The Human Genome

It is an awe-inspiring sight. Open up the folded figure that comes with this issue of *Science*. There you will see the human genome, chromosome after chromosome, with its major features color-coded and described. Black tick marks show the coding regions along orange, blue, pink, and purple genes, the colors reflecting the function of the corresponding proteins. All told, some 2.9 billion bases of the genome are represented on this beach towel–sized poster.

It took geneticists 7 years to find the gene involved in cystic fibrosis—but here you can locate it in a few seconds in the last third of chromosome 7. Look down toward the bottom

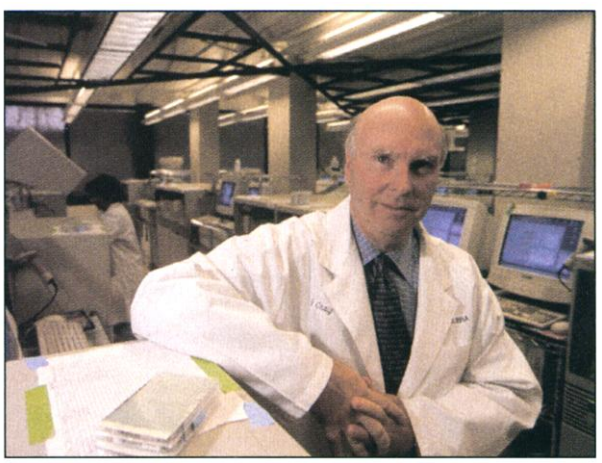
ILLUSTRATION BY CAMERON SLAYDEN

of the poster, on chromosome 17, to find BRCA1, one of the genes implicated in hereditary breast cancer. One quick look shows, too, that not all chromosomes are created equal. Number 19 is jam-packed with genes—23 per megabase, more than 1400 total, but chromosome 13 has relatively few, just five per megabase.

Thousands of scientists across the globe have labored for some 15 years to achieve this feat—the (almost) complete nucleotide sequence of human DNA, often called the book of life. Actually, two books exist, because the rival teams who compiled them were unable to mend their differences and pool their data. The genome sequence on the poster was compiled by J. Craig Venter and colleagues at Celera Genomics, a biotech company started just 3 years ago in Rockville, Maryland. The other, which appears in the 15 February issue of *Nature*, was produced by the International Human Genome Sequencing Consortium.

Both have yet to be finished, with all the i's dotted and the t's crossed. Small to large gaps exist in each draft, akin to a missing word or paragraph or page, but the gist of the story is still clear. Thus, even in this unpolished state, these two books offer the most comprehensive look at the human genome ever possible. To scientists like Richard Gibbs, who heads the sequencing effort at Baylor College of Medicine in Houston, that look is thrilling: "It's the same feeling you must get when you are on a satellite, and you are looking down at Earth." Even more exciting, says Celera's Mark Adams, is that these drafts are really just the beginning. The Celera paper "is mostly a presentation of how we got where we are," he points out, and it provides only a fleeting glimpse at the wealth of information contained in the sequence.

Just obtaining the sequence is a phenomenal achievement, one that many researchers did not believe possible 15 years ago. (*Science* has highlighted a few of the unsung heroes in this massive endeavor.) Until now, the largest genome ever sequenced was that of the fruit fly, with 180 megabases, which Celera and academic researchers knocked off in March 2000. The human is almost 25 times as big and is infinitely more difficult to decipher. In essence, Figure 1, even with almost 50 meters of chromosomes, is just an abstract of the book. Spelling out the entire sequence, all 3 billion or so chemical letters that make up the DNA along each chromosome, would fill tomes equivalent to 200 New York City phone books. Yet all it takes is Internet access to view those letters, one by one. With a few clicks of the mouse, one can now scroll through the book of life. Fifteen months ago, the true positions of barely 10% of those letters were known; now some 90% are represented in both the Celera and public databases, with varying degrees of certainty in the latter. "Having this enormous amount of sequence all laid out is just the coolest thing," says Robert Waterston, co-director of the Washington University Genome Sequencing Center in St. Louis.



J. Craig Venter, president of Celera Genomics and lead author of the paper published in *Science*.

This new text has enabled both groups to chart the genomic landscape with unprecedented precision and make their best guesses yet about the number and types of genes that humans share with other organisms or call their own. "There's a long list of things that blew my socks off," says Francis Collins, director of the National Human Genome Research Institute, which supported the lion's share of the U.S. Human Genome Project. Collins points to the number and source of human genes as just two surprises. As the sequence is filled in over the coming months and years, almost every conclusion drawn by the several hundred researchers who've scanned this text will need revisiting, they concede. But the discoveries made so far have already made even these drafts best sellers.

A new view

Perhaps most humbling of all is the finding by both Celera and the public consortium that humans have 32,000 genes, give or take a few thousand. That's only about twice as many as the nematode has, and the number "is a bit of an assault on our sensibility," Collins notes. Celera's scientists have detected 26,383 genes that are almost sure bets and another 12,000 distant possibilities; the consortium came in at 24,500, with another 5000 expected to show up as gene-prediction programs improve. Both are a far cry from the commonly cited number of 100,000 genes.

"It shows that it is better to draw conclusions based on data rather than conjecture," says Celera's Adams, who as late as May bet there were some 67,000 genes (*Science*, 19 May 2000, p. 1146). As the sequencers puzzled over what happened to the rest, reexamining evidence for the lower number, they realized that the oft-mentioned 100,000 arose from a back-of-the-envelope calculation by Harvard Nobel laureate Walter Gilbert in the mid-1980s; subsequent papers also predicted the total to be between 50,000 and 100,000 genes. Gilbert still stands by his count, and even those who have now predicted only about one-third that number are circumspect. There won't be fewer than 25,000, "but the top end of this number is still quite flexible," says bioinformatics expert Ewan Birney of the European Bioinformatics Institute branch in Hinxton near Cambridge, U.K. Adams agrees: "I'm sure in some cases we've underpredicted" the genes.

One reason for wiggle room is that gene-prediction programs work either by looking for a sequence that's similar to known genes or gene fragments or by homing in on a sequence of the right size that has the telltale beginnings and ends of a gene. What these programs miss is "the mythical stuff called

CREDIT: RICK KOZAK

SCIENCE NEWS ON THE WEB

Science's regular news section is not being published this week, but our daily news service, *Science*NOW, will carry expanded news coverage (sciencenow.sciencemag.org).

SEQUENCED ORGANISMS

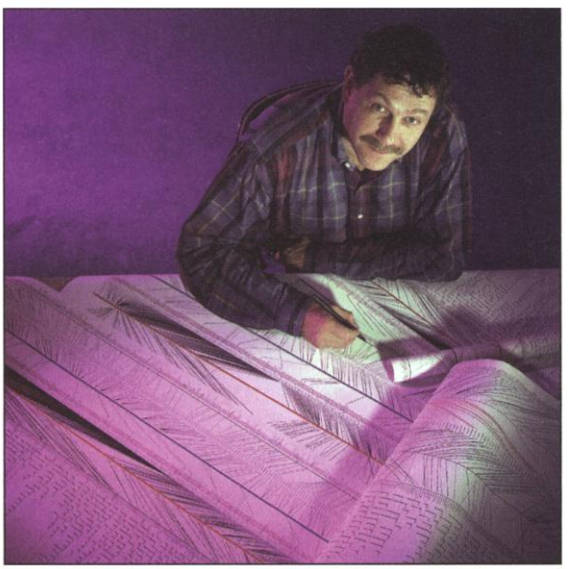
Organism	Genome size	Completion date	Estimated no. of genes
H. influenzae	1.8 Mb	1995	1,740
S. cerevisiae	12.1 Mb	1996	6,034
C. elegans	97 Mb	1998	19,099
A. thaliana	100 Mb	2000	25,000
D. melanogaster	180 Mb	2000	13,061
M. musculus	3000 Mb	–	unknown
H. sapiens	3000 Mb	–	35,000–45,000

dark matter" by the gene predictors, says Birney—genes that are not very active. Gene-prediction software relies on, among other things, catalogs of expressed genes known as expressed sequence tags. But genes that are rarely active would not be detected in most screens of expressed genes. "There could be lots of dark matter, because there is no way to know [how much there is]," says Eric Lander, head of the Whitehead/MIT Genome Center in Cambridge, Massachusetts.

The less mythical genes are showing, however, how fewer genes can yield an organism as complicated as a person. By comparing the human genome with expressed sequence tags and with other genomic and protein data, researchers have figured out that human genes do more work than those in other organisms do—and therein may lie the difference between us and them. Whether in human, worm, or fly, each coding region of a gene is about the same size. Yet human genes assemble these regions in a startling array of combinations. So rather than specify just one protein, as was long believed, each human gene can, on average, spell out three proteins simply by using different combinations of the coding regions, called exons, located within its boundaries. "We're [now] understanding what vertebrate innovation is about," Lander notes.

Proteins are turning out to be more complicated as well. Proteins consist of one or more identifiable domains, sections that have a particular shape or function. After looking at all the proteins potentially encoded in the genome, the public consortium concluded that although humans don't have appreciably more types of domains, they use those domains more creatively, "cobbling more of them together" than do worms or fruit flies, says Collins. Celera's team found this to be particularly true in certain classes, such as structural proteins involved in the actin cytoskeleton and proteins used in signal transduction and immune function.

Another surprise is "the whole architecture of chromosomes, the enormous differences," notes molecular biologist Leroy Hood at the Institute for Systems Biology in Seattle. Adams was particularly intrigued by the distribution of single-nucleotide polymorphisms (SNPs), places on the genome where a certain base varies among individuals. "In some regions, the SNP density is higher than you'd expect, and [elsewhere] it's lower than you'd expect," explains Adams. "There's something going on in the genome" that we don't understand, he adds, that determines why SNPs accumulate in some places but not in others.



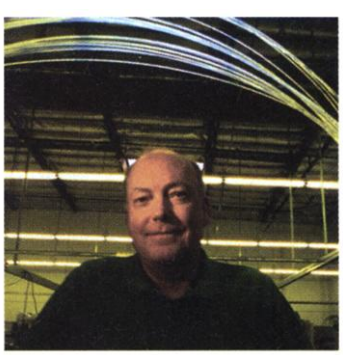
Eric Lander, head of the Whitehead/MIT Genome Center and lead author of the paper published in *Nature*.

Other features also vary across the genome. Regulatory regions called CpG islands that shut down nearby genes are denser in gene-rich regions than in the stretches of geneless DNA. Similarly, researchers are puzzling over why the rate of recombination, in which a pair of chromosomes swap equivalent bits of DNA, differs so dramatically. Parts of chromosome 13 are relatively stable, for instance, whereas chromosome 12 in men and chromosome 16 in women are enormously fickle.

Equally striking is how little of the genome actually codes for proteins and how those exons are distributed. Celera calculates that just 1.1% of the genome codes for proteins; the public figure is 1.5%. That's a sea change from when Fred Sanger, now retired and living outside Cambridge, U.K., did his pioneering work on DNA sequencing in the late 1970s. Then, "one imagined exons consecutively along the DNA," he recalls. That's how bacterial genes are arranged. But human genes contain intervening sequence, sometimes extending thousands of bases, between exons. Not only does this make for big genes, but it complicates the task of gene identification.

Moreover, genes themselves can be separated by vast "deserts" of noncoding DNA, the so-called junk DNA. The term is proving to be a misnomer, however (see p. 1184). Celera scientists estimate that between 40% and 48% of the genome consists of repeat sequences: DNA in which a particular pattern of bases occurs over and over, sometimes for long stretches of a chromosome. One of the more common repeats, called Alu's, cover 288 megabases in the Celera human genome—nearly 10% of the total. And the public consortium's analysis shows that older Alu's tend to concentrate in gene-rich areas, suggesting that those Alu's located near genes may serve some useful purpose and thus were retained by the genome. "It's like looking into our genome and finding a fossil record, [one that shows] what came and went," says Collins.

Among the most common DNA fossils are transposons—pieces of DNA that appear to have no purpose except to make copies of themselves and often jump from place to place along the chromosomes. They typically contain just a few genes—those needed to promote the transposon's proliferation. Both drafts confirm that transposons may also be a source of new genes. Celera found 97 coding regions that appear to have been copied and moved by RNA-based transposons called retrotransposons. Once in a new place, these condensed genes often decay through time for lack of any clear function, but some may take on new roles. And transposon genes themselves become part of the genome. Until recently, 19 of these transposon-derived genes were known. The public consortium just found 28 more. "It almost looks like we are not in control of our own genome," notes Phil Green, a bioinformatics expert at the University of Washington, Seattle.



UNSUNG HERO: MIKE HUNKAPILLAR

Mike Hunkapillar and his team at Applied Biosystems Inc. put the first automated sequencing machine on the market in the mid-1980s. In the late 1990s, Hunkapillar's group at PE Biosystems developed the lightning-speed PE Prism 3700 machine, which was used for all of Celera's sequencing and much of the public project's.

CREDITS: (TOP TO BOTTOM) SAM OGDEN; WILLIAM MERCER MCLEOD

Mysteries remain

For many years, these new texts are likely to suggest more questions than answers. Some questions, including gene number, arise because the incomplete sequence is hard to interpret. But continued sequencing by the public consortium should remedy that quickly, for both the public draft and the Celera version, as the company regularly incorporates new public data. "This is what scientists are supposed to do, look at the data" and revise their estimates as new information comes in, Adams says.

Other questions will persist despite an abundance of information. Both Celera and the public consortium, for instance, tried to determine whether sometime in its early history the human genome underwent a complete duplication similar to what is thought to have happened in plants. Such a duplication could explain why vertebrates have four times as many *HOX* genes, a group of key developmental genes, as do fruit flies. It might also explain why roughly 5% of the genome consists of stretches 1 kilobase or longer that have been copied and pasted, on either the same or a different chromosome, as the public consortium found. By contrast, large, duplicated segments make up less than 1% of the worm genome and less than 0.1% of the fly genome. Even so, the distribution of these human copies makes it hard to imagine that they resulted from a single whole-genome twinning event. "We can't entirely rule it out," says Adams, "but there's not a lot of evidence for a systemic duplication." Instead, duplication may have occurred in bits and pieces over millions of years.

Another head-scratching discovery, made by the public consortium, is that the human genome shares 223 genes with bacteria—genes that do not exist in the worm, fly, or yeast. Some researchers suspect that the ancient vertebrate genome took on bacterial genes, much the way pathogenic bacteria have taken in genes that confer antibiotic resistance. However, "it's not clear if the transfer was from human to bacteria or bacteria to human," Waterston points out.

All this from a first glimpse at the nearly complete genome. Although their analyses occupy several hundred pages in *Science* and *Nature*, both Celera and the public consortium came away knowing that they had only scratched the surface. "It's like a book in a foreign language that you don't understand," says Sanger. "That's the first job, working the language out."

–ELIZABETH PENNISI

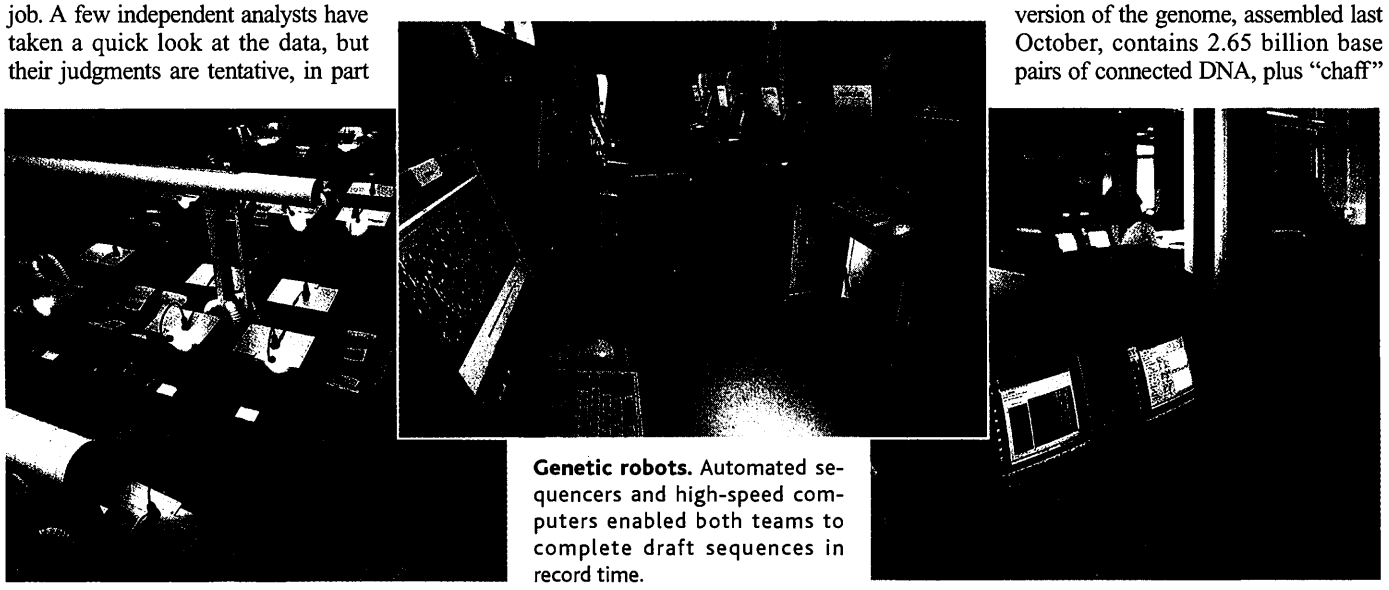
Comparison Shopping

Now that the human genome has come off the production line, researchers are eager to kick the tires and take it out for a spin. They actually have two versions to test drive, one produced with private money and the other with public funds. Naturally, people are asking how the two products compare. Getting an answer to that question, however, may not be straightforward.

Few scientists outside the groups that produced these draft genomes have examined the results side by side. Leaders of the two sequencing groups have written up their own evaluations; not surprisingly, each one concludes that its own team has done a superior job. A few independent analysts have taken a quick look at the data, but their judgments are tentative, in part because these genomes are fast-moving targets and are difficult to pin down. As additional data come in, both research groups are continuing to update their views of the human genome, touting the most recent improvements; the public consortium will continue to release updated drafts, but Celera's updates will be available only to its paying customers. The published reports appearing this week in *Science* and *Nature* represent a freeze of the data as they existed around the first week of October 2000. Given the extraordinary mass of data, it may take several months for molecular biologists to nail down the relative merits of each and get a good fix on their accuracy. Officials at the U.S. agencies that fund genome research are talking about holding a workshop to do just that, possibly on 3 April, but no meeting has yet been scheduled.

Anyone trying to evaluate the two products in the meantime needs to see the data in a format called a whole-genome assembly—a format that hasn't been released on the Web at this writing but will be available by the time the two papers are published. The assembly is a view of the genome that's meant to be as complete as possible: Redundancies in DNA sequence are supposedly removed, large chunks of contiguous DNA are assigned to specific chromosomes, and these chunks are meant to be in the right order and in the right back-to-front orientation.

J. Craig Venter and his crew at Celera Genomics in Rockville, Maryland, authors of this week's report in *Science*, say that their version of the genome, assembled last October, contains 2.65 billion base pairs of connected DNA, plus "chaff"



Genetic robots. Automated sequencers and high-speed computers enabled both teams to complete draft sequences in record time.

CREDITS: (LEFT TO RIGHT) WHITEHEAD/MIT GENOME CENTER; JOINT GENOME INSTITUTE; THE WELLCOME TRUST MEDICAL PHOTOGRAPHIC LIBRARY