Laboratory Workhorse Decoded

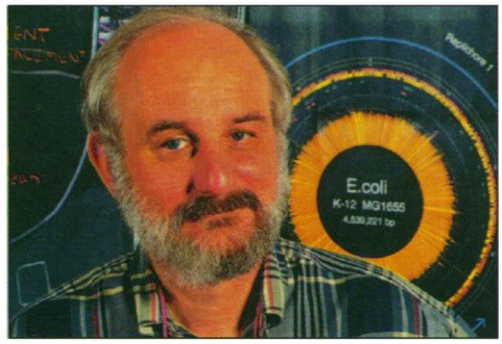
The genome of *Escherichia coli*, a research favorite and common pathogen, has finally been finished—twice—allowing researchers to tour the organism's molecular workings

It's rare for *Science* to devote 10 pages to a single scientific achievement. It's even more extraordinary for the journal's editors to accept an article presenting data that had been released months or even years earlier. But when the article marks the end of a 15-year race to determine the sequence of all 4.6 million base pairs in the genome of the bacterium *Escherichia coli*, exceptions can be made.

The *E. coli* genome, described on page 1453 by geneticist Frederick Blattner of the University of Wisconsin, Madison, and his colleagues, is neither the first nor the largest to be sequenced. But because of the rich history of the organism itself, it stands out among the dozen or so genomes now in the public record. *E. coli* "is the organism we have the most information about genetically and metabolically," says Francis Collins, director of the National Human Genome Research Institute (NHGRI) in Bethesda, Maryland.

For decades, microbiologists, biochemists, geneticists, and cell biologists have studied this easy-to-grow organism. More recently, a pathogenic strain in tainted meat has made *E. coli* a household word. Now, researchers have the full set of genetic instructions underlying the organism's structure and function. Already, partial *E. coli* sequences that Blattner and, independently, a Japanese team have been depositing in the public database have given researchers a taste of how valuable such data can be. Matches between new genes from other species, including mammals, and *E. coli* sequences previously filed have often provided researchers with a name and function for their discoveries.

Now that the genome is completed, there are more genes, known and unknown, to work with. *E. coli* has 4288 genes, about 40% of which are complete mysteries. With this genomic overview, researchers can begin to form a coherent picture encompassing all of *E. coli*'s biology. "Having this complete set of instructions gets us one step closer to understanding how a free organism functions," Collins points out. Moreover, comparing this genome to other microbial genomes that are rapidly being finished (see sidebar) should help explain how these organisms have evolved.



B. WOLFGANG HOFFMANN/UNIV. OF WISCONSIN

Genome pioneer. Fred Blattner first thought of sequencing the *E. coli* genome almost 15 years ago.

Long haul

Blattner first thought about sequencing the *E. coli* genome in 1983 after tallying the genes that had already been sequenced. Together they totaled 2.3 million base pairs—half the length of *E. coli*'s single chromosome. "I sort of convinced myself that [sequencing *E. coli*] was a feasible idea" and proposed doing it in a *Science* editorial, he recalls.

E. coli was the obvious choice for a sequencing effort, says Frederick Neidhardt, a microbiologist at the University of Michigan, Ann Arbor. "By the late '50s and '60s, so much information had been learned about metabolism by using *E. coli* that it meant there was an advantage for all kinds of people to study it," he explains. "Your work had a better chance of being interpretable." Figuring out the microbe's genetic code would help integrate all those years of study, Blattner reasoned.

D. M. PHILLIPS/VISUALS UNLIMITED

A 100% solution. A full view of *E. coli* biology is expected.

Other efforts, including those that produced the complete sequences of the genomes of the yeast *Saccharomyces cerevisiae* and the bacterium *Bacillus subtilis*, have involved multiple labs on two or more continents. But Blattner decided to take on the *E. coli* genome on his own, a move that ultimately led to a race between his team and Japanese sequencers.

As a first step toward that goal, his lab began breaking the *E. coli* chromosome into small pieces that could be used to build a physical map of genetic landmarks along the DNA. He was about 85% done when a Japanese group beat him to the punch, however. In 1987, Yuji Kohara of Nagoya University and his colleagues published its complete set of clones, along with a detailed physical map. Disappointed, Blattner quickly finished off his own map, determined to beat the Japanese in sequencing the genome. The race had begun.

But the next few years were rough on both sides of the Pacific. "The *E. coli* genome project languished as much in Japan as it did in the United States, [held up] in part by the funding and in part by the technology," notes Craig Venter, president of The Institute for Genomic Research (TIGR) in Rockville, Maryland. At the time, automated sequencing was in its infancy, and the existing methods meant sequencing was as much an art as a tested technology. "The equipment was so primitive that the speed was slower and the accuracy was lower than now," says Takashi Horiuchi of the National Institute for Basic Biology in Okazaki, Japan.

Indeed, the first Japanese sequencing attempt resulted in 186 kilobases, but so much was inaccurate that eventually both the Japanese team and Blattner's would resequence that section. A second attempt by a different team did little better, and not until 1995 did Horiuchi and his colleagues come up with an effective sequencing strategy. By having each participating laboratory do a different step, "we improved our efficiency," says Hirotada Mori, a molecular biologist at the Nara Institute of Science and Technology who compiled the incoming sequence data. That enabled the Japanese team to complete 2.6 megabases within a year.

Meanwhile, Blattner was having problems of his own. In 1990, he had received one of the first grants awarded by NHGRI's predecessor, the Human Genome Center. By the end of 1994, his team had completed 1.4 million bases—a significant chunk, but far short of the full genome he had promised to complete by then. As a result, his grant wasn't renewed (*Science*, 13 January 1995, p. 172). Only after the *E. coli* community protested (*Science*, 31 March 1995, p. 1899) and

Microbial Genomes Come Tumbling In

"It's an avalanche," says Daniel Drell, project manager for microbial genomes at the U.S. Department of Energy in Germantown, Maryland. Barely 26 months ago, researchers at The Institute for Genomic Research (TIGR) in Rockville, Maryland, submitted the first complete genomic sequence of a bacterium, *Haemophilus influenzae*, to GenBank, a public database run by the National Library of Medicine. Since then, 10 more bacterial genomes, plus that of the eukaryote yeast, have been completed, and at least a half dozen more should be finished by year's end. And the avalanche is just beginning. All told, some 50 microbial genome projects are under way, with more in the planning stages.

These sequencing efforts are already changing the way microbiologists do their work. By comparing the genes of these many microbes, researchers can gain insights into how they evolved and which genes are likely to be essential for particular functions. "Having genomes to compare makes it easier to draw conclusions about how [a microbe] lives," explains Anthony Kerlavage, a computational biologist at TIGR. That information in turn can help pharmaceutical companies search for new antimicrobial medicines.

Indeed, the surge in interest in microbial genomes stems from the realization of just how much they have to offer. In the past, "the human genome was seen as the big prize," explains sequencer Bart Barrell of the Sanger Centre in Hinxton, United Kingdom. "But once TIGR had sequenced *Haemophilus*, people realized it was possible and that immense amounts of information could be gotten quickly [from microbial genomes]."

Some of the sequencing efforts have been aimed at laboratory workhorses, such as *Escherichia coli*, whose genome appears in this issue (see main text), and *Bacillus subtilis*, the complete genome of which was reported at a July meeting. Because these organisms have already been so well studied, knowledge of their complete genome sequences will be particularly useful to researchers trying to link genes to function. The other genomes finished to date provide a diverse representation of the microbial world, including archaea that live in extreme environments. Together these sequences will enable researchers to sort out evolutionary relationships.

But the potential usefulness of their genome sequences has won pathogens the lion's share of attention. *Haemophilus* causes ear infections and meningitis, for example. And just last month, TIGR's Jean-François Tomb and colleagues described the genome of *Helicobacter pylori*, a bacterium responsible for peptic ulcers and linked to stomach cancer. (The results appeared in the 7 August issue of *Nature*.) Because *H. pylori* is specialized for living in the stomach's acid environment, "the standard [antibiotic] treatments don't work too well," notes molecular geneticist Antoine Danchin of the Pasteur Institute in France. Consequently, he says, having its genome is critical for finding new surface proteins that might be vaccine candidates.

Also completed recently is the genome of *Borrelia burgdorferi*, the spirochete responsible for Lyme disease, which TIGR's Claire Fraser put up on the TIGR World Wide Web site on 18 July. And the list will not stop there. This year alone, TIGR expects to sequence as much microbial DNA as is in the entire *E. coli* genome, which took 6 years to finish. That DNA includes partial genomes of *Mycobacterium tuberculosis*, which causes tuberculosis, of the malarial parasite *Plasmodium falciparum*, and of *Salmonella* and other important human pathogens.

The Sanger Centre, meanwhile, expects to finish the sequence of a laboratory strain of *M. tuberculosis* by the end of the year, and it is also sequencing the pathogenic bacteria that cause leprosy and meningitis. In addition, Sanger just expanded a pilot project aimed at sequencing the malarial parasite to a full-fledged, $12.8 million effort, in which the center will complete 15 megabases—one-half the malarial genome—in the next 3 years, says Barrell. And just last week, the U.K. Biotechnology and Biological Sciences Research Council announced that it will support the sequencing by Sanger of the genome of a *Streptomyces* bacterium, which is not a pathogen but an important source of antibiotics. "The goal is to establish a pathogen-sequencing facility that would be able to sequence 25 megabases a year," says Barrell.

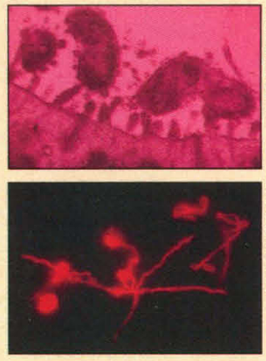
That's good news for microbiologists wanting to compare genomes. "Anything we can find out about one genome starts feeding back to the other organisms," says Guy Plunket III, a bacteriologist at the University of Wisconsin, Madison. Soon, for example, he and his colleagues will be able to compare the virulent mycobacterium strain being sequenced by TIGR with the lab strain being sequenced by Sanger—to find out what makes one so much more deadly.

TIGR's Fraser can't wait to have the genomes of several human pathogens in hand to look for unusual genes common to them all. "If we get a big enough set, we can begin to see if there are sets of genes responsible for the [pathogens'] specificity for the human host," she notes. The protein products of such genes might make particularly good vaccine or drug targets.

All of which puts a heavy burden on bioinformatics specialists who are developing the tools for making useful comparisons. So far, says Drell, "nobody is equipped to handle all the knowledge that's coming out of these sequencing projects." But Kerlavage says he welcomes the challenge, because "we're not working in a vacuum anymore." **–E.P.**

COMPLETED GENOMES

Organism	Genome size (megabases)	Tentative genes	Identified genes Number	%
Saccharomyces cerevisiae	12.1	6034	3089	63%
Escherichia coli	4.6	4288	2656	62%
Bacillus subtilus	4.2	~4000	~2320	~58%
Synechocystis sp.	3.6	3168	1402	48%
Archaeoglobus fulgidus	2.2	2471	1193	44%
Haemophilus influenzae	1.8	1740	1015	58%
Methanobacterium thermoautotrophicum	1.8	1855	816	44%
Helicobacter pylori	1.7	1590	907	57%
Methanococcus jannaschii	1.7	1692	776	46%
Borrelia burgdorferi	1.3	863	499	58%
Mycoplasma pneumoniae	0.8	677	333	49%
Mycoplasma genitalium	0.6	470	324	69%



Sequenced pathogens. *H. pylori (top)* causes ulcers, and *B. burgdorferi (bottom)* causes Lyme disease.

he reapplied and won new funding was he able to proceed.

Blattner was unaware of the progress the Japanese were making during this time until a September 1996 meeting in California, when both his team and theirs reported that they were about to make a final push to complete their work. The news precipitated "a hectic mad dash," recalls Guy Plunkett III, who had been with the Blattner team almost from the beginning. "To not even be the first to finish *E. coli* would have been heartbreaking." They did finish first, but not by much.

On 16 January, Blattner's group made its final deposit of sequence into GenBank, having completed 2.6 megabases in the previous 12 months. The Japanese team submitted its fully assembled genome on 23 January. It was the seventh genome to be made public and the largest to date for a prokaryote.

Although both genomes come from the same strain, K-12, of *E. coli*, they are not quite identical because the Japanese and U.S. groups used two different versions, or isolates. For example, there's a 496-kilobase-long stretch that's inverted in the Japanese isolate. Also, the completed genome from Japan includes sequences submitted by other groups, including Blattner's, and therefore it is a mosaic of isolates. Blattner, known for his meticulousness, had insisted that his team do the whole genome themselves, even to the point of redoing *E. coli* DNA already represented in GenBank. "What we wanted to present was a sequence from one particular isolate," says Plunkett. In this way, they were assured that all the genes in this genome actually work together to create a functioning organism.

NIBB

Speedy sequencers. This Japanese team, led by Hirotada Mori (second from lower left) and Takashi Horiuchi (to Mori's right), completed 2.6 megabases of *E. coli* sequence in a year.

Genomic revelations

By providing a panorama of the microbe's genome, the sequence is allowing researchers to spot new details. Indeed, describing those features became an important component of his effort, says Blattner. With computer programs, the sequencers first identified the main landscape feature—the open reading frames, which are long stretches of bases that code for amino acids and likely represent genes. This revealed that *E. coli* has some 4288 potential genes, encompassing about 88% of its DNA. A search through GenBank showed that 1827 of *E. coli*'s genes had already been characterized.

To piece together the identities of the rest, Blattner, Monica Riley of the Marine Biological Laboratory in Woods Hole, Massachusetts, Julio Collado-Vides of the University of Mexico in Cuernavaca, and their colleagues compared the *E. coli* sequences to those of known genes from other organisms or identified genes based on their neighbors. For example, biochemists had pinned down just two of the six genes needed to make a set of enzymes that *E. coli* uses to break down aromatic compounds. Because the genes needed for a particular metabolic pathway in microbes are often located together, his team found the rest by looking next to one of those genes. Those types of analyses enabled the Wisconsin team to identify hundreds of additional genes. "We were discovering new things every day," says Blattner.

For each gene newly deciphered, however, there remains another with an unknown function—some 40% of the total. "It shows how little we know, even with something like *E. coli*, about the biology [of these organisms]," says Venter. Adds Plunkett: "There are all these things waiting to be discovered.

Global resource

Genome enthusiasts expect those discoveries to start pouring in, in large part because of the experiments made possible by knowing *E. coli*'s genetic code. "There are a lot of experiments you can do that you couldn't do before," says Richard Roberts, a molecular biologist at New England Biolabs in Boston.

By examining the genome sequence, for example, Kenneth Rudd, a bacterial geneticist at the University of Miami School of Medicine, can determine whether the small proteins that he finds in *E. coli* are true proteins with possible functions of their own, or merely the breakdown products of larger proteins. "Without the complete genome sequence, we would just be waving our hands," he notes.

Other researchers are using the genome to guide them to proteins crucial to *E. coli* metabolism. Microbiologist Tyrrell Conway of Ohio State University in Columbus says that searching for genes known to be involved in sugar acid metabolism now takes "a matter of days" instead of months, enabling him to focus his efforts on assessing the function of that gene's protein product. In addition, when examining a cluster of *E. coli* genes known to control sugar metabolism, he homed in on another sugar acid pathway. Among those genes, he noticed an extra gene that had no known function. It turned out to code for an enzyme that initiates an alternative pathway that researchers had never known about before. "[The genome] will be a tool to really accelerate our understanding of how this organism makes a living," he predicts.

By comparing the *E. coli* genes to those of other organisms, the Blattner group also spotted clues to how particular genes originated. In some cases, the results were expected: Genes for the proteins that make up the whiplike flagella that microbes use for locomotion are alike in both *E. coli* and its close relative, *Salmonella*, suggesting they evolved long ago in a common ancestor. Other comparisons turned up surprises. For example, *E. coli* turns out to have not one, but two sets of enzymes for breaking down aromatic compounds, and the genes for the newly found second set are almost identical to the equivalent set in the soil microbe *Pseudomonas*, which is not closely related to *E. coli*. The question remains whether the genes sneaked into *E. coli*'s genome somehow from that soil bacterium.

Databases and computer programs already being developed to store and analyze the *E. coli* data and compare its genome with those of other organisms will speed such efforts. Rudd, notes, however, that researchers will have to follow up on the computer analyses in the laboratory, to verify that a gene does what the analyses say it does. He also worries that the *E. coli* community is not as well positioned to complete this next phase of the work as are the biologists interested in *B. subtilis* and yeast. They learned how to cooperate and coordinate efforts while sequencing the genomes of their organisms and can now put this lesson to use as they work to identify all the proteins encoded by those genomes.

Because the *E. coli* genome was sequenced by two competing groups, no such infrastructure exists for completing the next step. "It shouldn't be that everyone is knocking out the [same] 50 most exciting genes" to find out what they do, Rudd says.

Efforts are under way, however, to improve communications among *E. coli* researchers, if not coordinate their efforts. The half dozen or so *E. coli*–related sites on the World Wide Web link to one another, and Blattner's *E. coli* Web page, once completed, will likely be a cyberspace meeting place for researchers studying the microbe. And even if a full-blown, coordinated attack on the inner workings of *E. coli* never develops, the genome will still hold its value, says TIGR's Venter. "It's scientific information that will be used for centuries."

–Elizabeth Pennisi