A REPORT FROM THE AMERICAN ACADEMY OF MICROBIOLOGY

Large-Scale Sequencing:
the future of genomic sciences?

AMERICAN
SOCIETY FOR
MICROBIOLOGY

This report is based on a colloquium, sponsored by the American Academy of Microbiology, convened September 2008 in Washington, DC.

The American Academy of Microbiology is the honorific leadership group of the American Society for Microbiology. The mission of the American Academy of Microbiology is to recognize scientific excellence and foster knowledge and understanding in the microbiological sciences. The Academy strives to include underrepresented scientists in all its activities.

The American Academy of Microbiology is grateful for the generosity of the following organizations for support of this project:

The Gordon and Betty Moore Foundation
U.S. Department of Energy

The opinions expressed in this report are those solely of the colloquium participants and do not necessarily reflect the official positions of our sponsors or the American Society for Microbiology.

Large-Scale Sequencing:
the future of genomic sciences?

Prepared by: Margaret Riley, Ph.D., and Merry Buckley, Ph.D.

Large-Scale Sequencing:
the future of genomic sciences?

Large-Scale Sequencing:
the future of genomic sciences?

Colloquium Participants

Lisa Alvarez-Cohen, Ph.D.
University of California, Berkeley

Colleen Cavanaugh, Ph.D.
Harvard University

Michael P. Cummings, Ph.D.
University of Maryland

Rob Dorit, Ph.D.
Smith College

Aled Edwards, Ph.D.
University of Toronto, Canada

James E. Galagan, Ph.D.
Broad Institute, Massachusetts Institute of Technology

George M. Garrity, Sc.D.
Michigan State University

Paul S. Keim, Ph.D.
Northern Arizona University

Cheryl Kuske, Ph.D.
Los Alamos National Laboratory

Jeffrey G. Lawrence, Ph.D.
University of Pittsburgh

Ruth E. Ley, Ph.D.
Cornell University

Mary Ann Moran, Ph.D.
University of Georgia

Gary J. Olsen, Ph.D.
University of Illinois

Martin F. Polz, Ph.D.
Massachusetts Institute of Technology

Tatiana Rynearson, Ph.D.
University of Rhode Island

Gretta Serres, Ph.D.
Marine Biological Laboratory, Woods Hole

David A. Stahl, Ph.D.
University of Washington

William Whitman, Ph.D.
University of Georgia

Sponsor Representative Participants

Dan Drell, Ph.D.
U.S. Department of Energy

Lita Proctor, Ph.D.
National Science Foundation

Executive Summary

Genetic sequencing and the various molecular techniques it has enabled have revolutionized the field of microbiology. Examining and comparing the genetic sequences borne by microbes—including bacteria, archaea, viruses, and microbial eukaryotes—provides researchers insights into the processes microbes carry out, their pathogenic traits, and new ways to use microorganisms in medicine and manufacturing.

Until recently, sequencing entire microbial genomes has been laborious and expensive, and the decision to sequence the genome of an organism was made on a case-by-case basis by individual researchers and funding agencies. Now, thanks to new technologies, the cost and effort of sequencing is within reach for even the smallest facilities, and the ability to sequence the genomes of a significant fraction of microbial life may be possible. The availability of numerous microbial genomes will enable unprecedented insights into microbial evolution, function, and physiology. However, the current *ad hoc* approach to gathering sequence data has resulted in an unbalanced and highly biased sampling of microbial diversity. A well-coordinated, large-scale effort to target the breadth and depth of microbial diversity would result in the greatest impact.

The American Academy of Microbiology convened a colloquium to discuss the scientific benefits of engaging in a large-scale, taxonomically-based sequencing project. A group of individuals with expertise in microbiology, genomics, informatics, ecology, and evolution deliberated on the issues inherent in such an effort and generated a set of specific recommendations for how best to proceed.

The vast majority of microbes are presently uncultured and, thus, pose significant challenges to such a taxonomically-based approach to sampling genome diversity. However, we have yet to even scratch the surface of the genomic diversity among cultured microbes. A coordinated sequencing effort of cultured organisms is an appropriate place to begin, since not only are their genomes available, but they are also accompanied by data on environment and physiology that can be used to understand the resulting data. As single cell isolation methods improve, there should be a shift toward incorporating uncultured organisms and communities into this effort.

Efforts to sequence cultivated isolates should target characterized isolates from culture collections for which biochemical data are available, as well as other cultures of lasting value from personal collections. The genomes of type strains should be among the first targets for sequencing, but creative culture methods, novel cell isolation, and sorting methods would all be helpful in obtaining organisms we have not yet been able to cultivate for sequencing. The data that should be provided for strains targeted for sequencing will depend on the phylogenetic context of the organism and the amount of information available about its nearest relatives.

Annotation is an important part of transforming genome sequences into useful resources, but it represents the most significant bottleneck to the field of comparative genomics right now and must be addressed. Furthermore, there is a need for more consistency in both annotation and achieving annotation data. As new annotation tools become available over time, re-annotation of genomes should be implemented, taking advantage of advancements in annotation techniques in order to capitalize on

the genome sequences and increase both the societal and scientific benefit of genomics work.

Given the proper resources, the knowledge and ability exist to be able to select model systems, some simple, some less so, and dissect them so that we may understand the processes and interactions at work in them. Colloquium participants suggest a five-pronged, coordinated initiative to exhaustively describe six different microbial ecosystems, designed to describe all the gene diversity, across genomes. In this effort, sequencing should be complemented by other experimental data, particularly transcriptomics and metabolomics data, all of which should be gathered and curated continuously.

Systematic genomics efforts like the ones outlined in this document would significantly broaden our view of biological diversity and have major effects on science. This has to be backed up with examples. Considering these potential impacts and the need for acquiescence from both the public and scientists to get such projects funded and functioning, education and training will be crucial. New collaborations within the scientific community will also be necessary.

Given the proper resources, the knowledge and ability exist to be able to select model systems, some simple, some less so, and dissect them so that we may understand the processes and interactions at work in them.

Introduction

Until recently, studying microbial life and all its diversity was like exploring a new territory without the benefit of a map. By feeling around the terrain of this uncharted land, examining one organism at a time, microbiology made incremental progress in cataloguing the bacterial, archaeal, viral, and microbial eukaryotic life on this planet. We learned enormous detail about a very small fraction of microbial diversity, and we treated each species in isolation, as if it naturally lived on its own.

Today, genetic sequencing, whole genome sequencing, and the techniques enabled by sequencing are revolutionizing the process of discovery in microbiology and moving the field forward more quickly than ever before. These tools, and the techniques they have given rise to, including genomics, proteomics, transcriptomics, metabolomics, molecular phylogeny, and others, are putting powerful information about the relationships among microbes, their key pathogenicity resources, and their metabolic potential in the hands of scientists.

Current approaches to genome sequencing are organism-driven; that is, they are geared to determining the genome of a single organism or group of related organisms. The choice of organism is almost always driven by one of two motivations: either it possesses a special property or it is responsible for an interesting process (for example, the organism may be the causative agent of a disease or have a particular metabolic role), or the organism represents a taxonomic group otherwise unrepresented in the current sequence database. In any case, genome sequences provide clues into the metabolic potential of particular organisms, insight into their ancestry, the tempo and mode of evolution, and fodder for devising ways to manipulate them, either to inhibit the biological processes they catalyze (in the case of pathogens) or to facilitate them (in the case of those with industrial or biotechnological utility).

Organism-driven genomics approaches were appropriate when genome sequencing required a large expenditure of time, effort, and resources. Today, a broader view is required, since sequencing technology is changing so quickly that single bacterial

or archaeal genomes can be produced in a matter of hours and at a fraction of earlier costs. The ease with which sequence data can now be acquired allows researchers to examine a variety of microbial-mediated processes, such as bioremediation, environmental adaptation, speciation, or evolution. In this document, we propose to expand the scope of sequencing efforts to the microbial communities cooperating to sustain entire ecosystems. Although the scale of the focus has increased, the target of the sequencing effort remains the same: to characterize the genetic basis of microbiologically-mediated processes.

A Coordinated Large-Scale Sequencing Project

This report summarizes discussions at a colloquium convened to deliberate on the possibilities and benefits of engaging in a large-scale, taxonomically-based sequencing project to explore microbial diversity. The goal of the meeting was to review the primary issues inherent in a large-scale sequencing effort and to generate recommendations about how best to proceed. Participants also discussed where the field of microbial genomics is heading and identified some of the opportunities, challenges, and benefits of large-scale sequencing.

The scientific impact of a coordinated, large-scale sequencing effort could be quite significant. For starters, obtaining genome sequences of representatives of the most deeply divergent lineages of bacteria and archaea will allow better understanding of the history and diversity of life on the planet. Although these domains comprise the greatest diversity of life on this planet, we have yet to achieve an understanding of their taxonomic space, and there remains a fundamental need for genome information.

Equally important is the comparison of genomes of very closely related strains to inform population genomic studies of the evolutionary processes that happen at the tips of the tree of life. The rates and bounds of gene transfer, for example, which is a major force in shaping genomes, remain largely unknown. Significant advancements in medicine also await discovery through sequence-based population genetic studies. Pathogens are formed at the tips of the trees when harmless variants pick up pathogenicity determinant genes via lateral gene transfer. Genomics will be the key to understanding the genetic background of potential pathogens and the kinds of genes they receive via lateral gene transfer.

Understanding speciation and ecological specialization requires a broad look at genomes at varying levels of differentiation. Species criteria can be seen as thresholds that are crossed at different times in the evolution of lineages. In fact, this may be the only way to explore the nature of the microbial species concept. To make progress on this critical issue requires taxon sampling at different phylogenetic levels.

With the exception of the human microbiome project, microbial genome sequencing has largely been carried out to answer specific or immediate problems. A coordinated, large-scale effort would allow questions to be answered that lie beyond the scope of investigator-initiated research programs and catalyze the development or refinement of new technologies required to complete the project in a timely fashion. A systematic approach to sequencing also eliminates potential duplication of

effort, balances data-gathering that is performed by more than one method, creates a greater value, and is ultimately cheaper by preventing redundancy.

The subsequent sections outline recommendations for large-scale microbial sequencing efforts directed toward cultivated isolates and single cells, as well as a community-scale approach to characterize a set of defined ecosystems of varying complexity. Genome sequences from both cultured and uncultured organisms are extremely valuable, but, initially, there should be more emphasis on cultured organisms since they are readily available and are often accompanied by important metadata. As single cell approaches in genomics improve, there should be a shift toward uncultured organisms and communities. These proposals entail sequencing thousands of genomes, including type strains and user-proposed genomes. These expectations may need to be adjusted and scaled upwards as sequencing technologies improve. In later years, as sequencing becomes less costly, phylogenetically diverse eukaryotes should be targeted with particular attention.

Sequencing Cultivated Organisms

Sequencing genomes of organisms that are already in culture would provide a broad understanding of the tree of life and allow researchers to link genomic data with existing phenotypic information. This effort will provide a scaffold or map of knowledge about the microbial world upon which future information can be hung and placed in the context of a coherent set of "knowns." This contextual information would enable scientists to interpret the diversity, functions, and richness of newly discovered organisms and could change the format of biological education.

Sequencing bacterial and archaeal 16S genes and their corollaries in eukaryotes, 18S genes, has provided a great deal of information about phylogenetic diversity, but extensive sequencing of entire genomes from the culture collections is needed to provide a baseline of information to expand the knowledge of microbial diversity. It is important to leverage the existing resources in this sequencing effort by sequencing genomes of organisms for which phenotypic data are already available, but sequencing efforts should also include cultured isolates about which we know little or nothing. Each of these efforts would provide a baseline reference library for metagenomic sequencing projects.

Genome sequences for cultivated microorganisms would also provide information and insights about the biochemical diversity that gives rise to novel products, enzymes, and industrial biological chemicals for use in medicine, manufacturing, and other applications.

An organized effort is needed to cover the tree of life in an unbiased way. Although individual-based efforts are effective at concentrating resources on particular phylogenetic groups, a separate, more disseminated approach is needed to capture diversity across the broader spectrum. The specific number of full genome sequences that are needed to reveal the diversity of the microbial world is open to discussion, but a sequencing effort like that described here should include the major branches of all three domains (bacterial, archaeal, and eukaryal) and ensure deep coverage of the individual branches.

Criteria for Selecting Organisms to Sequence

The sequencing effort should include characterized isolates currently in culture collections where we already have biochemical data and pursue and locate other cultures of lasting value from personal collections. The genomes of type strains should be sequenced where metadata and phenotype data are available. Other obvious choices for genome sequencing include organisms for which a good deal of physiological and functional data is known. In a sequencing effort like the one described here, the cost of the actual sequencing would probably amount to only a small fraction of the total costs of the program, since culturing, describing, and the other aspects of producing organisms can be time-consuming and expensive. Considering that the cost of describing an isolate far outweighs the cost of sequencing its genome, it may be advisable to sequence any strain for which a minimum of baseline data are available.

Considering that the cost of describing an isolate far outweighs the cost of sequencing its genome, it may be advisable to sequence any strain for which a minimum of baseline data are available.

User demand for an organism will drive the upfront work of sequencing, like DNA preparation, as well as much of the follow-up work. Hence, some metric of user demand should be incorporated into the selection of strains for sequencing.

Some other considerations for selecting species for sequencing include:

- Branches of the phylogenetic tree that include representatives with **important industrial, pharmaceutical, or medically significant species** should be deeply covered.
- There is also value in picking **exemplary phylogenetic clades** to analyze in detail. For example, the genus Clostridium contains many species of interest for medical, environmental, and bioenergy problems. In-depth sequencing will define the family better and provide a parts list of that group that will have value for all the above applications.
- When given the choice, it may be advisable to sequence **minimally cultured isolates**, since microorganisms tend to lose phenotypic characteristics and genes quickly under the selective pressure of passage through cultivation.
- The **ecological context and numerical representation** of an organism are other factors to take into account. Sequences that represent numerically dominant organisms are of greatest help to metagenomics studies and understanding ecosystem function and biogeochemistry. The function of an organism in biogeochemical cycles is another consideration that will require sequencing of diverse organisms.
- The **culture collections administered by the World Federation of Culture Collections** should be formally included in this sequencing effort for a number of reasons. Collaboration with pre-existing culture collections would ensure a high standard of quality (that would be hard to achieve through user-driven efforts) and facilitate connections with international sequencing efforts, although clear criteria for data release must be negotiated. Culture collections possess expertise in maintaining cultures and isolating DNA for sequencing. Also, isolates from culture collections have been characterized to some extent, providing information that can be used to interpret sequence-based discoveries.
- Although culture collections hold many of the strains that should initially be sequenced, to obtain more complete coverage of the microbial domains, it will be necessary to go beyond the culture collections and **collect cultures and DNA**

from individual investigators. Since this may be a difficult task, a coordinated and systematic effort in collaboration with the sequencing centers may be used.

- **Quality control of cultures and DNA preparation are critical**, but difficult, tasks in genome sequencing. A series of guidelines and quality control steps are needed to ensure the results of sequencing cultivated isolates are reliable.

Cultivation: Methods & Facilities

A large-scale sequencing effort should encompass two parallel processes with equal emphasis:

- Sequence cultured microbes to fill out missing information in the tree of life, and
- Continue efforts to capture not-yet-cultured organisms to fill in areas of the tree that lack representatives.

Creative culture methods, novel cell isolation, and sorting methods would all be helpful in isolating organisms we have not yet been able to cultivate for sequencing. Some approaches may lead to culturing the desired organism; others may not. Enrichment, for example, can also be effective. Organisms responsible for anaerobic ammonia oxidation were only 80% enriched when they were isolated and identified. However, it is preferable to isolate the desired organism in a pure culture. One approach to targeting organisms for culture would be to describe the rRNA phylogeny for a given community, then seek out representative organisms to map onto that tree.

Since cultivation of novel organisms is difficult and often has an unpredictable outcome, researchers at academic institutions may be hesitant to allocate significant resources to this effort. Granting agencies tend to discourage projects that stand a chance, however slim, of failing to produce results. A central culturing facility could incorporate some of the more novel and high throughput culturing methods (gel microdroplet).

Sequencing Coverage

Genome sequences of cultivated isolates should be completed and closed. Approximately 20X coverage for a 5 mb genome would be sufficient. The rates of sequence production should be determined in collaboration with genome centers.

Phenotypic, Ecological, and Other Data That Should Accompany Sequencing Data

The data that should be provided for strains targeted for sequencing will depend on the phylogenetic context of the organism and the amount of information available about its nearest relatives. For type strains, it would be ideal to secure a great deal of information, but no strain should be excluded from consideration solely because

we lack metadata for it. For non-type strains, population structure and environmental context (including habitat, niche) would ideally be available and environmental MLST (multi-locus sequence typing) data should be available to provide context. However, this will not be possible for all groups.

Currently, a great deal of the phenotype and environmental data needed to buttress a large-scale sequencing effort are lacking. We need to invest in the annotation and discovery of functions of the genomes that are generated by the effort described here. Physiology studies are needed to go with the sequencing effort.

In an ideal world, every bacterium, archaeon, and eukaryotic microbe could be grown to high density in a pure culture.

Single Cell Sequencing (Unculturable Individuals)

In an ideal world, every bacterium, archaeon, and eukaryotic microbe could be grown to high density in a pure culture. Unfortunately, microbes do not comply with our fondest wishes, and even the most studied environments and niches are home to numerous "unculturable" strains. Microbial communities of the human gut, for example, have been the focus of scores of culturing efforts, but molecular studies reveal (1) that there are few representative cultivated strains from these communities, and (2) an entire phylum present in many people's guts that has yet to yield to cultivation. Researchers seeking the genome sequence of uncultured microorganisms are left with few choices.

Single cell approaches, cell sorting (especially sorting and capture to physically separate cells), and targeted metagenomics methods are needed to bridge this gap and help fill in the empty spots in microbial phylogenetic trees. Reference genomes acquired in this way can also assist in the interpretation of metagenomics data and extract more information on human health and disease from studies of the human microbiome. Different organisms will probably require different approaches, depending, in part, on their abundance.

Single cell technologies required for these efforts include:

- Micromanipulation, optical tweezers for separation of individual, morphologically distinct cells;
- Encapsulation and sorting after growth;
- Chemostats to create low nutrient environments;
- Tagging techniques (e.g., hybridization and antibodies) combined with sorting (e.g., flow cytometry);
- Magnetic separation;
- Germ-free organisms can be used to enrich from certain parts of communities;
- Stable isotope labeling to separate DNA from cells with a specific metabolic capability; and
- Single cell genome amplification (needs to be improved).

Annotation Efforts

Annotation is a critical part of making genome sequences into resources, but it represents a significant bottleneck right now. Certainly, more genome sequences are needed; therefore, a parallel effort in scaling up the analysis, manipulating, and understanding these sequences must also be undertaken. A major effort is needed in functional annotation to understand the roles encoded by the DNA sequences we already have. Additional initiatives are important to attract new intellectual capital to the problem. Unfortunately, there are no sources of ongoing money to do this.

Annotation may be defined as:

- Labeling and identifying features on genome sequence,
- Providing metadata that allows interpretation of the sequence information, and
- Complementary function information derived by sequence homology comparisons and experimentation.

Hence, annotation should be interpreted not just as "gene calling," but also as characterizing the variability in abundance, sequence, expression patterns, and so forth.

Annotation needs to be open-ended so we can identify new functions and pathways, not just rediscover old ones.

Annotation Needs

Moving forward in microbial genomics, there are a number of needs with respect to annotation that remain to be addressed. Perhaps chief among these is the need for consistency in annotation, documenting the annotation, and in archiving annotation data. A set of standard annotation platforms and a central annotation resource (which can be disseminated but operate from a central server) with defined methods and standards, and guidelines for using the resources would be ideal. This resource should be centrally curated so that the information is updated and kept accurate. Annotation needs to be open-ended so we can identify new functions and pathways, not just rediscover old ones. Also, it is useful to document the annotation, i.e., know what a gene has been determined NOT to be. This type of data should be included in annotation databases.

Currently, functional annotation is carefully tailored to the gene of interest and performed on a case-by-case basis. High-throughput methods for experimental analysis of function (e.g., enzyme assays) are needed to improve annotation. Function-based screening, for example, in which cloned genes are expressed in different microbial hosts and the hosts are then put on a diverse set of substrates, could produce annotation results quickly and efficiently.

There are many different annotation pipelines in use today that result in widely divergent annotations. A comparative, systematic study is needed to identify the best annotation methods. Researchers need to establish a gold standard for annotation. This is not a difficult task, but it is not being tackled in an organized, logical way. One approach to determine which pipeline/software is best would be to select a group of organisms and genes to annotate and then functionally test the results.

Organizing Annotation Data

Annotation data should be organized in a dynamic fashion, which would enable easy revisions and expansion. Metagenomic data are often queried and compared to find correlations among genes in different types of environments. Metadata should be linked to genome annotations (and be searchable) to facilitate these comparisons. Metadata could include the history of the organism, citations, functional verifications, crystal structures, physiology, growth conditions, and other facts.

Keeping Annotations Up-to-Date

Under the current funding paradigm, annotation funding is usually piggy-backed on sequencing funds. As a result, annotation often comprises a one-time effort, annotations are not updated, and new information relevant to a given annotation is not conveyed to the broader scientific community. Sequencing can be completed, but the work of annotation never ends. Genomes should be re-annotated over time so we can make full use of these sequences, take advantage of advancements in annotation techniques, and increase both the societal and scientific benefit of genomics work. By coupling genome projects with community-based annotation, the scientific community could curate and update annotations so that new information, such as refined functions, will be readily accessible.

There are currently automated annotation systems to which a researcher can upload sequences for re-annotation. These resources need to be improved continuously to update the algorithms, but there is currently no financial support for maintaining accurate annotations, so re-annotation is done on a case-by-case basis, and the information is not disseminated. To handle revised annotations, Genbank might become the repository for the primary sequence information, and a new database will contain the updated information. This might happen best as parallel versions of the annotation (such as done in the gene ontology annotation).

An annotation wiki tool would allow anyone in the microbiology community to contribute to the annotation of microbial genomes. The resource could be organized like Wikipedia, where many people add content based on their expertise, there are mechanisms of review and quality control, and there is an infrastructure to allow for these contributions that has few barriers to participation. An existing wiki system could be adapted to the specific needs of the large-scale annotation and has already been implemented for *E. coli*.

The driving motivation for collecting microbial sequence data is to learn about the phenotype, function, or environment of an organism. Unfortunately, errors, emissions, or dated information are common in genome annotations, a fact that diminishes the value of the information and faith in the data. When researchers are given the opportunity to correct or add information to annotations, the responsibility is shared among those who value the data most.

The benefits of a wiki system for annotations include:

- The information already known for many species, strains, genes, proteins, etc., could be linked to genomic information (since it is only in this context that the strings of A's, C's, G's and T's have meaning);
- The broader community would maintain ownership of the genomic data, as well as the associated benefits and responsibilities; and
- Data could be kept dynamic and current, since anyone at anytime could update it or give alternative interpretations.

A Community-Scale Sequencing Initiative

Although genomics, proteomics, transcriptomics, and other techniques now offer powerful means for addressing gaps in our knowledge about microbial life and exchanges, many of the current metagenomics projects amount to mere snapshots of communities. Given the proper resources, the knowledge and ability exist to be able to take a set of model systems and understand the processes and interactions therein. The time is right to go above and beyond simple descriptions of microbial life and begin to tackle the complex genetic and chemical interactions at work in microbial ecosystems.

This section presents a proposal for a five-pronged, coordinated initiative to exhaustively describe six different microbial systems. The project outlined here seeks to apply what is learned about genome sequences from pure cultures and single cell approaches to the members and interactions of actual communities. In this way, genomes acquired from pure cultures and single-cell approaches can be used as phylogenetic reference points or anchors to inform and interpret community data. Genome sequencing for its own sake produces haphazard results. This project will maximize the value of genomics as a tool, not the end goal, by coupling it with other approaches, including transcriptomics and physiology.

Mission-based science is valuable for making discoveries, but questions, however broad, should guide this initiative, and questions are answerable only if we draw boundaries around what we want to know. Some of the questions this work will address include:

- How do systems respond to transient or long-term perturbations? What role do organisms play in that response, and how does genomics illuminate those responses?
- What is the prevalence and what are the impacts of lateral gene transfer?
- Specific physiological and ecological questions about genomes and products of genomes. We need to interrogate the dynamics of the conversation at the genome and information level.
- Do very rare members of communities play important functional roles?
- We tend to think of bacterial genomes as "mom and pop stores" that specialize in particular interactions. Why are there not more "Wal-Mart" genomes that can access scores of different functions?
- What are the rules of assembly and rules of stability in communities? This involves knowing the distribution and abundance of every participant and its transcriptome. We need anchors for physiology (such as the American Type Culture Collection) to understand what the genes are doing.

It will be critical to collect and sequence each ecosystem thoughtfully and systematically, collecting as many sequences with as complete a view of the system genome as possible. Thoroughness will pay off, especially, for example, if future advances in recognizing viral hosts by genome signatures allow us to pull out new information from old sequence data. Sequence data are forever, and our ability to interpret them will constantly improve.

The Benefits of a Large-Scale Community Sequencing Initiative

A large-scale community initiative of the scope and design proposed here would yield numerous advancements for biology, medicine, and education. The opportunity to better assess unexplored microbial diversity ranks chief among the justifications for this genomics initiative. Researchers in microbiology have long lamented the existence of "orphan" 16S sequences: unique signatures for which no other information—neither sequence nor phenotype—is available. A large, centralized effort like the one described here would capture many of those orphan genomes by:

- Sequencing community mixtures to enough depth to capture the entire genome,
- Spending resources on obtaining the samples, metadata, and
- Providing funds to an investigator to increase depth of coverage of their metagenome so that the scientific community captures the genome of the missing organism.

Ecology is not always a predictive science, but the outcomes of this effort should be predictive and offer many new insights into the rules of ecology. For example, understanding ecosystems requires knowing how function is distributed globally and locally. Through in-depth exploration of complex natural communities, we can begin to grasp how redundant function is within ecosystems. Biogeography is an important component of this as well, since it remains unknown to what extent microbes are impacted by anthropogenic change. Also, by targeting entire communities from diverse environments, the project should inform our understanding of microbial interactions. Understanding community assembly will necessitate studying the source communities from which the target community receives input.

The ecology of climate change may offer a nice hook to engage the public and emphasize the societal benefits of this project. Simply put, microbes run the planet, and we are changing the planet. It is imperative that we harness the agility of the microbial world for the benefit of ecosystems and the global climate.

The project will also offer insight into infectious disease. Pathogenicity is not solely a function of virulence factor genes; the perturbations of the ecosystem that allow pathogens to emerge play an equally important role. This project can serve to educate the medical community about the intersection of infectious disease and ecology.

The educational aspects of a large-scale genomics program should be planned from the beginning and integrated into the project, not left to the last minute. It could be worthwhile to involve high school and middle school students in the project by demonstrating for them the processes of sequence generation, extracting DNA from various environments, and isolating and culturing organisms from system samples.

A large-scale community initiative of the scope and design proposed here would yield numerous advancements for biology, medicine, and education.

Whatever the approach, getting teachers and students involved in the project early is very important.

Experimental Design, Data Collection

This proposal outlines a plan to sequence the genomes of representative organisms which, together, comprise entire ecosystems. The goal is to describe all the gene diversity, across genomes, in ecosystems with different degrees of complexity. Sequencing should also be complemented by other experimental data and by "–omics" data, especially transcriptomics and metabolomics data. Moreover, there should be an iterative process between isolation and metagenomics. Researchers need to define genome projects more richly than they currently do. Genome projects should collect and analyze genome sequences, transcriptomes, proteomes, and, eventually, other features, like small functional RNAs. The definition of what comprises a genome will probably evolve over time, as new sorts of information become available.

The goal is to describe all the gene diversity, across genomes, in ecosystems with different degrees of complexity. Sequencing should also be complemented by other experimental data and by "–omics" data, especially transcriptomics and metabolomics data.

The goal of this initiative is to understand biological processes. This will involve sophisticated sampling to capture the complexity of the microbial world.

Microbial communities representing different levels of diversity and complexity and consisting of combinations of microbial eukaryotes, bacteria, archaea, and viruses should be selected for this study. The following steps will be employed for each community:

1. **Extract DNA from the entire community.** Metagenomic sequencing will be performed. Aside from providing a preliminary gene inventory, these data will provide a ribotyping profile of the community and identify both the major and minor taxonomic groups present.

2. **Determine at least one complete genome sequence from single cell isolates for all organisms found above threshold abundance.** Genomes will be determined from single cells, so the capacity to culture the organism is not required. Attempts can be made to propagate representatives from the ecosystem. The number of genomes to sequence will be determined according to the abundance of the taxonomic groups, so that the most abundant groups have the most representatives sequenced. Cells will be chosen by FISH to ensure both appropriate sequence representation among abundant genomes and inclusion of genomes from minority taxa. Single cell genomics plays an important role in this proposal, since it captures the population biology of the gene, whereas bulk metagenomic sequencing only captures the average gene. Single cell culturing is another possible component of the project. Sufficient material can be collected for single cell genome sequencing by isolating single cells using dilution to extinction, followed by division of that single cell so that a total of at least 64 cells can be amassed. The unused cells would be available for archiving. Since some cells might only grow in co-culture, this dilution model may need to be modified.

3. **Survey culture collections for organisms with close relationships to members of the ecosystem being analyzed.** Those organisms found with a significant body of experimental data associated with them should be sequenced to provide

phylogenetic reference points or "anchors" to both cultured and culturable strains. Having cultured organisms as "anchors" is a plus, but not a prerequisite for moving forward with the project.

4. **Assess the complete transcriptome of the community through mRNA sequencing.** Sequencing mRNA should accompany the initial metagenomic data collection.

5. **Assess the transcriptomes of communities perturbed in defined, relevant ways.**

Ecosystems Selected for Study

The six ecosystems selected for this project should represent six different levels of complexity. The least complex system should contain around 10 different genome sequences; the most complex system gets 1 million genomes sequenced. The simplest system could be an artificial community comprised of eight bacteria and two viruses, for example. This is a proof-of-concept project designed to characterize an ecosystem with relatively modest complexity. The consortia could be assembled in the laboratory so that they may be replicated in other laboratories and followed to see the effect of various controlled perturbations over time. These communities would be simple enough that we could sequence everything that is there through time and see exactly what changes occur at the community genome and transcriptome level.

The more complex systems selected for this project should entail the following numbers of members and genomes:

1. 1,000 genomes will be determined from an ecosystem with ~50 species present at 0.1%.
2. 10,000 genomes will be determined from an ecosystem with ~500 species present at 0.01%. This project will catalyze an increase in the flexibility and capacity of all data pipelines, and allow exploration of more complex ecosystems.
3. 100,000 genomes will be determined from an ecosystem with ~5,000 species present at 0.001%. Here, collection methods will be expanded to include strains collected from a series of sites, all representative of the ecosystem in question; for example, soil samples may be collected from a number of sites. Variability in ecosystem composition will be assessed by comparing species composition and abundance between isolates.
4. 1,000,000 genomes will be determined from an ecosystem with >25,000 species present at 0.0001%. This is the most ambitious project, to be undertaken when the others have catalyzed improvements in the DNA preparation, sequencing and annotation pipelines. This project will produce ~ 3,000,000 MB of DNA sequence data, or approximately 1,000 human genomes-worth of information.

There are many candidate ecosystems for study at each level of complexity, including:

- Communities exploiting environmentally extreme niches, such as those within hydrothermal vents or hyper-saline lakes,
- Host-symbiont assemblages,

- Multi-species biofilms,
- Communities confined to host intestinal, cloacal or naso-pharynx tissues,
- Defined layers of fresh- or salt water columns,
- Marine or fresh-water sediment, and
- Soil communities.

Model systems might be very useful for this project, particularly those for which a great deal about the community is already known. Examples include model systems used in evolution experiments, carefully selected sites, or animal-associated communities (especially endangered animals). Anaerobic digesters represent another good choice, since they are well understood and distributed across the world. The corn rhizosphere might also be a useful system to target, since the results can be used to better understand plant-microbe interactions and assist in efforts to optimize bioenergy production, biocontrol, or fertilizer reduction.

The project should include at least one experimental system that can be manipulated.

Knowing how much natural noise exists in the system at baseline and the sources of variation and variability are critical for evaluating whether responses to perturbations are significant and what they mean. This means understanding things such as natural noise and variability in gene order, genetic polymorphisms, and expression variability at the population level. It's important to incorporate the notion of natural variability into the six "experiments," whatever they might look like.

At each step of the way, separate approaches will be required for eukaryotic, bacterial, archaeal, and viral components of the ecosystems. Viruses present a special problem because they are difficult to collect (particularly the lytic ones) and it is difficult to tell which hosts they infect. One idea to address this is to dedicate one of the six "model ecosystems" as a viral-focused project.

Annotation for the Initiative

Annotation, broadly defined, should be a central and public component of these community based genome projects. Because the sequence of the ecosystem is considered as a whole, the nature of sequence annotation will differ from the traditional conceptualization. Rather than a simple description of the functions and features found within a single genome, this systems-level annotation should also include:

- An indication of the frequency of features within isolates,
- Population-level sequence variation (SNPs, gene order, absence/presence, polymorphism, length variation, gene order) of features,
- Information regarding levels of transcription collected from the mRNA sequencing efforts,
- Variability in expression,
- Transcriptional plasticity,
- Physiology and metabolism of isolates,
- Contextual data, and
- Changes in transcription in response to relevant perturbation of environmental conditions.

A project team, following a general model that facilitates community access to the data, should manage annotation of each ecosystem sequence database. Although the project team should retain responsibility for curation of the sequence annotation, the integration of additional data from the community ought to be encouraged.

The large-scale sequencing project should include continuous curation as a central role of the program. Each of the six projects could have a dedicated annotation/curation/repository system, which might be more manageable than a single system. Alternatively, a centralized system might be more cost effective and have a better chance of garnering financial support for a longer time (that is, it would not die with the end of the sequencing project).

Implications for Training, Education, and Outreach Efforts

Systematic genomics efforts like the ones outlined in this document would significantly broaden our view of biological diversity and have major effects on science. Considering these potential impacts and the need for acquiescence from both the public and scientists to get such projects funded and functioning, education and training will be crucial. New collaborations within the scientific community will also be necessary.

There are three groups on which education efforts should focus: scientists, the public, and educators. As to the scientists, many of them who work outside the field of microbiology do not appreciate the magnitude of microbial diversity and might not appreciate the significance of research advancements in this field. We need to change this thinking so that institutions value culture collections and the effort it takes to generate and maintain them. There must be a significant outreach effort that speaks to scientists to convey the importance of microbes for the planet.

Many microbiologists do not know how to use current genomics tools or make the most of the genomic data we already have, much less the data that will be generated through these projects. Genomics tools are available on the internet, but they need to be supported by organized classes for target groups that lower the barriers for people to learn the skills necessary to use the tools. Modular courses, like the summer courses available at the Marine Biological Laboratory, could be used to educate the scientific community in the use of genomic analysis tools, the value of systematics, or how to culture and maintain culture collections. There is also a need to train people to be genome curators.

There is a need to capture the public enthusiasm for this sequencing project. Effective public relations efforts can take the focus on microbes away from pathogens and toward the beneficial things that microbes make or do for us, including natural products and antibiotics.

We need to train K-12 educators so they can include these topics in their curricula. Educators could be given information about the topics related to the sequencing project through NSF-type activities and trickle down from training of scientists.

As to the scientists, many of them who work outside the field of microbiology do not appreciate the magnitude of microbial diversity and might not appreciate the significance of research advancements in this field.

Providing an open annotation model and a more dynamic annotation network would encourage new collaborations among scientists to support this sequencing effort and instill a feeling of collective ownership of the process and data. Better interfaces are needed between the existing genome centers and scientists and between large-scale protein structure institutions and genome sequencing groups. Institutional and agency pipelines make it difficult to collaborate across agency lines.

Recommendations

- Efforts to sequence cultivated isolates should target characterized strains from culture collections for which biochemical data are available, as well as other cultures of lasting value from personal collections. The genomes of type strains should be among the first targets for sequencing.

- Some metric of user demand should be incorporated into the selection of strains for sequencing.

- Branches of the phylogenetic tree that include representatives with important industrial, pharmaceutical, or medically significant species should be deeply covered.

- Sequencing should also be complemented by other experimental data and by –omics data, especially transcriptomics and metabolomics data. Moreover, there should be an iterative process between isolation and metagenomics.

- A major effort is needed in functional annotation to understand the DNA sequence we already have. Annotation is a critical part of making genome sequences into resources, but it represents a huge bottleneck right now. Certainly, more genome sequences are needed, but a parallel effort in analysis, manipulating, and understanding these sequences must also be undertaken.

- There is a serious need for consistency in annotation and archiving annotation data. The implementation of a set of standard annotation platforms and a central annotation resource (which can be disseminated but operate from a central server) with defined methods and standards, and guidelines for people to use the resource would be ideal for the field.

- Genomes should be re-annotated over time so we can make full use of these sequences, take advantage of advancements in annotation techniques, and increase both the societal and scientific benefit of genomics work. By coupling genome projects with community-based annotation, the scientific community would could curate and update annotations so that new information, such as refined functions, will be readily accessible.

Recycled Paper

PENSARÉ DESIGN GROUP
WWW.PENSAREDESIGN.COM