The publicly and privately funded teams have both finished drafts of the human genome. Now comes the daunting task of developing tools to figure out just what these volumes say

Finally, the Book of Life and Instructions for Navigating It

Beaming at each other, longtime rivals Francis Collins and J. Craig Venter shook hands in the East Room of the White House on 26 June as they declared joint victory—and announced an implicit truce—in their race to decipher the "book of life." President Clinton presided over the event, attended by a stellar cast of genome scientists, a few members of Congress, and a handful of foreign ambassadors, to celebrate completion of the "first survey of the entire human genome ... the most wondrous map ever produced by humankind." In fact, neither one's team has completely deciphered the human genome—that is, determined the exact order of all 3.12 billion or 3.15 billion bases, depending upon whom you ask, that make up our DNA. But each has completed a version of this book, which, hyperbole aside, promises to propel biology and medicine headlong through the 21st century. What's more, the two former adversaries, who until recently have minced no words disparaging the other's work, said they hope to publish their work simultaneously in a peer-reviewed journal sometime this fall (see p. 2294).

This very public and very carefully orchestrated denouement—which required diplomatic skills akin to those behind the Camp David Peace Accord—brings to an end one of the most high-profile fights in recent biology, one that pitted a publicly funded consortium of scientists, led by Collins, against Venter's upstart company, Celera Genomics of Rockville, Maryland. With obvious relief, Collins and Venter agreed to forgo the barbs and share the credit for a biological tour de force that many scientists thought was impossible a mere 15 years ago.

So what, exactly, have they produced, and how will they fine-tune it so that everyone from workaday biologists to pharmaceutical giants can mine its gold?

The public consortium has finished a "working draft," which covers 85% of the genome's coding regions in rough form. Although the sentences on some pages are mixed up and some words are missing letters, the data are freely available in several public genome databases. A polished version will be out in 2003 or sooner, promises Collins, director of the National Human Genome Research Institute, which funds most of the U.S contribution to this international endeavor. By all accounts, Celera's version is considerably more polished, thanks to a bold new sequencing strategy, deep corporate pockets, and Venter's ability to pool the public data with his own proprietary data. Venter promises to make his draft freely available to academic researchers at the time of publication; it is available now to subscribers who paid to get a first peek.

All smiles. At a White House event, J. Craig Venter (left) and Francis Collins put aside past animosity to celebrate completion of two drafts of the human genome.

Both books are clearly works in progress, the public's more so. As Venter is the first to admit, sequence data by themselves are of minimal use, so both teams have been scrambling over the past few months to improve the computer tools and analysis, known as annotation, that will enable biologists to make sense of the billions of A's, T's, G's, and C's contained in both databases. Although such efforts are already under way and some ingenious new strategies are in the works, full annotation of the human genome will continue well into this century.

Before the announcement, speculation was rampant that Venter and Collins might collaborate on annotating the genome, turning the truce into a real partnership. President Clinton encouraged such hopes at the White House briefing when he said that both sides had agreed to hold a historic sequence analysis conference. At a subsequent press briefing, however, both Venter and Collins went out of their way to downplay such expectations, saying that they were exploring the possibility of a workshop to compare their approaches after publication. For now, on this ever-shifting stage, it looks as though the two annotation efforts will proceed independently—as with the sequence itself, undoubtedly speeded by the competition.

The books

These books, the starting points for annotation, are distinct, reflecting the different processes used to create them. From the outset, the publicly funded Human Genome Project worked by consensus, using a painstaking approach that wins kudos in terms of democracy but is not conducive to speed. Starting in about 1990, researchers across the globe divvied up the work, first making genome maps of increasing resolution, then improving the technology for sequencing, testing it on model organisms, and finally, in 1997, launching into full-scale sequencing of the human genome (*Science*, 12 April 1996, p. 188). Across the Atlantic, the Wellcome Trust set up the Sanger Centre in Hinxton, guaranteeing that the United Kingdom would be a big player in genome sequencing. From the outset, the goal of the public effort was to produce a "finished," highly accurate sequence—to the extent possible (there will always be some holes), a continuous stretch of A's, T's, G's, and C's, arrayed in the exact order in which they appear along the chromosomes. But in a nod to the limitations of their technology, the

CREDIT: MARTY KATZ

team decided to sequence just the regions of the genome known to contain most of the genes—not the entire genome—but to do so with fewer than 1 error per 10,000 bases. Completion of this estimated $3 billion project was slated for 2005.

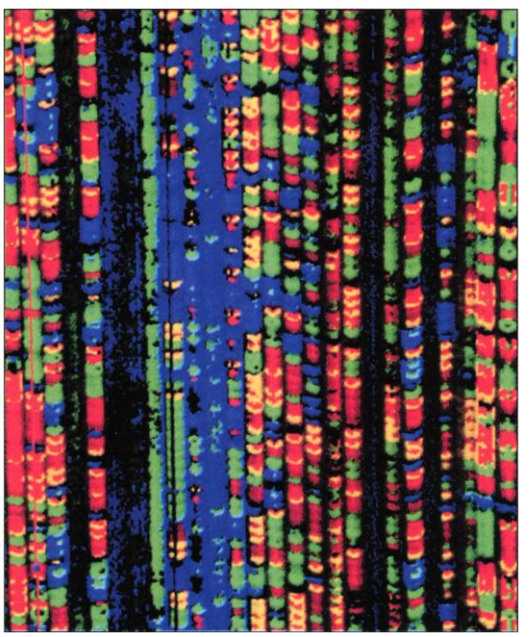
That changed when Venter, a former NIH scientist turned entrepreneur, threw down the gauntlet in 1998, declaring that his new company would single-handedly sequence the entire genome in just 3 years—4 years ahead of the public project. As a CEO, Venter had several tactical advantages over an NIH institute director. For one, he did not have to contend with peer review, nor did he have to strive for consensus. Instead, he adopted a radical sequencing strategy that depended upon some 300 of the fastest sequencing machines—made by PE Biosystems Corp., Celera's parent company—and one of the world's most powerful supercomputers (*Science*, 18 June 1999, p. 1906). What's more, Venter could build on—and later incorporate—the work of the public project.

Fearing that Venter planned to patent the sequence and sell it for profit—as well as hog all the credit—the public consortium rallied. The Wellcome Trust immediately increased its support for the project, promising that the Sanger Centre would do a third of the genome. The United States consolidated its sequencing effort and together, the two countries created five sequencing supercenters that drastically scaled up their efforts. (The five sequencing shops are the Sanger Centre; the U.S. Department of Energy Joint Genome Institute in Walnut Creek, California; Washington University School of Medicine in St. Louis; the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts; and Baylor College of Medicine in Houston.)

And in September 1998, the consortium announced a brand-new game plan: Instead of concentrating on finished sequence, it would produce a rough draft of 90% of the genome by spring 2001—about the same time as Celera's target date for producing the human genome. A year later, they moved up the completion date to spring 2000. The goal, said John Sulston, director of the Sanger Centre, was to get as much of the human genome sequence into the public domain as possible before Venter could lock it up.

To decipher the genome—actually, a mosaic of six to 10 anonymous individuals—the public consortium opted for a careful, if tedious, piece-by-piece approach. It's akin to ripping out a page of a book, shredding it multiple times, then taping it back together by looking for overlapping letters. But in this case, researchers start with a 150,000-base chunk of DNA (the page)—known as a BAC, or bacterial artificial chromosome, in which it is cloned—then chop it up into many smaller pieces, or subclones, that have overlapping ends. These pieces are then run through the sequencing machines and reassembled by a computer into ever longer pieces, called contigs. Then the group moves on to the next clone.

In reality, the process is more complicated. Each piece is sequenced not once but multiple times, because the more times a base shows up at the same position in these subclones, the more certain the computer is that the identification is correct. As a result, the whole genome is sequenced several times over—four times for the rough draft to have an error rate less than 1 in 100, and somewhere between eight and 11 times to reach the higher standard of no more than 1 error in every 10,000 bases.



Reading the code. A sequencing machine at Washington University spews out the genetic alphabet a letter, or base, at a time, each tagged a different color.

CREDIT: WASHINGTON UNIVERSITY SCHOOL OF MEDICINE

In this piece-by-piece approach, as a given BAC is worked through, its "sequence" will first show up as a series of short, unconnected strings of bases that may be in the wrong place on the BAC or even backward. Ideally, given enough time, as the computer slogs through additional sequence, it begins to fill in the holes, linking the small pieces of DNA together to form ever-longer stretches until most of the BAC is represented in the correct order. The next job is to align all the BACs, which also contain overlapping ends to aid in assembly—a task that sounds easy but is dauntingly difficult.

The rough draft has not yet achieved this level of completion. The current draft consists of BACs covering 85% (5% short of their announced goal) of the gene-containing regions of chromosomes. The BACs are in order, thanks to the efforts of Washington University's Robert Waterston and John MacPherson, who worked out how to put each BAC in its proper place. But the completeness of each BAC can vary from being quite jumbled to having just a few bases missing. Some 24% of the sequence is in finished, highly accurate form, said Collins at the briefing; another 22% is in near-finished form; and 38% is in draft form. Most of the remaining 15% is currently being sequenced, except for a pesky 3% that refuses to be cloned.

Celera, on the other hand, relied on the "whole-genome shotgun" strategy that Venter had pioneered for sequencing microbial genomes (*Science*, 28 July 1995, p. 496). Instead of going piece by piece, or shredding one page at a time, Celera shreds the entire volume—or more accurately, an entire set of encyclopedias—into millions of tiny overlapping pieces and then reassembles them with the aid of a superfast supercomputer. Although the company has not revealed its exact sequencing strategy for the human, it presumably resembles that used to sequence the *Drosophila* genome. It blasted the genome of one man first into 2000-base pieces, then into 10,000-base pieces, and again into 50,000-base pieces, covering the genome three times, between September 1999 and April 2000. Then, to fill in the gaps and increase accuracy, Celera sequenced parts of the genomes of three women and one additional man of diverse ethnic backgrounds, finishing that work by 23 June.

Celera also took advantage of the fact that the public consortium deposits its data nightly into GenBank. Each day, Celera scientists downloaded the human sequence data in GenBank, manipulated them so they looked like its own raw sequences, and fed both data sets into the company's supercomputers for comparison. By incorporating the public data into its analysis, Celera ensured that each base, in theory, had been sequenced six times or more, significantly boosting the odds that it is accurate—and shaving a year or two off its project, says Venter. Analyses of the recently completed *Drosophila* sequence data suggest that Celera can get reasonably accurate and assembled coverage of the genome by sequencing it just 6.5 times, rather than 10 times as was originally thought.

Celera then assembles these data into "scaffolds," which are sets of contigs whose locations along a chromosome are determined by matching up known DNA landmarks. Although there are likely to be some 200,000 gaps between and within scaffolds,

the Celera genome comes closer to covering all the gene-containing regions of the genome than does the public draft. Because the assembly is based solely on the overlaps—and not on the supposedly preestablished order of the pieces, as in the public project—more of the Celera genome is in the right place and in the right order. However, at this point, the human genome "will not be as good as the *Drosophila* genome" that was published in March, says Norton Zinder of Rockefeller University in New York City, who is also a scientific adviser to Celera.

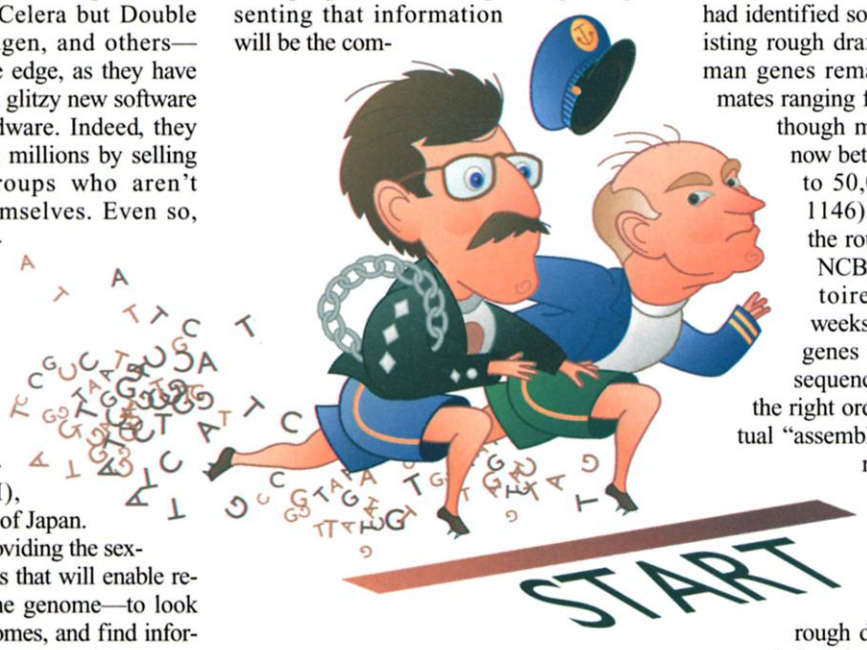
Charts, sextants, and compasses

Getting those billions of bases in order is just the first step. Next comes figuring out what they mean. With so much data now in hand, the race has shifted to developing ever slicker algorithms and more user-friendly packaging for the tools needed to analyze, or annotate, the genome. Here again, companies—and not just Celera but Double Twist, Incyte, Compugen, and others—would seem to have the edge, as they have more money to invest in glitzy new software and high-powered hardware. Indeed, they are banking on making millions by selling their analyses to groups who aren't equipped to do it themselves. Even so, new databases and computer programs are cropping up monthly in the public arena, some at GenBank and others at GenBank's European counterpart, the European Bioinformatics Institute (EBI), and the DNA Databank of Japan.

Together, they are providing the sextant, compass, and charts that will enable researchers to navigate the genome—to look for genes, compare genomes, and find information relevant to the stretch of sequence they want to study. "In the end it will not be the data that makes the difference; it will be the software," predicts J. Michael Cherry, a bioinformaticist at Stanford. "If [a company] can provide their customers with good tools to mine the data, they will do very well."

For both public and private annotation efforts, the basic task is the same; the products differ mostly in the bells and whistles they provide. The first priority of any annotation software is to pinpoint the genes. Only computers have the ability to scan billions of bases and pick out the potential genes. They do this by looking for characteristic sequences at the beginnings and ends of genes, or by comparing new sequence to known genes or bits of genes. Additional computer programs translate those genes into proteins and, based on similarities to other proteins, attempt to assign a function to each one. Still other programs, such as the National Center for Biotechnology Information's (NCBI's) BLAST, compare the new genome data to that from other organisms, such as the fruit fly or the nematode. At Celera, Venter's crew uses its supercomputer to routinely perform "all against all" searches—comparisons of the newly generated sequence with that in all available databases. The sequence similarities such searches turn up highlight regions, such as genes or regulatory DNA, that might be missed by other gene-hunting programs.

According to Zinder, when Celera publishes its version of the human genome, it will make available basic annotation—locations of the genes, with their coding and noncoding regions defined, and the predicted functions of their proteins—but not information based on genome-to-genome comparisons. Those comparisons and Celera's programs for manipulating and presenting that information will be the company's bread and butter, so Venter isn't cutting any corners or sparing any expense.



What race? Venter and Collins deny that they have been racing to finish the human genome; at any rate, both agree that the real work of deciphering it has only just begun.

Instead of sextants and compasses, Venter plans to have the genomics equivalent of the computer-linked Global Positioning System that guides his yacht. "It takes a lot more to navigate around the genome than to navigate around the world," he says. Scientists who had a preview of what's to come are enthusiastic. "Celera's annotation and database programs are excellent," says J. Troy Littleton, a neurobiologist at the Massachusetts Institute of Technology who worked with Celera to annotate the fly genome in November 1999.

At the same time, bioinformatics experts working with the Human Genome Project are scrambling to complete a set of navigating tools that they plan to provide online for free. True to the democratic and somewhat individualistic nature of the public endeavor, several annotation efforts have sprung up in conjunction with the major players in the project. One called the Genome Channel is an offshoot of the U.S. Department of Energy's genome effort; others, such as that at NCBI, GenBank's home, and EBI, were spawned to help users make sense of archived data. Also, because incoming sequence is immediately available, no matter how patchy and incomplete, EBI and NCBI have been working hard to make clear what's what and where to find the best sequence for the part of the genome being studied.

By late June, EBI's program, ENSEMBL, had identified some 38,000 genes in the existing rough draft; the total number of human genes remains a mystery, with estimates ranging from 28,000 to 120,000, although many genome scientists are now betting that the answer is close to 50,000 (*Science*, 19 May, p. 1146). Also, to compensate for the roughness of the public draft, NCBI plans to expand its repertoire of tools in the coming weeks. Because it's easier to find genes when the small chunks of sequence within each BAC are in the right order, NCBI will perform virtual "assemblies" that will clean up the rough draft electronically without generating additional sequence data.

With these tools in hand, asserts NCBI director David Lipman, the rough draft should be of sufficient resolution for most tasks biologists want to perform. Indeed, he says, a dry run using a subset of the data indicated that gene-finding programs do almost as well with rough-draft sequence as with finished sequences in finding at least some part of a gene.

Mary-Claire King, a human geneticist at the University of Washington, Seattle, concurs. "It's very rough, but very useful," she says. Increasingly, says King, gene hunters like herself determine the general location of a gene and then pull the sequenced version of that region out of GenBank to find the gene itself.

Setting a straight course

Over the next few months, these first computer-based expeditions will be overtaken by human explorations of the genome. Gene-prediction programs make mistakes, identi-

ILLUSTRATION: T. SMITH

fying fossil genes that are never expressed or fusing two genes together, for example. Protein classification programs also have trouble—one part of a protein may make it look like a transmembrane receptor while another part suggests it is a DNA binding protein. The human eye sees new patterns and possibilities in sequences that computer programmers never dreamed of. "You really need to look at the data," points out Gerald Rubin, vice president for biomedical research at the Howard Hughes Medical Institute in Bethesda, Maryland. "Any kind of [automated] annotation will not substitute for humans," says Rubin, who ought to know, as he and Venter arranged an "annotation jamboree," or research fest, to make sense of the *Drosophila* genome.

In November 1999, Celera brought together about 45 biologists and bioinformatics experts to take a first look at the newly assembled fly genome. The synergy that resulted led to many discoveries about the fruit fly and even some new ideas about how organisms in general evolve greater complexity (*Science*, 24 March, p. 2182). Venter is planning another jamboree, or likely several, over the summer and fall to annotate the human genome; for now, Celera is not saying whether the insights gained in those jamborees will be included in its initial publication.

Although a jamboree is great for a first pass, full annotation will take years, both Venter and Collins agree, and will increasingly depend on the contributions of bench biologists who are studying individual genes and proteins. For that reason, bioinformatics experts in the public consortium are focusing on ways to elicit continuing input from the biological community. EBI has embraced a strategy developed in large part by Lincoln Stein of the Cold Spring Harbor Laboratory in New York to keep nematode researchers involved with adding new results to the nematode genome database.

Called a Distributed Annotation System (DAS), it enables any researcher to add his or her two cents to the database, providing they follow the DAS format for presenting the information. EBI has the rudiments of the system in place; by August, Stein hopes to finish the final bit of software so the system can go online. The input won't come in time for the first publications, but the system may eventually become a powerful reservoir of biological knowledge.

Collins thinks even more is needed. "We need to figure out a way to capture community input in a way that doesn't contaminate the databases [with non–peer-reviewed information]," he says. Along those lines, Lipman envisions an army of curators who will scan the literature and cull important findings to add to the existing annotation.

Lipman already has a team that is developing a definitive reference list of genes—a task that is not as easy as it sounds. Many genes have multiple names and more than one predicted function. The genes in the database often come in multiple versions as well, of varying degrees of accuracy (*Science*, 15 October 1999, p. 447). For about a year, these curators have been resolving discrepancies and picking one reference sequence—hence the name RefSeq—as a prelude to more in-depth annotation of the human genome. When necessary, they call in outside troubleshooters to help. So far they have double-checked 1500 genes and expect that number to increase rapidly.

Eventually, Lipman would like to set up an electronic journal in which biologists would publish minireviews on their favorite gene families. These reviews would be hotlinked to the sequence and take advantage of "a model that's existed for a couple of decades," he explains, "[that] of combining databases and [scientific] literature."

But all of this will take time—and researchers are impatient. So both teams are turning to the mouse for help: They are planning to sequence its 3-billion-base genome and compare it to the human genome. "The mouse [sequence] will identify all the human genes like no prediction program could do," explains David Nelson, a biochemist at the University of Tennessee, Memphis.

Celera, in characteristic style, is blazing the trail. As soon as it finished sequencing one human genome in April, Celera began blasting through the mouse genome. By late June, it was halfway there, says Zinder, and would be finished by the end of the year. When Celera overlays the mouse sequence on the human genome, it expects to be able to find many of the 35% of the human genes missed by other approaches, as well as identify regulatory regions and other key pieces of DNA. In addition, after Celera knocks off the mouse later this summer, it plans to turn its sequencing prowess to the rat, and perhaps the zebrafish, the dog, or a primate.

That's about the same list as the public project proposes to sequence over the coming years. But true to consensus-building operations, the public consortium is still working out a sequencing strategy for the next year or so. In October 1999, it intended to start the mouse and had divided the task among 10 centers. But that work has barely started. Some researchers, like Doug Smith at Genome Therapeutics in Waltham, Massachusetts, are urging NIH to pick up the pace and sequence the mouse even more quickly than planned—especially as Celera intends to keep its valuable mouse data private.

Yet there's also a great need to finish the human genome. Finished sequence "will be critical" for a variety of experiments, says cell biologist Shirley Tilghman of Princeton University—for example, for making sense of very large genes or for figuring out how the shape and structure of the chromosomes influence gene regulation. As a result, Collins and his advisers have been debating for months whether to push through the rest of the human genome quickly or turn to the mouse. The emerging consensus seems to be to do both: continue sequencing human but devote substantial capacity to mouse, so that a rough draft will be available in 6 to 9 months.

Does all this mean that Celera's database is a must for genomics researchers? Opinion is divided. As long as Celera stays ahead and provides comparative analyses of an increasing menagerie of organisms, predicts Lipman, many genome researchers will likely ante up the funds to subscribe to it. Others disagree, saying that Celera's real advantage will be short-lived. "Once the human and mouse genomes are done and the genes have been identified by comparison of the two genomes, much of the excitement will pass," says Nelson. Adds Tilghman, "Why should I pay for something I can get for free?" **–ELIZABETH PENNISI**

CAN YOU FIND THE GENE CALLED GENETIC?

If DNA were really made up of letters and genes were words, this sequence—
asdfgeoglkodnjhqwerouieoptieswazxcvzcnmkjholoedogheoadsfkcoseafloraeadlk
—might look as follows during the different phases of sequencing:

Phase 1: ae l k a sdf tie kodj swa hq wero sea Uie op vz cn m kj hol o e d ogh eo sf k co fl or

Phase 2: Ae lk Asdf Kodnjtieswa ogl Hqwero sea Uieop zvz cnmkj holoed ogh eoad sfkco flor

Phase 3: asdfgeogl kodnjhqwero sea uieoptieswazxcnmk jholoed gheoadsfkco flor ae lk

Finished: Asdfgeoglkodnjhqwerouieoptieswazxcvzcnmkjholoedogheoadsfkcoseafloraeadlk