



ETHICS IN SCIENCE

Is Data-Hoarding Slowing the Assault on Pathogens?

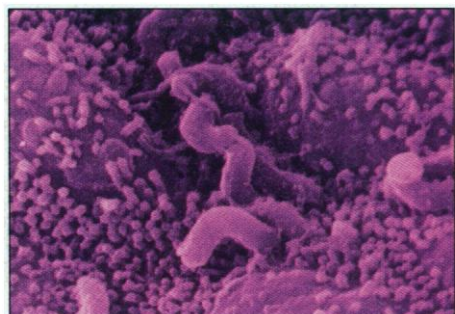
In May 1995, a new and virulent strain of tuberculosis appeared in rural Byrdstown, Tennessee, near the Kentucky border. It infected a 21-year-old worker at the Oshkosh B'Gosh clothing factory there and, in short order, every member of his family. By the time health experts had moved in to test friends and neighbors, TB had infected 75% of the man's co-workers and 80% of the people he had met in routine social encounters—more than 220 people in all. Fortunately, the organism—now known informally as the Oshkosh strain—was susceptible to antibiotics, so with standard therapy, the patients all recovered.

The Byrdstown outbreak didn't make the local news or even the federal government's *Morbidity and Mortality Weekly Report*. But it may make history: The Oshkosh bug may be the first virulent microbe to have its genome sequenced and made available over the Internet. Robert Fleischmann of The Institute for Genomic Research (TIGR), a non-profit genetics research group in Rockville, Maryland, has a \$3.2 million grant from the U.S. National Institute of Allergy and Infectious Diseases (NIAID) to do the job. He plans to complete it in 1998 and, after a 6-month delay, publish the data so that other scientists can see what makes this aggressive organism tick.

That 6-month pause has raised the ire of academic genome researchers, however. TIGR instituted the delay to check for errors and also to honor an agreement with Human Genome Sciences Inc. (HGS)—also in Rockville—TIGR's profit-making partner company. The agreement gives HGS scientists an early look at new discoveries—and time to patent them. This arrangement, and the far more secretive policies of many pharmaceutical firms, has rekindled a smoldering debate over who should control DNA sequence data, and how quickly it should be shared.

One contingent of researchers who receive public funding for sequencing says that new DNA data should be released immediately, even daily. They argue that failure to share data quickly leads to duplication and waste. They also feel that sequencing teams supported by government funds should not be allowed to lock up the data or give a favored colleague prepublication access to genetic information. And they have persuaded some sponsors to endorse a rule requiring groups funded to do high-speed sequencing

to share data as rapidly as possible, without filing patents. For their part, researchers at private companies oppose immediate data release, saying that it's like asking a pharmaceutical house to give away a formula for a new drug. And there's still a third group: researchers who receive money from both private and public sources, such as those at



Rogues' gallery. Companies are pursuing *Helicobacter pylori* (left) and *Staphylococcus aureus* (right).

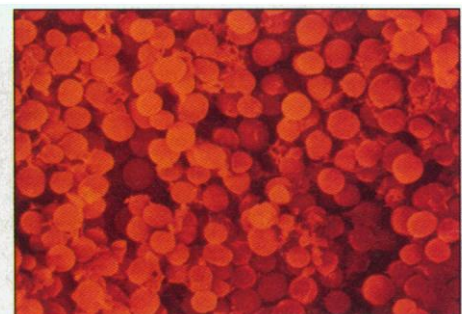
TIGR. Many—including TIGR's chief J. Craig Venter—say the demand for quick data release is based on the arrogant notion that sequencers are mere technicians. They claim that rapid data release could encourage second-rate research.

Pathogenic debate

The debate over access to DNA sequence data is raging among researchers studying species from viruses to *Homo sapiens*. But it is especially heated when it comes to the sequencing of pathogens, where holding back data, even for a year or less, arguably could cost human lives. Yet in this area, withholding sequence data is commonplace.

New software, the potential for enormous drug profits, and the lure of scientific discovery have triggered intense corporate interest in microbe DNA. HGS has a claim on *Haemophilus influenzae*, several companies are going after the ulcer bug, and many are pursuing *Streptococcus pneumoniae*. Every big pharmaceutical company, it seems, wants a genome it can call its own. One reason for the strong interest is that companies expect that it will be easier to use microbial genes in drugs that attack microbes than to create new drugs based on human genes. Academic centers and independent outfits such as TIGR and the Sanger Centre near Cambridge, U.K., also are competing furiously, spurred, if not by potential profits, by national rivalry and the prestige of having helped vanquish a disease.

The potential commercial market for pathogen sequence data was evident in August 1995, when Genome Therapeutics Corp. (GTC) in Waltham, Massachusetts, became the first private company to sell the genome of a bug it had sequenced: the bacterium that causes ulcers, *Helicobacter pylori*. GTC says it sold its data to the Swedish pharmaceutical company, Astra AB, for business deals worth \$22 million. According to GTC's vice president for research, Gerald Vovis, "Astra purchased the exclusive right to use that sequence database" and is not planning on making it public. He adds that it "wouldn't be unreasonable" to assume that GTC and Astra are try-



ing to patent the genome. Others, according to TIGR staffers, have paid for *H. pylori* data, including the British drug company Glaxo Wellcome and a group of European vaccine-makers. None is willing to disclose how much of the genome it has acquired.

By withholding DNA data, commercial drug companies may gain a slight advantage over competitors. But Jean-François Tomb, a researcher at TIGR, suggests that data-hoarding is slowing down the normal review process by which scientists check one another's results for variations and inaccuracies. He is heading up an *H. pylori* sequencing project that TIGR launched with its own funds in 1996. One aim, according to a TIGR staffer, was to share data with the research community "as a freedom-of-information kind of thing." Tomb says he would like to check his version of the *H. pylori* genome against another, so he recently asked GTC's sequencing expert Douglas Smith whether GTC would share its *H. pylori* data. The response, Tomb says, was "clearly no." Tomb says TIGR's data, which HGS has already reviewed and sought patents on, will soon be published. Tomb notes that, unlike TIGR, Astra will be able to make the comparison for "absolutely zero dollars."

Private sequencing projects have tackled many other organisms. One hotly pursued bacterium, for example, is *Staphylococcus aureus*, a bug that causes hospital infections and is increasingly resistant to antibiotics. GTC parlayed its *Staph aureus* and other genomic data

into a \$44 million commitment from the drug company Schering-Plough in December 1995. HGS also made a deal centered on *Staph aureus* data, receiving a commitment of at least \$9 million and a promise of royalties from the drug company Pharmacia & Upjohn.

While some duplication is normal in research, experts say it's getting out of hand in microbe sequencing. Tuberculosis, like *Staph aureus* and *H. pylori*, will be sequenced many times over in part because sequencers aren't sharing data, whether for business reasons or

because of interlab rivalries. The first group to take a stab at sequencing the organism was GTC. The Sanger Centre also embarked on a project to sequence a docile lab strain of TB known as H37Rv, releasing data on a daily basis over the Internet. Meanwhile, quite in-

Genomics's Odd Couple

It was an unlikely alliance from the start: When J. Craig Venter and William Haseltine hitched up in 1992, they differed both in goals and in style. Venter, an ex-government researcher, wanted to build a topnotch, nonprofit institute and publish articles. Haseltine, an Ivy League supercompetitor, hoped to create a new kind of pharmaceutical company based on genetics. Still, both men signed onto an ambitious joint project backed by \$125 million from the pharmaceutical company SmithKline Beecham (SB) to hunt for human genes.

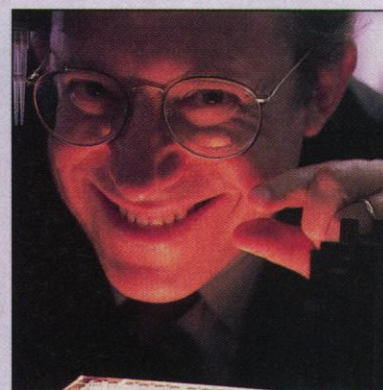
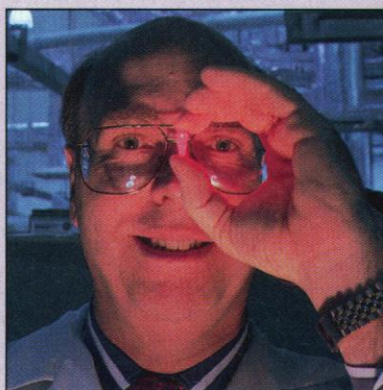
But the alliance between the "gene kings," as they were dubbed by *Business Week* in 1995, has worn very thin. Venter and Haseltine are clashing over how much control each can exercise over data The Institute for Genomic Research (TIGR) has obtained through research funded by SB and other sponsors. Says Venter, "The relationship has grown pretty distant ... there are no [longer] daily, weekly, or monthly interactions."

Three years ago, the gist of their legal collaboration seemed simple enough. Venter's nonprofit center—TIGR—would conduct basic genetic research. Haseltine's profit-making outfit—Human Genome Sciences Inc. (HGS)—would do some research, develop medical products, and manage finances, dispensing \$85 million from SB in quarterly payments to TIGR, which also would apply for government grants. In exchange, TIGR would allow HGS to preview its scientific findings and give HGS the commercial rights to all its discoveries.

As early as 1993, however, it was evident that Venter and Haseltine were growing apart. TIGR was supposed to give the collaboration a "quick start," says Haseltine, spitting out DNA fragments for HGS to evaluate. But according to Venter, HGS grew nervous about TIGR's plans to publish data, and, by 1993, HGS had launched its own more aggressive assault on the human genome, setting up a duplicate intramural sequencing program. After that, TIGR had "no incentive" to sequence human DNA for HGS, says Venter, so he steered his institute in a new direction: sequencing microbial genomes. But the change created fresh trouble with HGS, this time over the control of microbial sequence data.

Relations hit a low point in early 1995. TIGR had finished sequencing the genome of *Haemophilus influenzae*, a cause of serious bacterial infections in children. TIGR, which had financed the project with its endowment funds, was preparing to publish a scientific report. HGS objected, invoking clauses in its agreement with TIGR that entitle HGS to delay publication of TIGR's new scientific discoveries. (According to the agreement, HGS can delay data release in a series of steps for up to 18 months while HGS files for patents and develops medical products.)

Venter and Haseltine recall subsequent events differently. According to Venter, TIGR had to call in the legal guns to prevent HGS from interfering with the publication of *H.*



PHOTOGRAPHS BY PETER SIBBALD

Strained alliance. Fundamental tensions over control of data have sent J. Craig Venter (left) and William Haseltine (right) to their lawyers.

influenzae's sequence data. Venter says such clashes with HGS have "cost hundreds of thousands of dollars in legal fees that might have been spent on research." Haseltine, on the other hand, says TIGR and HGS "worked out a friendly accommodation" that allowed HGS to file for a patent on the organism's genes while Venter prepared a report for publication. Adds HGS Senior Vice President Bradley Lorimier, "We moved heaven and Earth ... to get that first bacterial genome patent" drawn up so that TIGR could publish the paper on time. It appeared in *Science* (28 July 1995, p. 496).

The bruises from that battle had not fully healed when Venter took steps late last year to put even more distance between TIGR and HGS. When TIGR, HGS, and SB created a restricted database in 1994 to share research with academics, they agreed that SB could bow out on a year's notice. Last fall, SB gave notice that it wanted out. Venter then said that TIGR would end the agreement early, open the database, and add full-length genes to it (*Science*, 29 November 1996, p. 1456). The resulting scuffle over who controls the database sent Haseltine and Venter back to their lawyers.

Haseltine immediately responded with a memo to HGS's partners, a copy of which he sent to *Science*, saying that HGS still would be entitled to preview and patent TIGR's discoveries. Venter and his attorney say they have taken the position in discussions with HGS that TIGR will allow HGS to preview discoveries for patenting, but not in a way that would clash with a research grant's terms. For example, the Department of Energy and the National Center for Human Genome Research ask that data be kept private for no more than 6 months. TIGR argues that if it accepts one of these grants, federal policy takes precedence. Haseltine disagrees; he says TIGR may not accept grants that are "not consistent with" HGS's proprietary rights in TIGR's work.

According to Venter, discussions about what can and cannot be published now are conducted through third parties. The gene kings, who have long been aloof, are now barely maintaining diplomatic relations.

—E.M.



dependently, TIGR decided to get into the TB-sequencing act and applied for an NIAID grant. The grant had already been approved when TIGR found out about the Sanger project. But rather than cancel the grant, NIAID and TIGR chiefs carved out their own niche by announcing plans to sequence a more virulent strain of TB—the Oshkosh bug.

Several other companies have sequenced—and kept secret—parts of the *Streptococcus pneumoniae* genome. This lethal pathogen attacks hundreds of thousands of people a year, causing an estimated \$4 billion in health costs. Incyte Pharmaceuticals Inc. in Palo Alto, California, announced in December 1996 that it will be receiving “genomic sequences” of *Strep pneumo* from the pharmaceutical company Eli Lilly and Co. The data will be added to other information from “one to two dozen” organisms in Incyte’s private data bank called PathoSeq. (In exchange, Lilly will get access to Incyte’s collection of human genetic data and possibly share royalties on products that result from the collaboration.) A Lilly staffer who asked not to be named says, “We’ve heard that at least six groups have done partial sequences” of *Strep pneumo*. So far, though, no DNA data have been published. TIGR and HGS have

Experts on the fruit fly (*Drosophila melanogaster*) and the nematode (*Caenorhabditis elegans*) say their fields, which until recently attracted little commercial interest, have a long tradition of sharing data. Gerald Rubin, a geneticist at the University of California, Berkeley, and a specialist on *Drosophila*, says his peers consider it “unethical not to give someone a published reagent ... definitely wrong,” and they view sequence data as just the starting point for a project.

But the tradition among scientists studying mammalian genomes, according to Rubin, is “completely different.” Sociologist Stephen Hilgartner of Cornell University in Ithaca, New York, agrees. In a study now in press, Hilgartner argues that some human gene hunters have been less generous with lab materials and willing to use subtle tactics to limit access to data.* He describes how some gene hunters and mappers have practiced “data isolation,” for example, making it hard for others to obtain bacterial clones of genes cited in their research by delaying their release or failing to identify them clearly.

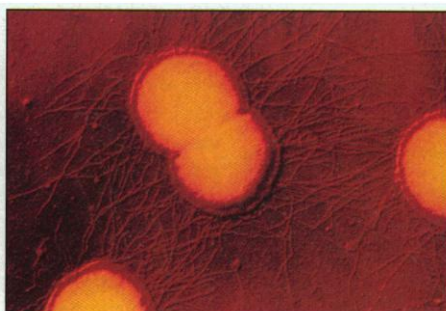
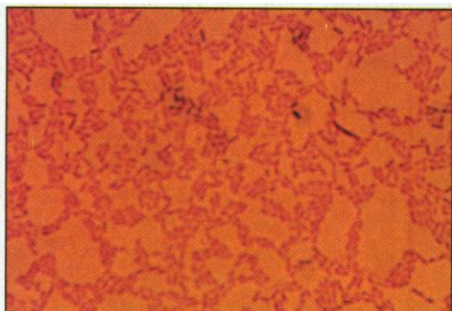
Venter agrees that refusal to share clones has been a problem. He was shunned by a research consortium in 1994, he says, when he offered to use TIGR’s sequencing equip-

the rich core of the human genome—the genes that code for proteins. When peer reviewers dismissed the project as infeasible, Venter left NIH and signed a deal to do the work with SB money.

The TIGR-HGS collection of ESTs, according to HGS chair and CEO William Haseltine, has now been used to identify about 100,000 genes. And over the past 3 years, HGS has unleashed a blizzard of patent applications at the Patent and Trademark Office. Not only is HGS seeking more than 200 patents on full-length genes—four of which have been granted—it has filed several massive patent applications containing tens of thousands of ESTs. The EST applications are still working their way through the patent office (see p. 780).

A religious campaign?

HGS’s vast, proprietary collection of genetic sequences quickly became a lightning rod for criticism of any attempts to lock up sequence data. Haseltine created a furor in 1994, for example, when he announced that academic researchers could use the database only if they signed an agreement to share proprietary rights to their work with HGS



Private pathogens. The potential for big drug profits has sparked corporate interest in *Haemophilus influenzae* (left), *Streptococcus pneumoniae* (center), and *Mycobacterium tuberculosis* (right).

nearly completed sequencing *Strep pneumo* as well, says TIGR project leader Brian Dougherty, who adds that the two organizations are discussing when and how the data will be released.

A communitywide problem

Some basic scientists who study the genes of small organisms say they are offended by the hoarding and duplication of sequence data on pathogens. “It’s a terrible waste of effort and money,” says molecular biologist Julian Davies of the University of British Columbia in Vancouver, who works primarily on pathogens. “It bothers me,” he adds, to see so much money going into duplicative work when first-rate academic projects are begging for help. But similar problems are cropping up in many other areas of genome research, and secrecy intensifies as research gets closer to the market.

ment to help locate a specific human gene.

But by many accounts, Venter was party to one of the biggest DNA-hoarding projects of all—a joint effort funded in 1993 by the pharmaceutical company SmithKline Beecham (SB) and managed by HGS, TIGR’s partner company. Its goal was to build a vast, private database containing informational “tags” from the ends of human genes. Venter was working at the National Institutes of Health (NIH) when he first proposed developing the database. The idea was to collect bits of sequence data short enough to be easily generated by robots, but long enough to be unique. These “expressed sequence tags,” or ESTs, he said, were the shortest route to

(*Science*, 7 October 1994, p. 25). The debate that ensued spurred a series of efforts to push as much sequence data as possible into the public realm.

The first move came in September 1994, when SB’s competitor, Merck & Co., announced that it would finance a sequencing project run by geneticist Robert Waterston at Washington University in St. Louis that would duplicate some of the work already done by HGS and TIGR. Merck won’t disclose the cost of its Merck Gene Index, but at the conservative estimate of 30 cents per base pair, the company has spent about \$50 million to sequence over 156 million base pairs of DNA. The entire collection has been deposited in the National Library of Medicine’s database, GenBank. Anyone can tap into the files over the Internet and pluck out sequences for study. A Department of Energy (DOE) group called IMAGE, run by Greg

* Hilgartner’s essay will appear in *Private Science: Biotechnology and the Rise of the Molecular Sciences*, edited by Arnold Thackray, University of Pennsylvania Press, 1997.

Lennon at the Lawrence Livermore National Laboratory in Livermore, California, distributes the related clones, which can be used to search for detailed biological information.

Like many other academics, Rubin views the Merck Gene Index as a godsend. Had the pharmaceutical firms kept all the human data to themselves, "it would have been a disaster," Rubin asserts. "I am extremely grateful to companies like Merck that have made available their precompetitive information. ... It furthers my research and that of many people."

In November 1995, the Wellcome Trust announced another gift to public databases: It pledged to give the Sanger Centre \$75 million over 7 years to begin sequencing the complete human genome. In February 1996, the Howard Hughes Medical Institute in Bethesda, Maryland, awarded a \$2.3 million grant to Waterston's group to create a complete gene index for the mouse, a valuable tool for gene-comparison studies. And the National Center for Human Genome Research (NCHGR), part of the NIH, followed Wellcome's move in April 1996, with \$22 million in support for five U.S. pilot projects that have now begun sequencing the human genome at an accelerated pace. In another effort still awaiting final approval, Wellcome is expected to announce that it will offer \$25 million in grants for the sequencing of microbes. In all cases, sponsors have insisted that the data be made public rapidly.

To reinforce this ethic, several research sponsors have adopted a series of increasingly pointed guidelines for grantees. In 1992, NCHGR and DOE jointly announced a policy novel to biomedical research: It asked grant applicants who were likely to generate "significant amounts of genome data or materials" to specify exactly "how and when" they would make the results available to the public. The policy also says grantees should not retain work for more than 6 months "from the time the data or materials are generated," whether or not they were part of a published study.

The Wellcome Trust and the Sanger Centre, joined by NCHGR's director Francis Collins, built on these principles in February 1996. At a meeting in Bermuda of newly funded sequencing teams, Sanger Centre director John Sulston proposed that everyone agree to release raw data on a daily basis, or "as soon as possible," without seeking patents on the raw data. There was no audible dissent, according to geneticist David Bentley of the Sanger Centre, who was present. Bentley has

defended the policy, in a Policy Forum in *Science* (25 October 1996, p. 533), as a way to limit duplication, stifle "inappropriate" attempts to garner early patents, and avoid giving any group preferential access to data.

Collins endorsed the policy again when he announced the NCHGR's grant awards in April. And he added a new touch, asking grantees not to seek patents on "raw genomic sequence" data, because this "could have a chilling effect" on future investments in gene technology.

Still, it is not yet clear just how these principles will translate into action. First, the new rules have not met with universal praise. Venter and his TIGR colleague Mark Adams, for example, recently attacked their underlying assumptions in print, arguing that the rules would encourage sloppiness and discourage researchers from trying to publish journal articles (*Science*, 25 October 1996, p. 534). They argue instead for release "as soon as ... data have passed a series of rigorous quality control checks and have been annotated." Also, the anti-patenting rule clashes with a 1980 federal law, called the Bayh-Dole Act, that encourages federal grantees to patent their discoveries.

And the issue of when and how to share sequence data is especially complicated when it comes to labs that take both private and public funds. TIGR's allegiance to HGS already has caused many headaches over data release (see sidebar). GTC also exists uneasily in two worlds: In addition to its private

income, it received at least \$37 million in grants from the U.S. government between 1990 and 1995. Yet it has released only random genomic data from parts of the microbial genomes it set out to sequence. Vovis explains that the federal grant was

mainly a "technology demonstration" project, one that was never meant to yield complete genomes. But as the company notes in its annual report, the grants helped defray the company's overhead research costs.

The debate over who should control DNA data, which has been going strong for at least 5 years, could easily continue for as many more. It is hard to predict whether the campaign for daily release of genomic data will prevail, or the patent seekers will come out ahead in the end. But one thing is clear: The amount of genomic sequence available in public databases is growing at a breathtaking pace. Venter, for one, fondly wishes that, as a result, the "whole argument" about who owns the genes "will just go away." But nobody is betting yet that it will go quietly.

—Eliot Marshall

INTELLECTUAL PROPERTY

Companies Rush to Patent DNA

Getting rich on human genes has become a fantasy for many investors in the 1990s. Big, savvy pharmaceutical companies and brash biotech start-ups are spending huge sums of money in the hope of gaining exclusive property rights to uncharted areas of the human genome. But who ends up getting rich may have more to do with their skill at navigating patent law—and with the unpredictable decisions of federal judges—than the importance of the biology they have discovered.

Although agencies around the world have been awarding patents based on DNA for more than 15 years, it's still not entirely clear which discoveries are patentable and which are not. One major unresolved issue is just how much biological data on the function of a DNA sequence is needed to win a patent. Applications based on whole genes whose function is well known stand the best chance of being awarded patents. But some less-than-complete gene sequences also have been patented in the past, when their commercial use was well defined.

This question has been brought to the fore by a mass of recent patent applications that try to lay claim to thousands of genes by patenting DNA fragments that can be used to reconstruct whole gene sequences. Even if these fragments, called "expressed sequence tags," or ESTs, are ultimately deemed unpatentable—and many experts now believe they will be—the filings could still cloud the legal picture for many years. The reason: These applications will create a priority date for the discovery of many genes, making it hard for later gene hunters to argue that they have made a truly novel discovery. This uncertainty about who can claim priority has been deepened recently by moves to place vast amounts of sequence data in public databases (see p. 777).

Opening the floodgates. While the policy on gene fragments may be in a muddle, the notion that a whole gene can be privately owned was firmly established in 1980; when the U.S. Supreme Court ruled that Ananda Chakrabarty, a molecular biologist then working for General Electric, could patent a genetically engineered organism. Chakrabarty had spliced a gene for an oil-dissolving enzyme into a microbe, creating a bug that could clean up oil spills. The U.S. Patent and Trademark Office (PTO) initially rejected the application on the grounds that life

"I am extremely grateful to companies like Merck. ... It furthers my research and that of many people."

Gerald Rubin