Q&A

MICROBIOLOGY

Metagenomics

Philip Hugenholtz and Gene W. Tyson

Ten years after the term metagenomics was coined, the approach continues to gather momentum. This culture-independent, molecular way of analysing environmental samples of cohabiting microbial populations has opened up fresh perspectives on microbiology.

Why the 'meta' in metagenomics?

Genomics determines the complete genetic complement of an organism by high-throughput sequencing of the base pairs of its DNA. The most prominent example was the Human Genome Project, which involved the sequencing of 3 billion base pairs. But the genomes of hundreds of organisms from all three domains of life (archaea, bacteria and eukarya), as well as those of quasi-life forms such as viruses, have now been sequenced. Metagenomics, by contrast, involves sampling the genome sequences of a community of organisms inhabiting a common environment. Metagenomics has also been more broadly defined as any type of analysis of DNA obtained directly from the environment — for example, after the appropriate procedures, screening such DNA for particular enzymatic activity. To date, the approach has been applied exclusively to microbial communities.

Why do we need metagenomics?

Microbiology has traditionally been based on pure cultures grown in the laboratory. But most microorganisms cannot be grown in this way and we have been ignorant of their existence. This cultivation bottleneck has skewed our view of microbial diversity and limited our appreciation of the microbial world. Metagenomics provides a relatively unbiased view not only of the community structure (species richness and distribution) but also of the functional (metabolic) potential of a community.

What environments can be analysed?

In principle, any environment is amenable to metagenomic analysis provided that nucleic acids can be extracted from sample material (Fig. 1). Simpler communities are more tractable to a particular technique called shotgun sequencing (Box 1, overleaf) — this was a rationale for one of the earliest studies, which targeted a biofilm in acid drainage from mines that consisted of only a handful of dominant microbial populations. Most interest, however, has centred on the marine environment: the largest metagenomic study to date is the Global Ocean Sampling Expedition, which follows the voyage of Darwin's ship HMS *Beagle*. Metagenomics is now also being adopted in medicine. Of particular note is an international initiative, the Human Microbiome Project, which aims to map human-associated microbial communities (including those of the gut, mouth, skin and vagina).

What surprises have there been?

A strength of metagenomics is its potential for serendipitous discovery. An example is the discovery of proteorhodopsin proteins, light-driven proton pumps that were first identified in environmental DNA from bacterioplankton. Proteorhodopsins have since been found to be widely distributed and highly expressed in diverse microbial groups from aquatic habitats, and they may represent a major source of energy flux in the photic zone of the world's oceans. A more recent discovery is that of archaeal ammonia oxidizers. It was thought that bacteria were solely responsible for aerobic ammonia oxidation, although their numbers often could not account for the observed rates of ammonia oxidation in many habitats. The fortuitous discovery of

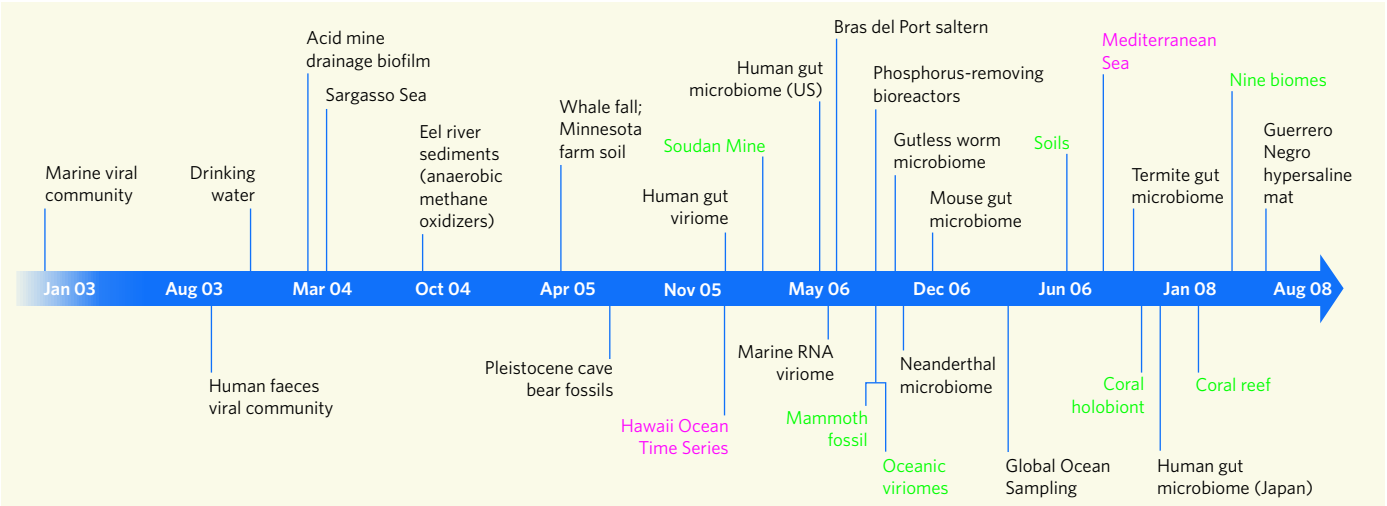

Figure 1 | Timeline of sequence-based metagenomic projects showing the variety of environments sampled since 2002. The oceanic viriomes (all viruses in a habitat) (August 2006) were from the Sargasso Sea, Gulf of Mexico, coastal British Columbia and the Arctic Ocean. The nine biomes (March 2008) were stromatolites, fish gut, fish ponds, mosquito viriome, human-lung viriome, chicken gut, bovine gut and marine viriome. The different technologies used are dye-terminator shotgun sequencing (black), fosmid library sequencing (pink) and pyrosequencing (green). (Graphic based on data sets represented at www.genomesonline.org.)

an ammonia monooxygenase gene next to an archaeal marker gene (encoding small-subunit ribosomal RNA) spawned a rash of papers implicating archaea as the main source of ammonia oxidation in many marine and terrestrial ecosystems.

Can whole genomes be reconstructed from an environmental sample?

Yes: the genomes of dominant species can be fully reconstructed from environmental samples using random sequencing. For example, complete or near-complete genomes have been assembled from microbial populations present in biofilms in acid mine drainage, in activated sludges and in marine samples. Having the complete or near-complete genome of a dominant population provides the gene inventory for the organism and allows its metabolic potential to be determined, including inferring the absence of metabolic pathways as well as their presence. A key feature of genomes obtained from environmental sources is that they are composites of the population from which they were derived, and encompass the genetic microheterogeneity present in that population.

What have we learned about microbial evolution?

Metagenomics provides the first broad insights into coexisting (sympatric) populations, as every sequence read is derived from a different individual within a given community. In communities in which deep sequence-read coverage of individual populations is possible, metagenomics provides an exquisite view of the evolutionary processes shaping these organisms. For instance, data from archaeal populations in acid mine drainage were used to show that genetic recombination occurs at a much higher frequency than previously predicted, and is the primary evolutionary force shaping these populations. And data from the Pacific and Atlantic oceans revealed that the greatest variation within populations of *Prochlorococcus* (the most abundant photosynthetic organisms in the ocean) occurs in genomic 'islands'. These islands are discrete regions in the genome that are believed to be hotspots for genomic innovation and derived in part by genes laterally transferred by viruses. Potentially the most important finding relates to the controversial topic of defining microbial species. Several metagenomic studies have shown clear discontinuity of microbial diversity, suggesting that some microbial species at least are defined by discrete 'sequence space'.

Are some environments too complex for metagenomics?

Certainly it seems unfeasible (at least for now) to obtain complete genomes of organisms from habitats with complex microbial communities because of the sheer amount of sequencing required. For example, it was estimated that a minimum of 6 billion base pairs would be required to obtain the genome sequence of the most dominant population in a soil sample, and many times that to obtain genomes from less dominant populations. But it is possible to extract biologically meaningful information from completely unassembled sequence data using a gene-centric approach.

... and the gene-centric approach is?

Rather than considering a community from the point of view of its component organisms (genomes), gene-centric analysis considers a community from the viewpoint of its component genes. Genes that are found more frequently in one community than another are assumed to endow a beneficial function on that community (Fig. 2). For example, proteorhodopsins are very common in marine surface waters compared with other habitats, and enzymes that break down cellulose are more common in the termite hindgut than in other habitats. Over-represented genes of unknown function (the majority) can provide foci for research that improves the odds of making biological discoveries. Gene-centric analysis can also be taken up a level to see if over-represented genes are part of higher functional units such as metabolic pathways. For example, genes involved in photosynthesis are generally over-represented in ocean surface waters compared with other habitats. Conversely, however, genes of low relative abundance supplying essential functions will not be detected by this approach.

Are we starting to get a handle on genetic diversity in the environment?

Scarcely: we are still far from measuring the full extent of genetic diversity encoded by microbial life. The Global Ocean Sampling Expedition has generated the largest metagenomic data set so far, comprising 6.12 million predicted proteins from 7.7 million shotgun sequences. The predicted proteins from this data set represented all previously known families of microbial proteins and added 1,700 new ones (each with more than 20 representatives). However, the rate of discovery of new protein families, containing two or more representatives, was linear with the addition of new sequences. This is not surprising in light of recent estimates of microbial diversity obtained using marker genes, such as those encoding ribosomal RNAs, which project the number of microbial species globally to be in the hundreds of millions to billions.

Can metagenomics go beyond functional prediction?

Strictly speaking, no. Metagenomics only looks at the gene sequences that encode proteins or functional RNAs. To go beyond this DNA-based view, expressed RNA transcripts and translated proteins must be obtained from environmental samples and examined directly. This is already happening. RNAs can be reverse-transcribed into DNA and sequenced, and the mass spectra of protein fragments can

Box 1 | The nuts and bolts of metagenomics

Metagenomics begins with the extraction of genomic DNA from cellular organisms and/or viruses in an environmental sample. For 'traditional' dye-terminator sequencing, the DNA is sheared into uniform lengths and inserted into a vector — a known DNA fragment that can be moved between organisms. The vector is then replicated in a bacterial host (typically *Escherichia coli*), which produces many clones of the genome fragment suitable for sequencing. Current sequencing technologies produce 'reads' of fewer than 1,000 bases (Fig. 3), so thousands of reads are required to re-cover whole genomes. As with single-genome (isolate) projects, a range of sizes of DNA can be cloned.

Initially, metagenomic studies focused on screening libraries of large-insert clones (fosmids and bacterial artificial chromosomes) for clones containing genes of functional or evolutionary interest, and these clones were then fully sequenced to provide contextual data. This is still a useful approach, but such directed sequencing has largely been replaced by 'shotgun' sequencing of randomly sampled, anonymous DNA, which is cheaper and has higher throughput.

As with isolate genomics, metagenomic data-processing usually involves the assembly of short, overlapping sequence reads into a consensus sequence, and prediction of which stretches of sequence encode genes. In metagenomics, however, complex communities may not result in any assembly, as no population may have been sampled a sufficient number of times to result in overlapping reads.

The amount of an organism's genome recovered from an environmental sample depends on how many sequence data are obtained and the relative abundance of the organism in its community. For example, assuming completely random sampling, a population with an average genome size of 3 million base pairs, and comprising 0.1% of a community, would require 3 billion bases of sequence data to obtain a 1× coverage (each base of the genome is represented on average by one read). This is beyond the range of dye-terminator sequencing, but new, highly parallelized sequencing technologies, such as 454–Roche pyrosequencing and Illumina sequencing, may be up to the job (Fig. 3).

A desirable extra step in metagenomics is to identify the owner of each anonymous DNA fragment, a process called binning, or classification. Binning methods are still in their infancy, and usually only longer fragments (5 kilobases or more) can be classified reliably. **P.H. & G.W.T.**

be determined and identified by reference to their gene sequences. These approaches have given rise to the fields of metatranscriptomics and metaproteomics.

Can metagenomics be used for viruses?

Certainly. Viral communities have been the subject of several metagenomic investigations and were among the earliest to be studied using the method (Fig. 1). These studies, and viral sequences cropping up in metagenomic data in general, all point to a central role for viruses in microbial evolution and ecology. Initially, only double-stranded DNA viruses were accessible through cloning, but the newer cloneless sequencing technologies allow access to all types (such as single-stranded and RNA viruses).

... and for eukaryotes?

Yes. But because of the cost, sequencing projects have tried to avoid them: eukaryotes in general have much larger genomes and a higher proportion of DNA that doesn't code for protein. Metagenomic analyses of insect–microbe symbioses, for example, have tried to exclude the eukaryotic host DNA. As sequencing costs continue to fall, particularly with the development of higher-throughput technologies, eukaryotes should become a tractable component of a metagenomic analysis.

What will eukaryotic metagenomics tell us?

A common misconception is that animals and plants constitute the bulk of eukaryotic diversity. In fact, most of the evolutionary diversity exists in the under-studied microbial eukaryotes, which are typically unicellular and as difficult to culture as their bacterial and archaeal cousins. For this reason, metagenomics will be an excellent way to study microbial eukaryotic biology (and thereby eukaryotic biology as a whole). However, this has not yet happened because most microbial eukaryotes also have large genomes. Even for plants and animals, a metagenomic approach may be warranted. Basing our understanding of a species on the genome sequence of one or a few individuals misses a lot of genetic information, as is beginning to be appreciated for the human population. In the case of large multicellular eukaryotes such as humans, the equivalent of metagenomics is to sequence the genomes of many individuals.

Will computational capacity keep pace?

That remains to be seen. Sequence data are increasing at a rate higher than increases in computational power. Even more problematic than simple storage of sequence data are the 'all-versus-all' sequence comparisons required to best interpret metagenomes, which raise the computational requirements exponentially. Unless there is a radical breakthrough in computing, for example if quantum computers become viable, then all-versus-all comparisons of metagenomic data will not be feasible in the near future. On the upside, it is unnecessary to compare all data with each other to extract biological insights, and the biology can drive the selection of a relevant subset of available metagenomic data.

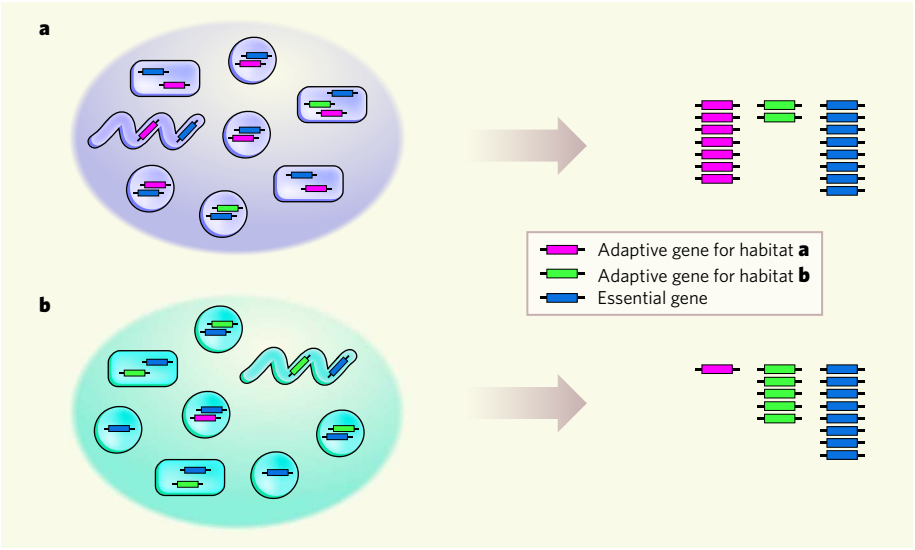

Figure 2 | Gene-centric analysis. Genes from communities **a** and **b** are assigned to their respective gene families (red, green or blue) and counted to highlight functions that are putatively beneficial (red over-represented in community **a**, and green over-represented in community **b**), or essential or having a generalized function (blue; high counts in both communities). The organisms that contribute the genes are largely ignored in this approach, and the underlying assumption is that high counts indicate functions that are important for survival in a given habitat. (Graphic by S. Tringe.)

What other bottlenecks are there?

The gap between characterized and hypothetical proteins identified in metagenomes is widening at an alarming rate. Next to computational resources, uncharacterized gene products are likely to be the biggest bottleneck for the foreseeable future. This means that our understanding of microbial ecosystems will be partial at best, being based on what we can infer from our existing knowledge of biochemistry. Before the advent of metagenomics, however, we were completely oblivious to what we didn't know (unknown unknowns); now we have the blueprints, although we can't read many of the instructions (known unknowns).

Platform	Million base pairs per run	Cost per base (US¢)	Average read length (base pairs)
Dye-terminator (ABI 3730xl)	0.07	0.1	700
454–Roche pyrosequencing (GS FLX titanium)	400	0.003	400
Illumina sequencing (GAii)	2,000	0.0007	35

Figure 3 | Comparison of the cost and throughput of sequencing technologies. New technologies (454–Roche pyrosequencing and Illumina sequencing) generate far more sequence data per run, at a much lower cost than conventional dye-terminator sequencing, but the reads are shorter. Improvements in these technologies are already producing longer reads, and single-molecule sequencing (an as-yet-unproven emerging technology) holds the promise of longer reads still. The dye-terminator approach may soon be obsolete.

Who can use metagenomics?

An increasing number of labs: metagenomics is becoming a basic technology for understanding the ecology and evolution of microbial ecosystems, upon which hypotheses and experimental strategies are built. And with new sequencing technologies producing more than 100 megabases of data for less than US$20,000 (Fig. 3), metagenomics is now within the reach of many researchers. ■

Philip Hugenholtz is in the Microbial Ecology Program, DOE Joint Genome Institute, Walnut Creek, California 94598, USA.
Gene W. Tyson is in the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
e-mails: phugenholtz@lbl.gov; gtyson@mit.edu

FURTHER READING

Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428,** 37–43 (2004).
Edwards, R. A. & Rohwer, F. Viral metagenomics. *Nature Rev. Microbiol.* **3,** 504–510 (2005).
DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311,** 496–503 (2006).
Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
Yooseph, S. *et al.* The *Sorcerer II* Global Ocean Sampling Expedition: expanding the universe of protein families. *PLoS Biol.* **5,** e16 (2007).
Warnecke, F. *et al.* Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450,** 560–565 (2007).