FOREWORD

# METAGENOMICS OR MEGAGENOMICS?

Metagenomics is an emerging and exciting new field that integrates biology and technology. Here, Jo Handelsman previews the selection of articles in this Focus issue.

*Jo Handelsman*
*Department of Plant Pathology, 1630 Linden Drive, University of Wisconsin, Madison, Wisconsin 53706, USA.*
*e-mail: joh@plantpath.wisc.edu*

Metagenomics is a youthful field that has barely scratched the surface of the microbiological landscape, but even in its infancy two features are evident. First, all metagenomics is based on the genomic analysis of microbial DNA that is extracted directly from communities in environmental samples, a process that circumvents culturing. Second, metagenomics is conducted on a grand scale.

The reviews in this special issue highlight the gargantuan size of most metagenomes and the DNA libraries that are constructed from them. The metagenomes of most microbial communities are several orders of magnitude larger than the human genome (3 Gb), which until recently was considered the most complex problem in genomics. Now, sequencing and mapping the human genome looks like child's play compared with the staggering challenge of sequencing, and perhaps reconstructing, the metagenome of the Earth. The complexity of the communities from which metagenomic libraries are built presents both an intellectual attraction and challenge. Soil, for example, might contain 40,000 species and virtually limitless variation among strains. Even simple communities, such as those associated with marine sponges, contain unanticipated complexity because of the genetic variation within a single species population. All microbial communities, whether species-rich or species-poor, contain uneven species representation, making it difficult to obtain DNA from all the members of the community. Genetic variation within species, species richness and unevenness of species all conspire to obscure the total diversity within communities, necessitating construction of large DNA libraries to provide an accurate representation of the community members.

Analysis of these large libraries requires ingenuity and brute force, as illustrated by each of the following reviews. In the first, Edward DeLong describes the types of libraries that have been constructed and then focuses on the analysis of those constructed from marine samples. Analysis of metagenomic libraries of seawater communities led to the discovery of the first bacterial rhodopsin (or proteorhodopsin), which functions as a light-driven pump when expressed in the foreign host *Escherichia coli*. This discovery arose from the use of 'phylogenetic anchors' — genes that indicate the type of organism that contributed the fragment of DNA to the library. The complete, new rhodopsin gene was identified by sequencing of a 130-kb library clone. Further study revealed the finding that proteorhodopsins at different depths in the ocean absorb light of different wavelengths, which match depth-specific light availability. The massive sequencing effort that Craig Venter's group invested in the Sargasso Sea microbial community led to further discoveries about the diversity of proteorhodopsins. The detection of microbial diversity was dependent on a massive DNA-sequence dataset that contained at least a million bacterial genes. Edward DeLong also points to the necessity for careful statistical analyses of these large datasets to avoid erroneous conclusions based on contaminants, as seems probable in the Sargasso Sea sequencing orgy.

Next, Rolf Daniel plunges us into the murky depths of soil communities. These complex communities present tough challenges for constructing, sorting and analysing metagenomic clones. The chemical constituents of soil interfere with the isolation of DNA in a sufficiently pure form to be cloned. To accommodate the thousands of species present, the libraries are enormous, creating the corollary challenge of sorting among the clones to identify those of interest. The sequence information from these complex libraries can overwhelm the current

capacity in bioinformatics computing power and swamp the databases. The soil is the focus of the most intensive application of functional metagenomics, which entails identifying clones that express a new trait. Daniel reviews the high-powered selections, enrichments and screens that have been applied to metagenomic libraries constructed from soil, which have revealed new enzymes, antibiotics and degradative pathways. Soil microbiologists are motivated to expand the technologies for functional analysis. Cultivated soil organisms have been the richest source of secondary metabolites on Earth, tantalizing us to explore the as yet uncultured community for useful or biologically informative functions. Studies that are based on functional screens have the potential to add new functional information to the databases because the analyses are not predicated on recognizing functions in new sequences based on previously identified sequences.

In an elegant review on progress in archaeal metagenomics, Christa Schleper *et al.* further develop the principle of scale. They discuss the population-level diversity within the sponge symbiont *Cenarchaeum symbiosum*, which was uncovered by a substantial sequencing effort. The study of archaeal metagenomes in soil has led to the suggestion that terrestrial Archaea use ammonia as their primary energy source. Schleper *et al.* stress the need for even larger sequencing efforts to accommodate the complexity of microbial communities, coupled with 'second generation' metagenomic libraries that would focus more narrowly on particular metabolic groups or functions, to provide the basis for testing hypotheses that are generated by the nucleotide-sequence data.

Eric Allen and Jillian Banfield contribute insights into population-level heterogeneity, metagenomics and how each of these inform our understanding of ecology and evolution of microbial communities. They discuss their reconstruction of genomes in the simple community found in acid mine drainage — a veritable tour de force. This genomic analysis led to proposed metabolic life strategies, some of which were substantiated through cultivation strategies based on the genomic information. Cultivation of microorganisms based on metagenomic information brings to fruition an idea that was launched in this field years ago.

Expanding on the theme of functional analysis, Marc Dumont and Colin Murrell provide a cogent presentation of stable isotope probing as a means to link microbial identity with function. Stable isotope probing involves supplying a community with a substrate that is labelled with a stable isotope, such as $^{13}C$. The labelled DNA can be separated from the rest of the DNA and used as the basis for determining which members of the community are using a substrate of interest. Dumont and Murrell report the first successful application of stable isotope probing to metagenomics, which led to enrichment of DNA from organisms that used methane. Innovations such as these offer fertile ground for analysing complex communities in soil or other habitats. Methods that divide the community into smaller pieces based on metabolic capacity or another significant characteristic will simplify and clarify future genomic analyses.

In a departure from the prokaryotic cell, Robert Edwards and Forest Rohwer turn our attention to the complexity of viral communities in the environment. Viruses are the most abundant biological entities on Earth. Their genetic diversity precludes the use of a universal tool, such as the 16S-rRNA genes that have been so informative about microbial communities. Viruses do not share any common features that can be exploited for genomic analysis except for their size. Therefore, Rohwer's group has isolated viral particles by filtration and constructed massive metagenomic libraries from these collections. A stunning feature of the viral metagenome is novelty. Most of the genes found in viral genomes that have been isolated directly from the environment are new. For example, 65% of sequences from a marine viral community and 68% from an equine faecal sample had no significant similarity to genes in public databases. The high species richness among viral types makes reassembly of viral genomes a challenge. New computational tools will hasten the reassembly of these petite, but highly diverse, genomes.

Patrick Lorenz and Jürgen Eck paint a promising portrait of the centrality of metagenomics to the future of industrial applications. They provide us with a survey of the private companies currently engaged in research using metagenomics to discover new biocatalysts and other products. The sizeable effort directed toward discovery of diverse products indicates the potential for metagenomics. Expectations for discovery are more likely to be met if effort is placed on improved expression systems and innovative screens to find useful clones.

Metagenomics is becoming synonymous with 'megagenomics' — genomics on a massive scale. The communities are large and complex, necessitating large libraries to achieve coverage. The rate of discovery is high because many (in some cases most) of the genes lack homologues in existing databases. This, in turn, calls for enormous sequencing efforts so that the patterns of gene representation, variation and genomic context can be discerned. In functional metagenomics, the frequency of clones expressing an activity is also low, which calls for a large-scale approach to library screening.

The promise of metagenomics will be realized as the field matures and gathers momentum. Improvements in DNA preparation, sequencing technology and clone storage are needed. Significant discovery will arise from innovations in functional screens and bioinformatics. Metagenomic analysis will advance, perhaps transform, our understanding of the structure and function of microbial communities. But metagenomics is a field for the lion-hearted and the bold — it requires large-scale experiments, intellectual risk and the integration of technical innovation and biological intuition that will challenge the best in microbiology. It will be worth it. ■

"Metagenomics is becoming synonymous with 'megagenomics' — genomics on a massive scale."