# Do-It-Yourself Gene Watching

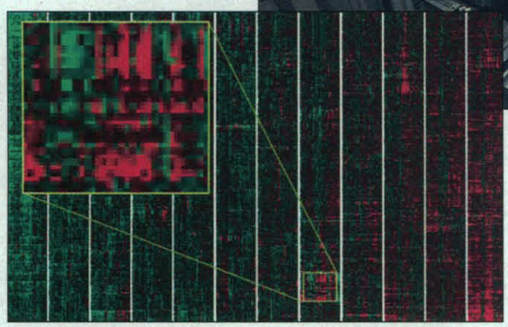
**The growing use of relatively inexpensive microarrays to monitor the expression of thousands of genes at once is creating a flood of data on everything from strawberry ripening to viral pathogenicity**

Next week, students will begin arriving at the Cold Spring Harbor Laboratory on Long Island to begin "our most oversubscribed laboratory course on record," says David Stewart, director of meetings. Sixteen people paid $1955 each to learn how to build and use a machine for genetics research—a device that deposits thousands of pieces of DNA in precise microarrays on glass slides. For another $30,000, four will actually take the machine home. "We were somewhat amazed," says Stewart, surrounded by boxes of parts waiting to be assembled. The course is new and it wasn't even advertised, yet eight times as many people signed up as could be accepted.

Microarrays are hot. People who never thought they would do large-scale gene studies suddenly are eager to try their hand at monitoring thousands of genes at once. They are watching patterns of gene expression change as strawberries ripen, viruses cause disease, and tuberculosis infects host cells (see sidebar). And they are cataloging the genes that are overexpressed or suppressed when normal cells become cancerous. The National Institutes of Health (NIH) is supporting this trend, funding its own microarray studies and providing grants to institutions to buy the technology. All this is generating a flood of data that traditional journals find hard to accommodate and digital databases don't yet know how to handle.

The basic idea behind this surge of interest isn't new: Researchers have been using microarrays since the early 1990s to study gene expression en masse. What is new is the relatively low cost of entry into the field. Over the past year or so, inexpensive, do-it-yourself techniques like the one being demonstrated at Cold Spring Harbor have become widespread, replacing or complementing the high-tech "GeneChip" technology that was once about the only game in town.

The GeneChip system, made by the Affymetrix Corp. of Santa Clara, California, paved the way, and is still the system of choice for many pharmaceutical companies and academic labs that can afford it. Affymetrix uses a photolithographic method borrowed from the electronics industry to deposit probes for thousands of different genes on a single wafer the size of a dime. Each probe is a short stretch of synthetic DNA called an oligonucleotide that replicates a unique sequence identifying a gene. These "oligos" are laid down in precise, sequence-specific arrays. To determine which genes have been expressed in a sample, researchers isolate messenger RNA from test samples, convert it to complementary DNA (cDNA), tag it with fluorescent dye, and run the sample over the wafer. Each tagged cDNA will stick to an oligo with a matching sequence, lighting up a spot on the wafer where the sequence is known. An automated scanner then determines which oligos have bound, and hence which genes were expressed.



**Switched on.** Jeffrey Trent is using a machine built in his lab to look for differences in gene expression patterns in melanoma cell types—producing huge data sets (*inset*).

Affymetrix sells a variety of standard kits for yeast, *Arabidopsis*, mouse, rat, and human genes, among others, which are listed at $500 to $2000 per chip. (The chips are good for one use.) The company donates equipment to collaborators at major genome centers, but few labs get free chips and few can afford the estimated $175,000 it costs to install an Affymetrix setup. Several researchers claim that, until recently, it was also hard to get GeneChip arrays because supplies were short.

Among those responsible for lowering barriers to the field are the three scientists who will be teaching the Cold Spring Harbor course, all from Stanford University: geneticist Patrick Brown, his former grad student Joseph DeRisi, and bioinformatics expert Michael Eisen. Brown, along with an engineering student named Dari Shalon, devised a cheap way of generating microarrays in the mid-1990s to study patterns of gene expression in yeast. It's simple but effective: Instead of using expensive and time-consuming photolithography to lay down oligo arrays, the Stanford team uses metal rods like fountain pens to deposit carefully selected cDNAs at known locations on a microscope slide. These cDNAs act as probes for genes expressed in a test sample.

Shalon left Stanford to found a company based on this concept, Synteni Inc. of Palo Alto, California. Last year, Incyte Pharmaceuticals, also in Palo Alto, acquired Synteni for $80 million. Incyte now processes microarray chips for a fee, much as film is processed. But Brown and his lab took a different tack: They give the technology away.

Last year, DeRisi launched a Web site that explains exactly how to build a microarray machine with off-the-shelf parts (see sidebar, p. 446). And Eisen has given away gene-clustering software that identifies patterns in microarray data. Brown, meanwhile, has become a big proselytizer, inviting dozens of collaborators into the field. Kenneth Burtis, a *Drosophila* expert at the University of California, Davis, who followed DeRisi's lead and built his own arrayer, says, "Joe's take on it was: 'People don't realize this isn't rocket science, and they shouldn't be afraid of it.' That's the way I got swept up in this."

Many other researchers are building ma-

CREDITS (LEFT TO RIGHT): J. TRENT; RICK KOZAK

## An Array of Uses: Expression Patterns in Strawberries, Ebola, TB, and Mouse Cells

When scientists began using microarray devices to study gene expression in the early 1990s, many focused on the same humble organism: brewer's yeast. Since then, they've widened their horizons. Experiments are under way studying how genes are turned on and off in complex plants, pathogens, model animals such as the nematode and mouse, and human cancer cells. Some of these projects were on display last month at a meeting of microarray users organized by *Nature Genetics* in Scottsdale, Arizona,* where the following examples were highlighted.

• Just about everyone likes strawberries, but no one has identified the genes involved in fruit development, according to Asaph Aharoni, who decided to look for the answer in a gene expression study. Aharoni, a biologist at the Center for Plant Breeding and Reproduction Research in Wageningen, the Netherlands, focused on a group of 1800 genes from strawberries. Using microarray technology developed at Stanford (see main text), he printed strawberry cDNAs—probes for expressed genes—on slides and monitored which genes were being expressed in fruit, from green to fully ripe. Aharoni found 200 genes whose expression varies with development, including a late-stage cluster that is turned on during membrane breakdown. Now he aims to look at genes affecting hormonal control.

• Kevin Anderson and Chunsheng Xiang of the U.S. Army Medical Research Institute of Infectious Diseases in Frederick, Maryland, investigated a more sinister organism: Ebola virus. They were curious about what makes the Ebola-Zaire strain a feared killer and the Ebola-Reston strain—which turned up in a Virginia primate lab in 1989—not a known threat to humans. Using a cDNA array of 1400 human genes, Xiang compared the gene expression profiles of normal human monocyte cells and cells that had been infected with two strains of virus. The Ebola-Zaire strain produced a "remarkably different" pattern from the Reston strain, according to Anderson. It strongly induced genes that produce immune-system regulators called cytokines and chemokines, along with inhibitors of apoptosis. Anderson says this may suggest how the deadly Zaire strain spreads rapidly.

• A large research team is building a comprehensive collection of full-length mouse genes under the leadership of Yoshihide Hyashizaki at the RIKEN Genomic Sciences Center in Tsukuba, Japan. Speaking for RIKEN, Yasushi Okazaki presented data from the most recent addition to this massive database—a map of gene expression in the mouse body. Okazaki and colleagues, with help from Stanford, have arrayed 20,000 mouse cDNAs and recorded distinctive gene expression patterns for the heart, liver, tongue, kidney, lung, spleen, placenta, and other tissues.

• A group of British researchers used the genomic sequence of *Mycobacterium tuberculosis*, finished just last year, to look at which genes are turned on in this lethal bug during infection. Joseph Mangan of St. George's Hospital Medical School in London used an array of TB genes to see which were most highly expressed as the organism invaded the scavenger cells called macrophages. Among the genes that appear most active during early infection are a group involved in capturing iron, suggesting that the organism competes with the host for iron, and in a "dormancy" response that may help TB evade immune attack.

–E.M.

* The Microarray Meeting, 22 to 25 September, Scottsdale, Arizona.

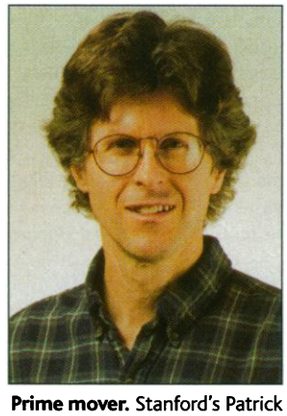
chines, and several companies are now selling machines like Stanford's at roughly twice the price of the do-it-yourself model. Geoffrey Childs and Aldo Massimi at the Albert Einstein College of Medicine in New York City, and Vivian Cheung at the University of Pennsylvania, Philadelphia, designed and built microarrayers from scratch. Others, including a team at Rosetta Inpharmatics Inc. in Kirkland, Washington, and at the Hewlett-Packard Co. of Palo Alto, have developed ink-jet oligo printers, but these are not generally available.

Affymetrix, meanwhile, has taken steps to increase its production of GeneChip arrays and offer terms more agreeable to academics. In September, the company also moved into the spot microarray world, acquiring a small company that sells these machines, Genetic Microsystems of Woburn, Massachusetts. DeRisi views this move as an attempt to swallow the competition, but Affymetrix's vice president of marketing, Thane Kreiner, describes it as a way to give clients a technology that "complements" the GeneChip, although the company insists that GeneChip arrays yield higher quality data.

All of this points to a boom in microarray experimentation by "mom-and-pop" genetics labs. What is the attraction? Simple, Brown says: "As people look at large-scale pictures of the expression programs of genomes, they've begun to realize that there's at least as much information in genomes entirely devoted to [controlling] where and at what level the genes are expressed" as to defining proteins. Gene expression, he points out, is what really distinguishes one cell type from another. "And suddenly, that's just an open book."

### The vanguard

Among the sponsors of this technology is National Cancer Institute (NCI) director Richard Klausner. NCI was an early collaborator on GeneChip technology and has been funding large-scale studies of gene expression in cancers since 1996. Now NCI is backing low-cost microarrayers as well. On 21 September, the institute awarded $4 million to 24 institutions, including cancer clinics, to help them set up microarray facilities. "It is absolutely imperative that cancer researchers have open access to this technology," Klausner said in a prepared statement.



**Prime mover.** Stanford's Patrick Brown.

Klausner and others are hoping that the ability to monitor gene expression will enable them to "produce a snapshot of the genes that are active in a tumor cell." This thrust was advocated by an advisory panel chaired by Eric Lander of the Massachusetts Institute of Technology and Arnold Levine, president of The Rockefeller University in New York City, both of whom are themselves major users of the Affymetrix technology. Lander, for example, has recently been developing tools for cataloging leukemias by their gene expression signatures (see Golub Report, p. 531). And Levine recently published a study of gene expression in colon cancer.

Several lab chiefs at NIH also began collaborating on microarray studies with Brown, Eisen, and Stanford geneticist David Botstein in the mid-1990s. Now they're hooked. Jeffrey Trent, intramural research chief at the National Human Genome Research Institute (NHGRI), built a Stanford-style arrayer 3 years ago on NIH's campus in Bethesda, Maryland, and has been using it to study genes involved in melanoma. Like other devotees, Trent believes that GeneChip ar-

CREDIT: HHMI

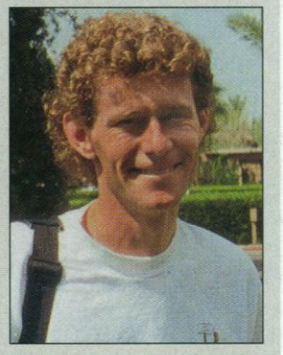
## Companies Battle Over Technology That's Free on the Web

The microarray revolution reached a flash point at Stanford University on 17 April 1998. That's when Joseph DeRisi, then a grad student in Patrick Brown's lab, posted a document called the "MGuide" on the Web. It isn't a radical tract; it's just a "lighthearted" manual, DeRisi says, telling the reader how to build a microarray robot and listing all the necessary parts, suppliers, and prices (cmgm.stanford.edu/pbrown/mguide/index.html). The estimated cost: $23,500.

The Brown-DeRisi machine employs a cluster of metal pens to print thousands of tiny DNA spots on glass slides, which can be used to perform rapid studies of gene expression. Researchers like the design because it allows them to do gene expression studies for a fraction of what it would cost to obtain the equipment for similar studies from large commercial enterprises, such as Affymetrix Inc. of Santa Clara, California, maker of GeneChip systems (see main text).

Affymetrix has never challenged the MGuide. But the company has been engaged in a furious legal battle with business rivals using the same technology, including a competitor that was born in Brown's lab called Synteni Inc. of Palo Alto, California.

Synteni is the brainchild of Dari Shalon, a former grad student of Brown's and co-inventor with Brown of the microarray gadget described in the MGuide. Brown, Shalon, and Stanford filed for a U.S. patent in 1994 and received one in 1999. Stanford gave Shalon exclusive rights to commercialize the arrayer, and in 1994 he founded Synteni. He then sold the company and its patent rights to Incyte Pharmaceuticals of Palo Alto for $80 million in January 1998. (Shalon is now director of Harvard University's Center for Genomics Research.) Incyte uses the technology to provide gene expression monitoring services to clients but doesn't sell machines. In May 1998, it closed a big deal to supply data to Monsanto.



**Giving it away.** Joseph DeRisi posted instructions on the Web for building an arrayer.

Right after Incyte went into the business, the legal battle over microarrays began to heat up. Affymetrix, which had filed broad patents on microarray concepts and systems between 1989 and 1996, sued Incyte in January 1998 in the U.S. District Court in Delaware for patent infringement. Incyte responded with a countersuit for infringement against Affymetrix. In September 1998, Affymetrix upped the ante: It asked the Delaware court for an immediate injunction to stop Incyte from "making, selling, or offering to sell their Gene Expression Microarray products and services." Incyte, meanwhile, appealed to the U.S. Patent and Trademark Office for an extraordinary "interference proceeding," arguing that its patents voided key claims advanced by Affymetrix.

Both maneuvers failed: The court dismissed Affymetrix's injunction request, and the Patent Office ruled that the evidence did not support Incyte's argument that the rival patents were void. It's now up to the courts to decide which company is violating the other's turf; the trial is scheduled to begin in September 2000 in the U.S. District Court in San Francisco. This battle ultimately will include other contenders, as well, including Hyseq Inc. of Sunnyvale, California, and Edwin Southern of Oxford University in Oxford, U.K., who hold broad patents on gene array technologies.

Meanwhile, DeRisi says that thousands visit the MGuide, and several dozen labs around the world—including in China, Japan, Australia, and Eastern Europe—download updated versions of the manual, presumably because they use it. Could this giveaway of the technology that companies are battling over draw legal attacks as well? Brown and DeRisi don't think so. They note that Stanford supports free use of the technology for research. Anyway, where patent issues are concerned, Brown says, "I don't want to have anything to do with them if I don't have to." DeRisi adds: "I've looked at the legal documents; I can't understand what they're talking about."

**–E.M.**

rays and microarrays are powerful because of their huge data output. Big samples make it easier to spot patterns, such as common sets of genes expressed in different kinds of cells. The Stanford "mantra" is quite simple, says Eisen: "More data is good." Eisen's software sorts through the color-coded microarray. readouts, clustering genes that exhibit similar patterns of expression in various cells.

Trent and his colleagues at NHGRI made their own slides to monitor the expression of more than 8000 human genes from 31 melanoma tumors. Offering a visitor a glimpse of the results last month, Trent pulled out a sheet with colored dots grouped in what he calls "Eisenized" clusters. Along the top are names of the melanoma cell types; down the side, in fine print, are the names of human genes whose fragments were deposited on the slides.

To generate the data for this chart, Trent tagged cDNA from cancerous and normal control samples with red and green fluorescent dye, respectively, then washed the samples over the slides. Genes strongly expressed in the cancer cells as compared to a reference standard gleamed a lurid red when excited by a laser, while those underexpressed showed up in green. Genes expressed in roughly equal proportions came out yellow. Eisen's algorithm grouped genes with similar expression patterns across the range of cell types in colored blocks on the chart, on the assumption that the function of these genes is similar as well. Genes of known and unknown function turn up in clusters, so researchers tentatively assign functional labels to unknown genes based on their cluster mates. Trent acknowledges that this approach is "speculative," but it is a first step, he believes, in developing new, molecular definitions of high- and low-risk types of melanoma.

A short distance from Trent's lab on NIH's Bethesda campus, an NCI team led by Edison Liu and Louis Staudt is using a locally made arrayer to investigate breast cancer, leukemia, and lymphomas. Staudt described some of this work at a meeting of microarray researchers in Scottsdale, Arizona, last month, comparing it to astronomy. His lab is doing "discovery" research, he explained. Like Galileo, he suggested, NCI scientists have a new instrument so powerful it will let them see patterns that just weren't visible before. Staudt warned, however, that there are professional risks in this venture. Galileo was denied tenure, he joked, because he was handed "a pink slip saying [his telescope] wasn't hypothesis-driven"—something for which microarray studies are sometimes faulted.

Staudt and colleagues have created what they call the "Lymphochip," an array with 18,500 carefully selected genes involved in the development of the immune system's antibody-producing B cells. "We had absolutely no trouble getting the technology up and running," says Staudt, who's working with Stanford to create a shared gene expression database. Already, he says, it looks as though microarray profiling "will be a very useful tool" for "subdividing disease categories and giving them a molecular identity."

Ash Alizadeh, one of Staudt's collaborators at Stanford, described how he used the Lymphochip to look at gene expression profiles in 50 cases of diffuse large cell lymphoma, long considered a "wastebasket category" of poorly defined illnesses. After linking genetic profiles to case outcomes,

he identified two distinct subgroups—"diseases within a disease," Staudt says. One gene expression profile appears to carry a good chance of survival; the other does not. If such results hold up, genetic profiling could be useful in diagnosing and treating lymphoma.

**Data overload**

With such tools coming on line and interest in expression studies on the rise, the volume of data in this field is likely to grow exponentially in the next few years. Already, Brown and others have been talking about new ways of storing, sharing, and publishing these huge files. Each experiment produces a flood of data: Trent's melanoma expression data, for example, would produce a print-out about 10 meters long if printed at full length—too big to publish in a journal.

For the moment, Brown says, microarray users are storing results in their own Web-accessible files and opening them to the public when they publish a journal article. Personally, Brown would be happy to skip the journal-controlled part of this process and put the data right out on the Web. That's why he's enthusiastic about NIH's plan for online publishing, PubMed Central (*Science*, 3 September, p. 1466).

One problem—where to archive data—may be solved soon. At the Arizona microarray meeting in September, David Lipman, director of NIH's National Center for Biotechnology Information, announced that NCBI staffer Alex Lash is heading up the design of a new database for the field, to be called the Gene Expression Omnibus (GEO). It will connect sets of experiments that appear "relevant to each other," so that a user could quickly find all the experiments involving certain gene families and look for common themes. "We're working on fields and data structures and will load samples this fall," Lipman says. He hopes GEO will be running by spring.

Yet to be resolved, however, is how to make results comparable. GEO will ask researchers submitting the data to define the experimental "platforms" they use. That may be simple for people using arrays or array services such as those provided by Affymetrix and Incyte. But there are no standards for homemade devices, and small differences in experimental conditions may lead to discrepancies in results.

But Lipman isn't rushing to impose standards on the young field. Brown thinks that's the right course: It would be a mistake, he says, to try to impose rules on the field while investigators are still in exploratory mode, pointing their microarray telescopes at the universe of genes. Better to let standards evolve gradually, as the data start pouring into GEO in 2000. **–ELIOT MARSHALL**

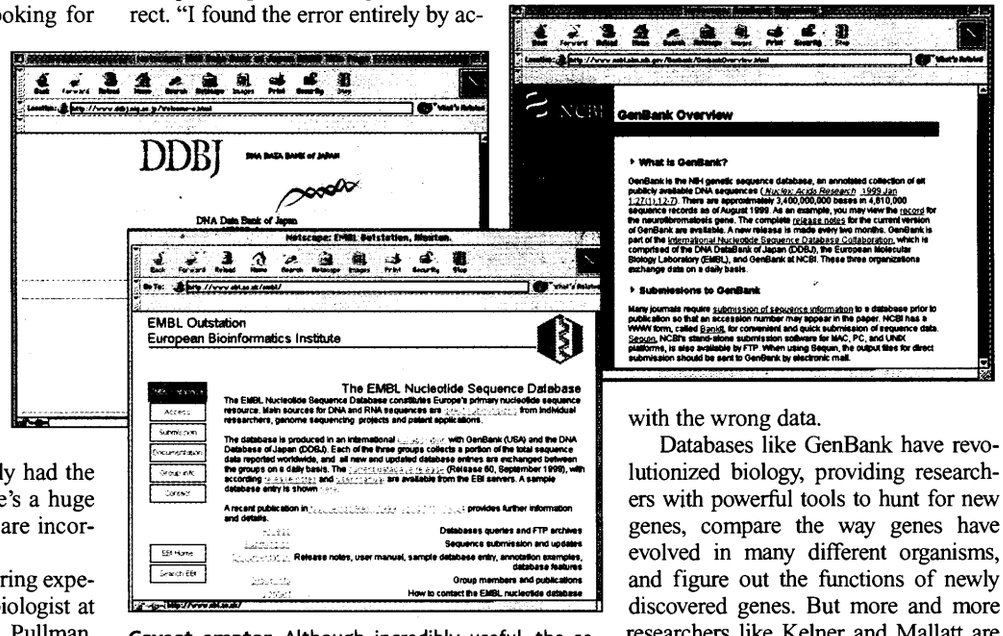
NEWS

# Keeping Genome Databases Clean and Up to Date

**As the size of GenBank and the number of other biological databases grows so does the need for ways to update and coordinate the information they contain**

Last year, Michael Kelner thought he had finally gotten his hands on the elusive front end of a gene he'd been working on, on and off, for months. But when he searched GenBank, a public archive that contains every published DNA sequence, looking for similar genes that might hold clues to his gene's function, he knew something was wrong: He turned up more than 100 matches, or "hits," from a wide array of genes from many different organisms. Kelner, a molecular pathologist at the University of California, San Diego, soon realized that the sequence all these genes had in common was a contaminant, introduced by the commercial kit he had used to clone his gene. And the fact that it turned up in so many genes in GenBank suggests that many other scientists unknowingly had the same problem. As a result, "there's a huge number of public sequences that are incorrect," he says.

John Mallatt had a similar sobering experience recently. An evolutionary biologist at Washington State University in Pullman, Mallatt was trying to determine evolutionary relationships between various organisms by comparing the sequences of specific genes. After several months' work, he realized that the GenBank sequence he had been relying on for one of the ribosomal RNA genes of *Xenopus*, a species of frog, was incorrect. "I found the error entirely by accident as I stumbled on [a report in the scientific literature] with the corrected sequence," he recalls. "It took me about 10 hours, crawling through the correct and incorrect sequences base by base, to fix it and enter the correct sequence into my phylogenetic analysis." Mallatt subsequently discovered that GenBank contains both sequences, but there is no indication which is the correct one. Because *Xenopus* is one of the few amphibians whose genes have been sequenced, it is widely used for evolutionary studies, so it's likely that other researchers have completed and published phylogenies with the wrong data.




**Caveat emptor.** Although incredibly useful, the sequence data archived in DDBJ, EBI, and GenBank are not completely error-free.

Databases like GenBank have revolutionized biology, providing researchers with powerful tools to hunt for new genes, compare the way genes have evolved in many different organisms, and figure out the functions of newly discovered genes. But more and more researchers like Kelner and Mallatt are discovering that this mother lode of information contains some fools' gold that