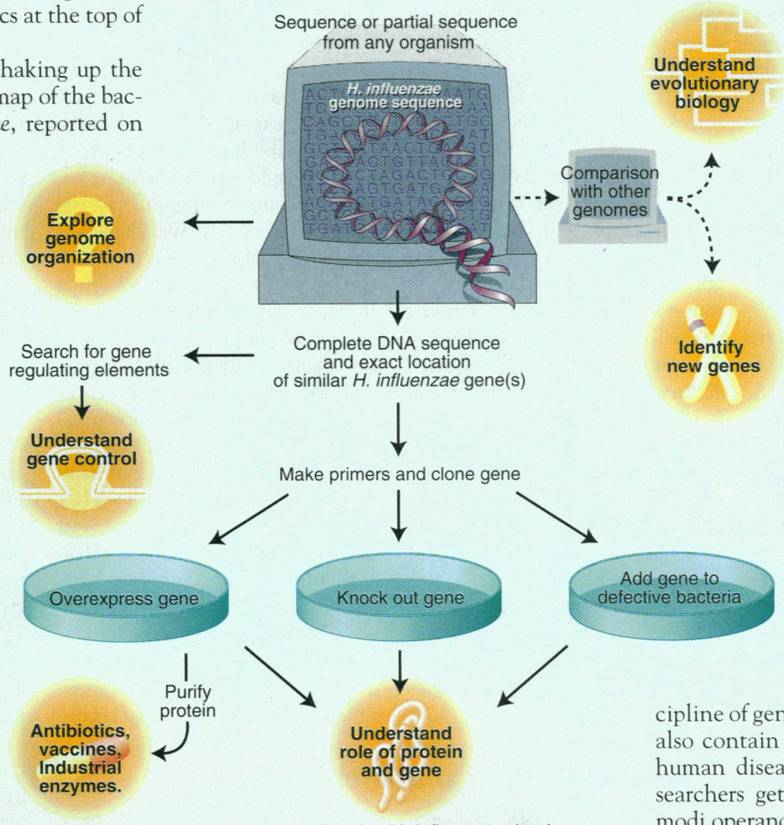
Bacterial Genome Sequence Bagged

The publication of the first complete genome of a free-living organism, the bacterium *H. influenzae*, opens the way to a wealth of fundamental and practical information

These days the human geneticists tend to attract all the glory. That's not surprising given their many high-profile discoveries of the human genes that underlie serious diseases such as cystic fibrosis and breast cancer. But if a report in this week's issue of *Science* has its predicted impact, microbial genetics is about to join human genetics at the top of the cool science hierarchy.

The achievement that's shaking up the field is the ultimate genome map of the bacterium *Haemophilus influenzae*, reported on page 496 by a team led by Craig Venter of The Institute for Genomic Research (TIGR) in Gaithersburg, Maryland, and Nobel laureate Hamilton Smith of Johns Hopkins University in Baltimore. The map consists of the complete 1.8-megabase genomic sequence of *H. influenzae* Rd—a benign laboratory strain of a bacterium that in its wild form can cause ear infections and meningitis—together with a catalog of the genes' locations and many of their functions. Although a handful of tiny viral genomes have already been sequenced, viruses can grow and reproduce only with the help of the cells they infect. In contrast, *H. influenzae* is the first free-living organism whose genetic blueprint has been determined. As a result, it will provide researchers with a guidebook to all the information needed to sustain life.



Versatile. The diagram illustrates that the *H. influenzae* database can be put to uses ranging from understanding evolutionary biology to working out the functions of specific genes. A link to the TIGR database of genome sequence data can be accessed from the *Science* home page (http://www.aaas.org/science/science.html) by clicking on "Beyond the Printed Page."

"It's a tremendous accomplishment," says genome scientist Robert Weiss of the University of Utah, Salt Lake City. "It's going to be just fascinating to see the whole genome and all that it encompasses." Frederick Blattner of the University of Wisconsin, Madison, who leads the effort to sequence the genome of the bacterium *Escherichia coli*, a one-time favorite to be the first fully sequenced free-living organism (*Science*, 13 January, p. 172), is just as enthusiastic. "I'm really thrilled. This is epic-making," he says.

For Venter himself, the recognition is especially sweet. His earlier work in building up a database of gene fragments was largely shunned by the genome community until it proved to be extremely valuable in tracking down disease genes (*Science*, 16 December 1994, p. 1800). And his project to sequence *H. influenzae* failed to win funding by the National Institutes of Health. Now Venter argues that the techniques he used to sequence the organism in record time will help speed the sequencing of the human genome (see box). Those techniques allow massive segments of DNA—in the case of *H. influenzae* the whole 1.8-megabase genome—to be broken into random fragments or "shotgunned," sequenced, and then reassembled in one go.

And the *H. influenzae* sequence is just the beginning of a wave of bacterial sequences that is destined to change the tranquil backwaters of microbiology into a surging torrent of new discoveries. In as yet unpublished work, the Venter group has already determined the sequence of *Mycoplasma genitalium*, and at least two more bacterial sequences are expected within the year.

These discoveries, say Weiss, Blattner, and others, will help researchers answer such fundamental questions as how microbes and higher organisms evolved. And they can have a practical payoff as well, helping to identify genes that transform harmless bacteria into killers or enable them to thrive in the searing heat of deep-sea vents. That information could aid in the design of new antibiotic drugs or in identifying enzymes of industrial importance.

Not that the *H. influenzae* genome isn't a breakthrough in its own right. It will provide the first information on how a complete genome is organized, giving a boost to the fledgling discipline of genomics. The sequence may well also contain some bacterial equivalents to human disease-causing genes, helping researchers get a grip on the human genes' modi operandi.

Mining the microbial genome

The Venter and Smith groups have already begun to sift through the data they have accumulated on *H. influenzae* for clues to what makes the organism tick. Its genome contains 1743 genes, with—as is typical for bacteria—little noncoding DNA.

More than 40% of those genes—736 in all—proved to have no counterparts of known function among the genes that have been sequenced from other organisms. This thick seam of previously unknown genes is expected to provide rich pickings for microbial geneticists. Now that every gene's sequence is known, "knocking out" the mystery genes will be a trivial procedure in *H. influenzae*, says Hopkins's Smith. By studying

Homing In on the Human Genome

"Lo and behold, the two ends joined. I was as stunned as anyone," is how molecular biologist Robert Fleischmann describes the moment when the final piece of sequence data was slotted into the circular *Haemophilus influenzae* genome. Fleischmann is a key member of the team, led by Craig Venter of The Institute for Genomic Research (TIGR) in Gaithersburg, Maryland, and Hamilton Smith of Johns Hopkins University in Baltimore, that is now reporting the first complete genome sequence of a free-living organism (see main story and p. 496). Their success, says Venter, could help complete another major sequencing effort: the Human Genome Project (HGP), whose goal is the determination of the exact order of all 3 billion bases in the human genome.

The success with the *H. influenzae* sequence has "raised the ante worldwide for sequencing the human genome."

—Craig Venter

In the 5 years since it began, the HGP has focused mainly on developing the physical and genetic maps needed to guide the sequencing effort. But that all began to change late last year, spurred by a proposal made by John Sulston, director of the Sanger Centre in Cambridge, United Kingdom, and Robert Waterston of Washington University in St. Louis (*Science*, 2 June, p. 1270). They suggested that the human genome could be completely sequenced, using technology already available, as early as the year 2001 and at a cost of 10 to 12 cents per base, if the sequencing accuracy were dropped from the original target of 99.99% to 99.90%.

Venter is now arguing that the new techniques worked out for *H. influenzae*, once adapted for the human genome, will make that goal even more feasible. The success with the bacterial sequence has, he asserts, "raised the ante worldwide for sequencing the human genome." And Venter has won some cautious support. "For larger genomes, it's anyone's guess whether this approach will pay off," says George Church, a sequencing technology expert at Harvard University, "but I'm optimistic."

The approach developed by the TIGR-Hopkins group differs from the method favored up until now. Indeed, Venter says, his team's application for National Institutes of Health (NIH) funding for the project was turned down because of doubts that its approach was up to sequencing the 1800-kilobase *H. influenzae* genome. In the end, the Venter team completed the work in about 13 months at a cost of 50 cents per base. (NIH officials declined to discuss the reasons Venter's application was turned down. But Robert Strausberg of NIH's National Center for Human Genome Research, who heads up sequencing technology, says that he is "comfortable" with the decision.)

In conventional sequencing, the genome is laboriously broken down into ordered, overlapping segments, each containing up to 40 kilobases of DNA, which are then shattered—or shotgunned—into smaller pieces. After these smaller fragments are sequenced, they are ordered according to how their sequences overlap, and the original segments are used to reconstruct the genome.

In contrast, the TIGR-Hopkins team shotgunned the entire 1800-kilobase *H. influenzae* genome. What makes that approach difficult is the computational power needed to reassemble the approximately 24,000 fragments once they have been sequenced. But the team's own software program, called the TIGR Assembler, met the challenge and reassembled the whole genome.

Larger contiguous pieces of genome have been sequenced—for example, whole yeast chromosomes (*Science*, 16 June, p. 1560) and a 2.2-megabase stretch of the genome of the nematode *Caenorhabditis elegans*. But these efforts required massive collaborations and took years. The success of the *H. influenzae* project, and of a subsequent one on the genome of the microbe *Mycoplasma genitalium*, has given Venter confidence that similar techniques will work in sequencing the human genome, even though it's 1500 times bigger than the *H. influenzae* genome.

Finishing the human sequence in 5 years would still require a 20- to 100-fold increase in output: to roughly 600 megabases per year from all the labs that join the effort. Venter proposes to achieve this goal partly by shotgunning the human genome in large clones such as Bacterial Artificial Chromosomes (BACs), which have a capacity of 80 to 350 kilobases. Doing away with the extra overlap that's needed to line up smaller clones would increase efficiency by about 25%, he says. For even greater efficiency, Venter suggests that the ends of each large clone be sequenced before the whole clone is tackled to ensure no more than a 2-kilobase overlap with any clone already sequenced.

But Waterston believes it's not the size of the clone but the speed of sequencing that's the major issue. "Whether we use BACS or [smaller] clones is not going to fundamentally impact the change of scale. ... Either way we are going to have to sequence [many] megabases a week," he says. And so far, he says, no one has accomplished that.

Nonetheless, both sides are ready to pit their techniques against large tracts of human genome. Venter is firming up plans with TIGR's Mark Adams and Helen Donis-Keller, also of Washington University, to sequence 100-kilobase fragments of the end regions of chromosome 7 that teem with repetitive elements. Sulston, Waterston, and Bruce Roe of the University of Oklahoma, Norman, meanwhile, are starting on chromosome 22.

Adding spice to the competition is the fact that NIH wants to test the sequencing strategies, too. It has put out two requests for applications for up to $20 million in funding. "We expect an enthusiastic response," says Strausberg, and with the first application deadline only 1 week away and the second 25 days after that, it undoubtedly means that Venter, Sulston, and Waterston—and as Waterston puts it, "just about everyone else"—are currently working feverishly to complete their proposals.

–R.N.

the impact of the knockouts on the bacterium's biochemistry, researchers should be able to identify many of the proteins the genes encode and also establish their functions.

The remaining 1007 *H. influenzae* genes do have counterparts of known function in other organisms, and that helps to provide information on the genes' functions. These included, for example, genes that code for metabolic enzymes, transcription factors, and many of the other accouterments of a living cell.

For the TIGR-Hopkins team, however, the genes that were not found in the *H. influenzae* genome proved just as interesting. For example, *H. influenzae* turned out not to have three enzymes of the tricarboxylic acid (TCA) cycle, a key pathway for energy production. "I would have expected virtually all bacteria to have a complete TCA cycle," says Smith. He concedes he has little clue about what this means, but suggests it could explain why the bacterium requires huge amounts of the amino acid glutamate to grow in the

laboratory. Glutamate may be needed to replace compounds missing because of the defective TCA cycle.

And in a second paper on page 538, Smith, Venter, and their colleagues describe work on an intriguing feature that *H. influenzae* shares with a few other bacteria: The bacterial cells can recognize and take up the DNA that other *H. influenzae* cells leave behind when they die. It's not clear why they do this, although the new work shows that the bacterial DNA is designed to maximize the chances that it will be taken up. The Smith-Venter team found that the *H. influenzae* genome contains 1465 identical copies of a recognition sequence that marks the DNA for uptake by cells of the same species. The fact that the bacterium devotes so much of its genome to this recognition sequence suggests that DNA uptake is a significant survival advantage to the bacteria, says Smith.

Other researchers are testing the *H. influenzae* database's mettle in identifying the genes that contribute to the organism's pathogenicity in its wild state. Pediatrician and molecular biologist Richard Moxon, Derek Hood, and Michael Jennings of Oxford University in the United Kingdom are particularly interested in identifying the bacterial enzymes that synthesize lipo-oligosaccharides (LOS), toxins located on the bacterial surface that trigger many of the life-threatening symptoms of bacterial infection, such as septic shock. By trawling the database, the team has identified 18 genes that might be needed for LOS synthesis and is now knocking them out to see if they do in fact play that role. "We made more progress in 6 months, frankly, than we had in 3 years," says Moxon.

The Moxon team is also using the *H. influenzae* database to investigate another pathogenic feature of bacteria. To pass from person to person—say in a sneeze—and survive, a population of bacteria must be capable of rapidly adjusting to new environments. Six years ago, Moxon and his Oxford colleague Jeffrey Weiser provided a clue that helps explain how bacteria do that. The researchers found that the three *H. influenzae* LOS genes that had been identified at that time carry certain repeated sequences that make them very susceptible to certain copying errors during DNA replication. As a result, the bacteria in a population synthesize a mix of LOS proteins with slightly different properties, ensuring that at least a fraction of the population is suited to any environmental change.

In their recent search of the *H. influenzae* genome, the Moxon team found that other virulence genes may make use of the same adaptive mechanism. They identified the same repeated sequences at eight new locations, at least seven of which are associated with suspected virulence genes, including one that resembles adhesin, a protein that enables bacteria to adhere to host cells. "Until we actually looked into the genome, we really didn't have any idea how powerful it was," says Moxon. "It knocks you over; we wouldn't have been able to find [the repeats] so quickly in any other way."

Genetics, genomics

This rich haul of data is just an indication of the bonanza that lies ahead for microbiologists. The Venter team has already completed, although not yet published, the sequence of *M. genitalium*, a bacterium that is associated with reproductive-tract infections and is renowned for having the shortest genome of all free-living organisms. The institute also expects to have the sequence of *Methanococcus jannaschii*, a bacterium that thrives at temperatures as high as 100°C, by year's end, and the sequence of another thermophile, *Methanobacterium thermoautotrophicum*, is expected from the biotech firm Genome Therapeutics Corp. of Waltham, Massachusetts, by the same date.

MARTY KATZ

Genome gang. This photograph shows the group that sequenced the *H. influenzae* genome, with team leaders Craig Venter at front left and Hamilton Smith at front right.

In addition, since the completion of the *H. influenzae* genome was announced in May at the annual meeting of the American Society for Microbiology (*Science*, 2 June, p. 1273), the TIGR-Hopkins team has received proposals to collaborate on the sequencing of 30 different microbial genomes, including the pathogens that cause tuberculosis and syphilis. "People thought [that bacteria] were multiyear, multimillion-dollar projects," Venter says. "We've shown that it can be done in less than a year and for less than 50 cents per base, [and] it's opened the floodgates."

As the number of fully sequenced genomes increases, the rate of discovery will rise exponentially. "The interesting stuff comes from the comparisons," says Blattner, who expects to complete the *E. coli* genome by 1997. Comparing the genomes of virulent and harmless strains of bacteria will further aid microbiologists in their search for disease-causing genes.

And the full sequences of the thermophiles *M. jannaschii* and *M. thermoautotrophicum* will help answer another slew of questions. "How do the enzymes function at [high temperatures]?" asks Moxon. "How on Earth are they repaired? Why don't they just crumple up?" Those questions are not just of academic interest. The bacteria may well contain enzymes of industrial importance that might be used for high-temperature catalysis.

By comparing genomes from different groups of organisms, geneticists—or genomicists—also hope to gain a much-needed fresh insight into evolution. In recent years, a fierce and sometimes bitter debate has split the field of microbial evolution wide open (*Science*, 3 July 1992, p. 32, and 27 May 1994, p. 1251). One side contends that micro-organisms called archaeons such as *M. jannaschii* and *M. thermoautotrophicum* and micro-organisms called eubacteria such as *H. influenzae* and *E. coli* belong on two distinct evolutionary branches. In this view, the archaeons are closer to higher organisms, including humans, than they are to eubacteria. Their major opponents believe that archaeons and eubacteria belong on one evolutionary branch, and that only some higher organisms arose from archaeons.

The debate has only been fueled, rather than quelled, by the partial data that are currently available on the organisms' gene sequences. But "the debate will disappear when we have the whole genomes," predicts evolutionary biologist Ford Doolittle of the Canadian Institute for Advanced Research in Halifax, Nova Scotia, who is a member of one of the first teams to start sequencing an archaeon, *Sulfolobus solfataricus*. The study of whole microbial genomes, says Doolittle, will also help answer questions about the origins of the genes of humans and other multicellular organisms.

Moreover, because so much of the work can be done on the computer, "it's going to empower the small investigator in ways that they had never dreamt," predicts the director of the Department of Energy's genome project, David Smith. Says Francis Collins, director of the National Center for Human Genome Research in Bethesda, Maryland, the *H. influenzae* sequence "is a milestone that people should look at very closely. Having the complete sequence[s] of organism[s] as a starting point is going to transform the way people approach biology, first in microorganisms, and later in higher organisms."

–Rachel Nowak