Exploring unseen communities

Advances in sequencing technology and tools for analysis are allowing researchers to unravel the environmental diversity of microbes faster and in greater detail than ever before. Nathan Blow reports.

This year marks the tenth birthday for metagenomics — the cloning and functional analysis of the collective genomes of previously unculturable soil microorganisms in an attempt to reconstruct and characterize individual community inhabitants. Since the term was coined by Jo Handelsman and her colleagues at the University of Wisconsin in Madison, its scope has expanded greatly with descriptions of the microbial inhabitants of environments as diverse as the human gut, the air over New York, the Sargasso Sea and honeybee colonies. And within these communities researchers are now uncovering a wider range of microorganisms, thanks in large part to advances in DNA-sequencing technology.

"We can look at the metagenomic analysis so much more deeply, at such a better cost," says Jane Peterson, associate director of the Division of Extramural Research of the National Human Genome Research Institute in Bethesda, Maryland, which recently launched a five-year initiative to explore the human microbiome.

Although sequencing technology is creating opportunities for metagenomics research, all these new data are straining downstream analysis. "Computational analysis of metagenomic data still has quite a few outstanding questions," says Isidore Rigoutsos, manager of the bioinformatics and pattern-discovery group at IBM's Thomas J. Watson Research Center in Yorktown Heights, New York. The assembly and prediction of gene function for high-complexity microbial communities still poses challenges[1], for example (see 'Benchmarks and standards').

454 LIFE SCIENCES

The 454 Life Sciences GS FLX sequencing system is used in many metagenomics projects.

Maybe it is the promise of rapidly improving sequencing technology or the new environments being explored, but Peterson says that she has seen a growing interest in large metagenomics projects — particularly the Human Microbiome Project, which aims to unravel the microbial communities associated with various parts of the human body, including the gut (see page 578). "People somehow identify with the Human Microbiome Project. It is interesting how this project, especially as it is studying the gut, has really caught a lot of people's attention."

Over the past few years, the race to sequence DNA faster and more cheaply has been taking

BENCHMARKS AND STANDARDS

The complexity of microbial communities can vary drastically, from a couple of microorganisms to thousands or even millions, making the reconstruction of whole genomes from some samples tricky. "If the community is low in complexity, it should allow one to reconstruct genomes with high accuracy," says Isidore Rigoutsos, manager of the bioinformatics and pattern-discovery group at IBM's Thomas J. Watson Research Center in Yorktown Heights, New York. But when it comes to highly complex communities, things are less straightforward.

Rigoutsos and his team have tested several genome assemblers and gene-prediction tools on simulated metagenomic data sets with varying degrees of complexity. Knowing the composition of the community allowed the team to benchmark and evaluate the tools.

"We found that as the complexity increased, many of the computational tools had an increasingly hard time," says Rigoutsos. For most high-complexity samples, he says, the genome assemblers could not generate larger contigs, and several contigs that were assembled were actually chimaeric mixtures of sequences.

For metagenomic analysis, smaller contigs and single reads make assigning the sequence to a specific microorganism difficult. "We want to be able to assign a read of less than 1,000 nucleotides," says Rigoutsos, which might allow researchers to determine species composition from high-complexity samples without the need to generate larger contigs.

Rigoutsos and his colleagues have made three simulated data sets available to researchers interested in testing assembly and prediction programs.

The problem of data analysis is not restricted to metagenomics — a growing number of researchers are using next-generation sequencing platforms and generating the quantity of data that in the past might only have been possible at large genome centres. Several companies are developing software to address this issue.

CLC bio in Cambridge, Massachusetts, offers the CLC Genomics Workbench, which provides reference assemblies of data from various next-generation sequencing systems as well as mutation detection. A future version of the program will incorporate algorithms for the *de novo* assembly of Sanger as well as next-generation sequence data. Meanwhile, Geospiza in Seattle, Washington, and GenomeQuest in Westborough, Massachusetts, are developing software to analyse data generated by Applied Biosystems SOLID next-generation sequencing platform.

The combination of assembly software and data sets to benchmark results should help solve some of the complexity problems associated with metagenomics. "If you sequence sufficiently, even 200 base-pair reads are enough," says Rigoutsos. But he adds that the real question is how many 200 base-pair reads will be needed before we can truly understand complex communities.

Others are finding that with enough reads, fewer than 200 base pairs might be sufficient. Jens Stoye from Bielefeld University in Germany has compared a data set of 35 base pair reads generated on the Genome Analyzer from Illumina in San Diego, California, with a 454 data set for the same low-complexity sample. Although 99% of the Genome Analyzer's sequence data were discarded, because the system generates up to 50 million reads he could assign the species in the sample with the same efficiency from both data sets. N.B.

centre stage. Several next-generation DNA sequencing systems are now available, boasting gigabase outputs for a variety of genetic applications. But when it comes to sequencing environmental samples that contain many different microorganisms in varying amounts, the next-generation options have their limitations.

"I would say the only next-generation sequencing technology suitable for metagenomics at the moment is the 454 system," says Stephan Schuster a biochemist at Pennsylvania State University in University Park. Schuster is not alone — almost all metagenomic studies currently being reported rely on either 454 technology or conventional Sanger sequencing. The main reason is simple: read length.

Long-term tool

Developed by 454 Life Sciences in Branford, Connecticut, the 454 system relies on an emulsion polymerase chain reaction (PCR) step that is coupled to pyrosequencing. Individual fragments of DNA, 300–500 base pairs long, are attached to beads *in vitro* and amplified with PCR to generate millions of identical copies on each bead. Fragments are then sequenced by use of a massively parallel reaction format in 1.6 million wells on a picotitre plate. Using this system, researchers can generate around 250 base pairs of sequence per reaction while performing 400,000 reads in a single instrument run.

Other next-generation sequencing systems typically generate fewer than 50 base pairs per reaction, relying instead on more reads to generate a greater number of base calls. These shorter read lengths are suitable for applications such as candidate gene resequencing, transcriptional profiling and microRNA discovery. But the 454 system's longer read length has attracted metagenomics researchers, and is enticing the community away from Sanger sequencing, which provides reads of 700–1,000 base pairs per reaction.

Forest Rohwer, a microbiologist at San Diego State University in California, sees the advantages of 454 when analysing sequence data from environmental samples. "We know that when you are below 35 base pairs it is hard to get information out, at 100 base pairs you tend to lose information, but if you get to 200 base pairs you start to gain a lot of information," he says. Although he also adds that it is not clear at the moment how much could be gained from even longer reads.

"The information content might be a little different because the Sanger reads are longer. At present, 454 is between 200 and 400 base pairs, depending on who you talk to, so you are not likely to cover an open reading frame, but you do get a significant amount of data," says Karen Nelson, an investigator at the J. Craig Venter Institute in Rockville, Maryland. But she notes that studies using 16S ribosomal RNA pose a problem for 454 read lengths.

K. NELSON

Karen Nelson is one of many investigators working on the Human Microbiome Project.

16S ribosomal RNA is sequenced to give a sense of species abundance and composition in a community without reconstructing entire genomes from the inhabitants. It is around 1,500 base pairs long, which can be sequenced in two or three Sanger reads. But, according to Nelson, with the 454 system, researchers will have to focus on a variable region of the gene to identify the species or come up with other metrics, instead of relying on full-length reads.

Nelson thinks one approach might be to combine Sanger capillary sequencing with 454, so that Sanger methodology could be used to sequence the ends of inserts from large-scale cosmid or fosmid libraries, which would act as scaffolds for the placement of the shorter, but more numerous, 454 reads.

But the length of 454 runs could soon be less of an issue. "We are getting very close to entire exons and whole genes with our next version," says Michael Egholm, vice-president of research and development at 454 Life Sciences, noting that the company plans to introduce a 500-base-pair capacity instrument later this year.

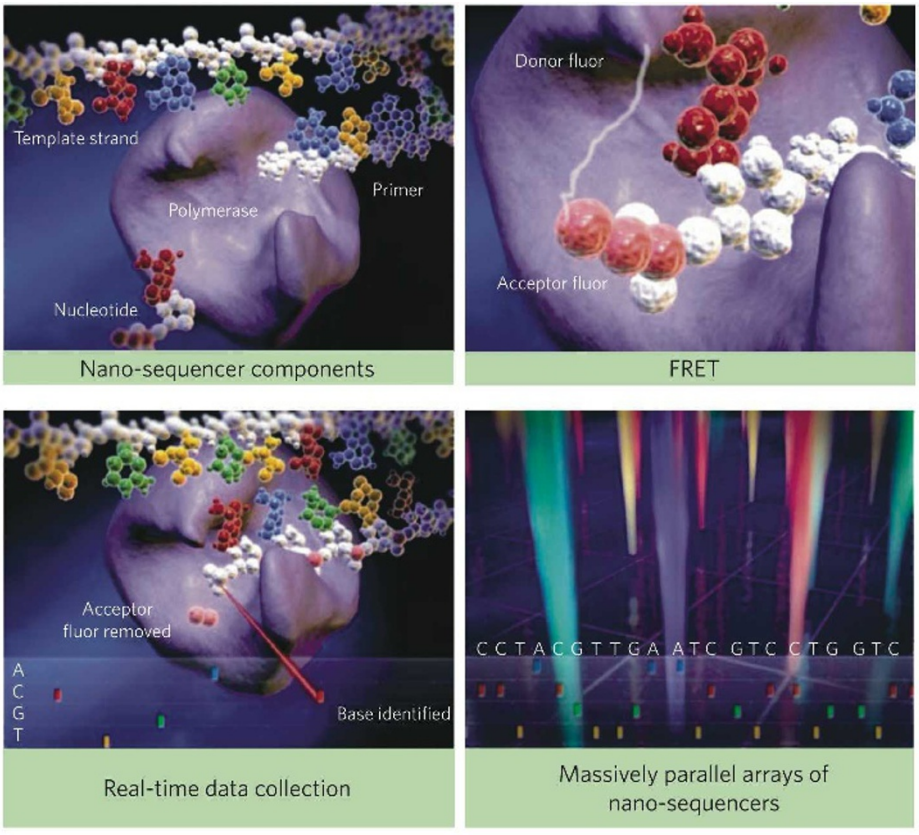
Diverse results

The widespread use of next-generation sequencing technology to explore microbial communities has brought something else to light — greater microorganism diversity. Next-generation systems bypass the cloning of DNA fragments before sequencing, a necessary step for most Sanger sequencing, and this has resulted in the discovery of new microorganisms that previously had been missed because of cloning difficulties. The lack of cloning has also made determining relative numbers of microbes in the community easier.

"As there is no cloning, we have very low bias," says Egholm. Schuster and others have also noted less bias when applying 454 sequencing to their environmental samples. "What is very comforting to see is that a lot of tests have been done by different groups finding the same results with 454 sequencing and qPCR," says Schuster.

Egholm and his colleagues are using their sequencing platform for several large-scale metagenomics projects — the biggest of which actually originated by chance. "The biggest metagenomic project on Earth might be our Neanderthal genome project," says Egholm. They are using 454 to sequence the complete genome of a Neanderthal, which Egholm says they hope to release by the end of the year. But 95–98% of the DNA in the Neanderthal sample comes from the environment rather than from

VISIGEN BIOTECHNOLOGIES



VisiGen Biotechnologies is developing a FRET-based approach to single-molecule sequencing.

a Neanderthal. This means that to get the 1× coverage, or roughly 3 billion base pairs, of the genome, the team must sequence somewhere between 70 billion to 100 billion base pairs of these environmental samples.

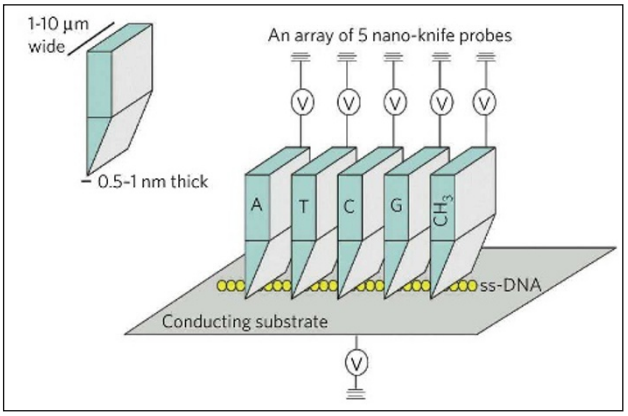
As Egholm's team begins to comb through those environmental data contaminating the Neanderthal samples, he says the initial results have been surprising. "Whether these are recent bacteria, or ones that ate the poor guy when he died, we cannot be certain yet, but I can tell you from the first few contigs where we got multi-kilobase lengths — they matched nothing in GenBank."

The next, next generation

"The biggest change in metagenomics will come from 'third generation' sequencing systems or single-molecule sequencing," says Schuster. For the metagenomics community this next-next-generation promises longer reads than Sanger sequencing, even higher throughput, lower costs and better quantitation of genes.

Isidore Rigoutsos is testing genome assembly.

VisiGen Biotechnologies in Houston, Texas, is working on a method for sequencing single molecules in real-time that uses Förster resonance energy transfer (FRET). In this system, the polymerase is engineered to contain a donor fluorophore, and each nucleotide has a differently coloured acceptor fluorophore attached to the gamma phosphate. When one of the nucleotides is incorporated by the polymerase into a native strand of DNA, a unique FRET signal is given off, and the pyrophosphate-containing fluorophore is then released leaving the synthesized strand of DNA ready for the next incorporation event. By imaging the colour changes as each incorporation occurs, VisiGen hopes to sequence single molecules in real-time and apply this on a massively parallel array with an output of up to one million bases per second.



REVEO

Reveo is developing a single-molecule sequencing technology that physically probes DNA.

Reveo in Hawthorne, New York, is developing a method to sequence DNA using nano-knife-edge probes, which pass over DNA that has been stretched and immobilized in a channel 10 micrometres wide. By using four different nano-knife-edge probes, each 'tuned' to a different frequency, Reveo hopes to non-destructively sequence DNA while eliminating the need for a costly imaging component. And Pacific Biosciences of Menlo Park, California, recently announced its single-molecule sequencing technology based on zero-mode waveguides[2].

These new single-molecule sequencing methods show the promise of generating much longer read lengths in the future, says Schuster. VisiGen's method predicts sequence reads of 1,500 base pairs — the size of 16S ribosomal RNA — whereas the Pacific Biosciences approach could produce reads as long as 10 kilobases.

And single-molecule sequencing is becoming a reality: in March 2008, Helicos BioSciences of Cambridge, Massachusetts, sold the first single-molecule sequencing system. The HeliScope uses a sequencing-by-synthesis approach in which the DNA is first fragmented into pieces 100–200 base pairs long. Adaptors of known sequence are attached to the ends of the fragments so that they can be captured on

THE HUMAN ENVIRONMENT

"I don't know why the Human Microbiome Project bubbled to the top now instead of previously," says Jane Peterson from the National Institutes of Health (NIH). "Certainly there have been metagenomic studies done with the old technologies." But Peterson thinks recent reports exploring the human gut microbial community, along with new, advanced sequencing technologies, might have been enough to pique the interest of the reviewers who decided to fund the Human Microbiome Project (HMP).

The project is a 5-year, US$115-million effort to study the microbial communities inhabiting several regions of the human body, including the gastrointestinal and female urogenital tracts, oral cavity, nasal and pharyngeal tract, and skin, and how those communities influence human health and disease. The effort is viewed by many researchers in the metagenomics community as particularly timely as most agree improvements in sample collection standards and analysis tools are much needed. Karen Nelson from the J. Craig Venter Institute, who is also a participating investigator in the HMP, says these issues can finally be addressed with a project of this scope as all samples will collected and treated in the same way, and standards will be put in placefor annotation of the metagenomic data.

The project will fund research in several areas, although the construction of a data resource for sequencing DNA samples from as many as 250 individuals, and projects aimed at demonstrating how changes in the human microbiome are related to health and disease are centrepieces of the programme. Other project initiatives include the development of new metagenomics technology to isolate bacteria that are currently cannot be cultured, development of new bioinformatic programs and tools for analysis of large genomic data sets, data analysis and coordination centres, and analysing and understanding the ethical, legal and social issues of the project.

The HMP's initial sequencing efforts began this year. Peterson says that towards the end of the last fiscal year, the NIH Roadmap office had funds available for the HMP. Through existing relationships with large sequencing groups, the HMP was able to start quickly and generate preliminary sequencing data, which are being used for the other demonstration projects. This initial effort will result in the sequencing of 200 new bacterial organisms, recruitment of patients for metagenomic studies, and some 16S ribosomal RNA metagenomic sequencing to assess microbial diversity at the various sites.

Peterson says that the protocols for sampling and recruitment have provided some early challenges. The HMP hopes to sample the same sites on all 250 individuals, but with so many sites, standardizing sampling can be tricky. "The protocol for the different sites has to be well worked out," she says, "the oral community has to be happy with the requirements that the skin community brings to the table."

Although it is just at the beginning, the HMP is scheduled to award its first rounds of grants to researchers this autumn. Peterson says that in the future the project's standardization efforts might not be restricted to the United States. "We are also forming an international consortium to coordinate international projects." N.B.

the surface of a coated flow cell. Once attached, a mix of labelled nucleotides and polymerase is flowed through the cell and the surface is imaged at different times to determine whether labelled nucleotides have been incorporated. The key to the technology is a method to effectively cleave off the labelled nucleotides following incorporation, permitting additional rounds to be performed.

"The moment metagenomics goes down to the single-molecule level, it will be possible to assess even very low abundance messages at a very large sequence interval," says Schuster.

Information overload

"People struggle at the moment with extracting basic information from metagenomic data," says Peer Bork a bioinformatician at the European Molecular Biology Laboratory in Heidelberg, Germany. Bork and others say that most bioinformatic analyses being done on metagenonomic data sets involve relatively basic procedures such as assembly and gene annotation.

There are several options for such procedures. Some use web-based tools, such as the Rapid Annotations using Subsystems Technology (RAST) server developed by Argonne National Laboratory and the University of Chicago, in which a metagenomics data set is submitted and an analysis file is returned. Others, such as the Metagenome Analyzer (MEGAN) program, developed by Schuster and his colleagues, run on a desktop computer.

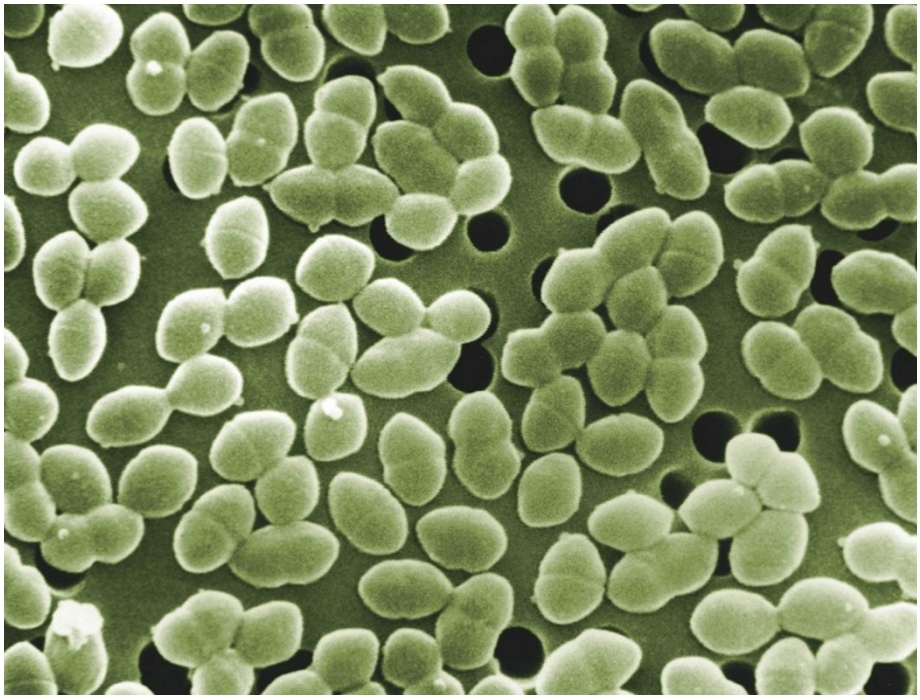
"I think being able to handle these large data sets and developing tools for visualization are critical, as they will allow researchers to write meaningful publications," says Schuster. He explains that MEGAN uses a binning format based on the National Center for Biotechnology Information's taxonomy database to display how often a certain taxon occurs. Other efforts, including the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis, or CAMERA, aim to bring together bioinformatics resources and a data repository to assist with analysis of the data sets.

F. ROHWER

Forest Rohwer studies phage distribution in environmental samples.

One issue with the various tools available, notes Rohwer, is that none has been adopted as standard. "It is still pretty much a cottage industry," he says.

"Standardization of annotations and gene-prediction quality during the early stages of metagenomic studies is something that needs to be addressed," agrees Nelson. She adds that it is now much easier to generate data than to interpret what they mean. "In a number of situations we are dealing with unknown species that have not had their genomes sequenced, so there will not be a reference genome to align to."



CDC

***Enterococcus* species are an example of bacteria found in the human gastrointestinal tract that will be examined during the Human Microbiome Project.**

Rigoutsos agrees. "How can you tell the phylogenetic provenance of a sequence segment if the databases contain no examples of it?" he asks. He sees the situation as being akin to the early days of human genomics when some gene-prediction programs relied on database searches to draw conclusions. His group has developed PhlyoPythia, a software tool that also takes a binning approach to classifying sequence contigs assembled from metagenomic data sets.

Annotation is not the only standardization issue. "We need to do a better job of collecting information when we take samples," says Rohwer. The time, place and collection method can profoundly affect the microbial composition in a sample. Whether the sample is collected during the day or at night, acquired from a person's right or the left arm, or even if two soil samples are collected 5 millimetres apart, these seemingly small differences can lead to very different communities of organisms. Rohwer thinks that collecting a standard set of information for each sample would make future comparisons between different data sets easier and so provide greater biological insight.

There is hope that some of these issues will be addressed as members of the metagenomics community become increasingly involved with large-scale, multi-institution projects (see 'The human environment'). "The Human Microbiome Project is going to be one good example where standards are in place for data acquisition, generation and analysis," says Nelson.

Dynamic future

With bioinformatics tools and sequencing poised to go even faster at a lower cost, researchers are eyeing the next level of metagenomic analysis — a move from simply cataloguing the microbes in a community to understanding the interactions and dynamics of the organisms.

"I think a systems-level approach that gives a holistic view of these communities will open different research possibilities," says Rigoutsos, whose group is now working on studying metagenomes for biofuel applications.

Rohwer hopes to use metagenomics to manipulate a system and then trace how the community goes through changes at a global level. He thinks that with sequencing speed and cost declining so rapidly, these experiments, which only a few years ago would have been impossible, are now within reach.

Bork also wants to move towards analysing global communities. "Most groups concentrate on the early parts, but I think it is time to ask the bioinformatics people to develop methods beyond there, to get ecology concepts projected on those molecular data."

Nelson says that all the technology development and interest in metagenomics is spurring the field along nicely. "I think the technology is very promising and it will get better. It is a great time to be doing this." ■

Nathan Blow is the technology editor for *Nature* and *Nature Methods*.

1. Mavromatis, K. *et al. Nature Methods* **4,** 495–500 (2007).
2. Korlach, J. *et al. Proc. Natl Acad. Sci. USA* **105,** 1176–1181 (2008).