

Researchers Find Pathogens in Unlikely Places Using High-density Microarrays

Lawrence Berkeley's Gary Andersen and David Rasko of the University of Texas Southwestern Medical Center discuss the surprises found using microarrays to detect pathogens in the environment

By Stacey Ryder

BERKELEY, February 1, 2006 — Scientists at Lawrence Berkeley National Laboratory are using two new microarrays to detect dangerous pathogens in air, soil, food and water, before they spread to human populations. In just nine hours, the group's 16S rRNA gene (16S phylogenetic) array identifies up to 8,900 distinctive organisms in a single experiment, while their multiple pathogen ID array (MPID) tests for more than 140 genetic

regions that make bioterror agents particularly dangerous.

The team leader, Dr. Gary Andersen, said that his group typically discovers about twice the number of organisms found when using conventional methods of cloning and sequencing. The group uses microarrays to increase sampling size and to more accurately determine the distribution and complexity of bacterial species in the environment.

Andersen's research group has already used an earlier, prototype version of the phylogenetic array to identify more than 14 orders of prokaryotes and three phyla of eukaryotes in British air samples. The array detects the organisms with probes complementary to variable regions of a universally conserved gene sequence—the small ribosomal subunit.

In a separate proof of concept study, Andersen's group used their MPID



Gary Andersen is a Byrd Professor in the Department of Biochemistry and Molecular Genetics at the University of Virginia. He received his Ph.D. from the Rockefeller University and performed his post-doctoral training at Cold Spring Harbor (CSH). Dutta's laboratory studies the regulation of the G1 to S-phase transition in normal and cancer cells with the goal of learning how to restore control to the cell cycle, preventing unrestrained growth of cancer cells.

array to detect Category A biothreat agents; including anthrax, plague and tularemia, in complex environmental samples. The group designed the array to identify the presence of virulence and pathogenicity genes from 18 potential bioweapons. They are now working on a multi-year project funded by the U.S. Department of Homeland Security to develop this tool for biodefense.

Andersen's laboratory is also examining the microbial composition of aerosols to identify which microorganisms, including pathogens, are normally present in the environment.

"I guess what blew our socks off in the beginning were the pathogens we would see in, for instance, the air," said Andersen. "We'd often see things like *Clostridium botulinum*. That pathogen is anaerobic, so we were a little surprised."

Andersen recently spoke to David Rasko, a senior research scientist at the University of Texas Southwestern Medical Center at Dallas, about the advantages of using microarrays to look at microbial ecology and some of the surprising findings Andersen's team has made.

Rasko, who is also an adjunct collaborative investigator at The Institute for Genomic Research (TIGR), studies the genomic evolution of closely related pathogenic species.

The two discussed:

- the sensitivity of microarray studies compared to plating and surveying studies of microbial ecosystems
- using microarrays to track the spread and dispersal of pathogens across geography
- the potential for tracking the path of infectious diseases using microarray technology

Microarrays vs. Plating and Surveying

Rasko: We've known for quite some time that using traditional plating and survey methods, you miss many species

and see underrepresented diversity. Your chip gives a better idea of that diversity. What percentage do you think is still unresolved in those environments that you've tested?

Andersen: Well, it's hard to say. Less than 1 percent of bacteria can be seen by culture. The route we've chosen is identification of bacteria by conserved genes, such as the 16S small subunit ribosomal gene. With this method, we can find greater diversity and species richness in the natural environment.

This has been going on for a number of years and it's allowed for the discovery of completely novel organisms that have never been observed in culture. But the number of species in different environments depends on your definition of a species.

For microbial ecologists like myself, the definition is 97 percent identity at the 16S gene level, or 70 percent DNA relatedness at the DNA/DNA hybridization level to be called a species. For organisms with important phenotypes, such as causing disease or catalyzing important reactions, we tend to use a much higher percent identity to call a species. As we sample new environments, we uncover new species. But as we sample many related environments, we are seeing less and less novel organisms being identified.

I think where the 16S chip helps is in determining the distribution and complexity of bacterial species in the environment. In our array experiments we typically identify about twice the number of organisms found by the cloning and sequencing. Most of this is due to sampling size. Cloning is expensive and time consuming. Arrays are less expensive, so many sampling points can be identified.

We typically use anywhere from 400 to 800 16S rRNA clones to compare with an array hybridization. Even though that's a large number, we're still missing a lot. If we increased the number, we'd probably see more, approaching the number we'd see on the array.

There are two reasons we can identify these low abundance organisms using our array. One is that we use multiple probes for each organism, giving us a greater degree of sensitivity. Taking advantage of the sequence information from multiple laboratories helps us select the most specific probes to identify a particular grouping of organisms we call a taxon.

The other is that we use the perfect match/mismatch probe pair to reduce the effects of cross-hybridization. Nonspecific hybridization has to be taken into account. You have many very similar 16S molecules and you're trying to find the differences between them.

Rasko: Do you think the lack of identification using traditional cloning and plating type methods is based on low abundance numbers of the diverse population?

Andersen: We noticed that it especially has to do with the way clone libraries are sequenced. When an organism is present in high abundance it dominates the clone libraries and crowds out the low abundance organisms.

One difference with the array is that instead of just pulling out individual 16S amplified molecules, we're putting the entire mixture on the array, allowing us to see the less abundant organisms.

We've split samples to compare our microarrays with the clone libraries, so we know we're not introducing any amplification bias when we compare microarray results with clone library results. But some probes just hybridize stronger than others.

So, we've done Latin square experiments where we take 16S targets of known organisms. In a rotating mixture we have different defined concentrations and we measure the hybridization intensity on the array. For an individual target, we see that as we increase or decrease the concentration, we increase or decrease the hybridization intensity. But if you compare several different targets at one concentration, you'll find that there will be quite a bit of differ-

ence in their sequence hybridization as measured by average difference hybridization score.

One thing that tells us is that we can't use our array for measuring absolute relative abundance of an organism unless we've already calibrated it to a

RNA molecules without cDNA synthesis. The protocol was developed for mRNA, but we noticed it actually works even better for our rRNA.

We've used this method to measure the rise and fall of metabolic activity in multiple organisms in field experi-

we will be able to have everything on one array and run one sample to answer both those questions simultaneously.

With our latest generation microarray, which we're using with air samples, there are certain pathogenic organisms that we knew we couldn't identify by 16S because they are so similar to other environmental organisms. We put these pathogen-specific regions on the array so we could distinguish those pathogens from other environmental organisms.

For example, there's just one base pair difference, or in some cases even no base pair differences, between *Bacillus cereus* and *B. anthracis*. The 16S can tell you that there are some organisms from the *Bacillus cereus* group there, but we have more specific probes for the anthracis pathogen.

Tracking Pathogen Movement

Rasko: Do you think you can use this microarray technology to identify movement of bacteria, fungi or viruses on a global level? For example, if you have something that gets aerosolized in California after weeks, months or years, would you expect that novel species to show up in Washington, D.C., or Florida or New York?

Andersen: In a way, we're doing those experiments now. We're currently measuring the bacterial populations in urban aerosols in a number of cities across the United States and over different seasons, for a U.S. Department of Homeland Security project. In the process of doing this we've identified a number of local reservoirs of bacteria that are entrained into air. We've also observed long-range dispersal of certain organisms.

By looking at multiple locations over several years, we're for the first time getting an idea of the scope of variation of organisms at a given location and what factors influence the microbial composition.

For the Department of Homeland Security, this will be especially important in view of their BioWatch program for the biosurveillance of major cities for potential bioweapon release. One

“Cloning is expensive and time consuming. Arrays are less expensive, so many sampling points can be identified.”

known concentration. The strength of what we do, though, is we can measure the relative increase or decrease in population size as measured by hybridization intensity over time or treatment.

The Role of 16S rRNA

Rasko: Different species have different numbers of 16S genes. Does the copy number affect your hybridization?

Andersen: There's actually a significant advantage to the different copy numbers. We get an estimate of metabolic activity of the organisms in the environment by knowing how much rRNA used for protein synthesis is present. In addition to using conserved primers for a 16S gene amplification, which will tell us which organisms are present, we can directly measure rRNA in a sample without the need for amplification to infer which organisms are most actively expressing proteins.

The typical active bacterium contains about 20,000 copies of rRNA. As a rule, organisms with more copies of the gene or with a higher metabolic rate make more rRNA molecules. Those with fewer copies of the gene or with a lower metabolic rate make fewer copies. There's no easier way of getting this information than by measuring the rRNA hybridization directly on our array.

The breakthrough that made this possible was the development of a protocol by Affymetrix to directly label

ments. For example, we were recently able to identify a syntrophic relationship between several Archaeal species and Proteobacterial metal reducers in a subsurface uranium-contaminated site. This is something we would only see using this high-density phylogenetic array and looking at the rRNA directly.

Rasko: Do you think that a version of your array could be used in a very similar way to identify different pathogens in respiratory diseases? Could you see the 16S and directed approaches working hand in hand or do you see them as competing technologies?

Andersen: I actually can see them working hand in hand. We've already experimented with a targeted approach for identifying pathogens in the environment. In a previous paper, we developed an array that targeted unique regions in the pathogen genomes to identify 18 specific pathogens. We had primers that were specific to each one of the pathogens. We used those primers to amplify 100 to 200 bp fragments. Multiple probe pairs were then used for what ended up being very highly specific and sensitive detection.

I think combining these two methods can give you a general overview of what's in a sample. If you're interested in specific pathogens, you can target them with a number of specific probe sets. As the number of probes that you can put on an array increases, I think

question that comes up is whether a pathogen detected by a sensitive detector could have occurred naturally.

That's what we're looking at now. What are the background organisms that would typically come up in a sample? We can see, for instance, different classes of organisms that seem to be more predominant based on prevailing wind patterns and other meteorological conditions. To some extent, that influences long-range dispersal as well.

Rasko: So, theoretically, you could identify the avian flu if it entered the United States in California and track its progress across the country through a methodology like this?

Andersen: You could, yes. You'd have to use much more extensive sampling than we're doing now and use probes for the specific organism you're looking for. But yes, that would be an ideal way for measuring the spread and dispersal. One of the reasons it hasn't caught on rapidly is probably the type of arrays. We use very large arrays, so there's a cost component there. We can do hundreds of samples, but I think it would require thousands or tens of thousands of samples. More inexpensive versions of the array might be necessary for monitoring these specific pathogens.

Surprises in Air Samples

Rasko: From your papers, you've already identified some things that you wouldn't have thought would be in the air or the soil or the water samples. Is there anything that really blew your socks off?

Andersen: All the time, actually! We see many examples of organisms that were identified as deep-sea or from the Yellowstone Hot Springs or the Arctic ice, in the air or other environments. I think there are a number of reasons for that.

An organism could truly be ubiquitous in a number of locations, but I think a more likely explanation is that there are subtle differences in the 16S sequence.

So, it's probably a variant of a species or a related species that has adapted to these different environments. It shares many of the properties, but not all. So, a deep subsurface vent organism, or a related organism with similar 16S, could also be in the middle of the cold Arctic Ocean.

I guess what blew our socks off in the beginning, were the pathogens we would see in, for instance, the air. We'd often see things like *Clostridium botulinum*. That pathogen is anaerobic, so we were a little surprised, but it makes sense when you think about it. A lot of these organisms are spore-forming. They're probably not active in the air, but they're dispersing.

Limitations of 16S rRNA Microarrays

Rasko: What do you think is the biggest limitation to chips in general?

Andersen: For the array that we're developing, I'd say one of the biggest limitations is the resolution of organism identification using the 16S rRNA. But for chips in general, we have a problem with data overload from the integration of our multiple experiments.

Let me first talk about our specific limitations using 16S rRNA. It gets back to our question of what is a species. The 16S gene is somewhat limited in the resolution and identification of particular organisms. It's about 1,500 bp long, and the difference between closely related species is at best a few regions within the gene with a few base pairs of sequence variation.

In addition to the *B. anthracis*/*B. cereus* example, there's *Yersinia pestis*/*Yersinia pseudotuberculosis*. The organisms have very different disease lifestyles, but a one base pair difference. The resolution of the 16S rRNA gene is not appropriate for individual pathogen species identification.

However, there are other candidate regions that can increase the resolution. For instance, there's the large subunit, or 23S rRNA gene. It has all the advantages of the 16S in that it's present in

every organism and there are conserved primers for amplification, but it's twice as long with a greater amount of sequence variation between species.

There are a number of other conserved molecules or genes you could use to increase resolution as well, so we need to develop a database. For 16S we have over 200,000 sequences in the database. Until we can get something approaching that for these other genes, they won't be as useful.

The other thing is that, although I'm really keen on using the large subunit at some point, it's 3,000 bases long. Until you can use one forward and one reverse reaction for sequencing a clone with a 3,000 bp gene, I don't think making a database for it is going to be that popular. It's just going to be too expensive.

The other part of data overload is the informatics. We have 500,000 data points for each array and multiple arrays for each experiment. I have a small lab without much bioinformatics support, so we get overwhelmed quite quickly.

We can see what's in a sample and compare one place to another at a gross level, but it takes us a lot of work to get into more detailed questions. I think if we had a better way of handling these large amounts of data, we would be able to get to these questions a lot quicker.

AFFYMETRIX MICROARRAY BULLETIN

Editorial Staff

Wes Conard, *Editor-in-Chief*

wes_conard@affymetrix.com

Tommy Broudy, *Managing Editor*

thomas_broudy@affymetrix.com

Rachel Shreter, *Editor*

rachel_shreter@affymetrix.com

Kamalia Dam, *Associate Editor*

Stacey Ryder, *Associate Editor*

Daniel Noble, *Copy Editor*

Michelle Majewski, *Contributing Designer*

Contacts

■ Gary L. Andersen, Ph.D.
Scientist

Center for Environmental Biotechnology
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Mail Stop 70A-3317
Berkeley, CA 94720
GLAndersen@lbl.org

■ David A. Rasko, Ph.D.
Senior Research Scientist
University of Texas Southwestern Medical
Center at Dallas
Department of Microbiology
6000 Harry Hines Blvd. NA6.138
Dallas, TX 75235
david.rasko@utsouthwestern.edu

Companies

■ Affymetrix Inc.
<http://www.affymetrix.com>

Further Reading

■ DeSantis TZ, Stone CE, Murray SR, Moberg JP, Andersen GL. Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. *FEMS Microbiol Lett.* 2005 Apr 15;245(2):271-278.

■ Hu P, Brodie EL, Suzuki Y, McAdams HH, Andersen GL. Whole-genome transcriptional analysis of heavy metal stresses in *Caulobacter crescentus*. *J Bacteriol.* 2005 Dec; 187(24):8437-8449.

■ Radnedge L, Agron PG, Hill KK, Jackson PJ, Ticknor LO, Keim P, Andersen GL. Genome differences that distinguish *Bacillus anthracis* from *Bacillus cereus* and *Bacillus thuringiensis*. *Appl Environ Microbiol.* 2003 May;69(5):2755-2764.

■ Wilson WJ, Strout CL, DeSantis TZ, Stilwell JL, Carrano AV, Andersen GL. Sequence-specific identification of 18 pathogenic microorganisms using microarray technology. *Mol Cell Probes.* 2002 Apr; 16(2):119-127.