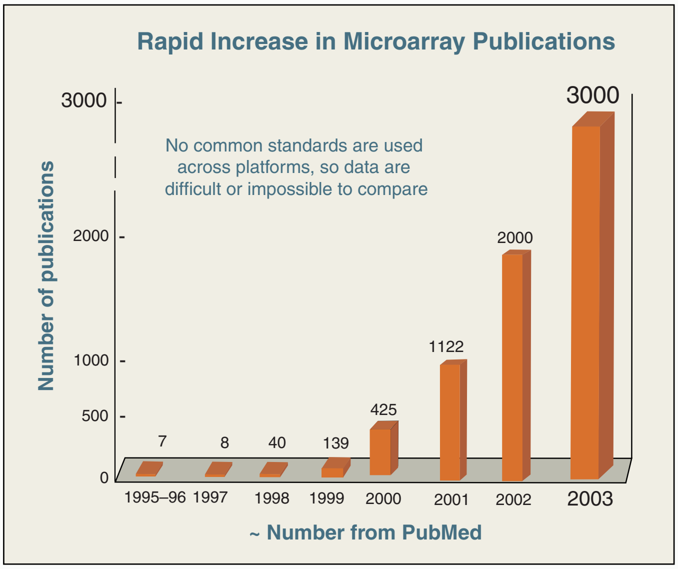
NEWS

Getting the Noise Out of Gene Arrays

Thousands of papers have reported results obtained using gene arrays, which track the activity of multiple genes simultaneously. But are these results reproducible?

When Margaret Cam began hunting for genes that are turned up or down in stressed-out pancreas cells a couple of years ago, she wasn't looking for a scientific breakthrough. She was shopping. As director of a support lab at the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), she wanted to test-drive manufactured devices called microarrays or gene arrays that measure gene expression; she had her eye on three different brands. These devices are hot, as they provide panoramic views of the genes that are active in a particular cell or tissue at a particular time.



Hot technology. The number of studies using microarrays to analyze genes being turned on and off in concert has exploded in the last decade.

Gene array studies are increasingly being used to explore biological causes and effects and even to diagnose diseases. But array experiments are expensive, and Cam wanted to be sure that her colleagues would get high-quality, repeatable, credible results.

She was taken aback by what she found. Not only was she unable to pick a clear winner, but she had a hard time figuring out whether any of the arrays produced trustworthy results. As she delved deeper, she found that the devices produced wildly incompatible data, largely because they were measuring different things. Although the samples she tested were all the same—RNAs from a single batch of cells—each brand identified a different set of genes as being highly up- or down-regulated.

The disharmony appears in a striking illustration in Cam's 2003 paper in *Nucleic Acids Research*. It shows a Venn diagram of overlapping circles representing the number of genes that were the most or least active on each device. From a set of 185 common genes that Cam selected, only four behaved consistently on all three platforms—"very low concordance," she said at an August forum in Washington, D.C., run by the Cambridge Healthtech Institute, based in Newton Upper Falls, Massachusetts. Using less rigorous criteria, she found about 30% agreement—but never more than 52% between two brands. "It was nowhere near what we would expect if the probes were assaying for the same genes."

Cam's findings caused "one's jaw to drop," says Marc Salit, a physical chemist at the National Institute of Standards and Technology (NIST). This was not the first paper to highlight such inconsistencies, but Cam's little diagram had an impact: With support from commercial array makers and academics, Salit is now coordinating an effort at NIST to understand exactly what is measured by these devices.

A few ex-enthusiasts think the promise of gene arrays may have been oversold, especially for diagnostics. Take Richard Klausner, the former director of the National Cancer Institute (NCI) now at the Bill and Melinda Gates Foundation in Seattle, Washington. "We were naïve" to think that new hypotheses about disease would emerge spontaneously from huge files of gene-expression data, he says, or that "you could go quickly from this new technology to a clinical tool." His own experience with arrays indicated they were too sensitive and finicky: The more data he gathered on kidney tumor cells, the less significant it seemed.

But those who have persevered with gene expression arrays attribute such problems to early growing pains. They claim that experienced labs are already delivering useful clinical information—such as whether a breast cancer patient is likely to require strong chemotherapy—and that new analytical methods will make it possible to combine results from different experiments and devices. Francis Barany of Cornell University's Weill Medical College in New York City insists that arrays work well—if one digs deeply into the underlying biology.

Imperfections

Digging into the biology is just what Cam did after her experiments produced reams of discordant data. She and colleagues in Marvin Gershengorn's group at NIDDK wanted to pick out a set of key genes active in pancreatic tumor cells undergoing differentiation. From there, they meant to go on to examine how islet cells develop. "We were very surprised," she recalls, when they couldn't cross-validate results from studies done with Affymetrix, Agilent, and Amersham arrays. So she began pulling the machines apart.

Cam soon ran into a barrier: Manufacturers weren't eager to share information about the short DNA sequence probes each kit uses to spot gene activity. Each commercial system uses a similar approach. Short bits of DNA from known genes are printed as probes on arrays. When an experimental sample is exposed to the array, RNAs made by genes cling specifically to the probes that have a complementary sequence, triggering a fluorescent signal that can be read by an optical scanner. The more RNA on a probe, the stronger the signal. The activity of thousands of genes can be tracked simultaneously this way.

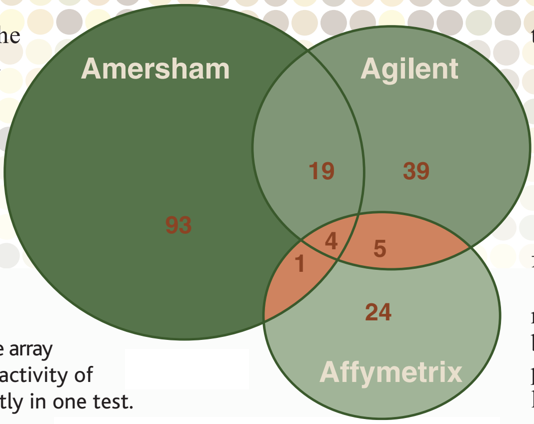
Although manufacturers identified which genes the probes targeted, they would not reveal the actual nucleotide sequence of each probe. This made it difficult to know exactly what the probes were picking up. Recalls Zoltan Szallasi of Harvard's Children's Hospital in Boston, "For the first 6 years researchers were actually flying blind." That changed in 2001, he says, when the companies began sharing data.

Cam says, "We managed to get partial sequences [of array probes] from Agilent," along with "full sequences from Affymetrix

CREDIT: ADAPTED FROM NIH

Published by AAAS

and Amersham." Her team analyzed the probe and gene matchup in detail and found that supposedly different probes were responding to pieces of the same gene. Targeting different parts can be a problem because genes often contain many components that can be spliced into variant RNA packages. The result, several experts say, is that probes can over- or underestimate gene activity.



Little overlap. Three array systems rated the activity of 185 genes differently in one test.

Sorting out the confusion is tough because the probes have not been designed to be specific to gene-splice variants, and no one has even created a master list of variants. Cam is encouraged that companies are beginning to make arrays specific to different splice variants. "That should reduce the ambiguity."

Another confounding factor, Szallasi claims, is promiscuous matches: Probes often respond not only to gene products that exactly fit the sequence but also to those that "cross-hybridize" with near matches. "Every manufacturer claims to have avoided this problem, but there must be a reason why microarray probes targeting almost the same region of a given gene give wildly different intensity signals," he says.

Moreover, many probes just don't correspond to annotated sequences in the public database, RefSeq, Szallasi says; removing these problematic probes significantly improves study results. But the best way to build confidence in gene array results and novel analytical methods, he argues, is to validate probe-gene matches using the more rigorous and time-consuming polymerase chain reaction methods of sequence testing. Szallasi has been doing just that as part of an effort to help collaborators at Harvard and at Baylor College of Medicine in Houston, Texas, merge their gene expression data sets. He's also been trying to get other researchers in Boston to share information on validated matches.

Emerging proof

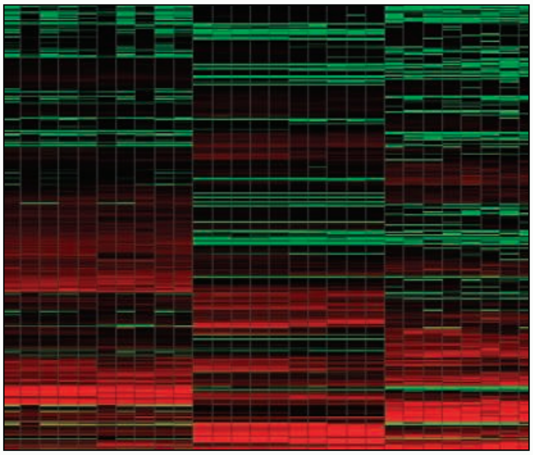
The difficulty in comparing gene array results, say Szallasi and others, raises questions about how much confidence to have in the thousands of papers already published and whether it will ever be possible to merge existing data and find significant results. Relatively few labs have tried to replicate large gene expression studies, even those using identical test devices, says NCI's Richard Simon, a cancer clinician who directs gene array studies. People don't want to waste hard-to-obtain tissue on such work, and they'd rather not spend money on replicating others' findings. Simon argues that the correct test of comparability in clinical medicine is not "whether you come up with the same set of genes" in two studies, but whether you come up with an accurate and consistent prediction of patient outcomes.

He and others note that gene arrays have already proved their mettle in two clinical areas: breast cancer and lymphomas. Molecular geneticist Todd Golub of the Broad Institute in Cambridge, Massachusetts, says his group has independently validated gene expression results of Louis Staudt of NCI and Pat Brown of Stanford University that identify subcategories of lymphoma that have relatively poor or good outcomes. And Lyndsay Harris, a breast cancer researcher at Harvard's Dana-Farber Cancer Institute, says many of her colleagues have confidence in gene expression data that identify a high-risk form of breast cancer associated with cells in the basal epithelium, a strategy that Charles Perou, now at the University of North Carolina, Chapel Hill, pioneered.

In basic research as well, Golub agrees with Simon that broad themes, not specific genes, should be the focus of comparison studies. He looks for a "biologically robust" response in patterns of gene activity—such as activation of coordinated cell signals—as a sign that two experiments have detected the same result. Spotting a signal in the noise is like "recognizing a face, regardless of whether you're wearing bifocals, or sunglasses, or no glasses." Eventually, Golub says, biostatistical methods can probably be used to define such "signatures" in a flexible way to recognize different expression patterns as alternative forms of the same result.

Trials have begun to test some of the newer interpretations of cancer pathology based on gene expression, such as efforts to profile high-risk breast cancer at the Netherlands Cancer Institute (*Science*, 19 March, p. 1754). But many champions of gene-expression tests contend that they are not yet ready for "prime-time" clinical use.

Staudt thinks the time will come when "every cancer patient gets a microarray-based diagnosis." But before then, "we still have to show how reproducible the results are." He is part of an NCI-sponsored consortium that is attempting to correlate results from his own group, obtained from one type of device (spotted arrays of lengthy cDNAs), with those from a type of mass-produced device (printed arrays of short oligonucleotides) made by Affymetrix. Already, they have achieved "very good prediction" of tumor type in retrospective studies of 500 samples. Now they plan to test the model prospectively.



Map of discordance. An experiment at NIH found that three commercial devices rated different genes as being turned on (red) and turned off (green) in a single batch of pancreatic cells.

Seeking harmony

Researchers have now got "all the major journals" using a single format to report array data, says Alvis Brazma of the European Bioinformatics Institute in Hinxton, UK, a cofounder of the Microarray Gene Expression Data Society. But eliminating discordance in the hardware may not be so easy, says Ernest Kawasaki, chief of NCI's microarray facility: "If I had all the money in the world, I would say the best thing is to start over from the beginning"—presumably with a set of validated gene expression probes and shared standards.

That kind of money isn't available, but Salit says NIST recently agreed to spend $1.25 million a year for 5 years to tackle the problem of "discordance." Salit is coordinating a group that includes microarray manufacturers and a coalition of academics—the External RNA Control Consortium—to develop a set of standards that can be used to calibrate gene arrays and ensure that results can be translated meaningfully from one lab to another. If it succeeds, "the pie is going to get bigger" because "everybody's results will improve."

–ELIOT MARSHALL

CREDITS (TOP): ADAPTED FROM *NUCLEIC ACIDS RESEARCH*; (BOTTOM) ADAPTED FROM *NUCLEIC ACIDS RESEARCH*