Metagenomics

Z L Sabree, M R Rondon, and J Handelsman, University of Wisconsin-Madison, Madison, WI, USA

Glossary

16S rRNA gene Highly conserved RNA involved in polypeptide synthesis that is commonly used for making phylogenetic inferences.
BAC Bacterial artificial chromosome.
contig A set of overlapping sequences derived from a single template.
cosmids Vectors containing *cos* sites for phage packaging that are used to clone large DNA fragments.
environmental DNA Genetic material extracted directly from a microbial community.
environmental gene tag Short sequence obtained from metagenomic sequencing.
fosmid Large-insert cloning vector based on the F-plasmid of *Escherichia coli*.
genome The genetic complement of an organism.
metaproteome The collective protein complements of a microbial community.
microbiome The collection of microorganisms residing in a particular habitat.
mini scaffold A single paired-end read.
read The output of an individual DNA sequencing reaction.
scaffold Contigs linked by overlapping regions.

Abbreviations

BAC	Bacterial artificial chromosome
CTAB	hexadecyltrimethylammonium bromide
EGTs	environmental gene tags
METREX	Metabolite-regulated expression
SIGEX	Substrate-induced gene expression

Defining Statement

Metagenomics is the study of the collective genomes of the members of a microbial community. It involves cloning and analyzing the genomes without culturing the organisms in the community, thereby offering the opportunity to describe the planet's diverse microbial inhabitants, many of which cannot yet be cultured.

Introduction

Prokaryotes are the most physiologically diverse and metabolically versatile organisms on our planet. Bacteria vary in the ways that they forage for food, transduce energy, contend with competitors, and associate with allies. But the variations that we know are only the tip of the microbial iceberg. The vast majority of microorganisms have not been cultivated in the laboratory, and almost all of our knowledge of microbial life is based on organisms raised in pure culture. The variety of the rest of the uncultured microbial world is staggering and will expand our view of what is possible in biology.

The challenge that has frustrated microbiologists for decades is how to access the microorganisms that cannot be cultured in the laboratory. Many clever cultivation methods have been devised to expand the range of organisms that can be cultured, but knowledge of the uncultured world is slim, so it is difficult to use a process based on rational design to coax many of these organisms into culture. Metagenomics provides an additional set of tools to study uncultured species. This new field offers an approach to studying microbial communities as entire units, without cultivating individual members. Metagenomics entails extraction of DNA from a community so that all of the genomes of organisms in the community are pooled. These genomes are usually fragmented and cloned into an organism that can be cultured to create 'metagenomic libraries', and these libraries are then subjected to analysis based on DNA sequence or on functions conferred on the surrogate host by the metagenomic DNA. Although this field of microbiology is quite young, discoveries have already been made that

challenge existing paradigms and made substantial contributions to biologists' quest to piece together the puzzle of life.

Building Metagenomic Libraries

DNA has been isolated from microbial communities inhabiting diverse environments. Early metagenomic projects focused on soil and sea water because of the richness of microbial species (e.g., 5000–40 000 species/g soil) as well as the abundance of biocatalysts and natural products known to be in these environments from culture-based studies. While soil has been most sampled for metagenomic libraries, aquatic sediments, biofilms, and industrial effluents have also been successfully tapped, often because of their unique physicochemistries, for various biological activities. Metagenomic libraries constructed from DNA extracted from animal-associated microbial communities have also been the source of a number of novel biocatalysts (i.e., hydrolases, laccases, and xylanases), antibiotic resistance genes, and inter/intraspecies communication molecules. See **Figure 1** for an overview of metagenomic library construction and screening.

Preparing Metagenomic DNA

The physical and chemical structure of each microbial community affects the quality, size, and amount of microbial DNA that can be extracted. Accessing planktonic communities requires equipment that is capable of handling large volumes of water to concentrate sufficient microbial biomass to obtain enough DNA to build libraries. Contaminating chemicals and enzymes often remain in the water, making it relatively easy to isolate DNA without abundant contaminants. In contrast, inorganic soil components, such as negatively and positively charged clay particles, and biochemical contaminants, such as humic acids and DNases, make DNA extraction from soils, and subsequent manipulation, challenging. The process for removing contaminants determines both the clonability and the size of the DNA because many of the processes that effectively remove contaminants that inhibit cloning also shear the DNA. Physical disassociation of microbes from the semisolid matrix, typically termed 'cell separation', can yield a cell pellet from which DNA, especially high molecular weight (>20 kb) DNA, can be obtained. Immobilization of cells in an agarose matrix further reduces DNA shear forces and, following electrophoresis, facilitates separation of high molecular weight DNA from humic acids and DNases. Various commercial kits can be used to extract DNA from soil and other semisolid matrices. Applying multiple extraction methods to a DNA sample can yield minimally contaminated DNA. For example, a FastDNA Spin (Qbiogene) preparation followed by a hexadecyltrimethylammonium bromide (CTAB) extraction yields high quality DNA. Due to shear forces, low molecular weight DNA is typically isolated, but the physically vigorous nature of some of these methods can facilitate lysis of encapsulated bacteria, spores, and other microbial structures that are resistant to more 'gentle' lysis methods, thereby providing access to a greater, more diverse proportion of the microbial community for cloning.

Cloning Vectors and Metagenomic Library Structure

The choice of cloning vector and strategy largely reflects the desired library structure (i.e., insert size and number of clones) and target activities sought. To obtain a function encoded by a single gene, small DNA fragments (<10 kb) can be obtained and cloned in *Escherichia coli* into standard cloning vectors (e.g., pUC derivatives, pBluescript SK(+), pTOPO-XL, and pCF430). Various enzymes, such as amidases, hydrolases, cellulases, and antibiotic resistance determinants, have been identified in functional screens of metagenomic libraries harboring inserts smaller than 10 kb. Conversely, to obtain targets encoded by multiple genes, large DNA fragments (>20 kb) must be cloned into fosmids, cosmids, or bacterial artificial chromosomes (BACs), all of which can stably maintain large DNA fragments. Two vectors, namely pCC1FOS and pWE15, have been used for cloning large DNA fragments from various microbial communities. The pCC1FOS vector has the advantage that, when in the appropriate host (e.g. *E. coli* Epi300), its copy number can be controlled by addition of arabinose in the medium to increase DNA yield. Microbial sensing signals, antibiotic resistance determinants, antibiosis, pigment production, and eukaryotic growth modulating factors have been identified from metagenomic libraries constructed with pCC1FOS. Additionally, the presence of considerable flanking DNA on fosmid or BAC clone inserts facilitates phylogenetic inference about the source of the fragment.

Community Complexity and Metagenomic Library Structure

The determination of target insert size, cloning vector, and minimum number of library clones is governed by the type of genes that are sought and the complexity of the microbial community. Shotgun sequencing is usually conducted on small-insert clones, whereas successful functional studies can be performed on small- and large-insert clones. Small-insert metagenomic libraries constructed in plasmids that stably maintain up to 10 kb of DNA require 3–20 times more clones compared to libraries constructed in fosmids (30–40 kb inserts) or BACs (up to 200 kb inserts) to obtain comparable coverage of the same microbial community.

Community coverage is possible only in relatively simple microbial communities like the acid mine drainage, which contains only about five members. This

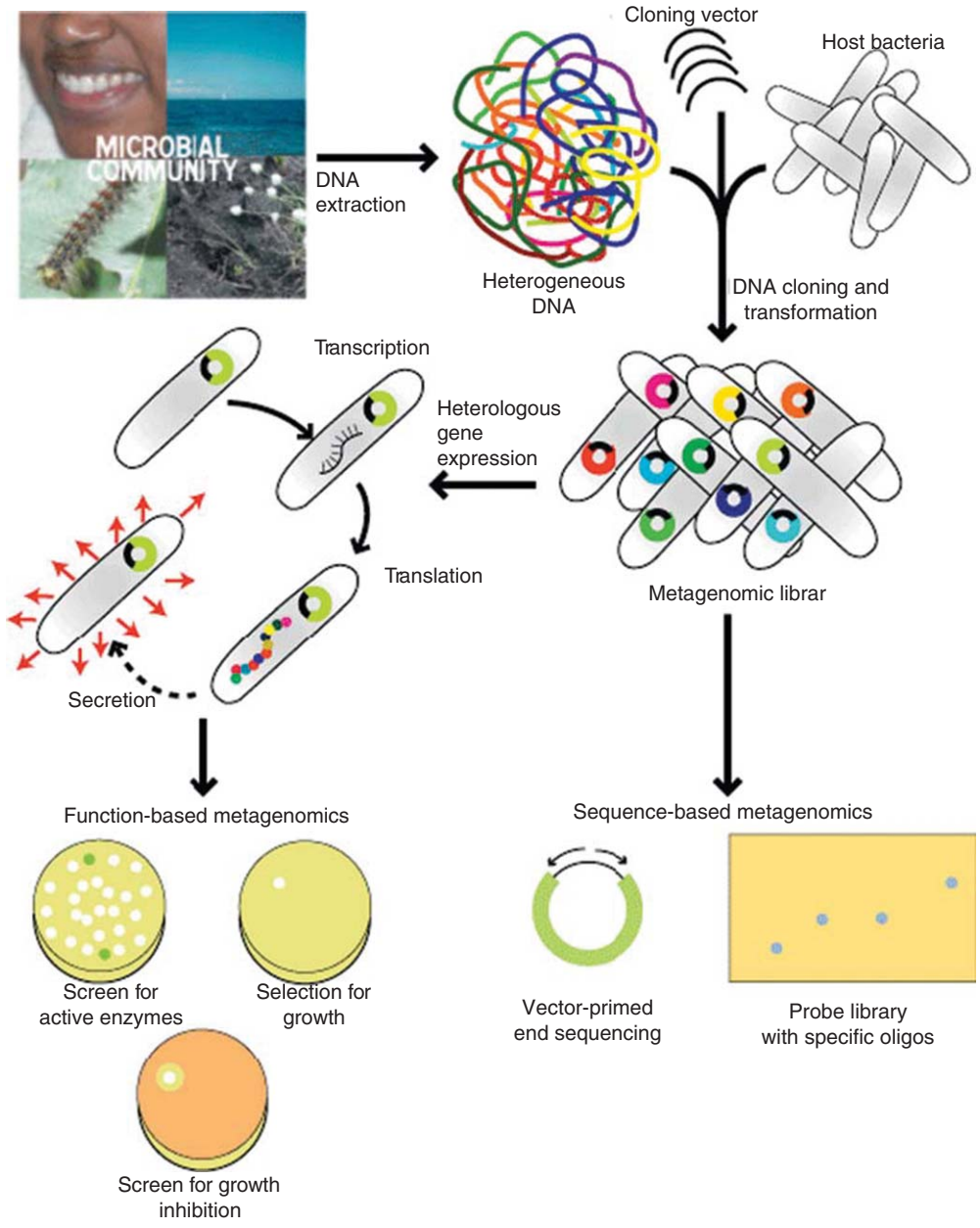

Figure 1 Metagenomics. Metagenomics is the study of the collective genomes of the members of a community. DNA is extracted directly from the community, cloned into a surrogate host, and then studied by sequencing or screening for expression of activities of interest. Many microbial communities have been tapped for metagenomic analyses. Following construction of a metagenomic library, two approaches can be taken to access the genomic information. Functional metagenomics requires that the host bacterium can express the recombinant DNA in either screens for active enzymes or antibiotic production or selections for growth under growth-suppressive conditions (e.g., nutrient deficiency or presence of an antibiotic). In sequence-based metagenomics, cloned DNA is randomly sequenced using vector-based primers or a specific gene is sought using complimentary oligonucleotides (oligos) to hybridize to arrayed metagenomic clones.

community was sampled and sequenced deeply with a high-density, large-insert metagenomic library, making it possible to reconstruct the genome of one member. In contrast, the metagenome of a very complex community such as that in soil can only be sampled, not exhaustively sequenced. With today's technology, a community can be sequenced to closure more quickly and cheaply with small- than large-insert clones, but future technology development may change this. In contrast, in activity-based analysis, large inserts are preferable because the probability that the target activity is encoded by any one clone is positively correlated with the size of the insert, and if the activity is encoded by a cluster of genes, they are more likely to be captured in a large insert.

Selecting and Transforming a Host Organism

Development of microorganisms to host metagenomic libraries has trended toward well-characterized and easily cultivated bacteria, with most work being conducted in *E. coli*. *E. coli* offers many useful tools for metagenomics, such as strains that harbor mutations that reduce recombination (*recA*) and DNA degradation (*endA*) and facilitate blue/white recombinant (*lacZ*) screening. Electroporation is the primary and most efficient method for introducing metagenomic DNA, especially large-insert libraries, into *E. coli*. Additionally, some vectors can be transferred from *E. coli* to other bacterial species by conjugation. This has been an important feature of activity-based analysis of metagenomic libraries. Libraries have been screened in *E. coli* with many successful outcomes, but the barriers to expression of genes from organisms that are distant from *E. coli* have led to screening of some libraries in other hosts. A number of clones containing interesting activities have been detected in only one host species and not in others. This suggests that the host background affects the expression of the genetic potential of the metagenomic library, and therefore the suitability of a particular host for the targeted screen must be considered. Sequence-based analysis of metagenomic libraries is conducted entirely in *E. coli*.

Storing Metagenomic Libraries

Library storage conditions should preserve clone viability as well as the original diversity of the library. *E. coli*-based metagenomic libraries can be prepared in liquid culture supplemented with vector-selective antibiotics and 10–15% glycerol and stored frozen in pooled aliquots. Pooled libraries are revived in fresh media supplemented with antibiotics, and brief incubation (1–2 h) at 30–37 °C on a rotary shaker is sufficient to initiate growth. Extended incubation could hinder recovery of the full metagenomic library due to overgrowth of some clones to the exclusion of others. The revived library is then suitable for screening or DNA preparation.

Sequence-Based Metagenomic Analysis

Sequence-based metagenomics is used to collect genomic information from microbes without culturing them. In contrast to functional screening, this approach relies on sequence analysis to provide the basis for predictions about function. Massive datasets are now catalogued in the 'Environmental Genomic Sequence' database, and each sequencing project is more informative than the last because of the accumulated data from diverse environments. As patterns emerge in the environmental sequences, sequence-based methods will be increasingly more informative about microbial communities.

Some studies use a gene of interest or 'anchor' to identify metagenomic clones of interest for further analysis. A metagenomic library is constructed and screened using PCR to amplify the anchor. Anchors are often a ribosomal RNA gene, but can also be a metabolic gene (e.g., a polyketide synthase). The clones that contain the anchor are then sequenced or further analyzed to provide information about the genomic context of the anchor. Thus, researchers can quickly focus on a clone of interest. For example, a marine picoplankton metagenomic library was screened by hybridization with a 16S rRNA gene probe and the 16S rRNA genes in the positive clones were then sequenced to provide a picture of the diversity of the community members.

Recently, the tremendous advances in DNA sequencing technology have made it feasible to sequence large libraries without preselection for clones containing a particular anchor. This has led to the accumulation of massive amounts of sequence data from uncultured microbes in several environments. Here we discuss a few examples of sequence-based metagenomic projects. These projects are also detailed in **Table 1**.

Anchor-Based Sequencing

Some early projects used rRNA or other genes as anchors to identify useful and informative clones from metagenomic libraries. For example, this technique was used to identify clones containing DNA from planktonic marine *Archaea*. Previously, 16S rRNA gene sequences had been recovered from several environmental samples, suggesting the presence of *Archaea* in nonextreme environments such as soil and open water, yet no cultured representatives of these clades were known. Stein and colleagues took a direct cloning approach to isolate genomic DNA from these organisms rather than attempting to culture them. They filtered seawater and prepared a fosmid library from the collected organisms. The library was probed to find small-subunit rRNA genes from archaeal species. One clone was found and completely sequenced. Several putative protein-encoding genes were identified, including some that had not yet been identified in *Archaea*.

Similarly, Béjà and colleagues constructed several BAC libraries from marine samples. These libraries were also screened by first probing for the presence of 16S rRNA genes. One 130-kb BAC clone was sequenced and found to contain a 16S rRNA gene from an uncultivated gammaproteobacterium. Surprisingly, this clone contained a gene encoding a protein with similarity to rhodopsins, which are light-driven proton pumps found thus far only in Archaea and Eukarya. This new type of rhodopsin, called proteorhodopsin, suggests that bacteria may also use these proteins for phototrophy. This discovery highlights the

Table 1 Sequence-based metagenomics

Year	*Environment*	*Total amount sequenced*	*Number of reads*	*Vector*	*Assembly*	*Comments*	*Reference*
Anchor-based projects							
1996	Ocean	~10 kbp	36	pFOS1	Not attempted	Shotgun sequence clones form 38.5-kb cosmid insert	Stein *et al.* *Journal of Bacteriology* **178**: 591–599
2000	Ocean	128 kbp	Not reported	pIndigoBAC536	Entire BAC insert assembled	Found proteorhodopsin gene	Béjà *et al.* *Science* **289**: 1902–1906
2002	*Paederus* beetles	110 kbp	Not reported	pWEB	110 kbp assembled from several overlapping cosmids	Found putative pederin cluster is probably bacterial in origin	Piel *PNAS* **99**:14002–14007
Viral Metagenomics							
2002	Ocean	ND	~1934	pSMART	Contigs assembled at low stringency	Used specific techniques to enrich for viral DNA	Breitbart *et al.* *PNAS* **99**:14250–14255
Community Sequencing projects							
2004	Acid mine drainage	76.2 Mbp	103 462	pUC18	85% of reads in scaffold 2 kb or longer Combined length of 1183 scaffolds is 10.82 Mbp	Assembled two near-complete and three partial genomes	Tyson *et al.* *Nature* **428**: 37–43
2004	Sargasso Sea	1360 Mbp 265 Mbp	1 660 000 325 561	pBR322 derivative	64 398 scaffolds 217 015 miniscaffolds 215 038 singletons (data for Weatherbird only)	Larger Weatherbird Sample yielded assembly, smaller Sorcerer II did not	Venter *et al.* *Science* **304**: 66–74
2005	Minnesota soil Whale-fall	100 Mbp 3 × 25 Mbp	149 085 Not reported	Lambda ZAP	Analyzed without assembly for environmental gene tags and metaproteome	Assembly not possible for soil sample due to complexity	Tringe *et al.* *Science* **308**: 554–557
2006	Human gut	78 Mbp	139 521	pHOS2	14 572 scaffolds for 33 753 108 bp 40% of reads not assembled for an additional 44 Mbp	0.7X coverage of *B. longum* genome and 3.5X coverage of *M smithii* genome	Gill *et al.* *Science* **312**: 1355–1359

expansive potential and novelty of bacterial genes in the environment and further supports the hypothesis that much of microbial diversity remains undescribed.

Genes besides 16S rRNA genes can also be used as anchors. To identify the source of production of the antitumor compound pederin, Piel constructed a cosmid library from pederin-producing *Paederus* beetles and screened the library for pederin synthesis genes. A locus encoding the putative pederin synthesis genes was identified, and the locus seems to be bacterial in origin, suggesting symbiotic bacteria within *Paederus* beetles are the true producers of pederin. This analysis shows that metagenomics can be used to isolate genes of potential medical interest and confirms that it is the bacteria, not their insect hosts, that produce this antitumor compound.

Viral Metagenomics

The world is thought to contain 10^{32} uncharacterized viruses, and characterizing them by metagenomics may be more efficient and informative than finding a suitable host for each of them. Moreover, since most of their hosts have likely not been cultured, metagenomics represents the only way to access this viral diversity. Cloning viral genomes from an environmental sample faces additional challenges that are not necessarily barriers in metagenomics based on cellular organisms. For example, the viral DNA must be separated from abundant free DNA in the environment, including organismal DNA; viral genes that kill the host must be inactivated or they will prevent the cloning of the viral DNA; chemically modified viral DNA can be unclonable; and ssDNA and RNA viral genomes are not amenable to traditional cloning. Generally, researchers have purified intact virus particles by physical separation and then extracted the DNA from those particles. In one example, Breitbart and colleagues used a linker-amplified shotgun library method to analyze a viral metagenomes from the Pacific Ocean. They found that most viral sequences recovered were not highly similar to known viral sequences, suggesting that the global viral metagenome is still undersampled.

Community Metagenomics

The development of increasingly fast, accurate, and inexpensive sequencing technologies, coupled with significant improvements in bioinformatics, has made it feasible to conduct large-scale sequencing of DNA from multispecies communities. This development will advance our understanding of microbial diversity in nature. Freed from the constraint of cultivating microbes to access their genomes, researchers have accumulated vast quantities of microbial genomic information that could not have been gathered even a few years ago. Although whole genomes cannot currently be reassembled from shotgun sequencing of complex communities containing dozens, hundreds, or thousands of species, rapid advances in technology development make it likely that such feats are not far off.

The first environment to be the subject of a comprehensive sequence-based metagenomic effort was the acid mine drainage environment at Iron Mountain, California. Acid mine drainage results when pyrite dissolution facilitates microbial iron oxidation, which is accompanied by acid production. The simplicity of the acid mine drainage microbial community led to the assembly of five nearly complete or partial genomes without first separation of cells or cultivation of community members.

Approximately 76 million base pairs (Mbp) were sequenced from small-insert libraries from the acid mine drainage biofilm. The community consisted of three bacterial and three archaeal lineages, thus raising the possibility of genome reassembly for many if not all of the members. To achieve this, the sequences were first assembled into scaffolds. Then the scaffolds were sorted into 'bins' based on their G+C content. Binning is the first step in assigning the scaffolds to a unique genome. The scaffolds in each bin were then sorted based on the degree of coverage, assuming that a more abundant member of the community would contribute more DNA to the library, and therefore be more highly represented in the sequence data and that all genes within one organism's genome should be similarly represented. Thus, the high G + C bin was separated into a 10X coverage genome and a 3X coverage genome. These two genomes were found to represent a *Leptospirillum* group II genome and a *Leptospirillum* group III genome, respectively. A nearly complete, 2.23-Mb genome of the group II genome was assembled. Similarly, the 10X coverage genome in the low G + C bin represented a nearly complete genome of the archaeon *Ferroplasma* type II. Partial genomes were also identified for a *Ferroplasma* type I strain and a 'G-plasma' strain.

In addition to providing a model for genome reassembly, the acid mine drainage study led to tremendous insight into the habitat and physiology of the community members. The representation of various functions was recorded, and a chemical model for the flow of energy, nutrients, and electrons was developed and tested.

Another pivotal study in sequence-based metagenomics involved the large-scale sequencing of libraries constructed from the DNA of microbes living in the Sargasso Sea. More than one billion base pairs of DNA were sequenced, representing approximately 1800 genome equivalents. Clones were end-sequenced to provide paired-end reads and assembled into scaffolds, mini scaffolds (with a single paired-end read), or left as single reads. Again, the sequences were sorted into bins, based on depth of coverage, oligonucleotide frequencies, and similarity to known genomes. From these data, several nearly complete genomes were assembled, as well as ten megaplasmids. Sequence analysis

predicted that 1 214 207 novel proteins were encoded in the environmental DNA. The analysis highlighted the presence of related strains of a given species in the sample, raising the complicating factor of strain heterogeneity in metagenome analysis.

When the community is sufficiently complex to prevent reassembly of genomes, other strategies must be undertaken to find patterns. For example, one study compared four metagenomic libraries, one from soil and three from oceanic 'whale-fall' samples (decomposing whale carcasses on the ocean floor), all of which were too complex for genome reassembly or even construction of contigs or scaffolds. Using paired-end sequences from small-insert libraries, 'environmental gene tags' (EGTs) were identified, 90% of which contained predicted genes, thus allowing for a global predicted metaproteome analysis without assembly of the DNA into larger scaffolds. This approach may be useful to find patterns in predicted protein functions and metabolic properties in many different environments.

In a powerful study, sequence-based metagenomics was used to investigate the microbiome of the human ecosystem. Since many of the microorganisms living in the human gut have not yet been cultured, metagenomics allows a direct approach to this microbial community. The data revealed nearly complete genomes of strains of *Bifidobacterium longum* and *Methanobrevibacter smithii*, an archaeon frequently found in the gut ecosystem. Gut microbes contribute many functions to human metabolism, including glycan degradation and fatty acid synthesis, and a number of these functions were identified and their diversity assessed by metagenomics. Scale-up of this approach could yield near-complete genomes of many of the important microbes in our own gut. Comparative studies among individuals, during infant development and after antibiotic ingestion, and among people consuming different diets are beginning to reveal both the idiosyncratic nature of each person's gut microbiota, which may provide a signature as unique as a fingerprint, and the microbial motifs that define health and disease.

Sequence-based metagenomics has the potential to revolutionize our understanding of microbial diversity and function on earth. However, even these initial studies have raised several questions of methods and technology. It is apparent that further advances in bioinformatics are needed to handle the vast quantities of data derived from these projects. In addition, the species richness of the communities, coupled with the genetic complexity of populations of each species, necessitates sequencing extremely large libraries to approach complete coverage. Finally, uneven species distribution, leading to the overrepresentation of abundant genomes in libraries, makes the desired library size even larger in order to capture rare species. The acceleration of sequencing, techniques that remove the most abundant sequences, and computational tools will enhance sequence-based metagenomic analysis.

Function-Based Metagenomic Analysis

Functional metagenomics involves identification of clones that express activities conferred by the metagenomic DNA. Sequence-based metagenomics has revealed physiological and ecological capacity that extends well beyond that of the culturable minority. Activity-based metagenomics provides an opportunity to circumvent culturing and to survey a community's functions (**Table 2**). Function-based metagenomics, unlike sequence-driven approaches, does not require that genes have homology to genes of known function, and it offers the opportunity to add functional information to the nucleic acid and protein databases.

Screening Metagenomic Libraries for Novel Enzymes

Microorganisms have always been a prime source of industrial and biotechnological innovations, but until recently applications have been derived from cultured organisms. Metagenomics presents the possibility of discovering novel biocatalysts from microbial communities that either confounded cultivation or failed to yield new culture isolates upon repeated attempts. Assays that have historically been used to identify enzymes (e.g., amylases, cellulases, chitinases, and lipases) in cultured isolates have been applied successfully to functional metagenomics. Function-based metagenomic analysis of Wisconsin agricultural soil yielded 41 clones having either antibiotic, lipase, DNase, amylase, or hemolytic activities among BAC libraries containing 28 000 clones with an average insert size of 43 kb. The frequency of finding active clones in these libraries ranged from 1:456 to 1:3648, which is similar to the results from other metagenomic surveys for biocatalysts, thereby highlighting the need for robust assays for functional analysis.

Exploiting Environmental Physicochemical Conditions for Biocatalysts Discovery

One approach intended to increase the likelihood of finding certain activities is to build metagenomic libraries from environments that are enriched for bacteria with the desired function. For example, a search for cellulases focused on the liquor of an anaerobic, thermophilic, lignocellulosic digester. Four clones expressing cellulolytic activity were identified, all of which had activity optima at pH 6–7 and 60–65 °C – conditions similar to those in the digester.

Table 2 Functional metagenomics surveys

Target gene	*Source*	*Host strain*	*Cloning vector*	*Insert size (kb)*	*Number of clones*	*Active clones*	*References*	*Extraction method*
Cellulases	Feedstock	*E. coli*	N.R.	N.R.	N.R.	4	Healy *et al.* *Applied Microbiology and Biotechnology* 43: 667–674	Direct
Biocatalysts and Antimicrobials	Soil	*E. coli* DH10B	pBeloBAC11	27/44.5	3646/ 24 546	41	Rondon *et al.* *Applied and Environmental Microbiology* 66: 2541–2547	Direct
Tetracycline resistance determinants	Oral cavity	*E. coli* TOP10	pTOPO-XL	0.8–3	450	1	Diaz-Torres *et al.* *Antimicrobial Agents and Chemotherapy* 47: 1430–1432	Direct
Xylanases	Insect	*E. coli*	Lambda ZAP	3–6	1 000 000	4	Brennan *et al.* *Applied and Environmental Microbiology* 70: 3609–3617	Direct
Amidases	Soil	*E. coli* TOP10	pZerO-2	5.2	8000 / 25 000	5	Gabor *et al.* *Environmental Microbiology* 6: 948–958	Direct/ enrichment
Aminoglycoside resistance determinants	Soil	*E. coli* DH10B	pCF430/pJN105	1.9–65	1 186 200	10	Riesenfeld *et al.* *Environmental Microbiology* 6: 981–989	Direct
Signal molecules	Soil	*E. coli* Epi300	pCC1FOS/ pSuperBAC/ pCC1BAC	1–190	180 000	3	Williamson *et al.* *Applied and Environmental Microbiology* 71: 6335–6344	Direct
Catabolic enzymes	Aquatic	*E. coli* JM109	p18GFP	7	150 000	~35	Uchiyama *et al.* *Nature Biotechnology* 23: 88–93	Direct
Antibiotic desistance determinants	Oral cavity	*E. coli* TOP10/ Epi300	TOPO-XL/ pCC1FOS	0.8–3/~40	1260/600	90/14	Diaz-Torres et al. *FEMS Microbiology Letters* 258: 257–262	Direct
Signal molecules	Insect	*E. coli* DH10B	pBluescript II KS (+)	3.3	800 000	1	Guan *et al.* *Applied and Environmental Microbiology* 73: 3669–3676	Direct

Another study was predicated on the finding that wood- and plant-eating insects are rich sources of enzymes involved in degradation of complex carbon polymers such as cellulose and xylan. These polysaccharides are the primary nutrient source for both the insect and the microbial community it harbors, and therefore should enrich for species that produce glycosyl hydrolases. Small-insert libraries constructed from termite and lepidopteran gut microbial DNA (1×10^6 clones) contained four clones with xylanase activity. These clones harbored genes that encoded xylanase catalytic domains with low sequence similarity (33–40%) to other glycosyl hydrolases. Not surprisingly, the enzymes most closely related to three of the four xylanases identified in this study were found elsewhere in gut-associated microbes. Prior knowledge of the habitat, based on either culture-based studies or metagenomic sequence analysis, facilitates shrewd choices that match habitat with the function that is sought.

Artificial Enrichment for Biocatalyst Discovery

Just as the environment can compel microbial communities to retain members with specific biochemical abilities, microbial communities can be actively manipulated prior to metagenomic library construction to enrich for desired activities. One study focused on the discovery of amidases that convert D-phenylglycine amide derivatives into key intermediates for the production of semisynthetic β-lactam antibiotics. Soil was added to minimal medium that had either D-phenylglycine amide or a mixture of various amides as the sole nitrogen source. Libraries were constructed in a leucine-auxotrophic *E. coli* strain and screened on the medium containing only phenylacetyl-L-leucine or D-phenylglycine-L-leucine, either of which would select for the growth of clones capable of hydrolyzing the amide compounds. According to DGGE analyses, enrichment cultures showed 64–77% lower bacterial diversity than in the original sample before enrichment, suggesting that the enrichment conditions enhanced growth of those bacteria that could utilize the amides as a nitrogen source and limited growth of the rest of the community, thereby reducing the diversity of the community. Four amidase-positive clones were identified from metagenomic libraries constructed from enrichment cultures and all had low to moderate homology to known enzymes. Two amidase-positive clones had low homology to hypothetical proteins. Following extensive amide substrate profiling, a single clone (pS2) was found to catalyze the synthesis of penicillin G from 6-aminopenicillanic acid and phenylacetamide. The pS2-encoded enzyme facilitated accumulation of twofold higher maximum level of penicillin G than *E. coli* penicillin amidase and performed better in amoxicillin and ampicillin production experiments. These analyses show that activity-based screening of metagenomic libraries can yield unique biocatalysts of ecological relevance and biotechnological importance, and some of these may be superior to those found in cultured organisms.

Rapid Discovery of Novel Antibiotic Resistance Genes

Genes that confer antibiotic resistance to bacteria are of great public health, pharmacological, and biological importance. Lateral gene transfer and broad antibiotic usage have resulted in wide distribution of antibiotic resistance genes in microbial communities. As a result, many antibiotic treatments are rendered less effective or totally ineffective against pathogens (e.g., methicillin-resistant *Staphylococcus aureus*). Characterization of the resistance genes in uncultured organisms may identify resistance determinants that will appear in clinical settings in the future. Characterization of resistance in environments that have not been influenced directly by human use of antibiotics could point to the origins of certain resistance determinants. A comprehensive understanding of resistance mechanisms in culturable and uncultured bacteria will improve drug design, by providing clues to how resistance can be combated. Additionally, antibiotic resistance and synthesis genes usually cluster together in microbial genomes. Therefore, antibiotic resistance genes can provide a signpost for biosynthetic pathways residing in nearby DNA.

Antibiotic resistance provides technical advantages over most other characteristics studied with metagenomics. Because the frequency of active clones is low, there is a substantive advantage to using a selection in which only the desired clones survive, rather than a screen, in which all clones need to be addressed to determine whether they express the desired function. When antibiotic-resistant clones are the target, metagenomic libraries are cultured in the medium containing the appropriate antibiotic and only those clones that contain and express cognate resistance genes grow. Using this strategy, a number of studies have identified clones that carry resistance to antibiotics such as tetracycline or aminoglycosides such as kanamycin, tobramycin, and amikasin. In one study, 1 186 200 clones from small- and large-insert soil metagenomic libraries were selected for aminoglycoside resistance and ten surviving clones were identified. Some of these clones carry resistance genes that are similar to known aminoglycoside resistance genes, some carry resistance genes that are distantly related to known resistance genes, and some resistance genes do not have detectable similarity to any known resistance determinants. The results of the antibiotic resistance studies provide a model for the rest of the metagenomic field. Two lessons are evident. First, the uncultured world contains some genes that are quite familiar based on studies of cultured organisms as well as more exotic ones that represent new classes of genes or proteins

for a known function. Second, selections provide the power to rapidly identify clones of interest, circumventing the tedious screening step that forms the basis – and bottleneck – of many function-driven metagenomic studies.

Intracellular Functional Screens

The daunting challenge of screening massive libraries with millions of members has been met with a novel type of screen, in addition to the selections described in the previous section. In an effort to devise screens that are both rapid and sensitive, the concept of 'intracellular' screens emerged in which the detector for the desired activity resides in the same cell as the metagenomic DNA. The first two intracellular screens, metabolite-regulated expression (METREX) and substrate-induced gene expression (SIGEX) (**Figure 2**), meet the criteria of speed and sensitivity and provide prototypes for sensitive screens that are capable of reporting poorly expressed gene products and can utilize technologies (e.g., fluorescence-activated cell sorting) that rapidly screen millions of clones.

METREX screening detects molecules that mimic quorum-sensing signal molecules (**Figure 2(a)**). These compounds stimulate transcription, regulated by the LuxR protein and initiated from the *luxI* promoter, which is linked to a reporter, such as the gene encoding green fluorescent protein. The typical quorum-sensing inducers are acylated homoserine lactones, but metagenomics has revealed other classes of molecules that can function similarly. Since the metagenomic DNA and the broad-specificity reporter plasmid are both present within the host cell, even low concentrations of poorly expressed metagenomic gene products can be detected. In one study, 180 000 clones from soil metagenomic libraries were subjected to the METREX screen; 11 clones that stimulated Gfp expression were identified. Additionally, two clones inhibited Gfp expression in the presence of 80 n mol l^{-1} *N*-(3-oxohexanoyl)-L-HSL, a typical quorum-sensing signal molecule isolated from cultured organisms, which binds to LuxR and stimulates transcription from the *luxI* promoter. Interestingly, only one of the Gfp-stimulating clones could be detected in an overlay-based Gfp-reporter screen, indicating that the intracellular screen detects clones that would be lost in a standard screen for quorum-sensing signal compounds in which the active molecule needs to diffuse out of the producing cell and into the cell containing the reporter. Most of these clones had no sequence similarity to genes known to direct production of quorum-sensing stimulating or inhibiting compounds.

SIGEX facilitates the rapid identification of promoters that drive transcription in the presence of a particular catabolite (**Figure 2(b)**). This can be used to identify pathways that are regulated by the metabolite. Degradative pathways are often regulated by the compound that is degraded, so SIGEX is of interest for rapid identification of new gene clusters for bioremediation. Of 152 000 clones from a metagenomic library derived from an aquatic microbial community, 33 clones were induced by benzoate and two by naphthalene. A wide variety of genes were resolved in these screens, reflecting not only the sensitivity of the assay but also the broad impact of aromatic hydrocarbons on bacterial community gene expression.

Further developments in high-throughput screening will enhance discovery in function-driven metagenomics just as advancements in sequencing and bioinformatics will accelerate discovery in sequence-driven metagenomics.

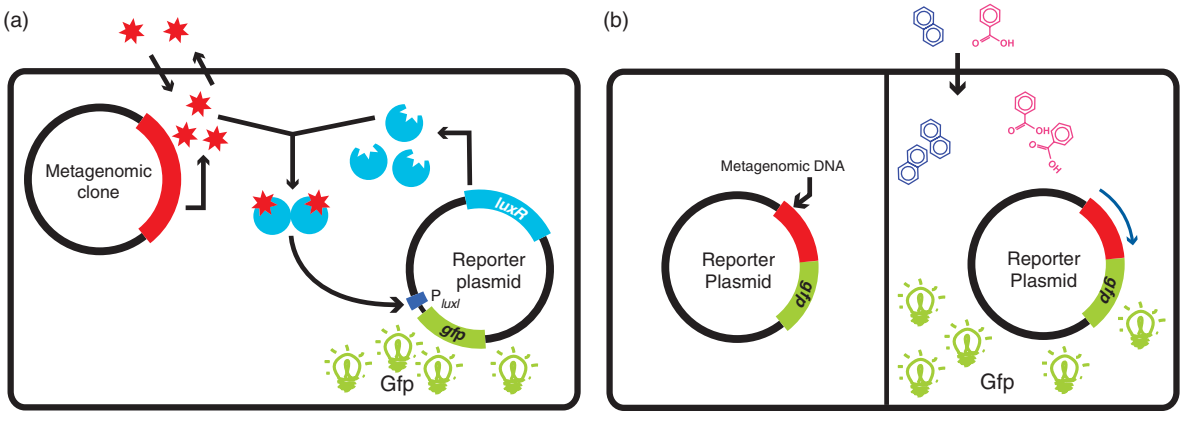

Figure 2 Intracellular screens for metagenomic libraries. (a) Metabolite-regulated expression (METREX): diffusible metagenomic gene products (red) stimulate dimerization of LuxR proteins (blue), which in turn induce expression of Gfp from the *luxI* promoter to produce light (green). Gene products interacting directly with the luxI promoter could also be detected. (b) Substrate-induced gene expression (SIGEX): Metagenomic DNA is cloned into a promoter-trap vector containing a promoterless Gfp-reporter downstream of the cloning site. Promoters in the metagenomic DNA that respond to specific catabolites induce expression of Gfp.

See also: DNA Sequencing and Genomics; Ecology, Microbial; Genome Sequence Databases: Genomic, Construction of Libraries; Phylogenetic Methods; Recombinant DNA, Basic Procedures

Further Reading

Casas V, and Rohwer F (2007) Phage metagenomics. In Hughes, KT and Maloy, SR (eds.) *Methods in Enzymology* 421: Advanced bacterial genetics: use of transposons and phage for genomic engineering pp. 259–68. Amsterdam: Elsevier.

Daniel R (2005) The metagenomics of soil. *Nature Reviews Microbiology* 3: 470–478.

DeLong EF (2005) Microbial community genomics in the ocean. *Nature Reviews Microbiology* 3: 459–469.

Frank DN and Pace NR (2008) Gastrointestinal microbiology enters the metagenomics era. *Current Opinion in Gastroenterology* 24: 4–10.

Gillespie DE, Rondon MR, Williamson LL, and Handelsman J (2005) Metagenomic libraries from uncultured microorganisms. In: Osborn AM and Smith CJ (eds.) *Molecular microbial ecology*, pp. 261–279. London: Taylor and Francis.

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* 68: 669–685.

Hugenholtz P, Goebel BM, and Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology* 180: 4765–4774.

Langer M, Gabor EM, Liebeton K, *et al.* (2006) Metagenomics: an inexhaustible access to nature's diversity. *Biotechnology Journal* 1: 815–821.

Lefevre F, Robe P, Jarrin C, *et al.* (2008) Drugs from hidden bugs: their discovery via untapped resources. *Research in Microbiology* 159: 153–161.

Li X and Qin L (2005) Metagenomics-based drug discovery and marine microbial diversity. *Trends in Biotechnology* 23: 539–543.

Rappe MS and Giovannoni SJ (2003) The uncultured microbial majority. *Annual Review of Microbiology* 57: 369–394.

Riesenfeld CS, Schloss PD, and Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics* 38: 525–552.

Rondon MR, Goodman RM, and Handelsman J (1999) The Earth's bounty: assessing and accessing soil microbial diversity. *Trends in Biotechnology* 17: 403–409.

Tringe SG and Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics* 6: 805–814.

Ward N (2006) New directions and interactions in metagenomics research. *FEMS Microbiology Ecology* 55: 331–338.