Minireview

Metagenomics for studying unculturable microorganisms: cutting the Gordian knot

Patrick D Schloss and Jo Handelsman

Address: Department of Plant Pathology, University of Wisconsin, Madison, WI 53706, USA.

Correspondence: Jo Handelsman. E-mail: joh@plantpath.wisc.edu

Published: 1 August 2005

Genome ***Biology*** 2005, **6:**229 (doi:10.1186/gb-2005-6-8-229)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/8/229

Abstract

More than 99% of prokaryotes in the environment cannot be cultured in the laboratory, a phenomenon that limits our understanding of microbial physiology, genetics, and community ecology. One way around this problem is metagenomics, the culture-independent cloning and analysis of microbial DNA extracted directly from an environmental sample. Recent advances in shotgun sequencing and computational methods for genome assembly have advanced the field of metagenomics to provide glimpses into the life of uncultured microorganisms.

The estimate that fewer than 1% of the prokaryotes in most environments can be cultivated in isolation [1] has produced a quandary: what is the significance of the field of modern microbial genomics if it is limited to culturable organisms? Until recently, this limitation meant that the genomes of most microbial life could not be dissected because more than half of the known bacterial phyla contain no cultured representatives, and the archaeal kingdoms are likewise dominated by uncultured members. The problem can be likened to the Gordian knot of Greek legend, which was impossible to unravel. The knot, which was constructed with interwoven strands with no ends exposed, served as a source of great pride of the citizens of Gordium where it was displayed. It was Alexander the Great who finally cut the massive knot and called the act his greatest victory. One strategy to expose the rest of the microbial world to the eye of the microbiologist - analogous to attempting to untie the knot - is to coax more bacteria into pure culture. The alternative approach - which could cut through it as Alexander the Great did - is metagenomics.

Metagenomics is the culture-independent analysis of a mixture of microbial genomes (termed the metagenome) using an approach based either on expression or on sequencing [2,3]. Recent studies in the Sargasso Sea [4], acid mine drainage [5], soil [6], and sunken whale skeletons [6] have used the shotgun-sequencing approach to sample the genomic content of these varied environments. In each study, environmental samples were obtained and the microbial DNA was extracted directly from the sample, sheared, cloned into *Escherichia coli*, and random clones were sequenced. In some of the studies sequence overlaps were then used to assemble contigs or scaffolds of genomic sequence. The Sargasso Sea study [4] resulted in nearly 2,000,000 random sequence reads, a massive total [7]; the acid-mine-drainage community sequence, more modest in size but impressive in the analytical insights gained, was based on 100,000 sequence reads (Table 1). Assembling so many sequence reads, while simultaneously accounting for heterogeneities between genomes, introduced unique challenges for each study. In the hyper-diverse soil metagenomic sequencing project, fewer than 1% of the 150,000 sequence reads could be assigned to a contig [6], whereas the acid-mine-drainage sequencing project successfully assigned 85% of the sequence reads to one of 1,183 scaffolds [5]. The genome sequences of uncultured microorganisms residing in mixed communities can now realistically be determined.

Table 1

Summary of metagenomic sequencing projects

Community	Estimated species richness	Thousands of sequence reads	Total DNA sequenced (Mbp)	Sequence reads in contigs (%)
Acid mine drainage	6	100	76	85
Deep sea whale fall				
Sample 1	150	38	25	43
Sample 2	50	38	25	32
Sample 3	20	40	25	47
Sargasso Sea				
Samples 1-4	300 per sample	1,662	1,361	61
Sample 5-7	300 per sample	325	265	<1
Minnesota farm soil	>3,000	150	100	<1

A simple oceanic community

The most extensive metagenomic sequencing effort has been the attempt by Venter *et al.* [4] to sequence the prokaryotic genomes in the water of the Sargasso Sea, a well characterized region of the Atlantic near Bermuda that has unusually low nutrient levels; this study has already spawned numerous other meta-analyses (for example, [8-11]). Among one billion nucleotides of sequenced DNA, Venter *et al.* [4] identified more than 1.2 million open reading frames (ORFs), including 782 that had significant similarity to rhodopsin-like proteins. This was a surprise because the rhodopsins were previously thought to be present in only a small group of organisms, and the Sargasso Sea study broadened the spectrum of species known to have them. One intriguing problem in metagenomics is that most ORFs cannot be assigned to gene families of known function [2]. In the Sargasso Sea sequences, for instance, 69% of the ORFs had no known function [4]. This analysis points to a major limitation in annotating sequences from uncultured microorganisms: if no relative of the organism being sequenced has ever been sequenced, then the likelihood of matching each of the newly identified genes to genes of known function is low. The choice of database used for comparison determines the answer, as demonstrated by the identification by Venter *et al.* [4] of 16S rRNA sequences from the Sargasso Sea by querying a database containing only 16S rRNA gene sequences from genome sequences of Bacteria and Archaea. As they limited the comparative database to cultured microorganisms, it was not surprising that they did not identify any 16S rRNA gene fragments from any phyla with no cultured representatives. A further limitation of this study was presented by Delong [12], who pointed out that the two genomes that Venter *et al.* [4] were able to complete were probably contaminants in the sea-water sample. Obvious examples of assembly error (for example, contigs containing bacterial 5S and 23S rRNA genes adjacent to an archaeal 16S rRNA gene) suggest an insidious assembly problem throughout the sequence collection [12]. Perhaps the next stage of the project will profit from the mistakes of this 'pilot' sequencing attempt [4].

An even simpler biofilm community

Although the nutrient-limited Sargasso Sea was selected for metagenomics because it was thought to contain a simple community [4], the community was not simple enough to allow assembly of most of the sequence reads into contigs. Tyson *et al.* [5] selected a far simpler community, that of a biofilm found in the very acidic waste water from an iron mine (termed acid mine drainage), which contains three bacterial and three archaeal lineages. By grouping the assembled contigs into 'bins' according to their GC content and the number of reads per contig, they were able to assign each bin to an organism. The near-complete genome sequences (ten-fold coverage) of *Ferroplasma* type II and *Leptospirillum* group II members enabled Tyson *et al.* [5] conceptually to model the metabolic processes that each genome contributes to the broader community.

This thorough sequencing and metabolic analysis provided the starting point for a 'proteogenomic' analysis. Protein was extracted from biofilms found in the acid mine drainage and digested with trypsin [13]. Applying shotgun mass spectrometry to the fragmented proteins, Ram *et al.* [13] obtained a sequence of part of the proteome. By combining the proteome and metagenome sequences [5], they linked one or more peptide sequences to approximately 49% of the ORFs from the five dominant genomes [13]. The most powerful outcome of this analysis was the identification from the *Leptospirillum* group II sequences of a novel acid-stable iron-oxidizing c-type cytochrome with an adsorption maximum wavelength at 579 nm (Cyt_{579}). Cyt_{579} is the primary iron-oxidizing enzyme in the microbial community and mediates the rate-limiting step in acid production. In this relatively simple community, the proteogenomic approach enabled Ram *et al.* [13] to quantify protein production from each ORF, validate the DNA-derived metabolic model, and identify a process that potentially acts as a keystone for the whole ecosystem.

First metagenomic analyses of complex microbial communities

A fundamental challenge in understanding microbial communities is to chronicle genetic conservation across time and location and to delineate the smallest complement of genes conserved in genomes across different communities [4]. Tringe *et al.* [6] tackled this problem by sequencing microbial communities sampled both from soil from a Minnesota farm and from three deep-sea communities living on sunken whale skeletons ('whale-fall') and comparing

them with the Sargasso Sea sequence collection. ORFs from each metagenomic sequence were assigned to clusters of orthologous genes (COGs [14]), operons, pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [15], and COG functional categories [14].

The relative enrichment found using each of the four annotation methods between the Sargasso Sea, deep-sea whale fall, and Minnesota farm soil was then determined, resulting, in essence, in an *in silico* subtractive hybridization. The over-representation of rhodopsin ORFs in the Sargasso Sea and ORFs encoding cellobiose phosphorylase in the Minnesota farm soil make biological sense, because marine microorganisms are more likely to use light-driven energy transduction systems and soil microorganisms are more likely to encounter plant-derived oligosaccharides such as cellobiose. The large number of ORFs of no known function that were over-represented in each community may indicate as-yet unknown functional systems. Generating copious sequence information from a community is intrinsically valuable, but this comparative analysis [6] is a worthy example of how metagenomics may move beyond descriptive, annotation-based analyses toward meaningful inference about ecological phenomena.

Dealing with complexity and contamination

Application of molecular biology methods to cultured organisms has led to striking insights into the life of microbes in mono-species culture. But genomics has failed to elucidate the functions of microbial communities, where most microorganisms on Earth spend most of their time and that provide the platform from microorganisms shape plant, animal, environmental and human health. Metagenomics, coupled with gene arrays, proteomics, expression-based analyses, and microscopy, will give insights into problems such as genome evolution and the membership of particular niches that are currently hindered by our inability to culture most microorganisms in pure culture [16]. To realize the full potential of metagenomics, however, a number of obstacles need to be overcome. Perhaps the most significant of these is the microbial complexity in most communities. The successful analysis of the acid mine drainage community was predicated on its simplicity. In contrast, the Minnesota farm soil probably contains more than 5,000 species and 10^4-10^5 strains, making it inevitable that the over 150,000 sequence reads could not be assembled into contigs [6] (Table 1). It is likely that 2-5 gigabase-pairs of sequence are necessary to obtain eight-fold coverage of the dominant species in the community, suggesting that inventive approaches are needed to enrich DNA sequences from less abundant organisms or from members that are unique to a community [3].

Another focus for improvement in metagenomics is the use of robust sampling and DNA-extraction procedures. Methodology that guards against contamination such as that revealed in the Sargasso Sea samples is essential. Making the metagenomic studies ecologically meaningful will require sampling strategies that account for spatial and temporal variability, thereby enabling comparisons between communities [17]. These comparisons will also require standardized and aggressive methods for extracting DNA. It is unfortunate that all of the large metagenomic sequencing projects used chemical extraction methods to obtain DNA, whereas the technique of 'bead beating', which applies high shear forces to cells, is more effective than chemical lysis methods at breaking tough cells (for example, [18]). The studies that used chemical lysis methods therefore include DNA from only a subset of the organisms that can be accessed by modern methods.

This is an exciting time for metagenomics, as many projects are underway to sequence the metagenomes of biologically interesting environments. The US Joint Genome Institute (JGI) has essentially sequenced the metagenomes of the microbial communities associated with two extinct ancient cave bears, which contained less than 2 and 6% cave bear DNA, respectively [19]. The JGI is also currently sequencing metagenomic DNA for more than ten studies through their scientific Community Sequencing Program [20], and the J. Craig Venter Foundation is sequencing the metagenomes of samples taken along a path intended to simulate the voyage of Darwin's ship *The Beagle*, as well as samples of New York City's air [21]. A future prospect is completing the human genome by sequencing the metagenome of the 10^{12} microbial cells that are associated with the human body [22]. Each of these studies will unearth secrets unique to the environment being examined, and comparison of results of these studies will provide a meta-understanding of the recurrent and unique themes in community structure and function.

References

1. Amann RI, Binder BJ, Olson RJ, Chisholm SW, Devereux R, Stahl DA: **Combination of 16S rRNA targeted oligonucleotide probes with flow-cytometry for analyzing mixed microbial populations.** *Appl Environ Microbiol* 1990 **56:**1919-1925.
2. Riesenfeld CS, Schloss PD, Handelsman J: **Metagenomics: genomic analysis of microbial communities.** *Annu Rev Genet* 2004, **38:**525-552.
3. Schloss PD, Handelsman J: **Biotechnological prospects from metagenomics.** *Curr Opin Biotechnol* 2003, **14:**303-310.
4. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson K E, Nelson W, *et al.*: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304:**66-74.
5. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428:**37-43.
6. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, *et al.*: **Comparative metagenomics of microbial communities.** *Science* 2005, **308:**554-557.
7. Handelsman J: **Metagenomics or megagenomics?** *Nat Rev Microbiol* 2005, **3:**457-458.
8. LeCleir GR, Buchan A, Hollibaugh JT: **Chitinase gene sequences retrieved from diverse aquatic habitats reveal environment-specific distributions.** *Appl Environ Microbiol* 2004, **70:**6977-6983.
9. McDonald AE, Vanlerberghe GC: **Alternative oxidase and plastoquinol terminal oxidase in marine prokaryotes of the Sargasso Sea.** *Gene* 2005, **349:**15-24.

10. Schloss PD, Handelsman J: **Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness.** *Appl Environ Microbiol* 2005, **71:**1501-1506.
11. Zhang Y, Fomenko DE, Gladyshev VN: **The microbial selenoproteome of the Sargasso Sea.** *Genome Biol* 2005, **6:**R37.
12. Delong EF: **Microbial community genomics in the ocean.** *Nat Rev Microbiol* 2005, **3:**459-469.
13. Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake RC 2nd, Shah M, Hettich RL, Banfield JF: **Community proteomics of a natural microbial biofilm.** *Science* 2005, **308:**1915-1920.
14. **COGs - Clusters of orthologous groups** [http://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Genomes2Other/genomes41.html]
15. **KEGG: Kyoto encyclopedia of genes and genomes** [http://www.genome.jp/kegg/]
16. Allen EE, Banfield JF: **Community genomics in microbial ecology and evolution.** *Nat Rev Microbiol* 2005, **3:**489-498.
17. **Breakthrough of the year: the runners-up.** *Science* **306:**2013-2017. [http://www.sciencemag.org/cgi/content/full/306/5704/2013]
18. Miller DN, Bryant JE, Madsen EL, Ghiorse WC: **Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples.** *Appl Environ Microbiol* 1999, **65:**4715-4724.
19. Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, Rabeder G, Krause J, Detter JC, Paabo S, Rubin EM: **Genomic sequencing of pleistocene cave bears.** *Science* 2005, **309:**597-599.
20. **DOE Joint Genome Institute** [http://jgi.doe.gov]
21. **Life in the Air.** *Science* **307:**1558d. [http://www.sciencemag.org/cgi/reprint/307/5715/1558d.pdf]
22. Relman DA, Falkow S: **The meaning and impact of the human genome sequence for microbiology.** *Trends Microbiol* 2001, **9:**206-208.