the needs of the metagenomics field are not entirely met by current funding mechanisms, and the most efficient way to boost the effectiveness of the field overall is to augment small- and medium-sized projects with a small number of large-scale projects.

The Global Metagenomics Initiative is envisioned to capture all three types of projects—small-, medium-, and large-scale. Familiar mechanisms are available for the first two, so this chapter will detail the characteristics of the large-scale projects; issues that should be considered in evaluating proposals for small and medium-sized projects have been discussed in the previous chapters, as have infrastructural needs that affect metagenomics research at all scales (the need for, software development, database curation, and access to sequencing capacity, for example).

Much as the Human Genome Project drove advances in methods and technology, the large-scale projects will lead the development of broad principles and new technologies and methods that are more easily conceived and validated in the context of a multidimensional and highly replicated study than in traditional single-investigator projects. The large-scale projects will also offer special opportunities for public outreach and training of a new generation of scientists. There is excellent precedent in the genomics field to suggest that large-scale projects provide benefits far beyond the data gathered. Providing a community data resource was the initial motivation, but the Human Genome Project and other model-organism genome projects have also spurred technological advances and inspired the development of new tools, common standards, and shared software resources. This chapter will argue that the potential value of large-scale metagenomics projects is substantial.

CHARACTERISTICS OF SUCCESSFUL LARGE-SCALE PROJECTS

A recent Institute of Medicine-National Research Council report examining large-scale projects in biomedical science set forth the following reasons for undertaking a large-scale project (Nass et al. 2003):

- "A major intent of such projects is to enable the progress of smaller projects."
- "Large-scale collaborative projects may also complement smaller projects by achieving an important, complex goal that could not be accomplished through the traditional model of single-investigator, small-scale research."
- "The objective of a large-scale project should be to produce a public good—an end project that is valuable for society and is useful to many or all investigators in the field."
- "Unconventional large-scale projects take advantage of economies of scale to produce relatively standardized data on entire classes or categories of biological questions . . . they may reveal novel areas of research for follow-up by smaller science projects, and they also provide essential tools and databases for subsequent research."

The committee believes that, if carefully chosen and planned, large-scale metagenomics projects will have all of these characteristics.

WHY METAGENOMICS NEEDS A "BIG SCIENCE" COMPONENT

Metagenomics has great promise, but is challenged by the extreme complexity of microbial communities, by the lack of sufficient data on many aspects of microbial communities (such as diversity and conservation of structure or function across geographic location) to support valid generalizations and, because of these factors, by the lack of unifying ecological principles that enable predictive modeling. Put simply, it is hard to derive general principles from very few specific cases.

Table 7-1 lists a number of challenges, each of which would require substantial investment to address in depth. The knowledge needed can be obtained best in concerted, multi-investigator efforts. Although many individual-investigator-led and small-group collaborations in metagenomics have been successful, none has been able to generate sufficient data to allow comprehensive understanding of a complex microbial community or to invest the time and effort needed for the development of new tools and methods. For example, the assembly of individual genomes from metagenomic sequence information has been achieved only in the acid mine drainage project. A large-scale project could bring to bear a multipronged attack on the challenge of assembly in a complex community: redundant, deep sequencing; whole-genome sequencing of numerous community members as scaffolds; cell-sorting and single-cell analysis techniques; and analytical tool development and conceptual advances. The progress made would be available to individual researchers applying metagenomics in a plethora of environments.

TABLE 7-1 Challenges Facing Metagenomics

Challenge	Questions to Be Answered	Possible Strategies
Complexity and unknown structure of microbial communities	How much sampling is enough? What is a representative sample?	Sample a complex community to completion, that is until few or no new species are collected with further sampling Develop new mathematical models that can predict species richness and community structure so that the representativeness of samples can be evaluated
Methodological biases	What taxonomic groups are not accessed with the methods used? What habitats are not accessible with current technology?	Apply multiple methods to the same samples to assess the biases of each Systematically survey diverse habitats and assess access to microbes and their DNA
Improved correlation of phylogenetic analysis and community function	What roles does each taxon play in community structure and function? Can generalizations be made about these roles? Do communities always have definable functions?	Develop mathematical tools to establish associations between phylogeny and function Develop ecological methods to remove specific community members and study the effects on structure and function Explore broader definitions of function
Habitat variation and conservation	On what scale should habitats be studied? What are the limits of habitats? In what ways is an example of a habitat representative of other examples of the same habitat? Are there core characteristics associated with every member of a type of habitat (that is, is there a set of traits required to live in soil)? Do all human guts share a core community? Which is more highly conserved, the taxa making up a community, or the community function? Or is there coconservation?	Conduct a worldwide sampling of many habitats of the same type and compare exhaustive descriptions of membership and function Develop statistical methods that identify similarities at both taxonomic and functional levels Compare variability between similar communities at different sites and in the same site at different times Implement clustering methods for enumeration and identification of community types and representative diagnostic taxa

continued

TABLE 7-1 Continued

Challenge	Questions to Be Answered	Possible Strategies
Metagenome assembly	What are the rules of metagenome assembly? What "binning" techniques are most useful in assigning sequences to taxa? How much assembly is necessary to make sense of a community? How can microdiversity (many similar genomes at one site) be handled?	Reassemble numerous metagenomes of various levels of complexity and extract common features and principles to construct a method and a set of rules for assembly Select a few communities whose metagenomes have been assembled and study their structure and function in sufficient detail to determine how much and in which ways assembly contributes to ecological understanding (what organism is doing what)?
Functional analysis	Are there rules that guide the choice of expression system for function-based analysis? Are there ways to increase the probability of finding a particular function (such as choice of habitat or expression system)?	Conduct global studies to correlate frequency of expression of particular characteristics with analysis parameters Develop new gene expression systems for all phyla of bacteria and archaea Map functional diversity to community type and map both to phylogenetic diversity Correlate functions with extensive physical and biological metadata

Similarly, a large-scale project could advance the coupling of large sequencing databases with functional analysis. Massive sequencing has been conducted on samples from the Sargasso Sea and the Global Ocean Survey, but the metagenomic libraries from these environments have not been subjected to functional-expression assays. Conversely, a number of functional-expression studies have been conducted on soils for which there is not a rich base of sequence information. A global project might tackle one of these habitats from many angles—sequencing, functional-expression analysis, genome reassembly, deep phylogenetic analysis, hybridization-based screening, and much more. A large-scale project could involve investigators in many disciplines such as genomicists, statisticians, geneticists, physicians, and sociologists. The genomicists would have different expertise: one might be an expert in the habitat itself who could establish the strategy for collecting relevant metadata, another might be skilled in handling sequence data, and another might be experienced in functional screening. Working together, such a team could make substantial progress in understanding the functional potential of the genetic repertoire of a microbial community, predicting function from sequence and developing new tools for functional screening and database mining and management. The resulting rich store of new knowledge would greatly improve the yield of information from smaller studies.

WHAT KIND OF LARGE-SCALE PROJECTS IN THE GLOBAL METAGENOMICS INITIATIVE AND HOW MANY?

Careful consideration must be given to the choice of projects for the large-scale portion of the Global Metagenomics Initiative. The challenges posed by metagenomics depend on the habitat being studied. No large-scale project would be able to address all the challenges. In broad terms, there are three types of habitats on Earth: unmanaged landscape and aquatic environments (such as seawater, soil, and sediments), managed ecosystems with a directed function (such as sewage treatment, bioremediation, and bioreaction), and host-associated habitats (such as the human gut, plant roots, and insect symbionts). Because the scientific knowledge and practical benefits to be gained differ among environments, the committee believes that three very different communities should be chosen for in-depth analysis.

Sampling challenges differ among the habitats because the sources of variability are different. The challenges associated with DNA extraction also differ. Host DNA is the most important contaminant in host-associated communities, whereas tannins, humic acids, polysaccharides, and other compounds are the dominant contaminants in environmental samples. Different organism genomes will be needed as scaffolds to facilitate assembly and for functional and evolutionary interpretations. To some degree, statistical methods will apply to all habitats, but the differences in community membership, size, structure, and complexity create different needs for analysis. Perhaps the most important difference in studies of diverse habitats is the type of metadata needed to make sense of genomic sequence data. A global effort is needed to develop standards of and methods for gathering metadata. In the human gut, for example, the host's diet, genotype, and age will probably be critical; in an environmental sample, global positioning, meteorological, chemical, and physical data are likely to be needed. Information about habitat will also often need to include historical trends in these variables. Interoperable but separate model community databases would be the most efficient framework in which to develop the specific tools necessary to analyze data from the different environments and thereby maximize the utility of the data. Consequently, the committee believes that the greatest gains would ensue from including one example of each of the three types of habitats in the Global Metagenomics Initiative's large-scale projects.

EXPECTED BENEFITS OF LARGE-SCALE METAGENOMICS PROJECTS

The large-scale projects will bring benefits to the field that cannot be achieved with small-scale research. The benefits can be described, broadly, as contributing to ecological theory and principles, understanding of specific habitats and functions, technical advancement of the field, and international collaboration and training.

Theory and Principles

Large-scale projects that engage researchers in many locations and disciplines could reveal the principles of microbial community ecology through intensive studies. For example, whereas a small-scale project might aim to study the distribution of cellulases in the rumen, a large-scale study might attempt to provide a nearly complete inventory of the members of the rumen, assemble some of the members' genomes, identify cellulases and other traits important to that community's function and the animal's feed efficiency, and assess the variation of all these characteristics among many animals and perhaps among ruminant species.

Some community behaviors will be peculiar to each community, but some will be governed by universal principles that can be derived by studying a few communities in great detail. Once those principles are derived, they can be tested with more focused experiments in small-scale studies to assess the degree to which they can be generalized. The proposal to create large-scale projects in the Global Metagenomics Initiative is driven in part by the need for these principles. Just as studies of different microbial communities face different technical challenges, they also raise different theoretical issues:

- Study of a community in a natural environment would act as "proof of concept" for using metagenomics to understand the interaction between microbial communities and geochemical processes, eventually helping to understand change in global elemental cycles.
- Study of a host-associated community would probe the interaction between a microbial community and the physiology and health of its host.
- Study of a managed-environment community would seek to understand the effects of environmental change or human activity on microbial

communities and would have the potential to develop enough understanding to manage or mitigate environmental damage or maximize efficiency and sustainability of a bioreactor.

Each large-scale project would provide a comprehensive dataset about a particular habitat or function that could be the basis for building general theories and principles. The teams leading the large projects would need to communicate often because comparison among the three kinds of habitats could further illuminate global principles about the microbial world.

Understanding Specific Habitats

The committee anticipates that the large-scale projects will focus on habitats whose study has obvious and immediate benefits to society. In addition to contributing to broad theory, the large-scale projects would result in a comprehensive understanding of critical habitats at many levels. Full genome sequencing of organisms from a wide variety of phylogenetic groups represented in the three habitats should be an early focus of the large-scale projects; the resulting genomes would be an important resource for researchers in small and medium-sized projects. The chosen habitats should be of clear interest to the general public, and frequent public updates should be an integral part of each project. The funding agencies should encourage the development of strong outreach programs to the communities where the studies are being conducted. Due to the decentralized nature of the Global Metagenomics Initiative and its projects' geographic diversity, this would have a broad impact on the public's understanding of metagenomics and microbiology generally and would present an opportunity to train a new generation of scientists skilled in outreach and communication of science to the public.

Technical Advancement of the Field

Large-scale projects would unite scientists of multiple disciplines around the study of a particular habitat. These multidisciplinary groups would have the resources to develop new technical approaches useful to all metagenomics studies. The projects would also serve as incubators and evaluators of novel technologies, more precise and automated measures of conditions, and community databases and would equip smaller-scale projects with the knowledge to design efficient sampling schemes, make informed choices about habitats to study, and identify fruitful strategies for identifying specific functions.

The large-scale projects would offer an incomparable opportunity to lead the development of standards for data acquisition, management, and release. Few projects can focus on scientific questions while evaluating sampling methods, experimental design, and data analysis. Such an integration of biology and evaluation of the outcomes of various approaches would be a central mission of the large-scale projects.

The size of the large-scale projects would provide economies of scale for "omic" analyses and the development of computational tools and provide guidance for future movement toward or away from centralized facilities for sequencing and data analysis. Furthermore, the large-scale projects would provide an interdisciplinary community to lead novel downstream metagenomic analyses, perhaps including uses for structural biology, high-throughput "omics," new modeling of the evolutionary history of the early biosphere, and assessment of the current patterns and rate of evolutionary change. No doubt, metagenomic data will yield major approaches and questions that we cannot envisage today; these breakthroughs are best stimulated by large-scale projects.

International Collaboration and Training

The large-scale projects would require and enable collaboration and coordination that are difficult to achieve with single-investigator projects. Because they would be international and involve many investigators, they will require carefully considered and executed management plans and funding dedicated to fostering communication and promoting successful collaboration through scientific discourse. The large-scale projects would provide a unique setting for training a new community of young scientists who are skilled in collaboration and the execution of large-scale science. The nature of modern biology necessitates that at least some students have the skills to provide future leadership to international and multi-investigator projects as these become more prominent in biological research.

Thus, the large-scale projects would provide the intellectual environment and resources for the training of a new cadre of scientists to populate the field of metagenomics. In training, just as in research, the field would benefit from a healthy balance of large-scale and small-scale projects.

LEARNING FROM PREVIOUS LARGE-SCALE GENOMICS PROJECTS

Several collaborative research projects comparable with the proposed global metagenomics projects yielded important transformative science, such as the human and *Arabidopsis* genome sequencing efforts. An examination of the history of these projects reveals factors that proved to be crucial to their success.

The Human Genome Project

The Human Genome Project (HGP) provides an excellent window into the processes and pitfalls of "big science." The HGP required the collaborative management of a large-scale, international, interdisciplinary research project involving input from several independent research teams. Two critical lessons of this highly visible, highly successful effort can be noted. First, there was a clear goal for the collaborative project that all collaborators could embrace—the sequencing of the whole genome. The goal was:

- Specific in stating what would be done (sequence the human genome).
- Publicly understandable in terms of the benefit to society (human health).
- Time-bounded (within 15 years).
- Finite and with a specific associated cost estimate ($200 million per year for 15 years; $3 billion total).
- Wild and audacious (the goal was substantially beyond the technology that existed when the project was proposed).

Several intermediate end points were set, allowing the public and policy-makers to monitor progress of the project—such as completion of the physical map (two key maps in 1992 and 1994), completion of individual chromosomes (1999 and 2000), a draft genome (2001)—and then the "final" genome (2003). Effort could proceed in parallel at organizations around the world that contributed to the overall effort. Common data standards helped to enable this, and the discrete nature of chromosomes helped to organize the effort. Rapid data release and globally available databases ensured open sharing of information.

Second, the HGP devoted substantial resources to consensus building and coordination. It was an international collaboration involving 20 groups and funding from the United States, the United Kingdom, Japan, France, Germany, and China. Sequence data were contributed by many centers. The direction of the HGP was set by the major funders—the National Institutes of Health (NIH), the US Department of Energy (DOE) and the Wellcome Trust. They established mechanisms to assist with the coordination of research, in particular to avoid unnecessary competition or duplication of effort, and to coordinate research with parallel studies in model organisms; to coordinate and facilitate the exchange of data and biomaterials; and to encourage public debate and provide information and advice on the scientific, ethical, social, legal, and commercial implications of sequencing the human genome. The methods used by the funders to achieve collaboration and coordination included open "Bermuda" meetings, periodic international meetings and regular telephone conferences; rapid and unrestricted data release (all genomic sequence data were made publicly available without restriction within 24 hours of assembly); and data integration using a common software platform.

The *Arabidopsis* Genome Project

The *A. thaliana* genome sequencing project provides a slightly different perspective on how to establish and maintain such extensive, international collaborative research efforts. The Multinational Coordinated *Arabidopsis thaliana* Genome Research Project was conceived in 1990 by a small group of investigators who believed that such a project would have a profound enabling effect on the field of plant biology.

The *Arabidopsis* Genome Initiative (AGI) was formed to establish standards for sequencing accuracy and guidelines for data release, and to allocate workloads for each participant. The AGI was made up of representatives of the six research groups involved in sequencing the *A. thaliana* genome and played a key role in the oversight of the project. The group communicated regularly to deal with issues as they arose. Some members of the AGI lobbied for immediate data release as was the practice in the HGP, but there was considerable disagreement among the participants and their funders on this point, and public availability of data ranged from deposition of a draft sequence in GenBank within 24 hours of its generation to data release only when a sequence was finished and annotated. Data release was a subject of continuing discussion throughout the project, and the participants finally agreed to disagree about it. The AGI also played an important role in the final stages of the project when it became necessary to reallocate genome regions to centers that had finished their initial assignments ahead of schedule. This helped to ensure a steady flow of data and in part contributed to the completion of the project nearly 4 years ahead of schedule.

The project benefited from the additional oversight provided by the Science Steering Committee, composed of members of the *Arabidopsis* research community and representatives of some sequencing centers. A US Steering Committee was also established to facilitate interactions between the participating US laboratories, to serve as an additional link to the international efforts, to provide guidance on database issues, and to generate annual progress reports to the *Arabidopsis* community.

One of the things that set the *A. thaliana* genome project apart from other sequencing projects was that the scientists on the steering committees, not the representatives of the funding agencies, were empowered to make decisions on the overall management of the project. Agency representatives were, however, invited to all the meetings as observers and helped to ensure that the sequencing groups met their obligations.

Another aspect of the project that required coordination was annotation and data analysis. Although each of the sequencing groups was involved in annotating its regions of the genome, different methods were used to generate the information. It was decided that the ultimate goal was to provide the scientific community with a unified set of genome annotations. Through the implementation of open communication and clear procedures, a plan for a joint annotation effort between the Institute for Genomic Research and the Munich Information Center for Protein Sequences was established.

LESSONS FOR METAGENOMICS

The HGP and AGI provide valuable lessons for implementing a successful Global Metagenomics Initiative. Both projects benefited from having a clear goal, broadly accepted scientific and public benefits, and continuing coordination and communication among scientists and funding agencies.

To succeed, the large-scale projects in the Global Metagenomics Initiative would need to replicate these qualities. The initial challenges will be to develop a consensus around the choice of three microbial communities for in-depth study, to set clear goals, and to map out a program that establishes priorities and intermediate milestones. It will be important to identify model communities whose understanding would be of immediate and obvious public benefit. The human microbiome (its health implications) or ocean microbes (their role in the global carbon cycle) are two examples. Fully characterizing these communities is as daunting a task as decoding the human genome appeared to be 25 years ago. Consensus-building, planning, and staging will be necessary.

A PRELIMINARY ROAD MAP

Phase I: Choosing Model Communities

The first challenge that the scientific community needs to meet is to develop a consensus around the desirability of launching a few large-scale metagenomics projects and then to delineate the principles for selecting and recommending model communities. The choices in metagenomics are more daunting than choosing which model organisms to sequence. The broad categories of microbial communities are quite diverse: natural environments from ocean to soil to extreme habitats, such human-made environments as bioreactors of many types, and a vast array of host-associated microbial communities, from insect symbionts to the human microbiome. Each of these categories of environments offers different opportunities and challenges for metagenomics study, and thus it would probably be desirable to draw an in-depth, large-scale project from each. The broad scope of potential metagenomics study projects shows that metagenomics has much to offer in furtherance of the missions of many funding agencies, including the NSF, NIH, DOE, and the USDA. The choice of habitats to explore should be the product of discussions among members of the scientific community in a process that we recommend be initiated and coordinated by the Microbe Project Working Group. One way to enable such efforts is to support cross-disciplinary, international workshops to debate the principles to apply in choosing model communities and to debate how to establish and maintain multidisciplinary and multinational research efforts.

Alternatively, a process could be modeled after the NIH director's roadmap meetings in which five different groups of scientists with diverse perspectives each spent a day at NIH discussing topics for the NIH's long-term planning, called the Roadmap Initiative. The consensus ideas were then posted on the web for public comment, comments were collected, and decisions on themes were made based on the collective input of the scientific community.

The large-scale projects in the Global Metagenomics Initiative will be chosen, in part, based on the rationale for the habitats of choice. Basing choice of the habitats on the following criteria will ensure that the desired outcomes of the Global Metagenomics Initiative are identified and satisfied by the mix of systems that are selected. Three large-scale model microbial-community projects will probably be appropriate.

In each of the three broad categories (natural, host-associated, and managed communities), a specific model community will need to be chosen. The following criteria characterize communities that will yield the most useful data:

- A community in which there is some fundamental understanding of the major functions and roles of the microbes and in which there would be a distinct benefit in improving that understanding, such as the microbial community colonizing the human gut or oral cavity.
- A community of moderate complexity that is well characterized by environmental or geological criteria and that can be systematically sampled over long durations, such as those found in seasonally variable, depth-stratified lakes, hypersaline ponds, and low-nutrient oceans.
- A community whose members can be well characterized by current sequencing technologies so as to make it possible to address fundamental questions of how the community is organized and stabilized, that is, of an appropriate level of complexity and where eukaryotes play a minimal role in community dynamics.
- A community whose variation based on physical/geo/chemical

characteristics can be resolved by reproducible sampling, such as a soil community colonizing a winter wheat crop.

- A community to which a particular treatment can be applied so that factors that shape community structure and function can be tested, such as intestinal tracts, bioreactors, streams, or soils.

Model metagenomics communities should be chosen to leverage past knowledge and current research. Several funding agencies target long-term investigation of particular communities or environments, including NSF's LTER Network,[1] NSF's and USDA's Microbial Observatories,[2] and NSF's NEON.[3]

Phase I would include one or more workshops to develop a consensus on at least three and perhaps up to 10 communities as possible objects of a large-scale project. The workshops would define a clear goal and end point for each project and elucidate the expected public benefit of achieving the goals. Examples of possible projects are listed in Table 7-2, but the committee emphasizes that these are not prescriptive; the choice of projects would be best determined with a great deal of community input.

Phase II: Planning and Initial Data-Gathering

Once the communities have been chosen, Phase II would begin with a peer-reviewed, competitive process wherein groups of scientists submit interdisciplinary proposals for planning projects. The proposals would be evaluated according to the criteria presented in this report and any further criteria developed in the Phase I workshops. Planning proposals would be expected not only to address scientific issues but to outline project management, including coordination, milestones, oversight, data management and release, intellectual property, training, and public outreach.

Project-planning awards will support a year of meetings of the international group to hone hypotheses, approaches, and methods and to support the gathering of the baseline information needed to pursue the chosen hypotheses. Baseline information might include low-depth sequence of the habitat, phylogenetic analysis of the community, an assessment of variation among samples or locations, complete genome sequencing of culturable members of the community, or development of hybridization arrays, expression systems, or high-throughput assays. Establishment of a strong bioinformatics team would also take place during the 1-year planning

[1] *http://www.lternet.edu/.*

[2] *http://www.csrees.usda.gov/fo/fundview.cfm?fonum=1460; http://www.nsf.gov/pubs/2005/nsf05600/nsf05600.htm.*

[3] *http://www.neoninc.org/.*

TABLE 7-2 The Global Metagenomics Initiative: Examples of Large-Scale Projects

Habitat	Approaches	Knowledge Derived
Microbial community associated with the human body	Sample the intestinal microbiota of people in many locations who consume diverse diets and have diverse genetics and lifestyles Characterize the metagenome of the community with phylogenetic techniques, functional analysis, and massive sequencing	Determine whether there is a "core metagenome" of the human gut, a core community that is found in every person Describe the extent of variation in communities at various location in the human gut and between individuals Develop correlations between physiological conditions (health and disease, diet, and lifestyle) and microbial community structure and function in the human gut
Microbial community in an unmanaged habitat (such as soil or seawater)	Conduct extensive sampling over space and time Conduct an extensive analysis of 16S rRNA sequences (>200,000) Produce extensive sequence information about the metagenome in soils under different regimes Conduct a function-based analysis of the metagenome of soil under each regime	Establish relationships between seasonal and daily cycles, structure, and function of microbial community Determine how environmental change affects substrate use, polymer degradation, and secondary metabolite production Describe the distribution of characteristics in the community
Microbial community associated with managed ecosystems that perform a service (such as bioremediation or sludge processing)	Conduct an extensive analysis of 16S rRNA sequences Construct extensive metagenomic communities from habitat in various locations; characterize with sequencing and functional analysis Sample over time, including when the community is fully operative and when it functionally collapses	Establish relationships between particular functions or members of the communities and community persistence and collapse Identify organisms, traits, or chemical conditions that prevent or reverse community collapse Develop and implement interventions for experiments and improved performance

phase, and the team would define the tools needed to test hypotheses (for example, the association of phylogeny and function, integration of medical records and genomic information, searches for rare motifs, or pattern-recognition algorithms).

Phase III: Implementation

Phase III proposals would be submitted after the planning period, when sufficient baseline information and preliminary development were achieved. Phase III proposals would present a strategy that included evidence that all the methods needed are in hand, that the variation is known so that sampling strategies can be developed, and that the experimental design is carefully matched to the questions asked. Deep sequencing in a habitat would occur during Phase III, as would site-to-site comparisons, testing of hypotheses that are central to developing principles of microbial ecology, and potential new downstream uses of the metagenomics data in later years. The project Web site would provide up-to-date information about the project and direct viewers to the sequences and metadata that have been released. Phase III projects should be designed for a 10-year period with periodic review to achieve the larger-scale goals.

CONCLUSION

Undertaking three model, large-scale metagenomics projects in which the chosen environments can be characterized at great depth from a variety of perspectives would profoundly advance the field. No investigator can bring to bear all the different approaches that will be necessary to begin to understand the complex physical, chemical, genetic, metabolic, and environmental interactions that are taking place in even a moderately complex microbial community. The insights derived and the tools developed by a large, multidisciplinary group would be immediately useful to the wider community of investigators. If the model communities are carefully chosen, such large-scale projects would have obvious, major societal benefits. The Human Genome Project captured society's imagination with the promise of a deeper understanding of the basis of human health. Well-chosen metagenomics initiatives could similarly inspire with the promise of understanding of the microbial communities that contribute not only to human health but to the health of the biosphere (see Box 7-1).

The projects outlined in Table 7-2 would furnish the statistical and biological power to support conclusions that cannot be drawn from smaller-scale projects that lack enough breadth of sampling to be representative or

BOX 7-1
Key Outcomes of Large-Scale Projects in the Global Metagenomics Initiative

- Broad principles and unifying theory for microbial-community ecology.
- Large-scale, intensive studies of important habitats or questions.
- Methods of broad applicability to metagenomics research, both basic and applied.
- Massive contributions to metagenomics databases.
- Standards for data acquisition, management, and release in the field of metagenomics.
- Lessons about economies of scale in metagenomics research.
- International cooperation and collaboration.
- Training for young people in the conduct and management of international, collaborative “big science” projects.
- An opportunity to share science with the public and train graduate students to do so effectively.

enough depth of analysis to assess the variation within and between sites with precision.

The large-scale projects would be “virtual” centers. They would include scientists at many locations in the world to maximize the scientific diversity of the project team. Communication would be achieved by frequent meetings in person and by videoconference or other technology that becomes available during the course of the projects. The projects would probably need to be sustained for 10 years; so changes in personnel and participating institutions during a project’s lifetime would be expected.

8

Recommendations

The scope of metagenomics is vast. Defining the metagenomic characteristics of microbial communities in the biosphere is a critical first step in understanding their contributions to the health of the planet, their roles in the well-being of humans, and the environmental consequences of human activities. Because so little is known about microbial communities, the potential for discovery is great in any habitat chosen for study. The committee identified eight potential opportunities in different application areas that can be addressed with metagenomics:

- **Earth Sciences:** The development of genome-based microbial ecosystem models to describe and predict global environmental processes, change, and sustainability.
- **Life Sciences:** The advancement of new theory and predictive capabilities in community-based microbial biology, ecology, and evolution.
- **Biomedical Sciences:** The definition, on a global scale, of the contributions of the human microbiome to health and disease in individuals and populations and the development of novel treatments based on this knowledge.
- **Energy:** The development of microbial systems and processes for new bioenergy resources that will be more economical, environmentally sustainable, and resilient in the face of disruption by world events.
- **Environmental Remediation:** The development of tools for monitoring environmental damage at all levels (from climate change to leaking gas-storage tanks) and microbially based (green) methods for restoring the health of an ecosystem.

- **Biotechnology:** The identification and exploitation of the biosynthetic and biocatalytic capacities of microbial communities to generate beneficial industrial, food, and health products (pharmaceuticals, antibiotics, and probiotics).
- **Agriculture:** The development of more effective and comprehensive methods for early detection of threats to food production (crop and animal diseases) and food safety (monitoring and early detection of dangerous microbial contaminants) and the development of management practices that maximize the benefit from microbial communities in and around domestic plants and animals.
- **Biodefense and Microbial Forensics:** the development of more effective vaccines and therapeutics against potential bioterror agents, the deployment of genomic biosensors to monitor microbial ecosystems for known and potential pathogens, and the ability to precisely identify and characterize microbes that have played a role in war, terrorism, and crime events, thus contributing to discovering the source of the microbes and the party responsible for their use.

Meeting these challenges will require progress on several fronts. Technological, methodological, computational, and conceptual advances will be needed to develop the potential of metagenomics fully.

Furthermore, as microbiologists turn their attention to the study of microbes in their natural environments, it is likely that many of biology's most basic organizing concepts will be affected by deeper understanding of life at the microbial level. Metagenomics will probably expand answers to questions like, What is a species?, What is the role of microbes in maintaining the health of their host?, How diverse is life?, and What ecological and evolutionary roles do viruses play? The metagenomics approach is uniquely well suited to gathering the information necessary to make progress on such basic conceptual questions.

FINDING 1

The opportunity intrinsic to a new frontier of science is accompanied by new challenges that were not anticipated by prior research. Metagenomics is no exception. Current metagenomics researchers face several difficulties and obstacles. Early metagenomics studies have been able to survey the metagenomes of complex microbial communities, but have been able to characterize in depth only the simplest communities. Generating massive sequence databases is not the limiting step; using the databases to determine the complete genomes of community members and to understand a community's metabolic capabilities or potential responses to environmental change is still beyond the field's capabilities in even moderately complex environments. A number of technological, methodological, and computational advances are needed for metagenomics to reach its full potential. Encouraged by the example of the human and other model organism genome projects, the committee believes that the best way to spur these advances is through a multi-scale approach that includes support for small, single-investigator projects; medium-sized, multi-investigator projects; and large-scale, multidisciplinary, multinational metagenomics projects.

The *small-scale projects* will ensure that creative contributions are solicited from a broad scientific community and engage many scientists in metagenomics. The *medium-sized projects* will provide centers of study that unite diverse techniques and disciplines to study numerous habitats encompassing diverse organisms, scientific questions, and technical challenges. The *large-scale projects* will characterize a few microbial habitats in great depth, using large multidisciplinary and multinational teams to address challenges in metagenomics that require massive datasets or highly diversified scientific approaches and engaging more investigators than would typically participate in a medium-sized center. The large-scale projects will cross national lines, facilitating study of many examples of a habitat worldwide, thereby generating sufficient data to develop generalizations about the communities that reside in that habitat.

The large-scale projects will also provide an excellent opportunity for young biologists to gain experience in participating in “big science” and global partnerships. And they offer unique opportunities for public outreach and stimulation of public interest in science because they will highlight the ability of metagenomics to explore a new biological frontier.

The medium-sized projects need to be funded through organizational models that recognize habitat differences, such as the National Science Foundation’s (NSF) Long Term Ecological Research (LTER) and National Ecological Observatory Network (NEON) programs. Similarly, the National Institutes of Health (NIH) recognizes very different human microbial habitats in humans, as does the US Department of Energy (DOE) in its different missions of bioenergy, carbon sequestration and bioremediation, and the US Department of Agriculture (USDA) in the different agricultural habitats of microbes. In addition to exploring a habitat in depth, medium- and large-scale projects would also likely develop different expertise, technology, and analytical methods to meet the challenges of their particular habitat type (e.g., one may take the lead in proteomics, another in chemical informatics, another in community signaling, and another in microhabitat sensors). In this way a suite of projects will provide more tools and knowledge to the metagenomics community than any single project could offer.

RECOMMENDATION 1

The committee recommends the establishment of a Global Metagenomics Initiative that includes a small number (perhaps three) of large-scale, comprehensive projects that use metagenomics to understand model microbial communities, a larger number of medium-sized projects, and many small projects. Large-scale projects will study microbial communities in great depth, exploring a habitat worldwide, with attention to variation, commonalities, and detailed characterization. Medium-sized projects will provide centers of excellence in metagenomics that can be somewhat more focused than the large-scale projects, but will include a multidisciplinary approach to the study of a community. The small-scale projects will be single-investigator initiated and will examine a slice of a community, a particular function in multiple communities, or a specific technical advance.

The communities chosen for the large-scale projects should have broad applicability and impact and represent a diversity of habitat types. The studies would establish methods, approaches, and conceptual insights that could be applied to ever more complex and dynamic systems. Large-scale projects would achieve a depth of analysis not possible with smaller-scale projects and provide a template for comprehensive system analysis. Large-scale projects would also provide a forum for developing and testing new experimental and analytical tools and for establishing standards of sampling and data quality. The large projects may also generate economies of scale, new mechanisms for data sharing or storage, and point to new models of collaboration among large research groups. Different communities will have different benefits, technical challenges, and conceptual frameworks. These differences necessitate studying more than one community in great depth, leading the committee to recommend that three large-scale projects be identified and developed.

To maximize the benefits and knowledge gained from the large-scale projects, they should represent a breadth of habitat type, including:

- A community in a natural environment, to understand the interactions between microbial communities and geochemical processes or global nutrient cycles.
- A host-associated community, to probe the interactions between a microbial community and the physiology and health of its host.
- A "managed-environment" community, to learn to predict and manage the effects of environmental change or human activity on microbial communities.

The development of the large-scale projects should be carefully staged in three phases, as follows:

Phase I: Choosing the model communities. During Phase I, input should be solicited from a wide array of metagenomics and microbiology researchers to choose model communities of broad public interest with potential for immediate contribution to important environmental and public-health challenges. Clear goals or end points for each project should be defined during this phase. Phase I would conclude with a peer-reviewed competition for planning grants to be awarded to multidisciplinary, international teams. The committee anticipates that at least three model communities would be needed to cover the range of microbial community types, but Phase I may identify more than three projects with sufficient merit to proceed to Phase II.

Phase II: Planning. Each successful team would gather preliminary data and develop roadmaps for the completion of its project, including establishment of a data management and analysis group, development and testing of necessary methods and technologies, and launch of a Web site to provide access to data and analysis tools and to support public outreach.

Phase III: Implementation. Intensive sequencing, functional analysis, proteomics, and many other approaches would be applied to model communities that successfully complete Phase II.

FINDING 2

The metagenomics approach is of potential value in fulfilling the missions of many federal agencies, including NSF, NIH, DOE, and others. Support for individual projects specifically tied to each agency's mission has been and will continue to be productive, but communication and coordination across the interested agencies would be extremely useful. In particular, developing a consensus around which model communities to include in a Global Metagenomics Initiative should include the scientific constituencies of all these agencies. Scientific societies also can play a critical role in ensuring broad participation.

RECOMMENDATION 2

The committee recommends that an interagency working group like the Microbe Project take responsibility for ensuring open communication about the metagenomics portfolios of relevant agencies and for facilitating the organization of workshops and meetings to bring together metagenomics researchers who are working on different types of communities. The involvement of scientific societies is strongly encouraged. The Microbe

Project would be an appropriate forum for planning and promoting a Global Metagenomics Initiative.

FINDING 3

Metagenomics will draw on expertise from people in many disciplines:

- Those with knowledge of microbiology, including microbial genetics, biochemistry, physiology, pathology, systematics, ecology, and evolution.
- Other biologists, including molecular and cell biologists and those with knowledge of host organisms, including humans and other mammals, plants, insects, and other microbial hosts that have important roles in nature or that are of economic importance.
- Those with knowledge of the environment, including soil and atmospheric scientists, geologists, oceanographers, hydrologists, and agriculture and ecosystem scientists.
- Those who stimulate microbial communities to achieve specific end points, including biological, chemical, and environmental engineers.
- Computational scientists, including those with knowledge of statistics, computer science, data mining and visualization, database development, modeling, and applied mathematics.
- Those with expertise in scaling information to large ecosystem parameters, and in evaluating the impact of global change and its interface with policy.
- Engineers, physical scientists, and chemists whose skills and insights are potentially field-transforming in their contribution to new methods, chemistries, devices and applications (within and beyond metagenomics) and the understanding of complexity, networks, and system structure.

The value of integrating experts from such a wide array of fields into metagenomics projects is very high. Both they and metagenomics researchers will require appropriate cross-disciplinary knowledge in order to gain the full benefit of their different expertise. To realize the potential of metagenomics, interdisciplinary projects will be necessary and they will be aided by new education and training programs.

RECOMMENDATION 3

The committee recommends establishing several types of training programs to encourage scientists to develop the skills needed for metagenomics research. The following mechanisms have been successful in providing cross-disciplinary training: interdisciplinary training to augment traditional graduate programs, summer courses patterned after the Cold Spring Harbor or Marine Biological Laboratories summer courses, and postdoctoral programs in which fellows undertake training in a new discipline. Support for faculty to attend metagenomics workshops or to spend sabbaticals in metagenomics research laboratories or facilities would also be beneficial in expanding appropriate training.

FINDING 4

The value of the Human Genome Project was multiplied because the data that it generated were rapidly made available in a public database. GenBank and its collaborators in Europe and Japan serve as repositories for nucleic acid sequence data. They ensure that the data are accessible to all and can be obtained from a single site. Similar accessibility would multiply the value of metagenomic data.

The analysis of metagenomic data will require the establishment of new databases in addition to the sequence archives. It is essential that the databases use common data standards and agree on the metadata that will describe metagenomic sequences. This will ensure that data can be exchanged between researchers. It will also facilitate comparative analyses of data and the development of software. Community databases like those established for the *Drosophila* and *Arabidopsis* genome projects are excellent models for the type of databases metagenomics will require.

Information from metagenomics studies will be fully exploited only if appropriate data management and analysis methods are in place. Furthermore, metadata—for example, data on sampling method, sample treatment, and precise description of the sampled habitat—are essential for the analysis of metagenomic sequence data. If metagenomic data are to be used to their fullest advantage, metadata infrastructure is urgently needed. No one metadata standard will be appropriate for all samples, which will come from extremely diverse environments, but there should be close collaboration and coordination among the communities of scientists developing metadata standards.

One major challenge faced by metagenomics databases compared with "conventional" genomics databases will be the demand for community input into the annotation process. Annotation is the process of assigning functional, positional, and species-of-origin information to the genes in a database. In conventional genomics, primary responsibility for annotating data falls on the authors. In metagenomics projects, in which annotations will change as additional data (or metadata) are collected by other groups, an annotation database must be able to accept and integrate both individual and large-scale (computational) annotations of metagenomic data continually. Furthermore, the sources of and methods for modified annotations should be transparent to database users.

RECOMMENDATION 4

The committee recommends the establishment of new databases for metagenomic data and the development of tools for the storage, analysis, and visualization of these data. Early attention should be given to the challenge of providing dynamic and traceable annotation in metagenomics databases. Also warranting high priority is the development of a consensus—in a process that includes the research communities and the database developers—on the metadata that need to be collected and on the data standards to be used. Maintenance and curation of metagenomics databases will greatly add to their value, but are expensive and will require consistent support. Funding for databases requires a different approach than that for research projects: the committee recommends the development of mechanisms for long-term funding, coupled with community oversight and evaluation.

The enormous amounts of data generated by metagenomics should be made available as rapidly as possible, and deposition into the international sequence archives should be required. Some projects, like those of the proposed Global Metagenomics Initiative, would be undertaken specifically to create a community resource and these should follow accepted standards, such as those of the Fort Lauderdale Agreement, in immediately releasing the data without constraints as to their use. Data from single-investigator projects should be released within a short time, for example, within 6 months of its collection.

FINDING 5

The analysis of genomics data is absolutely dependent on computer software. In general, metagenomics projects will require an even higher percentage of funds for bioinformatics and statistical support than have genomics projects, or most other kinds of biological research. It is common for software developed for a particular project gradually to find wide use in the community. Providing a mechanism whereby such analytical tools that have proved their value to the community can be brought up to robust, engineered, documented form would be very worthwhile.

RECOMMENDATION 5

Funding agencies should consider the development of a mechanism for identifying analytical tools that are finding wide use in the community and for providing for their development up to robust standards.

FINDING 6

Current metagenomics researchers face several difficulties, including inadequate characterization of many habitats and inadequate understanding of the scope and nature of variation in different microbial communities. Therefore, determining how best to sample and determining whether a sample is representative remain challenging. DNA extraction techniques to minimize contamination and to ensure that a community's genome is adequately represented have yet to be optimized. And expression systems for functional metagenomics are not yet sufficiently robust and flexible to express most genes in most metagenomes.

RECOMMENDATION 6

The committee recommends investment in the following because improvements would enhance the productivity of many metagenomics projects: new or improved technologies for appropriate habitat sampling, macromolecule recovery, and habitat characterization, depending on habitat; new approaches to deal with the unevenness of population sizes in communities and to target populations of interest within complex communities; development of measures of community diversity to supplement 16S rRNA gene surveys, including arrays and additional phylogenetically informative genetic markers; and development of diverse host species and expression strategies for functional-expression analyses.

FINDING 7

The more is known about microbes, the greater the value that metagenomic data will have. It is extremely important for basic microbiology research not to be neglected but to be strengthened and deepened. Advances in the culturing of currently unculturable bacteria and archaea, in sequencing of their genomes, and in genetic and physiological studies are key reference points for interpreting a community's metagenome. Active discussion involving metagenomics researchers and members of other subdisciplines of microbiology and their representatives in funding agencies will help to guide the fields in complementary directions.

RECOMMENDATION 7

The committee recommends that funding agencies consider the potential contributions of basic microbiology research to progress in metagenomics as they evaluate their overall research portfolios.

FINDING 8

Because metagenomics constitutes a revolutionary advance in the ability of scientists to study a previously invisible biological realm, results of metagenomics studies have great potential interest not only for scientists but also for the general public. Metagenomics presents an important opportunity to engage the public in the excitement and value of basic and applied scientific research. Outreach efforts will help to train a new generation of scientists who are skilled in communicating science to the public.

RECOMMENDATION 8

The committee recommends that education and public outreach be components of all metagenomics projects. Both large and small projects can be used as catalysts for teaching microbiology. Each large project should have a budget for developing materials that explain its scientific basis and implications in accessible and interesting ways. Metagenomics researchers should be encouraged to teach about their science in their local communities and metagenomics projects should include training scientists in effective outreach teaching.

9

Epilogue

Twenty years ago, the Human Genome Project, and the nascent genomic sciences more generally, were highly controversial. Many biologists thought that investing resources in such “molecular natural history” was economically wasteful and intellectually suspect. Now, practically all biologists are genomicists. If not directly pursuing genome sequencing and the other “omic” methods, biologists nevertheless often ground their particular genetic, biochemical, physiological, behavioral, or ecological studies in the work of someone who is. Genomics has been transformative in the deepest sense, not only answering many questions about how organisms function, develop, and evolve, but also driving a radical reformulation of the terms in which such questions are asked. Although initially many of us thought of genomics mostly as a more economical and efficient way (because of economies of scale) to recover and study the behavior of individual genes, in fact it has shifted focus to the collective and integrated activities of genes functioning together, to the networks of interactions between them, and to how these are integrated (and have evolved) in the highly complex and coordinated business of living and reproducing at the level of cells and organisms. As noted earlier, genomics and the associated high-throughput “omic” technologies targeting gene expression, protein synthesis (and modification), protein interactions and protein structure are all becoming experimental subdisciplines of a new concept-driven computational science called systems biology.

What then, will metagenomics have become, in 20 years? We believe that it too will be a concept-driven computational science with subdisciplines that have evolved from the fusion of “omic” approaches and more traditional disciplines, such as environmental and clinical microbiology, biogeochemistry, biological oceanography, soil sciences, and theoretical ecology. It will indeed be the systems biology of the most inclusive biological system we know about: the biosphere of the planet. These disciplines will in the process be transformed and many questions redefined and refocused, most often at a level below (genes and genomes) or above (communities and ecosystems) the organism and species levels at which microbial ecologists have traditionally concentrated their efforts. Although individual microbial cells will always be suitable units of study, the "species," because we have just begun to uncover the enormous genomic diversity within it, may no longer be a reliable or useful ecological unit. Instead, we will understand ecosystems in terms of the collective activities and interactions of the genes they contain, how these are distributed and expressed in space and time, and how they function together.

We can expect, in 20 years, enormous advances on three fronts—technical, computational, and biological—as well as a host of specific applications.

POTENTIAL TECHNICAL ADVANCES

Sequencing technology will have reduced the per-base price of finished sequence to fractions of a cent, and the cost of sequence-data acquisition will no longer by a serious consideration in studies of specific ecosystems. Sequencing methods now in use will have increased run lengths substantially but will themselves probably have been replaced with even more direct, and often also cloning-independent, approaches, perhaps single-molecule technologies now under development or others yet to be imagined. Single cell genome sequencing will be routine, and cell-sorting methods that readily permit recovery of even unique individual cells will be well advanced. Complete genome sequences, some produced by "traditional" methods based on isolates (or single cells) but others acquired metagenomically, will number in the thousands, perhaps even tens of thousands. There will be many "species" for which hundreds of individual isolates will have been sequenced.

Transcriptomic and proteomic applications to community samples will be comparable in their reliability and efficiency with such methods as are used in human genomics today. Incremental improvements in microarray sensitivity, specificity, and reproducibility will make it possible to assess community membership and abundance down to the "species" level, however that concept is then understood. New normalization protocols will allow a census of even the rarest members of a community, and whole-community RNA amplification will access their transcriptomes. We will be able routinely to classify or type ecosystems and monitor changes in their compositions and activities with arrays (and their future equivalents, which may be microfluidics-based) that are inexpensive and readily available commercially. Such monitoring will indeed be routine practice in many environment-based business and regulatory activities and in epidemiology. New "omic" methods and sciences will have been developed for characterizing communities and their genetic, physiological, biochemical, and biogeochemical activities.

Many currently unculturable organisms and consortia will have been "domesticated," by using knowledge of their individual needs and potentialities as derived from community metagenomics. As we come to appreciate the true extent of diversity (even within designated species) we will know that even such facilitated pure-culture or defined-culture studies will never be adequate for global understanding, but will provide excellent models of physiological interactions and the refinement of computational models for such interactions.

POTENTIAL COMPUTATIONAL ADVANCES

In 20 years, infrastructural accommodations will have been made for the almost unimaginable amount of metagenomic data that will have accumulated. For reasons elaborated in Chapter 5, the metagenomics databases are expected to dwarf genomic databases, no matter the predicted rate of growth of the latter. Although all sequences and trace data (or their future technological equivalents) will be available through GenBank or comparable public repositories there will be specialized (but fully public and interoperable) databases of all sorts. It will be possible to answer questions like those sketched in Box 5-1 by direct queries to the databases, which will also be rich in associated metadata. Just as much biological research is now conducted by computer scientists, much microbial ecology will be purely computational. Indeed, these downstream activities may be the dominant form of metagenomics employment; but metabioinformaticians will need even broader interdisciplinary training and collaborative links—in geochemistry, oceanography, earth and atmospheric sciences, biochemistry, microbiology, ecology, genetics and genomics, statistics, and computer science.

Although traditional microbial classification practices (phenotypic characterization and identification at the level of species and genus) may remain useful, the basis on which we predict properties of isolates will be sequence- and computation-driven and probabilistic. Equally often, investigations of community activities of any magnitude (from the tiny but complex ecosystem of a termite's gut to the Pacific Ocean) will be conducted at the level of genes and their interactions—understanding the "games being played," with decreased emphasis on phylogenetic identification of the "players."

POTENTIAL BIOLOGICAL ADVANCES

It is of course pure science fiction to predict what we will know about the biosphere 20 years from now and it is in the nature of a transformative science to be unpredictable. But it is of some value to guess at the kinds of breakthroughs in biological science that metagenomics will make possible.

Viruses

There are many more viruses (and possibly more kinds of viruses) than there are cells (or kinds of cells). In many ecosystems, viruses are the principal regulators of organismal abundance and may well be the principal agents of genetic exchange between organisms. Their genomes collectively harbor a vast number of genes about which we know almost nothing and that can be exchanged between viruses and cells in a mix-and-match fashion. In 20 years, we hope to have some good idea of the depth of this enormous gene pool and (through comparative genomics, ab initio structural modeling, and extensive structural genomics) a vastly better understanding of what many of the genes do for their viral or cellular hosts and what they might do for us. We will understand and be able to monitor the exchange of information between viruses in the environment and those infecting us and the animals and plants that we use. Our ability to monitor and predict the emergence of viral diseases will be much enhanced.

Cells and Their Genes and Genomes

We will have come to an understanding of the diversity of gene content within species, of how many strain-specific genes are involved in strain-specific biology, and of how many are "just passing through." We will have a vast inventory of gene sequences and, through structural genomics, a vast reservoir of genes with reasonably inferred functions even if the organisms of origin and the roles of the genes in their biology remain a mystery. We will be able to say whether adaptation to environmental change of any sort most often involves recruitment of preadapted lineages from elsewhere or cobbling together of novel lineages by exchange and assembly of genes already present.

Species

We will have enough information on the diversity of environmental gene sequences to allow us to redefine the species concept to a more consistent, accurate, defensible, and enduring concept that will have broad value across numerous disciplines and applications. We will have relegated so much of the task of identification of isolates and prediction of their properties to computers and sequence databases that it will be the predictions, not formal identification, that we care about. We will understand the various processes that might be termed "speciation" and have a good idea of their relative frequencies in nature. We will have redefined questions of diversity ("How many species are there in an environment or in the world?") in terms of the sequences of genes and the composition of genomes.

Biogeography

We will have mapped an enormous number and diversity of genes and genome compositions in space and time and will be able to retrieve and reanalyze this information and associated physical, chemical, and biological metadata. We will have substantial gene-expression and metabolomic data on the same sites and can begin to look at Earth as though it were an organism-like spatiotemporally defined entity with an evolved and homeostasis-promoting global "metabolism." Gene frequency and expression will make sense in that context even though Earth is not an organism. The question of whether "everything is everywhere" will be subsumed into this gene-level and genome-level analysis, which will be recast in terms of relative rates of divergence and dispersal of genes.

Community Structure and Function

Model-community projects undertaken in the next 5 years will have been completed and, in addition to a deep understanding of their target systems, will provide templates for other studies, smaller in scope but greater in number and ultimately interconnectable. Community structure will be understood and described ("profiled") in terms more of gene presence and abundance than of species presence and abundance, and we will have developed a typology or catalog of communities that will allow us to infer what sort of biogeochemistry is happening at any place and time and to monitor changes. Such profiling is already done with ribosomal RNA and a few other markers, but comprehensive functional gene (and gene-function) assessment will be vastly more subtle and informative. One safe prediction is that such profiling will be extensively applied and prove of great value in disease diagnosis and determination of nutritional status of humans (individual and communities) and of animals and plants that they use or care about. Probiotic therapies and regimens will become evidence-based and increasingly valuable, as will microbiome profiling in the detection of diseases that originate in the host.

Interactions Within and Between Communities

Gene frequency and expression data will, in 2027, have long been the basis for constructing community "interactome" maps, comparable in character but vastly more complex than maps now used by systems biologists to study individual organisms and their responses to perturbations. The combinations of genes and organisms that influence community robustness will have been identified and predictive principles of community behavior will have been derived. The development and implementation of such analytical models will allow computational microbial ecologists to predict responses (at the level of gene frequency, expression, and exchange) to environmental challenges of all sorts. Testing such predictions will lead to better models. Such reiterative approaches are already used, but models based on all genes rather than a few diagnostic markers will have immensely more explanatory and predictive power. The ultimate goal, perhaps in sight by 2027, would be a metacommunity model that seeks to explain and predict (and retrodict) the behavior of the biosphere as though it were a single superorganism. Such a "genomics of Gaia" would be the ultimate implementation of systems biology. The enormous challenge that creation of such a metamodel represents is matched by its importance for the future of the human species.

POTENTIAL ADVANCES IN EDUCATION AND PUBLIC UNDERSTANDING

By 2027, we will have many more mechanisms for communication than we have now, but all will be usable to teach the public about microbes through the excitement and "big science" appeal of metagenomics. Microbiology will be required in the K-12 curriculum and as a prerequisite for teaching certification, and metagenomics centers across the United States will have developed robust mechanisms for communication with diverse people, including those who do not have access to a university. The mechanisms might include distance-education courses, mobile microbiology units, press releases about milestones in projects, hosting of teachers in research laboratories, and teaching by metagenomics scientists in K-12 classrooms. Graduate students will be trained to teach microbiology in the classroom and in the larger community.

SOME POTENTIAL SPECIFIC APPLICATIONS

We see metagenomics as a new basic science with many eminently useful (and in tomorrow's world essential) applications, some accomplishable over the short term and probably most on the drawing board by 2027.

Earth Sciences

The biological forcing of elemental cycles is key to understanding a wide variety of Earth-system processes. Large-scale, ecosystemwide fluxes of energy and matter, however, are difficult to model accurately or to study in the laboratory. By 2027, Earth-system processes will have been examined in much greater detail with metagenomics coupled with other synoptic physicochemical and biological measurements. Microbial-community genomics will provide information important for understanding energy fluxes and biogeochemical mechanisms in the deep subsurface, modeling biologically mediated rock weathering and surface chemistry, and defining the key genetic and biogeochemical drivers of processes that influence greenhouse-gas production and consumption. The oceans, which harbor millions of microbes in each teaspoonful of seawater, will be modeled more fully as we become able to visualize the rich biological systems they encompass. In a practical sense, such processes as uranium immobilization or acid mine drainage cleanup, which involve coupled biological-geochemical interactions, will be enhanced and improved with new community-genomic datasets. Microbe-enabled oil recovery, subsurface methane production and consumption, and carbon storage and turnover are other critical interfaces between the microbial world and the Earth system. The new "whole-Earth catalog" of microbial genes and genomes provided by metagenomics will propel a new understanding and new technologies for more appropriate resource use and sustenance of the living Earth system. Predictive models of many vital biogeochemical processes will inform enlightened policy makers. We will be able to say, for instance, why it might or might not be a good idea to seed oceans with iron to increase carbon sequestration. Similarly, we will be able to model (and predict the extent of) methanogenesis in the permafrost as it thaws. Metagenomics-based environmental monitoring will be a thriving industry.

Life Sciences

Through a fine-scale and nuanced understanding of genetic and ecological processes, we will demolish many generalizations about microbes, replacing them with particularized knowledge. We anticipate that many basic concepts that have vexed biologists for decades (sometimes centuries), a few of which were alluded to earlier in this epilogue, will be recast in molecular terms. Taxonomy, the science of identification and naming organisms according to their relationships, will be radically transformed. The enormous combined genomic and metagenomics databases will enable us to predict the behavior of an isolate, a consortium, or a complex community on the basis of carefully targeted sequence or other molecular information.

Metagenomic methodology and concepts will have expanded well beyond the realm of viruses, bacteria, and archaea, to embrace the population biology and biogeography of microbial eukaryotes (protists, algae, and fungi). Indeed, the new research methodology and paradigm will have found uses even for macroscopic organisms, when it is population or ecological processes that are of interest. And with a proper appreciation of the roles of microbes in the balance of life, a new global systems ecology embracing all species, including humans, will have been born. This will mandate changes in how we teach biology at all levels. The teaching of microbiology, ecology, and evolutionary biology will all be profoundly affected by metagenomics, bringing the focus of a generation of students back "down to the ground," where problems can be directly addressed.

Biomedical Sciences

The full extent of interindividual diversity within the human microbiome will be understood, and changes in microbial-community composition that contribute to or are responsible for a number of acute and chronic diseases will have been elucidated. Microbiome-based diagnosis will be an essential component in treatment for many diseases. Preventive medicine will be a major component of health care and health industries with the development of rational probiotic therapy as a means of maintaining a "healthy" human microbiome. By understanding how the human microbiome differs in health and disease, physicians will be on a much better footing to understand and predict the incidence of chronic inflammatory and infectious diseases, both viral and microbial. Therapeutic interventions (in addition to probiotics) will be based on comprehensive knowledge of the effects of treatment (such as with antibiotics) on the microbiota as a whole. New antibiotics from currently unknown natural (and generally microbial) sources will have come on line, and new strategies (such as those described below) for forestalling the development and spread of antibiotic resistance will have been devised.

Agriculture

Microbial communities will continue to affect productivity in agriculture, both plant-based and animal-based. Metagenomics studies of gut populations in poultry, pigs, and other food animals will increase our knowledge of gut-microbe interactions, which will help to formulate more effective probiotic mixtures in the future. We expect a comparable impact on plant-based agriculture. The function of the crenarchaeotes and other microbes that colonize plant roots and their importance to carbon and nitrogen cycles will be better understood. We will understand how plants and their beneficial microbial partners deal with antagonistic microbes. Lessons will have been learned from the food crops that have been successfully cultivated over the centuries. Using metagenomic approaches, we will exploit the interplay of microbes and plants more intelligently for human benefit.

Bioenergy

Fossil fuels are a nonrenewable natural resource. It is projected that energy demand will increase by more than 50% by 2025 (US Department of Energy 2005). The US economy depends on oil imports, so there is an interest in augmenting domestic energy production. Corn serves as the major feedstock for ethanol production, and biofuel-producing companies are using specialist microbes to convert cornstarch to ethanol, a high-octane, environmentally friendly biofuel. Cellulosic ethanol—made from such agricultural wastes as corn fiber, corn stalks, and wheat straw and other biomass, such as switchgrass and miscanthus—uses as substrates products that are not usable by humans as food. Furthermore, cellulosic materials are inexpensive, renewable, and their efficient use will reduce the cost of ethanol production. Most of the known ethanol-producing microbes are incapable of using cellulose to produce ethanol, because they lack the enzymes required to break it down. In nature, however, several microbes are equipped with arrays of enzymes that act together to release glucose from cellulose. The glucose can then be fermented to ethanol. Metagenomics will enable discovery of new cellulosic enzymes and novel microbial strategies for hydrolysis of biomass. These discoveries will lead to engineering of enzyme complexes and novel pathways for enzymatic hydrolysis of cellulose and a concomitant increase in production of biofuels from cellulosic materials.

Bioremediation

Metagenomics will shape bioremediation in many interrelated ways. First, vastly increased understanding of how microbes form "bucket brigades" for the degradation of xenobiotic compounds will allow us to distinguish contaminated sites in which the native microbiota is competent to restore environmental health from sites in which intervention in the form of in situ bioaugmentation or intensive ex situ treatment at special facilities is needed. Second, metagenomics will facilitate sensitive monitoring of remediation activities of either sort. Third, it will identify key microbial processes and keystone species and indicate how community composition could best be complemented. Fourth, it will lead to the isolation of specific strains or consortia that could be used for such complementation. Fifth, a host of novel enzymes that might be useful in cellfree treatments of specific contaminants will be found. And sixth, where appropriate and permitted, the metagenomics database will provide a rich stock of genes for the construction of novel specialized strains for targeted use in bioeremediation.

Biotechnology

The biotechnology industry already employs hundreds of microbial enzymes and related products, and the global industrial enzyme market is currently in excess of $2 billion per year, primarily in technical (including scientific, pulp and paper), food, and agriculture and feed applications. The great majority of such enzymes are the result of traditional approaches: enrichment, culture, isolation, and enzyme purification. Collectively, the metagenomics database and the effort, now in full swing, to express, crystallize, and characterize structurally and functionally entire proteomes of many model organisms are likely to enhance the rate of discovery of such valuable catalysts by at least an order of magnitude—a revolution in green chemistry. Ironically, some of the key products of such activities to date have vital applications in the discovery process itself. For instance, the polymerase chain reaction—which is the basis of modern molecular environmental microbiology, DNA forensics, and molecular diagnosis—is based on genes cloned from thermophilic bacteria and archaea.

Biodefense and Microbial Forensics

The same methods that will allow us to assess community composition and activity will enable construction of biosensors for biodefense and microbial forensics. In 2027, the threat of terrorist or criminal use of pathogenic organisms and their toxins against human populations or agricultural (plant and animal) targets may still be of concern. However, society's ability to anticipate and respond to these threats will be markedly enhanced through the continued application of new technologies that will allow us to assess microbial community composition and activity in various environments. This will permit precise, rapid, and sensitive monitoring of air, water, and food supplies for potential biothreat agents with novel biosensors. We will be better able to identify the presence of a natural or engineered biothreat agent against a large natural microbial background, and we will be able to predict virulence properties and sensitivity to antiviral or antimicrobial drugs. Another anticipated outcome of research in biodefense will be a strong forensic capability to carry out attribution for acts of bioterrorism that use animal, plant, and foodborne pathogens and toxins. Such capability will provide the law-enforcement, intelligence, agriculture, public-health, and homeland-security communities with informa-

References

Akkermans, A. D. L., J. D. Elsas, and F. J. Bruijn. 1995. *Molecular microbial ecology manual.* Dordrecht: Kluwer.

Allen, E. E., and J. F. Banfield. 2005. Community genomics in microbial ecology and evolution. *Nat Rev Microbiol* 3 (6):489-98.

American Society for Microbiology and National Institutes of Health. 2007. *Basic research on bacteria: The essential frontier. Report on the American Society for Microbiology and National Institutes of Health workshop on basic bacterial research*. Washington, D.C.: American Society of Microbiology.

Angly, F. E., B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer. 2006. The marine viromes of four oceanic regions. *PLoS Biol* 4 (11):e368.

Arico, S., and C. Salpin. 2005. *Bioprospecting of genetic resources in the deep seabed: scientific, legal and policy aspects*. New York: United Nations University-Institute of Advanced Studies.

Berry, A. E., C. Chiocchini, T. Selby, M. Sosio, and E. M. Wellington. 2003. Isolation of high molecular weight DNA from soil for cloning into BAC vectors. *FEMS Microbiol Lett* 223 (1):15-20.

Brodie, E. L., T. Z. DeSantis, J. P. Parker, I. X. Zubietta, Y. M. Piceno, and G. L. Andersen. 2007. Urban aerosols harbor diverse and dynamic bacterial populations. *Proc Natl Acad Sci USA* 104 (1):299-304.

Brown, M. V., and J. A. Fuhrman. 2005. Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. *Aquat Microb Ecol* 41:15-23.

Campbell, C. L., P. D. Peterson, and C. S. Griffith. 1999. *The formative years of plant pathology in the United States*. Minneapolis, MN: American Phytopathological Society Press.

Coleman, M. L., M. B. Sullivan, A. C. Martiny, C. Steglich, K. Barry, E. F. Delong, and S. W. Chisholm. 2006. Genomic islands and the ecology and evolution of Prochlorococcus. *Science* 311 (5768):1768-70.

Crump, B. C., and J. E. Hobbie. 2005. Synchrony and seasonality in bacterioplankton communities of two temperate rivers. *Limnol Oceanogr* 50 (6):1718-29.

D'Costa, V. M., K. M. McGrann, D. W. Hughes, and G. D. Wright. 2006. Sampling the antibiotic resistome. *Science* 311 (5759):374-7.

DeLong, E. F., and D. M. Karl. 2005. Genomic perspectives in microbial oceanography. *Nature* 437 (7057):336-42.

DeLong, E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N. U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311 (5760):496-503.

Desantis, T. Z., C. E. Stone, S. R. Murray, J. P. Moberg, and G. L. Andersen. 2005. Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. *FEMS Microbiol Lett* 245 (2):271-8.

Desantis, T. Z., E. L. Brodie, J. P. Moberg, I. X. Zubieta, Y. M. Piceno, and G. L. Andersen. 2007. High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microb Ecol*. In press.

Dumont, M. G., and J. C. Murrell. 2005. Stable isotope probing—linking microbial identity to function. *Nat Rev Microbiol* 3 (6):499-504.

Eckburg, P. B., E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman. 2005. Diversity of the human intestinal microbial flora. *Science* 308 (5728):1635-8.

Edwards, R. A., and F. Rohwer. 2005. Viral metagenomics. *Nat Rev Microbiol* 3 (6):504-10.

Edwards, R. A., B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D. M. Peterson, M. O. Saar, S. Alexander, E. C. Alexander, Jr., and F. Rohwer. 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7:57.

Energy Information Administration. 2005. *International energy outlook 2005*. Washington, D.C.: U.S. Department of Energy.

Engebretson, J. J., and C. L. Moyer. 2003. Fidelity of select restriction endonucleases in determining microbial diversity by terminal-restriction fragment length polymorphism. *Appl Environ Microbiol* 69 (8):4823-9.

Falkowski, P. G., R. T. Barber, and V. V. Smetacek. 1998. Biogeochemical controls and feedbacks on ocean primary production. *Science* 281 (5374):200-7.

Fan, J. B., M. S. Chee, and K. L. Gunderson. 2006. Highly parallel genomic assays. *Nat Rev Genet* 7 (8):632-44.

Field, K. G., D. Gordon, T. Wright, M. Rappe, E. Urback, K. Vergin, and S. J. Giovannoni. 1997. Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Appl Environ Microbiol* 63 (1):63-70.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocyne, J. Scott, R. Shirley, L-I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M.D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, Lisa A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and J. C. Venter. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269 (5223):496-512.

Forney, L. J., X. Zhou, and C. J. Brown. 2004. Molecular microbial ecology: land of the one-eyed king. *Curr Opin Microbiol* 7 (3):210-20.

Fraser-Liggett, C. M. 2005. Insights on biology and evolution from microbial genome sequencing. *Genome Res* 15 (12):1603-10.

Fuhrman, J. A., I. Hewson, M. S. Schwalbach, J. A. Steele, M. V. Brown, and S. Naeem. 2006. Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci USA* 103 (35):13104-9.

Garcia-Martinez, J., and F. Rodriguez-Valera. 2000. Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine Archaea of Group I. *Mol Ecol* 9 (7):935-48.

Garcia Martin, H., N. Ivanova, V. Kunin, F. Warnecke, K. W. Barry, A. C. McHardy, C. Yeates, S. He, A. A. Salamov, E. Szeto, E. Dalin, N. H. Putnam, H. J. Shapiro, J. L. Pangilinan, I. Rigoutsos, N. C. Kyrpides, L. L. Blackall, K. D. McMahon, and P. Hugenholtz. 2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* 24 (10):1263-9.

Gardner, W. L., and W. B. Whitman. 1999. Expression vectors for Methanococcus maripaludis: overexpression of acetohydroxyacid synthase and beta-galactosidase. *Genetics* 152 (4):1439-47.

Gevers, D., F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F. L. Thompson, and J. Swings. 2005. Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol* 3 (9):733-9.

Gill, S. R., M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. 2006. Metagenomic analysis of the human distal gut microbiome. *Science* 312 (5778):1355-9.

Gray, J., D. T. Liu, M. Nieto-Santisteban, A. S. Szalay, D. DeWitt, and G. Heber. 2005. *Scientific Data Management in the Coming Decade*. Redmond, WA: Microsoft Corporation.

Green, B. D., and M. Keller. 2006. Capturing the uncultivated majority. *Curr Opin Biotechnol* 17 (3):236-40.

Guan, C., B. Borlee, J. Ju, L. Williamson, B. Shen, K. Raffa, and J. Handelsman. 2006. Signal mimics derived from a metagenomic analysis of gypsy moth gut microbiota. *Nat Biotechnol*.

Hazen, R. M. 2005. *Genesis: The scientific quest for life's origin*. Washington, D.C.: Joseph Henry Press.

Hellani, A., S. Coskun, M. Benkhalifa, A. Tbakhi, N. Sakati, A. Al-Odaib, and P. Ozand. 2004. Multiple displacement amplification on single cell and possible PGD applications. *Mol Hum Reprod* 10 (11):847-52.

Holben, W. E., and D. Harris. 1995. DNA-based monitoring of total bacterial community structure in environmental samples. *Mol Ecol* 4 (5):627-31.

Hood, L. n.d. Cited at: *http://www.systemsbiology.org/Systems_Biology_in_Depth*.

Janssen, P. H., P. S. Yates, B. E. Grinton, P. M. Taylor, and M. Sait. 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 (5):2391-6.

Laird, S. A., and R. Wynberg. 2005. *The commercial use of biodiversity: An update on current trends in dmand for access to genetic rsources and benefit-sharing, and industry perspectives on ABS policy and implementation.* Item 8 in *Convention on Biological Diversity UNEP/CBD/WG-ABS/4/INF/5*. New York: United Nations Environmental Programme.

Lambright, W. H. 2002. *Managing "Big Science": A case study of the Human Genome Project*. Arlington, VA: The Price Waterhouse Coopers Endowment for the Business of Government.

Levantesi, C., L. S. Serafim, G. R. Crocetti, P. C. Lemos, S. Rossetti, L. L. Blackall, M. A. Reis, and V. Tandoi. 2002. Analysis of the microbial community structure and function of a laboratory scale enhanced biological phosphorus removal reactor. *Environ Microbiol* 4 (10):559-69.

Ley, R. E., D. A. Peterson, and J. I. Gordon. 2006. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 124 (4):837-48.

Li, W., X. Zhou, and P. Lu. 2004. Bottlenecks in the expression and secretion of heterologous proteins in Bacillus subtilis. *Res Microbiol* 155 (8):605-10.

Liles, M. R., B. F. Manske, S. B. Bintrim, J. Handelsman, and R. M. Goodman. 2003. A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl Environ Microbiol* 69 (5):2684-91.

Lindell, D., M. B. Sullivan, Z. I. Johnson, A. C. Tolonen, F. Rohwer, and S. W. Chisholm. 2004. Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proc Natl Acad Sci USA* 101 (30):11013-8.

Liu, W. T., T. L. Marsh, H. Cheng, and L. J. Forney. 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol* 63 (11):4516-22.

Mahenthiralingam, E., A. Baldwin, P. Drevinek, E. Vanlaere, P. Vandamme, J. J. Lipuma, and C. G. Dowson. 2006. Multilocus sequence typing breathes life into a microbial metagenome. *PLoS ONE* 1:e17.

Matta, C. 2007. *The science of small things: The botanical context of German bacteriology, 1840-1910*. Ph.D. dissertation. University of Wisconsin.

Mazzola, M. 2004. Assessment and management of soil microbial community structure for disease suppression. *Annu Rev Phytopathol* 42:35-59.

Metcalf, W. W., J. K. Zhang, E. Apolinario, K. R. Sowers, and R. S. Wolfe. 1997. A genetic system for Archaea of the genus Methanosarcina: liposome-mediated transformation and construction of shuttle vectors. *Proc Natl Acad Sci USA* 94 (6):2626-31.

Metzker, M. L. 2005. Emerging technologies in DNA sequencing. *Genome Res* 15 (12):1767-76.

Munch, R. 2003. Robert Koch. *Microbes Infect* 5 (1):69-74.

Nass, S. J., B. Stillman, eds. 2003. *Large-scale biomedical science: exploring strategies for future research*. Washington, D.C.: The National Academies Press.

Newman, D. J., G. M. Cragg, and K. M. Snader. 2003. Natural products as sources of new drugs over the period 1981-2002. *J Nat Prod* 66 (7):1022-37.

NIH. n.d. *New Roadmap Emphasis Areas for 2008*. Office of Portfolio Analysis and Strategic Initiatives. Available from: *http://nihroadmap.nih.gov/2008initiatives.asp*.

Osborn, A. M., E. R. Moore, and K. N. Timmis. 2000. An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. *Environ Microbiol* 2 (1):39-50.

Pace, N. R. 1997. A molecular view of microbial diversity and the biosphere. *Science* 276 (5313):734-40.

Peck, R. F., S. Dassarma, and M. P. Krebs. 2000. Homologous gene knockout in the archaeon Halobacterium salinarum with ura3 as a counterselectable marker. *Mol Microbiol* 35 (3):667-76.

Pedros-Alio, C. 2006. Genomics and marine microbial ecology. *Int Microbiol* 9 (3):191-7.

Petit, J. R., J. Jouzel, D. Raynaud, N. I. Barkov, J.-M. Barnola, I. Basile, M. Bender, J. Chappellaz, M. Davis, G. Delaygue, M. Delmotte, V. M. Kotlyakov, M. Legrand, V. Y. Lipenkov, C. Lorius, L. Pépin, C. Ritz, E. Saltzman, and M. Stievenard. 1999. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* 399:429-36.

Priest, F. G., M. Barker, L. W. Baillie, E. C. Holmes, and M. C. Maiden. 2004. Population structure and evolution of the Bacillus cereus group. *J Bacteriol* 186 (23):7959-70.

Ram, R. J., N. C. Verberkmoes, M. P. Thelen, G. W. Tyson, B. J. Baker, R. C. Blake II, M. Shah, R. L. Hettich, and J. F. Banfield. 2005. Community proteomics of a natural microbial biofilm. *Science* 308 (5730):1915-20.

Rappe, M. S., S. A. Connon, K. L. Vergin, and S. J. Giovannoni. 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* 418 (6898):630-3.

Riesenfeld, C. S., R. M. Goodman, and J. Handelsman. 2004. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ Microbiol* 6 (9):981-9.

Rondon, M. R., P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, C. L. Tiong, M. Gilman, M. S. Osburne, J. Clardy, J. Handelsman, and R. M. Goodman. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* 66 (6):2541-7.

Sait, M., P. Hugenholtz, and P. H. Janssen. 2002. Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. *Environ Microbiol* 4 (11):654-66.

Samuel, B. S., and J. I. Gordon. 2006. A humanized gnotobiotic mouse model of host-archaeal-bacterial mutualism. *Proc Natl Acad Sci USA* 103 (26):10011-6.

Schloss, P. D., and J. Handelsman. 2006. Toward a census of bacteria in soil. *PLoS Comput Biol* 2 (7):e92.

Shendure, J., R. D. Mitra, C. Varma, and G. M. Church. 2004. Advanced sequencing technologies: methods and goals. *Nat Rev Genet* 5 (5):335-44.

Socolow, R. H. 1999. Nitrogen management and the future of food: lessons from the management of energy and carbon. *Proc Natl Acad Sci USA* 96 (11):6001-8.

Stephenson, K., and C. R. Harwood. 1998. Influence of a cell-wall-associated protease on production of alpha-amylase by Bacillus subtilis. *Appl Environ Microbiol* 64 (8):2875-81.

Thompson, J. R., S. Pacocha, C. Pharino, V. Klepac-Ceraj, D. E. Hunt, J. Benoit, R. Sarma-Rupavtarm, D. L. Distel, and M. F. Polz. 2005. Genotypic diversity within a natural coastal bacterioplankton population. *Science* 307 (5713):1311-3.

Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. 2005. Comparative metagenomics of microbial communities. *Science* 308 (5721):554-7.

Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444 (7122):1027-31.

Tyson, G. W., and J. F. Banfield. 2005. Cultivating the uncultivated: a community genomics perspective. *Trends Microbiol* 13 (9):411-5.

Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428 (6978):37-43.

US Senate, Committee on Foreign Relations. 1994. *The Convention on Biological Diversity (Treaty doc. 103-20) : hearing before the Committee on Foreign Relations, United States Senate, One Hundred Third Congress, second session, April 12, 1994*. Washington, D.C.: US Government Printing Office.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304 (5667):66-74.

Wang, G. Y., E. Graziani, B. Waters, W. Pan, X. Li, J. McDermott, G. Meurer, G. Saxena, R. J. Andersen, and J. Davies. 2000. Novel natural products from soil DNA libraries in a streptomycete host. *Org Lett* 2 (16):2401-4.

Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* 95 (12):6578-83.

Winstead, E. R. n.d. *Genome News Network's Genome Glossary*. Available from: *http://www.genomenewsnetwork.org/resources/glossary/#g*.

Worthington, P., V. Hoang, F. Perez-Pomares, and P. Blum. 2003. Targeted disruption of the alpha-amylase gene in the hyperthermophilic archaeon Sulfolobus solfataricus. *J Bacteriol* 185 (2):482-8.

Wu, L., X. Liu, C. W. Schadt, and J. Zhou. 2006. Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification. *Appl Environ Microbiol* 72 (7):4931-41.

Zengler, K., G. Toledo, M. Rappe, J. Elkins, E. J. Mathur, J. M. Short, and M. Keller. 2002. Cultivating the uncultured. *Proc Natl Acad Sci USA* 99 (24):15681-6.

Zhang, T., M. Breitbart, W. H. Lee, J. Q. Run, C. L. Wei, S. W. Soh, M. L. Hibberd, E. T. Liu, F. Rohwer, and Y. Ruan. 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Bio*l 4 (1):e3.

Appendix A

Statement of Task

The committee will convene a workshop and other appropriate information gathering activities in order to define the scope of metagenomics, understand how it is being used now in various disciplines, the technical approaches being used by different groups, and how metagenomics may develop in the future. The report will frame the key scientific questions that could be addressed using the approach of metagenomics. It will also identify the major academic, governmental and potential commercial stakeholders in the field of metagenomics, both nationally and internationally. It will include findings about obstacles or difficulties current researchers are encountering (e.g., lack of awareness of the field, infrastructural needs, lack of consistency and standardization in data annotation and management). The report will also make recommendations concerning 1) the most promising directions to pursue to advance the field of metagenomics, 2) possible mechanisms for addressing infrastructure needs including the annotation and sharing of data, and 3) improving communication and collaboration between groups applying metagenomic techniques to different microbial communities. The committee will not make budgetary or government organizational recommendations.

Appendix B

Committee Biographies

Jo Handelsman *(Cochair)* is a Howard Hughes Medical Institute Professor at the University of Wisconsin-Madison in the Departments of Plant Pathology and Bacteriology. She received her Ph.D. in Molecular Biology from the University of Wisconsin-Madison in 1984 and joined the faculty of the University of Wisconsin-Madison Department of Plant Pathology in 1985. Her research focuses on the genetic and functional diversity of microorganisms in soil and insect gut communities. The Handelsman lab has concentrated on discovery of novel antibiotics from cultured and uncultured bacteria and on the role of antibiotics and other small molecules in robustness and communication in microbial communities. She has also contributed to the development of functional metagenomics, which facilitates the genomic analysis of assemblages of uncultured microorganisms through expression of their genes in a surrogate host. In addition to her research program, Dr. Handelsman is nationally known for her efforts to improve science education and increase the participation of women and minorities in science at the university level. She served on the National Academies' panel that wrote the 2006 report, *Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering,* which documented the issues of women in science and recommended changes to universities and federal funding agencies. In addition to numerous scientific research publications, Dr. Handelsman is co-author of two books about teaching: *Entering Mentoring* and *Scientific Teaching*. Dr. Handelsman is an editor for Applied and Environmental Microbiology and the book series, *Controversies in Science and Technology*, and a member of the National Academy of Sciences Board on Life Sciences and the National Institute of Medicine Forum on Microbial

Threats. She is a National Academies Mentor in the Life Sciences, a fellow in the American Academy of Microbiology, co-founder of the Women in Science and Engineering Leadership Institute, President of the Rosalind Franklin Society, Director of the Wisconsin Program for Scientific Teaching, co-director of the National Academies Summer Institute on Undergraduate Education in Biology, and chair-elect of the Department of Bacteriology at the University of Wisconsin-Madison.

James M. Tiedje *(Cochair)* is a University Distinguished Professor of Microbiology and Director of the Center for Microbial Ecology at Michigan State University. He received his B.S. degree from Iowa State University and his M.S. and Ph.D. degrees from Cornell University. He has 30 years experience leading internationally recognized research on understanding the ecology, physiology, and biochemistry of microbial processes important in nature and of value to industry, especially to find ways to destroy hazardous wastes and to use DNA-based technologies to explore the unknown microbial world. He has received the Soil Science Research Award from the Soil Science Society of America, the Environmental Award from the American Society for Microbiology and shared the 1992 Finley Prize given by UNESCO for research contributions in microbiology of international significance. He is Fellow of the AAAS and Fellow of the American Academy of Microbiology. Dr. Tiedje is past chair of EPA's Science Advisory Panel (FIFRA), past chair of the Soil Biology Commission of the International Union of Soil Science, is past president of the International Society of Microbial Ecology, and is the past President of the American Society for Microbiology. He also serves on the Department of Energy's Biological and Environmental Research Advisory Committee. He was elected to membership in the National Academy of Sciences in 2003.

Lisa Alvarez-Cohen is the Fred and Claire Sauer Professor of Environmental Engineering in the Department of Civil and Environmental Engineering at the University of California, Berkeley. She received her B.A. from Harvard University and her M.S. and Ph.D. in environmental engineering and science from Stanford University. Her research interests are on the microbial degradation of environmental contaminants in natural and engineered systems with focuses on emerging contaminants and application of innovative molecular tools. Dr. Alvarez-Cohen is an associate editor of Environmental Engineering Science and recently co-authored a textbook entitled *Environmental Engineering Science*.

Michael Ashburner is Professor of Biology at the University of Cambridge. He is the former Joint-Head of the European Bioinformatics Institute (EBI). Dr. Ashburner received both his undergraduate degree and Ph.D. at the

University of Cambridge, both in genetics. He then went to the California Institute of Technology as a postdoctoral fellow with Hershell Mitchell. In 1979, he returned to the Department of Genetics in Cambridge where he has been based since, as Assistant in Research, University Demonstrator, University Lecturer, Reader in Developmental Biology and Professor (Ad hominem) of Biology (since 1991). Dr. Ashburner's major research interests are in the structure and evolution of genomes. Most of his research has been with the model organism *Drosophila melanogaster*, about which he has written the standard research text (*Drosophila: A Laboratory Handbook*, Cold Spring Harbor Press, New York, 1989, 2nd ed. 2005). He was a member of the consortium which recently sequenced the entire genome of this fly. His research has covered a range of subjects, from classical genetics, developmental biology, cytogenetics to evolution, at both molecular and organismal levels. Dr. Ashburner is a founder of FlyBase, a major database for researchers using *Drosophila* as a model organism, and of the Gene Ontology Consortium, a project to provide infrastructure for biological databases by a defined taxonomy of gene function. He is a Fellow of the Royal Society of London and of the Academia Europeae, a foreign honorary member of the American Academy of Arts and Sciences, a member of the European Molecular Biology Organization, and past president of the British Genetical Society.

Isaac K.O. Cann is Assistant Professor in the Department of Animal Sciences at the University of Illinois at Urbana-Champaign. He received his B.Sc. at the University of Ghana and his M.S. and Ph.D. at Mie University in Japan. He was a Postdoctoral Research Associate in Microbiology at the Department of Animal Sciences at the University of Illinois at Urbana-Champaign from 1994-1997, and was then a Research fellow at the Biomolecular Engineering Research Institute in Osaka, Japan. Following that he was a Senior Research Scientist at the Biomolecular Engineering Research Institute. Since 2001, he has been Assistant Professor of Microbiology in the Department of Animal Sciences at the University of Illinois at Urbana-Champaign. He is a faculty member of the Institute for Genomic Biology at University of Illinois at Urbana-Champaign. He is a member of the American Association for the Advancement of Science, the American Society for Microbiology, and the American Society for Biochemistry and Molecular Biology.

Edward F. DeLong is a Professor in the Division of Biological Engineering and the Department of Civil and Environmental Engineering at the Massachusetts Institute of Technology. He received his B.S. in bacteriology at UC, Davis and his Ph.D. at UC, San Diego in Marine Biology. He is a Fellow in the American Academy of Microbiology, a Fellow in the American Academy of Arts and Sciences, a Gordon and Betty Moore Foundation Marine

Microbiology Investigator, and an Honorary Professor at the University of Queensland, Brisbane, Australia. Dr. DeLong is an Editor of Environmental Microbiology, and on the Board of Reviewing Editors at Science Magazine. He serves on the Scientific Advisory panel for the Canadian Institute for Advanced Research, the Scientific Advistory Committee for the DOE Joint Genome Institute, and the Fachbeirat (Advisory Board) for the Max-Planck-Institut fur Marine Mikrobiologie. Dr. DeLong's lab is currently engaged in applying contemporary genomic technologies to dissect complex microbial assemblages. While biotic processes that occur within natural microbial communities are diverse and complex, much of this complexity is encoded in the nature, identity, structure, and dynamics of interacting genomes in situ. This genomic information can now be rapidly and generically extracted from the genomes of co-occurring microbes in natural habitats, using standard genomic technologies. Dr. DeLong and his research group have pioneered and applied these and related technologies, to better describe and exploit the genetic, biochemical, and metabolic potential that is contained in the natural microbial world. The central focus is on marine systems, due to their fundamental environmental significance to the oceans, as well their suitability for enabling new development of technologies, methods, and theory. Results from Dr. DeLong's efforts include the discovery of new groups of marine planktonic Archaea, the development of cloning large genome fragments for characterizing indigenous microbes, and the unanticipated discovery of new photoproteins (protoerhodopsins) among many different groups of marine bacteria, using genomics.

W. Ford Doolittle is the Director of the Program in Evolutionary Biology of the Canadian Institute for Advanced Research, holds a Canada Research Chair in Comparative Microbial Genomics, and is Professor in the Department of Biochemistry and Molecular Biology at Dalhousie University. He received his B.A. in Biochemical Sciences from Harvard College and his Ph.D. from Stanford University. He undertook postdoctoral work with Sol Spiegelman (University of Illinois) and Norman Pace (National Jewish Hospital and Research Center, Denver). Over the years, he has contributed to proof of the endosymbiont hypothesis, development of archaeal genetics, many aspects of prokaryote and eukaryote phylogeny, and the "introns-early" and "selfish DNA" theories. Dr. Doolittle joined the Department of Biochemistry at Dalhousie in 1971. His laboratory currently employs culture-dependent (multi-locus-sequence-typing) and culture independent (fosmid-based metagenomic) methods to study recombination and lateral gene transfer in natural populations of hyperthermophilic bacteria and halophilic archaea. These microevolutionary studies are complemented with phylogenomic bioinformatics approaches to assessing the role of lateral gene transfer in microbial macroevolution and its implications for phylogenetic reconstruction and classification.

Claire M. Fraser-Liggett is Director of the Institute of Genome Sciences at the University of Maryland School of Medicine. Until 2007 she was President and Director of The Institute for Genomic Research (TIGR), which has been at the forefront of the genomics revolution since she co-founded the not-for-profit institute in 1992. Starting with her work in 1995 on the first bacterial genome to be sequenced, she has become an international leader in the field of microbial genomics and forensics. Dr. Fraser-Liggett has been a member of National Research Council committees on countering bioterrorism and on domestic animal genomics, and has served on review committees of the National Science Foundation, Department of Energy and the National Institutes of Health. Dr. Fraser-Liggett has published more than 200 articles in scientific journals and books. Before becoming TIGR's president in 1998, Dr. Fraser-Liggett was the institute's vice president of research and director of its microbial genomics department. Prior to that, she worked as a researcher at the National Institutes of Health. She is a summa cum laude graduate of Rensselaer Polytechnic Institute and received a Ph.D. in Pharmacology from State University of New York at Buffalo. She has received numerous academic and professional honors, including the E. O. Lawrence Award from the Department of Energy and the Promega Biotechnology Research Award from the American Society of Microbiology. In addition to her leadership of TIGR, Dr. Fraser-Liggett also holds professorships in Microbiology and Tropical Medicine as well as in Pharmacology at The George Washington University School of Medicine.

Adam Godzik is Professor and Program Director of the Program for Bioinformatics and Systems Biology at The Burnham Institute and Bioinformatics Core Leader at the Joint Center for Structural Genomics, UCSD. Dr. Godzik is a physicist who is now applying tools of physics and computer science to analyze biological systems. He developed several protein structure and function prediction algorithms and led development of large biological databases, integrating results of experiment and theoretical analysis of individual proteins and entire genomes. He is a member of the Bioinformatics editorial board. Dr. Godzik received his Ph.D. in physics from the University of Warsaw, Poland, and before joining the Burnham Institute worked at EMBL in Heidelberg and the Scripps Research Institute in La Jolla.

Jeffrey I. Gordon is the Dr. Robert J. Glaser Distinguished University Professor and Director of the Center for Genome Sciences at Washington University School of Medicine. He received his A.B. in Biology from

Oberlin College and his M.D. from the University of Chicago. Dr. Gordon joined the faculty of Washington University in 1981, after completing his clinical training in internal medicine and gastroenterology. He has remained at Washington University for his entire professional career. From 1991 to 2004, he was Head of the Department of Molecular Biology and Pharmacology. In 2004 he resigned as Department Head to become the first director of a newly founded Center for Genome Sciences. This new Center represents an interdepartmental, interdisciplinary, and multigenerational intentional community of faculty, post-docs and students who are geneticists, population biologists and biostatisticians, computational biologists and computer scientists, systems biologists and engineers, and microbiologists and ecologists. The focus of the Center is on comparative genomics and biodiversity, plus systems biology (an emerging area that seeks to describe how complex networks of interacting genes, proteins and metabolites function to maintain normal cells, and how these networks adapt to perturbations, including those brought about by various disease states). Dr. Gordon has published over 350 scientific papers, and is named as inventor or co-inventor on 23 US patents. He has received a number of honors in recognition of his scientific contributions, including election to the National Academy of Sciences and the American Academy of Arts and Sciences.

Margaret Riley is a Professor of Biology at the University of Massachusetts, Amherst. She received her Ph.D. in population genetics from Harvard University and performed postdoctoral research in microbial population genetics with a Sloan Postdoctoral Fellowship in Molecular Evolution. She joined the faculty at Yale in 1991 and recently moved to UMass Amherst. She has a broad set of research interests that range from studies of experimental evolution of microbes to developing novel antimicrobials and redefining the microbial species concept. Dr. Riley studies the evolution of microbial diversity, with a particular emphasis on the ecology and evolution of microbial toxins. She is co-founder of Origin Antimicrobials, Inc., whose mission is to discover and refine novel antimicrobials to address the challenge of antibiotic resistance. Dr. Riley is the Director of the Organismic and Evolutionary Biology Program and the Director of the Museum of Natural History at UMass Amherst. From 1999 to 2002 she chaired the Gordon conference on molecular evolution and from 2003 to 2005 she chaired the Gordon conference on microbial population biology and evolution. She is a fellow of the American Academy of Microbiologists.

Molly B. Schmid joined the Keck Graduate Institute (KGI) in January 2005 as Jacobs Professor and Entrepreneur-in-Residence. At KGI, she teaches Risks and Rewards in Drug Discovery and Development and continues to explore her interests in chemical genetics and antimicrobial drug discovery. Formerly, she was Senior Vice President of Preclinical Programs at Affinium Pharmaceuticals (Toronto, ON); Senior Director, Functional Genomics & Bioinformatics at Genencor International (Palo Alto, CA); and Vice President, Research Alliances, with Microcide Pharmaceuticals (Mountain View, CA). From 1986 to 1994, she was Assistant Professor of Molecular Biology at Princeton University, where her lab investigated bacterial chromosome structure and function, and her research group discovered Topoisomerase IV in *Salmonella typhimurium*, as well as a genetic strategy for identifying new antimicrobial targets. She is a Fellow of the American Academy of Microbiology, a Searle/Chicago Community Trust Scholar and a Damon Runyon-Walter Winchell Fellow. She received her Ph.D. in Biology from the University of Utah, and her B.S. from SUNY Albany.