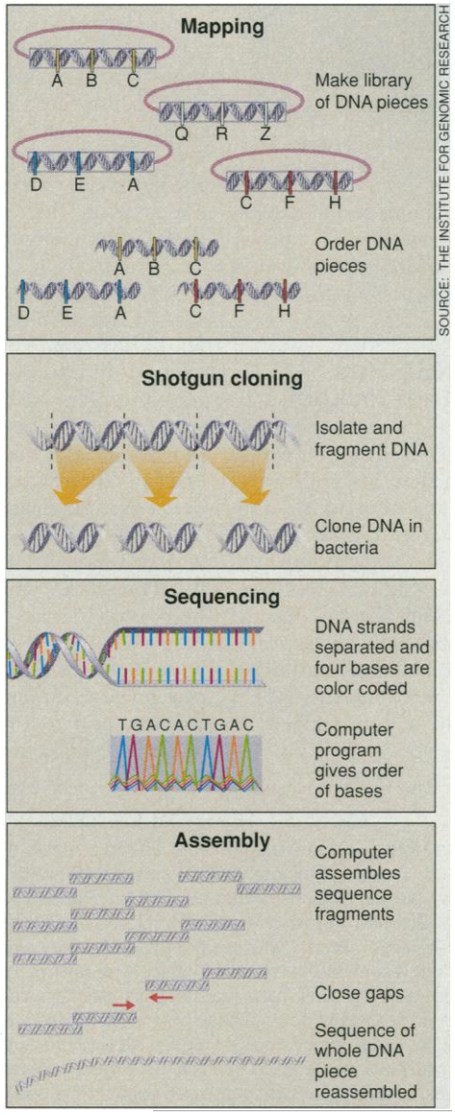
DNA Sequencers' Trial by Fire

With 97% of the human genome yet to be deciphered, research teams are contending for a place in the world's largest biology project; many have stumbled in the early going

When Stephanie Chissoe was a graduate student in biochemistry 8 years ago, determining the order of molecular building blocks in a small piece of DNA—such as the 50,000 bases in a viral genome—was a major project, enough to support a Ph.D. dissertation. But now that Chissoe directs a crew at one of the biggest human genome labs in the world—the Washington University Genome Sequencing Center in St. Louis—she expects to churn out twice that much data each day. Chissoe is part of a small but burgeoning workforce that is revolutionizing biology, changing not just the substance but the culture of science. In this new world of genetics, machines and robots do much of the lab work, and data accumulate faster than the mind can absorb. Where Ph.D.s or graduate students once held sway, skilled technicians, high-capacity sequencing machines, and computers now take up most of the lab space. Progress is measured not in journal articles but in the daily accumulation of newly sequenced DNA.

Welcome to life in the trenches in the biggest collective undertaking in biology: the attempt to sequence the entire 3 billion bases in the human genetic code by the year 2005. The U.S. agencies contributing to this international project—the Department of Energy and the National Human Genome Research Institute (NHGRI)—have already spent $1.5 billion and 7 years on what some regard as biology's equivalent of the Apollo Project. And, in spite of the progress evident in Chissoe's lab, it's turning out to be a tough slog.

Midway through the 15-year Human Genome Project, researchers are just beginning to tackle sequencing the genome on a large scale. Not one of the six pilot centers NHGRI funded in 1996 to encourage faster, cheaper DNA sequencing methods has achieved the production rates promised 2 years ago, and few can see a clear path to really high output. Costs are not well understood, and many of the centers seem to be changing procedures on a daily basis. "Our stated goal is 100 megabases [a year] by 1999 or 2000," says Mark Adams, who directs the human sequencing program at The Institute for Genomic Research (TIGR) in Rockville, Maryland. "I don't know how we're going to do that." The Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, expected to have 23 million bases done by May this year but has completed less than 9 million. The genome center at Baylor College of Medicine in Houston aimed for 15 million bases but has contributed about 8 million to GenBank, the public database for sequence data. And Stanford University has only about 900,000 bases in GenBank to show for its first 2 years of sequencing.



Megabase biology. Genome researchers aim to automate every step of DNA sequencing.

"[These numbers] reflect the true rigor of this first period," says Baylor's Richard Gibbs. Indeed, this initial phase of sequencing has provided some sobering lessons. Initially, for example, the Massachusetts Institute of Technology (MIT), TIGR, and just about everyone else overestimated the efficiencies to be gained by automation. A year ago, Eric Lander, who heads the Whitehead center at MIT, was convinced that robots could bring tremendous savings in labor costs, which represent 30% to 40% of the overall costs of sequencing. The estimate "was hopelessly overoptimistic," he now admits. Also, despite the millions of dollars the genome program has spent in past years on mapping—finding landmarks on human chromosomes that researchers can use to explore regions that have not yet been sequenced—sequencers say most maps still are not detailed enough for their purposes. Several are custom-designing new maps for their own labs, gearing up for the 97% of the human genome left to be sequenced.

To many on the front lines, these difficulties are not surprising, given the unprecedented scope and ambition of the undertaking. "You have a research problem coexisting with a production problem," explains Lander. Moreover, because biological assays are by nature extremely complex, "you can't fix the parameters the way you can in physics," explains Gibbs. Nevertheless, Gibbs is finding that output is becoming more predictable: "Everything is getting smoother and smoother." NHGRI certainly hopes that's the case, because it is about to shift the sequencing into high gear.

This fall, NHGRI intends to create a Cooperative Research Network of sequence production centers, funded to the tune of about $70 million a year through 2005—a jump from the current rate of about $40 million a year. The network will encourage interaction and sharing among laboratories, but rivalry won't end: At least eight groups, including the six existing pilot operations and two genome research centers, will be competing for a place in the network. The winners (the exact number hasn't been set yet) will receive 5-year grants to continue and expand sequence production.

During the past few months, NHGRI has been taking a hard look at which of the pilot groups have proved themselves capable of churning out their share of the 300 to 500 megabases of data a year needed from the United States to meet the international goal of finishing the human genome by the year 2005. "There's a lot more

Playing With the Numbers

Among sequencers, the centers competing to join the U.S. network that will decipher millions of bases of human DNA per year are sometimes jokingly referred to as the Liars' Club. The reason: Members of this club have made predictions about lowering costs and increasing productivity that thus far have been too good to be true.

The predictions began in 1996, after Francis Collins, director of the National Human Genome Research Institute (NHGRI) said that he would evaluate pilot projects seeking to become full-scale production centers by looking at the amount of money they spent for the number of bases submitted to GenBank, the public repository of DNA data. At the time, he noted, "if we can't get the price down to 20 cents a base or less, it's going to be hard to get this project done."

Two years later, most sequencers say the price hovers at about 50 cents a base. But Collins told the U.S. Congress in April that NHGRI has only 40 million high-quality bases to show for the $52 million spent these past 2 years on six pilot projects. That comes to about $1.30 per base, "and that's much more accurate" for human sequencing, says Craig Venter, director of The Institute for Genomic Research (TIGR) in Rockville, Maryland.

TIGR's Mark Adams and Stanford University genome sequencer Rick Myers argue that these early figures are meaningless because they're based on too little experience. "This is the wrong time to measure cost very carefully," agrees Maynard Olson, who runs a human genome production center at the University of Washington, Seattle. Eric Lander of the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, thinks some unusual scale-up costs shouldn't be counted—such as the unamortized purchase of new equipment. Others complain about the expense of redoing low-quality data establishing a new process. "The biggest issue isn't cost, it's scalability," Olson adds.

Club members like these also have good reasons why they have not scaled up as fast as they predicted they would (see main text). Current monthly production rates indicate, they say, that soon production will be on track. Yet even if results from the first 2 years of the 3-year trial are not exact, they reveal that "there are clearly differences in efficiency" among the centers, says Robert Waterston, who heads the genome center at Washington University in St. Louis. He adds: "I don't know that [the centers] are equally capable [of scaling up]."

Institution	Funding 1996–98	Output		Latest Month
	($ millions)	Megabases	$/Base	Megabases
Baylor	5.3	8.2	0.65	2.3
Stanford University	6.3	0.66	9.55	0.27
TIGR	10.1	8.87	1.14	2.4
Washington University	16.3	26.89	0.61	2.4
Whitehead Institute	10.6	8.7	1.22	1.1
University of Washington	3.3	2.5	1.32	–
University of Oklahoma	2.4	4.3	0.56	0.19
University of Texas SW Medical School	8.6	4.35	1.98	1.95

Notes: University of Washington principal investigator declined to provide data; information obtained from Web site. $/base figures are not strictly comparable (for example, Texas dedicated substantial sums to automation development; TIGR's grant also supports development of a bacterial artificial chromosome library; and Oklahoma did not incur costs for making sequence-ready maps). Oklahoma and Texas are not in the pilot project.

SOURCE: NIH AND GENOME CENTERS

Waterston may be right, but the indications are that NHGRI is not going to take a very hard line in evaluating performance. For one thing, cost accounting still seems badly confused. And, because some centers put more money than others into developing new technology, says NHGRI's Jane Peterson, it makes little sense to try to compare one center's cost per base to another's.

Some genome sequencers contend that too much emphasis is being placed on cost reduction in any case. Current costs are a far cry from 1988, when the cost per base was $3 to $5. For that reason, "you can't say people are not trying" to be more efficient, argues Ellson Chen, a sequencer at the Applied Biosystems Division of Perkin-Elmer in Foster City, California. He thinks that cost could be trimmed to 35 cents a base, "but it will take a lot of effort to get there. It may be easier just to raise that extra amount of [funding]."

Despite the confusion over costs and the slow progress in ramping up sequence output, NHGRI considers the pilot program a success. "If we hadn't done this," says Peterson, "a lot of the issues of scaling would not have come up as early as they have." –E.P.

to it than just quality and production," explains Jane Peterson, a program officer at NHGRI. Equally important is "what have they done that indicates they can scale up." Recently, NHGRI has indicated that it is ready to lower the bar it originally set for entry to the network. In August 1997, NHGRI's draft requirements said applicants would have to have sequenced 10 megabases of DNA per year. Now, all NHGRI wants is 7.5 megabases total by the deadline of 1 October 1998—and that can include sequences from bacteria, plants, mouse, or any other organism.

One thing is already clear, however: NHGRI needs all the capacity it can afford. "It's going to be a long, hard climb," says Phil Green, a computational biologist at the University of Washington, Seattle, who adds that he's "still optimistic we're going to get there by the year 2005."

Trial and error

To achieve higher production rates, NHGRI told the six pilot centers that they are free to choose any strategy they like. This multicenter approach stands in stark contrast to sequence production efforts in other countries. From the beginning, the United Kingdom's Wellcome Trust settled on a single site, the Sanger Centre near Cambridge, U.K.; that facility already churns out about 25 megabases of human sequence data a year. But by spreading its grants around, the NHGRI planned to test different approaches for sequencing before scaling up. Two years into this 3-year experiment, "there's been a tremendous convergence of methodology," says geneticist Maynard Olson of the University of Washington, Seattle, a pioneering strategist in this project. "The differences [are those] that only an expert can love."

"What you have is a multidimensional optimization problem," says Lander. Each of the pilot centers, along with a few other genome centers, has been tweaking its operations as well as its hardware to see what works best. They have found that every change in the process brings a period of adjustment and sometimes new problems. Ideas that look sensible on paper don't always increase output in the lab.

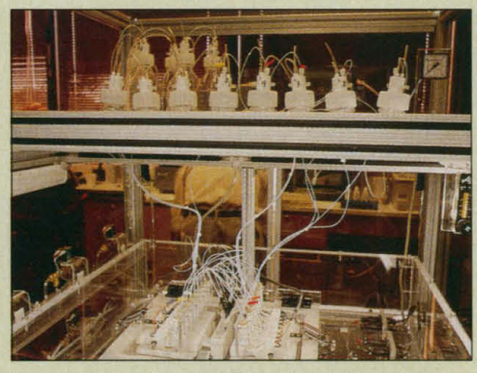
TIGR, for example, found that processing more samples at a time sounds good, but it

The Key to Success Is Finishing Well

"Sequencing is always a bit of an art," says Bart Barrell, who 20 years ago helped pioneer the technology used to determine the molecular makeup of DNA. For many like Barrell, now a sequencer at the Sanger Centre near Cambridge, U.K., the most challenging part of the job is "finishing"—the final stage in which fragments of raw DNA data are arranged into a completed sequence. Even as robots and computers take over the grunt work, this last step still requires an experienced eye and an innovative mind. Reliance on skilled finishers is "a very big issue" for labs seeking to scale up, says Robert Waterston of the genome center at Washington University in St. Louis.

Today, most big sequencing labs use a version of the "shotgun" method to decode DNA. A stretch of human DNA is first blasted into smaller pieces, which are cloned into BACs (bacterial artificial chromosomes) or PACs (phage artificial chromosomes) for fine analysis. During the shotgun phase, the material in the BACs and PACs is then chopped into smaller chunks of DNA about 30,000 bases long. Because these fragments overlap, some DNA is sequenced several times, making it possible at a later stage to arrange the fragments in order by matching their overlapping sections.

Each chunk is replicated many times over and allowed to react with fluorescing dyes that label each of the four nucleotide bases that are DNA's building blocks. Researchers feed the fragments into automated sequencing machines, which "read" the dyes and churn out red, green, yellow, and blue lines that crest and dip across the computer screen. When a color crests—usually just one crests at a time—the machine records it as one of the four nucleotides. In this way, a DNA sequence is deciphered automatically.



UNIV. OF TEXAS SOUTHWESTERN MEDICAL SCHOOL

Cheap labor. "MerMade" robot decides when targeted DNA primers are needed and makes them.

Not all "reads" are clear, however. The colored peaks sometimes overlap or exhibit other irregularities. Until just a few years ago, the standard way to resolve such uncertainties was to ask a specialist called a finisher to scan the color patterns and interpret them. It is hard, tedious work. Many sections of DNA commonly had to be redone.

Then Phil Green stepped in. A mathematician formerly at Washington University interested in DNA studies, he and his colleagues developed a computer program that analyzes machine readouts as a human would. Drawing on 20,000 known DNA sequences, Green created a system that estimates the chances that a particular color peak is really the base it seems to be. PHRED, as this base-calling program is called, quickly caught on. Green—now based at the University of Washington, Seattle—also introduced a companion program, PHRAP, which assembles short sequences into longer ones based on the overlaps. Because PHRED rated the quality of each base, it enabled PHRAP to do a better job of assembling long stretches of bases. These programs made a "huge" improvement, says Stephanie Chissoe, who manages sequencing teams at the Washington University genome center. She calls them "the single most important advance in finishing."

But the computer programs only go so far, sometimes leaving several stretches of sequence unconnected. When that happens, "somebody really skilled takes over," says Barrell, adding, "everything becomes a little problem, and you have to solve it." And closing the last gap can bog down the whole process. "You can sequence 100,000 base pairs in a day; however, putting it all together with no gaps can take 2 months or more," says Rhonda Brandon, a sequencer at The Institute for Genomic Research (TIGR) in Rockville, Maryland.

Sometimes it means starting over and resequencing some DNA. The labs at Washington University and the Sanger Centre use a program called FINISH that attempts to fill in the gaps and calls for more data if needed. FINISH may identify the area of DNA that needs to be redone, "but then [someone] carries out the reaction by hand," says Sanger's Jane Rogers.

Often, in areas rich with repeated bases, the DNA may bind to itself and kink in a loop that cannot be reached by base-identifying dyes, making it unreadable. Drastic measures may be required to unravel it. Sometimes new chemicals are used to tag the bases or make them accessible to the dyes, and other times new "primers"—short stretches of DNA that match either side of the gap—are deployed to hook onto the ends of the unreadable area and open it like a book.

Using special primers is expensive, however. To lower costs, Skip Garner in Glen Evans's group at the University of Texas Southwestern Medical Center at Dallas has built a machine that makes these primers, or oligonucleotides, in-house. Called MerMade, the machine is tied into the computer analyzing the DNA data. Once PHRAP has finished assembling the sequences, another program called PRIMO looks for gaps, decides what primers need to be made, and instructs MerMade to make them. With this machine, "we believe we can finish [an assembled stretch of DNA] 10 times faster," Evans says. Bruce Roe, who heads the genome center at the University of Oklahoma, Norman, says MerMade "can make 200 [primers] a day, for $2 a piece," in contrast to the retail price of $20 per primer.

Sometimes all the familiar tricks fail, and "we just sit there beating our heads against the wall," says Brandon. The finisher may have to leave aside a piece of DNA until some puzzle-solving technique—such as a new dye—comes along. Each innovation helps a little, and some can make a big difference. Take the morale boost Waterston's group got last fall from a new enzyme for treating DNA. Waterston's finishers had been stymied for more than 2 years by one gap in a sequence that they had been unable to fill. On one side of the gap was a string of about a dozen cytosine bases, and on the other was a string of guanine bases. The guanines and cytosines acted like magnets, sticking to each other and trapping 10 bases in between in an unreadable hairpin loop. "We had tried everything in our bag of tricks," Chissoe recalls. Anytime a company sent a new enzyme, Chissoe's colleague Elaine Mardis would see if it helped resolve this problem. In September, one called SequiTherm Excel II from Epicentre Technology prevented the kinking. The 2-year-old puzzle was solved.

It was the kind of triumph that only a finisher can appreciate fully. "Depending on how you look at [this work], it's frustrating or challenging," says Chissoe. She agrees with Barrell: "It's an art."

–E.P.

can increase the tracking problems that occur when a sequencing machine shifts from one sample to another, producing faulty data. And when Lander's group moved into a bigger space and doubled the number of automated sequencing machines, output took a nose dive. "Nothing worked," he recalls. Moreover, it wasn't just teething troubles. Lander has found that robots simply don't work well enough on the gap-filling "finishing" stage, which still requires the intuition and interpretive abilities of an experienced person (see p. 816). Lander now has eight people devoted to that task. Adams, too, says he didn't see a big jump in output with new machines. But in TIGR's case, that may be because the lab was going through a retraining exercise. Stanford had taken a different tack, using a "directed sequencing" strategy, which was meant to reduce the number of sequencing reactions required. But the resource-saving methods never got a real test, because the lab had trouble getting reliable raw sequence data.

All the centers are given a free hand to follow their own paths, but they must abide by a few rules. The most important is that finished DNA data must contain no more than one error in every 10,000 bases—an accuracy rate of 99.99%. This standard at first proved hard to meet, as several centers discovered after exchanging materials last year in an informal test.

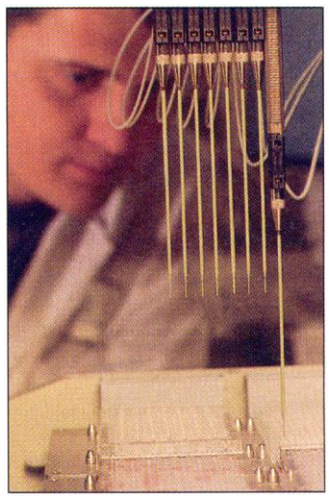
Some teams realized early on that they were having trouble, and, as a result, "ratcheting up was slower than we hoped it would be," says Gibbs. Stanford researchers had pinned their hopes on a strategy that required extra work deciding in advance which pieces of DNA to sequence, but might have saved time down the line. But when they implemented it, they ran into snags and have little finished sequence to show for 2 years' work. Bruce Roe of the University of Oklahoma, Norman, asked his group to go back and redo 3 megabases of data to ensure their quality. Both Stanford and Oklahoma have now changed their requirements to monitor for problems earlier in the process. "If you don't check [quality] at every step of the way, you're sure to have glitches," says Rick Myers of Stanford.

The challenge, of course, is to maintain high standards without sending costs through the roof. And people are trying all kinds of money-saving tricks. Consider the dyes used to tag the bases in DNA so that robot sequencers can identify them. One lab is trying to save money by making its own dyes. Details like these may seem trivial, but they can add up over the 7 years it will take to finish the genome. Small differences in the way duties are distributed also can add up to big disparities over time.

Indeed, despite the convergence in technologies Olson sees, the leaders of each of these centers are betting that the seemingly subtle differences will let their operations scale up more efficiently. "It's all in the details," says Adams. And the details suggest that they still have a lot to learn (see sidebar on p. 815). "We're realizing that you can't do it the way you make M&M's [candies] or cars," says Roe. Two years' experience, coupled with the daunting size of the unfinished task, has fostered an unprecedented degree of collaboration, he adds. "There's competition, but I feel there's room for everybody." Others, like Lander and Gibbs, worry that there may not be enough players in the sequencing game. The real question, Lander suggests, isn't how many centers will have to drop out but rather "will there be enough sequencing capacity?"

The human dimension

As NHGRI pushes to expand the production rate, some centers are experimenting with different—and they hope more efficient—approaches to organizing their personnel. At first, both the Sanger Centre and Washington University stuck with a traditional style, one in which small teams are given stretches of DNA 150,000 bases long to work on from start to finish. Members of each team clone short lengths of sequence into bacterial vectors, feed these into automated sequencers, and reassemble the finished data for release on the Internet. "You have contact with the project from beginning to end," says Sanger's Jane Rogers. Some routine tasks are centralized, such as preparing reagents.



S. KITTNER

Steady hand. This device adds reagent to DNA in precise quantities.

Recently, however, the St. Louis group has begun to organize researchers according to tasks, aiming for greater efficiency. For example, everyone involved in shotgun sequencing has been put into three teams, while data finishers have been put into six others, and the stretches of DNA are passed off from sequencers to finishers.

Some genome centers, such as the one at MIT, favored specialized teams from the start, moving deliberately toward a manufacturing approach. At MIT, each person specializes in one task or group of tasks and continues with that assignment indefinitely—day in and day out. That's Stanford's approach as well. "It ends up like an assembly line," says Myers. "You get a lot higher efficiency."

Adams also thought specialization was better at first, but has now changed his mind. A year ago, managers told Adams that it was hard to keep people motivated when they had to do the same job over and over. TIGR tried to break up the tedium by introducing more freedom. Now, everyone learns all aspects of the process and switches between mapping, sequence production, and finishing. "Compartmentalizing it can lead to very burnt-out people," notes TIGR's Rhonda Brandon. Introducing flexibility helps, but then the key is to keep track of what's being done, "so you don't do something twice," Brandon adds. Because the chances of duplication of effort increase as the number of people expands, Adams hopes to scale up without increasing staff. "I don't want a group of 100 people," Adams says. Instead, he hopes to automate more of the routine tasks and boost the efficiency of hardware, for example, by using sequencers that process 96 samples at a time instead of 48 or 64.

Over the next few years, any center that continues to participate in this giant biological experiment will have to face—and overcome—more setbacks like the ones MIT, Stanford, TIGR, and others have already confronted. "It's clear we have a big job ahead of us," says Myers. But he also points out that, bit by bit, costs have decreased and sequence output has increased.

Recent monthly output exceeded 2 million bases for both TIGR and Baylor; that's equal to about a quarter of their production for the past 2 years. Stanford has made big strides, finishing about 0.4 megabase of data in the past month, which brings its output to more than a megabase. These centers take heart from the successes of both the Sanger Centre and Washington University in scaling up to more than 20 megabases a year. "We feel we're fairly close in reaching our goals," says Washington University's Chissoe.

At the same time, as each new megabase of DNA is released, interest in genome science grows. "Many people call this the most important scientific undertaking in our time, perhaps in all time," says NHGRI's director, Francis Collins. NHGRI and the sequencers expect that the demand for sequenced genomes will continue to grow, even after the first human genome is finished. These centers will be needed to fulfill that demand. Genome sequencing has finally arrived, they claim, and it is establishing itself as an essential part of the scientific infrastructure. Says Myers: "A lot of us are in for the long haul."

–Elizabeth Pennisi