

Anthrax Genetic Discovery Provides Basis for New Biodefense Test

NMRC's Michael Zwick and Tim Read discuss how microarrays that resequence the entire anthrax genome may help the military detect novel anthrax strains

By Stacey Ryder

ATLANTA, February 1, 2006 —

Scientists at the Naval Medical Research Center (NMRC) have uncovered a way to quickly identify genetically engineered strains of anthrax bacteria and track their origins using resequencing microarrays. The group's research revealed that naturally occurring anthrax bacteria have undergone virtually no recombination and have a large excess of rare SNP variants; genetic characteristics that would make engineered strains easily distinguishable and traceable.

“The great advantage of resequencing is that, in many respects, it can be maximally informative,” said Michael Zwick the anthrax resequencing project leader. “If you have the entire sequence, you have much more information at your fingertips, not only to characterize the strain, but maybe also to figure out where it came from, to ask what other strains it is related to, and do things like detect the presence of genetically engineered sequences or identify strains that haven't been seen before.”

Zwick's group compared the genomes of 56 different strains—more than 3.1 megabases total—and found that the microarray has a seven percent lower discrepancy rate than traditional sequencing methods. The project took the group about 12 weeks to complete. Using traditional methods, sequencing this many genomes would have taken 10 times as long and required 10 times the resources. Zwick and colleagues published their work in the December 2004 edition of *Genome Biology*.



Michael Zwick is an assistant professor in the Department of Human Genetics at Emory University School of Medicine. As a commander in the United States Naval Reserve, he also conducts research at the Naval Medical Research Center, working for the U.S. Navy and the Department of Defense to develop and implement biodefense strategies. He has been a strong proponent of bringing new technologies, such as resequencing microarrays, to biodefense research with the goal of finding faster, more accurate ways to detect deadly pathogens.

Zwick recently spoke to Timothy Read, also at the NMRC and a co-author of the Genome Biology publication, about the potential of putting microarrays to work for biodefense.

The two discussed:

- the importance of resequencing pathogens
- the surprising lack of recombination found in the anthrax genome
- the shortcomings of algorithms and how they are evolving

n the effects of feature size on anthrax genome analysis

The Importance of Resequencing Pathogens

Read: How did you go from your work as a human population geneticist at Emory to resequencing biodefense pathogens at the Naval Medical Research Center?

Zwick: I was, and still am, a Naval officer in the reserves. As a reservist, my job was to work with the Naval Medical Research Center. There was a lot of interest in biodefense pathogens, particularly after 9/11. Around that time, anthrax was being found in the mail in the D.C. area and my colleagues at the Naval Medical Research Center were playing a very significant role in its detection.

It became clear to me that the ability to sequence these sorts of genomes rapidly was rate limiting. And I realized that chip technology might have some advantages for resequencing bacterial pathogens. So, I took up a project in my reserve time. I started working to develop a 30kb Affymetrix chip to resequence a number of regions from *Bacillus anthracis*.

Read: In many ways, it was an interesting sales job as well. I understand you were also looking to push the idea within the biodefense community, right?

Zwick: I was. At that time in the biodefense community, there was a lot of

emphasis on building assays to identify known polymorphisms, or specific variants that could be used to identify the presence or absence of a pathogen. Those sorts of approaches, while they can be very fast and reliable, tend to be somewhat labor intensive and they require some prior knowledge of the variants.

The great advantage of resequencing is that, in many respects, it can be maximally informative. If you have the entire sequence, you have much more information at your fingertips, not only to characterize the strain, but maybe also to figure out where it came from, to ask what other strains it is related to, and to do things like detect the presence of genetically engineered sequences or identify strains that haven't been seen before.

Read: What was it about the *Bacillus anthracis* genome that made it amenable to chip resequencing?

Zwick: Bacterial genomes, in general, have certain advantages. They tend to be relatively small and compact compared to eukaryotic genomes. They're usually very gene-rich.

The other thing about anthrax that caught my interest is that it was a relatively low GC content genome. We knew from our work on the human genome, where the GC content varied from 34 to 58 percent, that the resequencing arrays worked the best in genomic regions with relatively low GC content. The anthrax genome is around 34 percent GC content. So it actually seemed like an ideal system.

Having read about some of the limited sequencing that had been done in anthrax, I also knew that its genome didn't seem to be very variable. The paucity of variation was an interesting observation, and I realized that a lot of the variation might be SNP variation. I also felt this might reflect something interesting about the origins of anthrax.

I thought the chips might be very good at screening large regions that are not

very variable and also finding the SNPs with high reliability.

Read: How did you select what you wanted to put on the chip?

Zwick: I chose sequences arbitrarily from both of the plasmids in anthrax and the main chromosome. I select contiguous genomic regions that contained genes of interest. I chose genes that have interesting phenotypes or ones where previous surveys suggested that there was variation in the region.

Read: And, of course, one thing about bacterial genomes is, they're haploid.

Zwick: Exactly. So when I performed the human genome experiment in Aravinda Chakravarti's laboratory at Johns Hopkins, we used 32 regions from autosomes and eight from the X chromosome. The DNA was from males, so the X chromosomes were effectively haploid. Haploid base calling is easier and allows you to push the technology faster. So the fact that microbial genomes are haploid is a huge advantage.

Surprising Resequencing Results

Read: What were the main results that came out of that study?

Zwick: There were a number of aspects to the study that were pretty interesting. We did the study first to demonstrate that the data quality could be very high. We had a large amount of replication in the experiment, and we showed that the error rate, in terms of replication in base calling ability, was very low—equivalent to a Phred 60 score with error rates less than 10⁻⁷. Also, in collaboration with Jacques Ravel at TIGR, we were lucky enough to gain access to shotgun sequence that TIGR had produced. We were able to directly compare our base calls with around 398,000 bases shotgun sequenced at TIGR. And that discrepancy rate was closer to Phred 50 data quality.

I think the interesting thing in terms of the sequence quality is that the data clearly exceeds the Bermuda standard

for genome sequencing—to have a Phred 40. That’s where you want to be. If you’re going to use this as a genome sequencing tool, you want to exceed that commonly accepted threshold.

A surprising result was that if we looked at the patterns and levels of genetic variation, there was an excess of rare variants. If you characterized SNPs within functional classes, it was remark-

morphisms. They’re not designed to detect insertion-deletion variation. In principle, we can detect deletions, particularly in haploid organisms like this, because those bases will remain uncalled, but they’d have to be confirmed with an alternative sequencing technology.

It turns out these deletions are rare, so I think future algorithms could improve

A striking observation we made in the human genome work was that purine-rich sequences on chips had much lower mean intensity when you performed the experiment. As a consequence, you were less able to distinguish the proper base for base calling. I would have thought that this problem might have been much less in the anthrax genome, given the lower GC content.

I did a bit of analysis, where we actually looked at the combination of features that we called in all 115 chips and compared them to sizes that we failed to call in any of the 115 chips. I wanted to know whether there were any obvious differences in the sequence composition.

I found that bases that failed to be called had a much higher guanine composition, even though there weren’t as many G’s in the anthrax genome. We saw similar results with purine composition. I think that’s one of the critical issues with using resequencing arrays. In really purine-rich sequences your base calling rates may be much lower.

Read: The second-generation array went from 30kb to 300kb and you implemented a program called UniqueMER. Would you explain what it is and why you felt that it would be useful?

Zwick: When I designed the original human chips, I used a combination of programs. I used RepeatMasker, which uses a database of transposable elements to find transposable elements in the human genome. Then I used another program called Miropeats, which looked within a local region to identify nonunique sequences.

For chip resequencing, the goal is to have the sequences on the chip be as unique as possible. When I used Miropeats to design the 30 kb chip, we bumped up against a problem in that Miropeats, which was a genome finishing program, had a sequence limit size around 4 MB.

Building upon ideas that had been developed in human genetics, where

“I think our most surprising result was the apparent absence of recombination in the anthrax genome.”

ably similar to what we see in the human genome or in the *Drosophila* genome.

I think our most surprising result was the apparent absence of recombination in the anthrax genome. Markers that were either on one of the plasmids or on the main chromosome appeared to be in complete linkage disequilibrium.

That’s consistent with a model that suggests that anthrax arose from a single clone, and has spread throughout the world since then. We looked at only 30kb, so we’d obviously like to look at larger parts of the genome to confirm that, but it provides a clear, testable hypothesis that can be analyzed in future studies.

The implication for biodefense, as we suggested in the paper, is that if you collect a sample from the environment and you see evidence of recombination, there might be something novel about that strain since it appears that anthrax is at the extreme end of the distribution in terms of recombination rates among bacteria.

Read: So what sort of variations do you think we were missing using resequencing?

Zwick: The chips are designed primarily to identify single nucleotide poly-

on that. But, in general, insertion-deletion variation is the type of variation that we might not see. With insertions, we might have a clue that there’s a certain disruption in the reference sequence that might warrant going back and looking at it more carefully with a different sequencing technology.

With small deletions or insertions, it’s hard to know whether a base was not called because there’s a deletion there, or whether it was just something about the sequence composition that prevents the quality score from exceeding the threshold that we set.

Evolution in algorithms

Read: Why don’t resequencing arrays make calls at every site, and what are the properties of the sites that you don’t get calls for?

Zwick: One of the key aspects of resequencing arrays is that they are dependent on the underlying reference sequence. In effect, they query the identity of every single base, in accordance to a specific reference sequence. So if the underlying reference sequence is significantly disrupted, the oligos on the chip will not be complementary to the target DNA that you apply to the chip and you won’t be able to make base calls at those sites.

you could actually make an index of all the oligos of certain sizes in the human genome, I realized that such approaches would actually work much more easily with bacterial genomes.

I suggested this idea to Peter Chen, a computer scientist at Johns Hopkins, and he wrote a program called UniqueMER, which scans through the anthrax genome, or any bacterial genome, and identifies all the unique sequences. So you can make an index and then identify just those unique sequences that you want to place on the chip.

So it's a more general solution and it doesn't have the limits of microbial genome size that we were facing previously.

Read: You have really expert knowledge when it comes to processing of the data—how you get from the image to the sequence and how you put confidence in the data. Would you go over how the quality score for the resequencing array data is measured?

Zwick: Sure. When Dave Cutler and I began developing the likelihood framework for base calling out of chips, we were very influenced by Phil Green's work on Phred and his approach of automated and assigned quality scores to every base in a traditional DNA sequencing trace.

Our philosophy was that you never want to have to look at the chip to do base calling. You want it to be a computerized, automated algorithm that produces a Fasta file with associated quality scores for every base.

So that's what we developed and that's the origin of ABACUS, which stands for adaptive background genotype calling scheme. ABACUS is a likelihood model that takes information from the four features that query the forward strand and the four features that query the reverse strand, and combines them into a likelihood score that provides a number for every base, and then the user can set a threshold level for the

data. Bases that exceed that threshold will be called and bases below that threshold will remain uncalled.

Read: So basically, you can directly compare chip resequencing data quality to conventional genome sequencing data quality.

Zwick: Right, exactly. That is what we did for the haploid base calling in the anthrax paper.

“The quality score is simply the likelihood score—the level of support for the best hypothesis minus the level of support for the second best hypothesis.”

In the previous human genome work that I had done, we resequenced both haploid and diploid sites and compared diploid base calls at sites we knew were polymorphic and had genotyped using an alternative technology. The base calls were around 99.3 percent accurate. So the agreement was very good.

I should say ABACUS is a purely statistical model. It doesn't take into account sequence composition of the probes. It looks at the mean and the variances, and tests a series of hypotheses about what the possible genotypes could be. Then it assigns a quality score for each base. The quality score is simply the likelihood score—the level of support for the best hypothesis minus the level of support for the second best hypothesis.

Read: Would you talk a little about your efforts to make this quality score data available for resequencing arrays through NCBI?

Zwick: There are a couple of areas that ABACUS has spread to. First, ABACUS provides the foundation for the Affymetrix algorithm that's been implemented in their GDAS software.

There's also an independent open-source implementation called RATools, which should give you the same answer, although there are always bug changes and different parameters and that sort of thing as you go through time.

At NCBI, we've been in touch with the trace archive to develop a database to take the data from chips and put it into the trace archives so that as we generate data from studies, such as our anthrax paper, or any organism whether it's

Drosophila or humans, we have a repository where we can place the data and it can be made available to the broader community.

I think this is important because there are aspects of data analysis and base calling that other people might want to improve upon. Having the data in an open trace repository allows the maximum number of eyes to look at it.

Effects of Feature Size

Read: We're currently working with the 8 micron array. We have early access, thanks to Affymetrix, and we generated quite a lot of data for that. As companies like Affymetrix work to reduce feature size, how do you think that's going to affect the quality of the data and the confidence in the base calling?

Zwick: So, ABACUS, as it's conceived now, makes base calls by using the mean and the variance among features. So, in principle, as you decrease feature size, if you keep the scanner the same, you have fewer observations per feature. So you might expect base calling to get somewhat worse.

On the plus side, Affymetrix has improved its scanner. There's a new

upgrade that reduces the pixel size to, I think, around half a micron. So that should bump the base calling back up. We should be able to keep similar base calling rates at similar quality to what we've seen on the larger feature chips. Our experience on the 8 micron features has been that we're still calling more than 92 to 93 percent of the bases. And again, we're getting discrepancy rates that certainly exceed the Phred 40 threshold for the Bermuda standard and often approach Phred 50. Other current and future scanning technologies might allow even more high-quality data to be obtained from a single array.

Read: What are your thoughts about what happens as the feature size goes under 5 microns?

Zwick: I think it's highly desirable to shrink the feature size as small as possible, because what you'd really like to do is to be able to sequence an entire anthrax genome on a relatively small number of chips.

Even with 8 micron features, covering the 5 MB anthrax genome requires about 15 chips. If you shrunk the feature size to 5 microns, it would be substantially smaller, which would further drive

the price down. It's unclear how the 5 micron features will behave. Again, I think the scanner issue's pretty important because I think the better the scans are, the more data there is and the better algorithms, like ABACUS, will work.

Read: Thinking in terms of the future, does the Naval Medical Research Center have plans to resequence other pathogens? And what other technologies may we be using?

Zwick: We certainly do have plans to resequence other pathogens. I plan to continue to collaborate with you and to bring these sorts of ideas into the Department of Defense. I'd like to explore and push the technology to see if getting the whole sequence is possible and can be done in a cost-effective way.

We are exploring a lot of technologies. I think chips have a lot of advantages, particularly if you can shrink feature sizes. Affymetrix chips are very interesting, and we're exploring technology with companies like NimbleGen, who make very flexible, high-density arrays. We're also looking at an alternative sequencing approach with a company called 454 that uses massively parallel pyrosequencing chemistries.

The bottom line is, these different technologies and approaches have strengths and weaknesses that really complement each other depending upon the certain application.

Disclaimer: The views expressed in this document are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense or U.S. Government.

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*

FOR MORE INFORMATION

Contacts

■ Michael E. Zwick, Ph.D.
Assistant Professor
Department of Human Genetics
Emory University School of Medicine
615 Michael Street, Suite 301
Atlanta, GA 30322
mzwick@emory.edu

■ Timothy Read, Ph.D.
Scientist
Biological Defense Research Directorate
Naval Medical Research Center
12300 Washington Avenue
Rockville, MD 20852
readt@nmrc.navy.mil

Companies

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

■ 454 Life Sciences
<http://454.com>

Organizations

■ Naval Medical Research Center –
<http://www.nmrc.navy.mil/>
■ Emory University – <http://www.emory.edu>
■ The Institute for Genomic Research (TIGR)
– <http://www.tigr.org>
■ National Center for Biotechnology
Information (NCBI) –
<http://www.ncbi.nlm.nih.gov/>

Further Reading

■ Zwick ME, McAfee F, Cutler DJ, Read TD, Ravel J, Bowman GR, Galloway DR, Mateczun A. Microarray-based resequencing of multiple *B. anthracis* isolates. *Genome Biol.* 2004;6(1):R10.

■ Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA, Chakravarti A. High-throughput variation detection and genotyping using microarrays. *Genome Res.* 2001;11(11):1913-1925.

People

■ David Cutler, Ph.D.
Johns Hopkins Medical Institute
http://www.hopkinsmedicine.org/geneticmedicine/Research_Activities/bio_comp.html

■ Phil Green, Ph.D.
University of Washington
Genome Sciences
<http://www.gs.washington.edu/faculty/green.htm>