MICROARRAYS: The Search for MEANING In a Vast Sea of Data

BY KRISTIN COBB, PhD
When DNA microarray technology emerged more than a decade ago, it was met with unbridled enthusiasm. By allowing scientists to look at the expression of enormous numbers of genes in the genome at once, microarrays promised to revolutionize our understanding of complex diseases and usher in an era of personalized medicine. Advocates vowed that, someday, with just a finger prick, doctors would instantly know whether patients were having a heart attack, rejecting a transplant, or in the early stages of cancer based on their mRNA patterns, and would tailor treatment accordingly.

In the decade since their introduction, microarrays have permeated virtually all corners of biomedical research; have yielded some useful insights into basic biology and cancer; and are being used, in a preliminary way, to diagnose disease, guide treatment, and streamline drug discovery. But early enthusiasm has been tempered with a dose of reality. Progress has been slower than predicted. And some splashy results in high-profile journals have proven difficult to reproduce, casting a shadow over the real successes.

The shift in perception is palpable in the literature: a 1999 *Nature Genetics* article was entitled “Array of hope,” but a 2005 *Nature Reviews* article was entitled “An array of problems.” One recent paper called microarray studies a “methodological wasteland.”
All new technologies have growing pains, and early glitches with the technology itself are partly to blame. But the bigger problem is more fundamental. The huge promise of microarrays is that they give information about every gene, but this is also their huge curse—a crushing onslaught of data. A decade ago, these data were a mismatch with existing statistical tools. Today, there is still no consensus on how to analyze and interpret them. A 2005 survey of microarray users concluded that, “Data interpretation and bioinformatics remain the major hurdles in microarray technology.”

**EXPONENTIAL ADVANCE**

Microarrays capture a snapshot of which genes are turned on—or expressed—in a given cell at a given time. Before 1995, scientists could only explore the activity of a few genes at a time. Then two groups of researchers—at Stanford University (led by Patrick Brown, MD, PhD) and at Affymetrix—scaled this up thousands-fold with the invention of the microarray. The Stanford microarray is a glass slide coated with a grid of thousands of microscopic spots—each corresponding to a gene—that light up to show which genes are on, which are off, and to what degree they are being expressed.

“At the time it just blew away the next best thing that you could do. It was really a quantum leap ahead,” says Todd Golub, MD, director of the Cancer Program at the Broad Institute of Harvard and MIT.

Besides expression microarrays, genotyping microarrays are becoming increasingly popular—these reveal variation in the DNA code rather than in gene activity. Scientists are also working on microarrays that use antibodies to detect proteins, but these have even more technical challenges.

**ARRAY OF HOPE**

Microarrays are ideally suited to study cancer, a disease of multiple genetic mishaps. They may also yield improved tests for diagnosis and prognosis.

In a 1999 paper in *Science*, Golub automatically and accurately classified leukemia patients into the two main subtypes of the disease using only gene expression patterns. Though these two forms of leukemia were already well recognized and characterized, in principle this strategy could uncover previously unknown subtypes of cancer.

Indeed, in a 2001 *Proceedings of the National Academy of Sciences* (PNAS) paper, researchers identified five unique gene expression patterns in breast cancer and showed that these subtypes were five distinct diseases with different risks of progression.

“This surprised the lab scientists. They said, ‘Wow, look at this—breast cancer isn’t really breast cancer, it’s many types of breast cancer,’” says Gilbert Chu, MD, PhD, professor of medicine and biochemistry at Stanford. “But if you talk to anyone who’s been a clinician for many years, they already knew this. They’ve seen breast cancers that looked the same but in some cases vanished with chemotherapy and in others did not. So it’s not a surprise that the gene expression profiles are proving that these cancers are different.”

A natural extension of this work is to isolate the particular genes and expression patterns that are linked to prognosis. For example, Golub derived a 13-gene expression signature that correlated with survival in lymphoma patients; several other groups have isolated gene sig-
natures for breast cancer prognosis. These signatures can be used in prognostic tests that gauge if a tumor should be treated aggressively.

Prognostic genes may also point to novel drug targets. For example, in a recent paper in the Journal of Clinical Oncology, Elaina Collie-Duguid, PhD, research fellow at the University of Aberdeen in Scotland, identified one gene that had a 50-fold higher expression in lung cancers that were not responsive to chemotherapy compared with those that were responsive. The gene codes for a protein that prevents tumor cell death; blocking this protein might boost chemotherapy response.

Microarrays may also help tailor a treatment to the person, not just the disease, Chu says. He has identified preliminary gene signatures in the healthy cells of cancer patients that predict which patients will suffer serious side effects from radiation therapy.

In addition to oncology, microarrays are also being widely applied in heart disease and transplant research. Daniel R. Salomon, MD, associate professor of molecular and experimental medicine at the Scripps Institute in San Diego is working on developing a microarray-based test to quickly tell him if a kidney transplant patient is in acute rejection, chronic rejection, or good condition. And he envisions an even more sophisticated personalized medicine scenario: “What we ultimately want is where the doctor says to these patients, ‘I saw you on Thursday and you’re doing well but your gene expression analysis tells me you need more immunosuppression, so I’m increasing your dose. Come back and see me in four weeks and we’ll draw blood and check your immunosuppression again.’”

Microarrays are also critical in basic biological research that may ultimately have a clinical payoff. For example, by mapping the precise genetic program of embryonic development, Wing Hung Wong, PhD, professor of statistics and
of health research and policy at Stanford, may be drawing a blueprint for where and when to deliver genes for gene therapy.

**Array of Problems**

The initial successes in microarrays and their exhilarating promise set off a dizzying flood of microarray studies: fewer than 100 publications in 1999 grew to more than 6000 in 2004. Suddenly investigators were identifying a molecular signature for every disease.

But many publications have since been discredited or have simply fizzled out. Scientists say it’s hard to find studies that have led to anything concrete.

“The thing that’s surprising to me is that it’s taking so long to figure out whether and when the technologies work, and it’s taking so long in the face of such enormous enthusiasm,” says David Ransohoff, MD, professor of medicine and epidemiology at the University of North Carolina, Chapel Hill.

Of the many factors at work—including initial snags with the technology—scientists consistently point to data analysis and interpretation as the critical stumbling block.

“For the part of running the experiment, microarrays seem to be working pretty well,” says Stanford cardiologist and medical fellow Greg Engel, MD.

“The informatics part is a whole other area. The statistics and how you analyze the data are still a quagmire.”

“I think the greatest challenge at this time remains data interpretation,” Golub agrees.

When the data amounted to whether a single gene was on or off, biologists had little need for statisticians. But finding patterns in the activity of 36,000 genes is fundamentally a statistical problem. When microarrays were introduced, even the statisticians were stumped. Existing statistical tools were built to analyze data on a few variables measured on many samples. In microarrays, the situation is reversed: tens of thousands of genes are measured on just a few samples—a phenomenon statisticians are dubbing “p bigger than N” (p features on N samples).

To illustrate the difficulty, imagine that you randomly divide 50 people into two groups and start endlessly measuring their characteristics: age, hair color, favorite food, height, weight, and so forth. Eventually, you will find characteristics that are slightly imbalanced between the two groups just by chance. And the more variables you consider, the more differences you will find. But the pattern of characteristics that separates the two groups is an idiosyncrasy of the sample and has no larger meaning. The same thing happens when you compare 36,000 genes between two sets of 25 cellular samples—some differences in expression may reflect real biological changes, but many more will be false positives.

“The problem here is a deeply profound statistical one,” Chu says. “The very nature of microarrays is that they give you tons of data, and very unusual patterns can emerge that are not anything more than noise and statistical fluctuation.”

As Engel describes it, “What we’re all doing is we’re taking a statistical approach and we’re all trying it every which way. And you’ll even get a pattern. But is that pattern real? That’s the major issue for gene chips—is it real?”

Subtle statistical mistakes lead you to find patterns and get published in high-profile journals, Chu says. And it may take years and several expensive follow-up studies for anyone to realize that the finding is not reproducible, unless someone spots the error sooner.

Unfortunately, such sleuthing isn’t a job for the casual scientific reader, Chu says. As a perspective in *Nature Genetics* quips, this task requires “forensic statisticians.”
At the request of a colleague, Robert Tibshirani, PhD, professor of statistics and health research and policy at Stanford, set out to evaluate a 2004 paper in the New England Journal of Medicine that reported a novel gene signature for predicting survival in follicular lymphoma.

Using the data the authors provided online, Tibshirani spent two grueling weeks reconstructing the steps of their analysis and writing a computer program that reproduced their results. Then he put their approach to the test.

To help determine whether a pattern is real or just random noise, statisticians use a trick called split-sample validation: they fit a model only on a portion of the dataset (called the training set) and then test its discriminatory ability on the untouched data (called the validation or test set). If the model only fits noise in the training set, it will usually fall apart when applied to the test set. But even this isn’t perfect, because bias can be introduced in choosing the training set and specifying the model.

The authors of the lymphoma study had fit a model (a gene signature) using half the data and found that the model performed well when tested on the remaining half. But when Tibshirani simply swapped these training and test sets and applied the model-fitting program to the new training set, unexpectedly no models popped out. In fact, no models popped out, suggesting that their whole finding was spurious.

He also re-ran the computer program on the original training data with tiny changes in the choice of parameters. “Again, the whole thing fell apart like a house of cards,” he says. “I also had other colleagues look over my analysis, and they all agree with me: these data look like noise.”

“On the broader issue, I think probably a good portion of microarray analyses are wrong,” he says.

A 2005 Lancet paper confirms his suspicions. Stefan Michiels, PhD, and colleagues at the Institute Gustave-Roussy in France reanalyzed data from the seven largest published studies to report gene expression signatures for cancer prognosis. The papers were published in top peer-reviewed journals, including: Nature, PNAS, the Lancet, and the New England Journal of Medicine.

For each of the seven datasets, Michiels’ team randomly selected 500 training sets of different sizes; then they built and tested 500 models. What they found: The results were highly dependent on the choice of training set. Every different training set led to a different molecular signature. Moreover, in the majority of trials the signatures selected in the training set had poor or no discriminatory ability in the validation set.

Their conclusions: five of the seven studies did not classify patients better than chance, and the remaining two did only slightly better than chance.

“The original investigators may have consciously or unconsciously reported the best performing pair of training-validation data,” explains John P. A. Ioannidis, MD, PhD, professor and chair of the department of hygiene and epidemiology at the University of Ioannina in Greece, who wrote a commentary for the Lancet paper. “I suspect that they probably had some source of selection bias somewhere in the process,” he says.

To guard against bias, he recommends using a repeated sampling scheme like that in the Lancet paper or having independent groups do the training and validating steps.

The field of statistics moves more slowly than biology, Tibshirani says. So, as the microarray technology raced ahead of the analysis tools, non-statisticians made up their own statis-
biologists wouldn’t even want to talk to you if you were a statistician. But now the biologists all realize that statistics has something to offer. It’s really raised the profile of our field.”

BACK TO BASICS

Even if the statistical analyses are perfect, however, this does not guarantee a reproducible finding, Ransohoff cautions. Too often biologists and computational biologists overlook an even more basic problem: “Fancy math can’t undo biases that have been hard-wired into the data from fundamental errors in clinical study design,” he says.

“This is not fancy molecular stuff, its basic study design that goes back to the 19th century. If case and control samples are not maintained the same way, then we might develop molecular signatures that simply tell us what refrigerator the samples were stored in,” Ioannidis adds.

At the same time, journal editors are tightening standards and requiring authors to follow the MIAME (Minimum Information About a Microarray Experiment) guidelines and to make data available online. They should also encourage authors to provide a script of their analysis, like a statistician’s lab book, Tibshirani says.

“There’s implicit pressure to find positive results. And that’s not a good way to operate,” he says. “A script keeps you honest. It forces you to remember exactly what you did, maybe six months ago. Maybe you’ve forgotten that you’ve actually tried 25 models since last July.” A script also makes it easier for others to evaluate the approach.

Another solution is canned software—such as the packages that he’s developed and made freely available online, SAM and PAM (Significance Analysis of Microarrays and Prediction Analysis of Microarrays). These programs constrain people from simply making the choices that make their data look best. “You need a little bit of a straight-jacket almost,” he says.

Biologists are also realizing the importance of having statisticians on their microarray teams, Wong says.

“I can clearly detect a changed perception about statisticians,” he says. “Before, the
For example, a 2002 *Lancet* paper (by Lance Liotta and Emanuel Petricoin) announced the development of a highly accurate blood test for early stage ovarian cancer. Ovarian cancer is usually fatal because it is diagnosed too late, so accurate early detection would be a huge leap forward—exactly the incredible payoff that the “-omics” technologies have long promised to deliver. The test was based on proteomics—patterns from mass spectrometry, rather than microarrays—but of biostatistics and human genetics at the University of California, Los Angeles.

"Five years ago there was wide enthusiasm about microarrays, so people were probably a little bit too naive about the challenges that lay ahead," he says. "Now the pendulum appears to have swung back in the opposite direction, where people are much too negative about the promise of microarray data."

Indeed, the backlash has overshadowed some exciting successes. In 2005, the FDA approved the first microarray-based clinical test, AmpliChip (from Roche and Affymetrix). The test identifies genetic variations in the gene for cytochrome P450—an enzyme that metabolizes common drugs—and allows doctors to personalize drug choice and dosing accordingly.

A 21-gene expression test for breast cancer, Oncotype DX (Genomic Health), has been validated in large, independent studies. By distinguishing lower and higher risk tumors, Oncotype DX may spare up to half of women with a common type of early-stage breast cancer from unnecessary chemotherapy. A 2005 analysis showed the $3000 test to be cost-effective because of the averted chemotherapy.

Oncotype DX is now being tested in a major prospective clinical trial sponsored by the National Cancer Institute.

A 70-gene breast cancer test developed in the Netherlands, MammaPrint (Agendia), is undergoing a second round of validation studies. The jury is still out, but it is already being used in some clinical settings. A recent study in the *New England Journal of Medicine* found that though MammaPrint and Oncotype DX only overlap in one gene, they give similar results—they agreed about whether tumors were "high" or "low" risk in 81% of cases.

While these examples fall far short of a finger-prick test that instantly sizes up your current and future health, they show that microarray data are not an empty wasteland. Dismissing microarray technology now would be like stopping flight travel because the first few planes crashed, Horvath says. These early crashes led to strict and effective safety procedures for flight, and, similarly, early failures in the microarray field have led to stricter standards to ensure reproducibility, he says.

As a more mature field faces its second decade, it is also adopting a more realistic outlook. Microarray users acknowledge that an all-inclusive finger-prick test is unlikely to materialize anytime soon, but they have a more modest goal for their next decade: to streamline their search for meaning in a vast sea of data.

### Footnotes