Getting the Bugs Worked Out

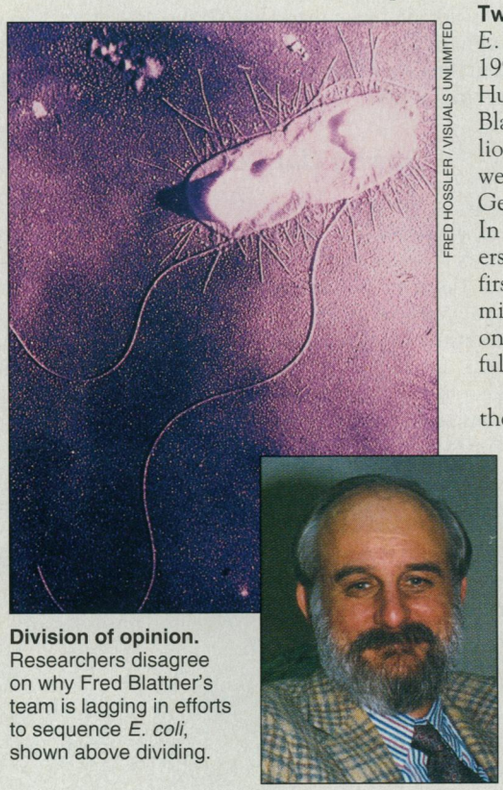
A couple of years ago, it looked as though *Escherichia coli* would be the first free-living organism to have its genome fully sequenced. Not now. What happened? That's a controversial question

When researchers set out to sequence the genome of the bacterium *Escherichia coli* three-and-a-half years ago, it was viewed by many as a prime opportunity to test the strategies and equipment that would be used for future large-scale sequencing projects. Indeed, although sequencing *E. coli* was certainly a major scientific project in its own right, to human geneticists it was also akin to a climb to the base camp before mounting a major assault on the Everest of genetics: sequencing all 3 billion base pairs in the human genome. And the *E. coli* project has turned out to be a long hard slog that holds some key lessons for further climbs in the Himalayas of genome research.

Since at least 1990, insiders anticipated that *E. coli*, the white rat of microbiology, would be the first free-living organism to have its genome completely sequenced. It seemed a safe assumption. For starters, a quarter of *E. coli*'s genes had already been sequenced before the genome project got under way. In addition, *E. coli* was the first and smallest of the "model" organisms to win funding from the Human Genome Project for a concentrated sequencing effort, in part because the bacterium's genome is expected to help solve scientific puzzles in subjects ranging from evolution to cancer.

Yet the way things have turned out, the one-time favorite is unlikely to be the first to cross the finish line. The *E. coli* project has missed its early deadlines by several years, and the honor of being the first fully sequenced free-living organism (the tiny genomes of several viruses have already been sequenced) is expected to go to a less well-known bacterium such as *Haemophilus influenzae*. "[*E. coli*] is clearly not going to be first, and at the rate it's going it may not even be in the top 10 genomes," says Craig Venter, director of The Institute for Genomic Research (TIGR) in Gaithersburg, Maryland. Because results of the *E. coli* project will be important in many fields of biological research, there is "an enormous amount of frustration in the community extending over years that *E. coli* hasn't been done," says Chris Fields, scientific director of the National Center for Genome Resources in Santa Fe, New Mexico.

The reasons why the *E. coli* project has fallen behind schedule are a matter of debate in the genome community. Some critics of the *E. coli* project, which is led by Frederick Blattner of the University of Wisconsin, say the project is irretrievably shackled by an early commitment to a technology that quickly became outmoded. If that analysis is correct, it has implications for the Herculean task of sequencing the human genome: Human genome sequencers have long worried that their best efforts will become unstuck if full-scale sequencing is launched before the technology is ready. Others, however, defend the *E. coli* project, saying that the early deadlines were unrealistic and that the project is now churning out high-quality sequences at an impressive rate. In Blattner's view, during the early stages of the project he and others were guilty only of an "honest optimism" that now seems "astonishing," considering that "we were trying to sequence almost 5 million base pairs, something that had never been done before."



FRED HOSSLER / VISUALS UNLIMITED

Division of opinion. Researchers disagree on why Fred Blattner's team is lagging in efforts to sequence *E. coli*, shown above dividing.

This division of opinion isn't just a matter of hallway gossip among molecular biologists; it could have a considerable influence on how the rest of the *E. coli* project is carried out. The Blattner team's NIH grant is about to run out with only about 40% of the work done, and the question of who should complete the *E. coli* sequence is up in the air. The project could remain with Blattner, go to private industry, or to other academic centers. For the moment, the only thing genome researchers agree on, says Robert Strausberg, chief of sequencing technology at the National Center for Human Genome Research (NCHGR) in Bethesda, Maryland, is that "the community is anxious to have *E. coli* sequenced as quickly as possible."

Two years and out

E. coli was deemed an official priority in the 1990 plan for the first 5 years of the U.S. Human Genome Project (HGP). A year later, Blattner's group received a 4-year, $7.8 million NCHGR grant ($2 million of which went towards "indirect" costs), becoming a Genome Science and Technology Center. In those heady early days, genome researchers thought the job could be done within the first 4-year grant. Many, in fact, were optimistic that the 4.7 megabase (a megabase is one million bases) *E. coli* genome could be fully sequenced in as little as 2 years.

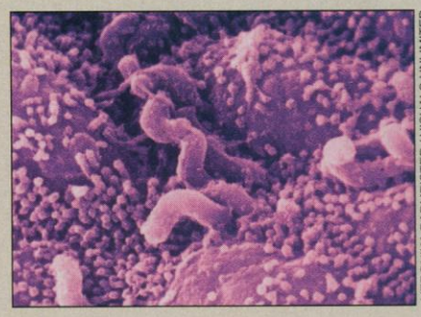
But early progress was far slower than those rosy initial estimates. And the downward spiral of expectations was reflected in the Human Genome Project's updated 5-year plan, drawn up in 1993 (*Science*, 1 October 1993, p. 43), which anticipated finishing *E. coli* by 1998. Meanwhile, the Blattner team had set itself initial goals of sequencing 1 megabase per year. If the team had managed to maintain that pace, by now 4 megabases—more than 80% of the *E. coli* genome—would be finished. In fact, Blattner and his co-workers have sequenced only 1.4 megabases. George Church of Harvard University has sequenced another 0.1 megabase, and a Japanese effort coordinated by Kiyoshi Mizobuchi of the University of Tokyo and Katsumi Isono of Kobe University has sequenced 0.19 megabase (see diagram on p. 174).

Another 1.6 megabases have been sequenced piecemeal over the past few decades by microbial geneticists interested in one or another of the 4000 genes strung out along *E. coli*'s single chromosome. But those sequences contain numerous errors and represent myriad different *E. coli* strains. As a result, many genome sequencers, including Blattner, believe that, at the very least, the

The Gold Bug: *Helicobacter pylori*

Last year, the race to sequence the first complete genome of a free-living organism entered the final straightaway. And although it had long been assumed that the *Escherichia coli* project would cross the finish line first (see main text), as the competing research teams headed home, knowledgeable railbirds began to think the winner would not be *E. coli* but another bacterium: *Haemophilus influenzae*. There were good reasons for the change. For a start, *H. influenzae*'s genome is less than half the size of *E. coli*'s; more important, *H. influenzae* is being sequenced by scientists at The Institute for Genomic Research (TIGR) in Gaithersburg, Maryland, an institution renowned for its bullish approach.

Then, on 9 December, a dark horse appeared out of nowhere and seemed to have crossed the finish line first. Genome Therapeutics Corp., a Waltham, Massachusetts, company, announced in a press release that it had sequenced the genome of *Helicobacter pylori*, the bacterium that causes most peptic ulcers. A letter accompanying the press release, which the letter says was sent to a "few knowledgable writers," refers to the sequencing of *H. pylori* as a "milestone" that will help in the development of drugs, vaccines, and diagnostics for peptic ulcers; the press release claimed that the world drug market for peptic ulcers reached $8 billion in 1992. Following the press release, an account of the company's claim appeared in the *Wall Street Journal*, in *BioWorld*, a daily biotech newsletter, and in *Newsday*, but Gerald Vovis, GTC's vice president for research, says GTC has no intention of going through peer review and publishing the sequence or depositing the data in public data banks.



VERONIKA BARMEISTER / VISUALS UNLIMITED

Twisted tale. *Helicobacter pylori.*

It looked as though the *E. coli* and *H. influenzae* projects had been nipped at the finish line. But many in the sequencing community say the race isn't over yet. In the first place, what GTC calls sequencing the genome is not what geneticists generally consider finishing the job. GTC has sequenced enough random pieces of *H. pylori* DNA to cover the entire genome five times—but it has not yet lined the pieces up in the correct order. While the fragments can be used to identify genes, most geneticists consider the assembly of fragments the most difficult part of sequencing, as well as a prerequisite for announcing the complete sequence.

Furthermore, some scientists say they can't evaluate the claim if the data aren't put through the usual procedures of peer review and publication. "It's a meaningless announcement ... if they don't make the data available," says microbial geneticist Jeffrey H. Miller of the University of California, Los Angeles. TIGR's Craig Venter called GTC's press release and the accompanying letter "science by press conference" and an example "of the worse part of the commercialization" of science.

In defense of the company's decision to announce the findings in a press release rather than a peer-reviewed publication, Vovis points out that, although GTC receives National Institutes of Health (NIH) funding, the *H. pylori* project was funded entirely by private investments in the company. "It is not our goal to use shareholder's money to enrich pharmaceutical companies," he says. And that's just what would happen if the sequences were made public, says Vovis, because, according to current patent law, until there are experimental data proving the usefulness of each *H. pylori* sequence—a process that could take years—it's not clear that GTC could win patent protection. One way to safeguard GTC shareholders' investments is to treat the *H. pylori* sequence as a trade secret, says Vovis.

In an ironic twist—given Venter's outrage—GTC is considering following a controversial practice made famous by TIGR and its sister company Human Genome Sciences Inc. (*Science*, 16 December 1994, p. 1800). In that strategy researchers are given access to gene sequences—but only if they sign an agreement that protects GTC's commercial interests.

To many geneticists, GTC seems to be trying to have it both ways. "A company is certainly entitled to keep its data [secret]," says Kenneth Rudd of NIH's National Center for Biotechnology Information. "But there's a little bit of wanting your cake and eating it too. Wanting credit for a scientific discovery and wanting proprietary protection ... if they are going to keep it proprietary they ought to keep it to themselves."

–R.N.

early sequences must be checked for errors and, in a large number of cases, resequenced. When resequencing is included in the calculation, at least 60% of the *E. coli* genome remains to be done.

Clinging to the old ways?

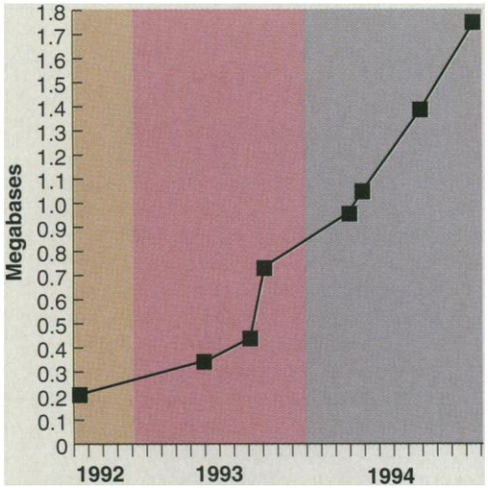
Blattner's critics argue that this failure to meet the targets is due to his group's methods. "[They] are essentially using yesterday's technology," says genome scientist Robert Weiss of the University of Utah in Salt Lake City. Blattner "has accomplished a tremendous amount—considering the way he's doing it. ... It's being done by the clone-by-clone approach, and that is very tedious," adds Venter, who admits his focus on Blattner's progress is not strictly disinterested, as TIGR has considered tackling *E. coli* itself.

The Blattner team concedes that its techniques now appear old-fashioned, but says that when they set out to catalog *E. coli*'s sequence, those techniques were at the cutting edge of sequencing technology. Furthermore, the early stages of the project did involve some arduous tasks. Before they could even start sequencing, Blattner and his colleagues had to finish the task of breaking the *E. coli* genome into an ordered set of about 400 overlapping λ clones, or DNA segments. That task was completed in 1991, and since then Blattner's group has concentrated on breaking each clone into random subclones of unknown order and then determining the sequence of bases in each subclone. The process is called "shotgun cloning."

Once the subclones' sequences are known, it's possible to put them in order again by looking for overlapping sequences. The researchers then have the almost-complete sequence of the original λ clone in hand, except for a few gaps where the subclones do not overlap, which are then filled in.

The whole process is so labor-intensive, says Blattner, that at a rough estimate probably only 50,000 to 100,000 bases a year could be completed if the job were done entirely by hand. By designing its own robotics and software, however, the Blattner team says it has speeded up the process 10-fold and has now reached its target rate of 1 megabase a year.

Critics, however, contend that Blattner's group by now should have turned to the state of the art in rapid sequencing technology: automated sequencing machines, or ABIs (for Applied Biosystems Inc.—now part of Perkin Elmer Inc.—the company that manu-



On the up and up. The rate at which *E.coli* DNA sequences are being released has risen rapidly.

factures the machines). These machines speed up sequencing by computerizing collection and processing of the data. They also use fluorescent labeling, which allows four sequencing reactions to be run on each of the 36 lanes of gel. In the most advanced labs, each ABI can sequence 30,000 raw bases a day, translating into about 3000 bases of finished sequence.

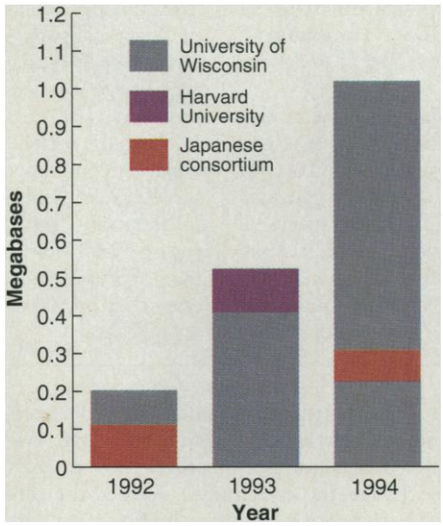
Although Blattner's critics say he's technology-shy, he argues that just the opposite is true: His group, he says, is the victim of its own technical trailblazing. When he started off, he says, sequencing machines were considered too expensive and too slow to be worthwhile. Now, he says, "the technology has evolved, and [our critics] are saying 'why aren't you using machines?' " NCHGR Director Francis Collins agrees with Blattner. "[*E. coli*] had the misfortune of being the one that everyone said we should start on even before the technology was quite up to the task," he says. Blattner's group has recently installed an ABI machine, but they cannot afford to install more because they have already spent all their equipment money souping up the old technology. Blattner and his colleagues have, however, applied to have their NCHGR grant renewed, and they plan to use it to buy two of the most up-to-date ABI machines.

But even with its older technology, Blattner's outfit now has a full head of steam, according to some experts in sequencing. Kenneth Rudd of NIH's National Center for Biotechnology Information, who catalogs the sequence data once it has been deposited in GenBank, says Blattner is "on quite a roll now. He's depositing data [in GenBank] as fast as I can analyze it." Church feels the Blattner team's reputation has been unfairly tarnished. "They do not deserve the reputation of being behind deadline," says Church. "There were a lot of 'if' and 'then' clauses in the [early] optimistic opinions that [*E. coli*] could be done in one year or two."

But even if Blattner's team has now gotten up to the speed it initially expected of itself, it is now being outpaced by other efforts. The TIGR team claims it has sequenced 99% of the 1.9-megabase *H. influenzae* (which causes ear infections, not flu) in less than 6 months; they say that they expect to publish the final and complete sequence and deposit it in GenBank early this year. That result will have been achieved, in part, through the use of special software that Venter and his colleagues have developed to shotgun the entire bacterial genome at once, without first creating a detailed map of the genome and breaking it into clones.

Researchers at Genome Therapeutics Corp. (GTC) in Waltham, Massachusetts, have also been pushing the state of the art. Last month, the company announced that it has used "multiplexing," a technique developed by Church, that allows as many as 40 samples of DNA to be analyzed on each gel lane, to sequence the complete genome of the ulcer-causing bacterium *Helicobacter pylori*. Some researchers are questioning that claim, however (see box on p. 173).

And the Department of Energy (DOE) is funding three teams to make use of whole-genome shotgun sequencing and multiplexing to sequence *Pyrococcus furiosus*, *Methanococcus jannashii*, and *Methanobacterium thermoautotrophicum*. Besides answering some of the fundamental questions the *E. coli* genome was expected to address, the full sequences of *P. furiosus*, *M. jannashii*, and *M. thermoautotrophicum*—all of which live at extremely high temperatures and some of which convert waste products into methane—should help identify industrially important enzymes (*Science*, 22 July 1994, p. 471), says DOE's Jay Grimes. Those complete genomes, which range from 1 to 2 megabases, are expected within a year and a half, says Grimes, who says DOE funded the work in part because other bacterial sequencing efforts were lagging. DOE, he says, wanted to "get microbial sequencing up and running so that complete sequences of organisms become available."



Three-way split. Three groups have deposited *E. coli* sequences in databases.

But although *E. coli* will not be the first completely sequenced genome, it could still be among the most valuable. The *E. coli* sequence is "in a different class in terms of how well annotated it is," says Church. Blattner's approach has been to plow through the *E. coli* genome, meticulously correcting errors in other peoples' sequence data, closing gaps, identifying genes and regulatory regions such as promoters, correlating the genes with known biological function, and publishing detailed accounts of the data.

Venter agrees that Blattner has done "a superb job" of annotating the *E. coli* DNA sequences, but argues that more rapid sequencing and good annotation are not mutually exclusive. TIGR expects to have both sequenced and annotated the 1.9-megabase *H. influenzae* within 1 year, he points out.

As the 4-year grant to the Blattner group winds down, those who aren't completely satisfied with the job his group has done are talking openly about the possibility that the project be finished up by others. Weiss, who leads the team that is being funded by DOE to sequence *P. furiosus*, says the key question is simply: "Who's got the capacity now for putting out sequence?" Weiss argues that, with its banks of ABI machines, "TIGR has had the most success" at high-speed sequencing.

Because NCHGR is currently considering Blattner's grant application, Collins says he cannot comment on Blattner's progress or his chances of being chosen to complete *E. coli*. But he adds that "in the hypothetical situation that Blattner is not funded to finish [*E. coli*]," one of NIH's options would be to issue a "Request For Applications ... and see who can do it the most quickly and the most cheaply." Such an invitation, he says, would be open to private-sector institutions like TIGR.

An alternative, says Collins, would be for *E. coli* sequencing to come under the auspices of one of the other Genome Science and Technology Centers, such as the one led by Robert Waterston of Washington University in St. Louis that last year used its 14 ABI machines to sequence 5 megabases of *Caenorhabditis elegans* and 1 megabase of the yeast *Saccharomyces cerevisiae*. What is certain, says Collins, is that the project "needs to happen. The whole biological community would be badly served if *E. coli* is not sequenced." And, for the moment, that seems to be the only point on which all genome researchers agree.

–Rachel Nowak

SOURCE FOR GRAPHS: KENNETH RUDD AND FRED BLATTNER