GENOMICS

Genome Centers Push for Polished Draft

The genome centers are working flat out, trying to overcome perhaps the hardest challenges yet in finishing the sequence of the human genome

GENOME PROGRESS

The rough draft of the sequence of the human genome provided a tantalizing peek at what's to come. The genome centers are on track to produce a polished draft by spring 2003, but they still must devise clever ways to sequence portions of the genome that knocked computers to their knees. At the same time, researchers comparing drafts of the mouse and human sequences are uncovering a surprising architecture of the genome, causing them to question earlier notions.

▶ **PROGRESS**

ARCHITECTURE

COLD SPRING HARBOR, NEW YORK— Faced with stiff competition from a private company, the publicly funded Human Genome Project 3 years ago pulled out all the stops to finish a rough draft of the sequence. Since then, even though there's been no competition, the consortium's pace hasn't flagged. It will produce a polished version of all 3 billion bases of the human genome by spring 2003—right on schedule, and right on time for the 50th anniversary of the discovery of the structure of DNA, Jane Rogers of the Wellcome Trust Sanger Institute in Hinxton, U.K., reported at the Genome Sequencing and Biology meeting held here earlier this month.

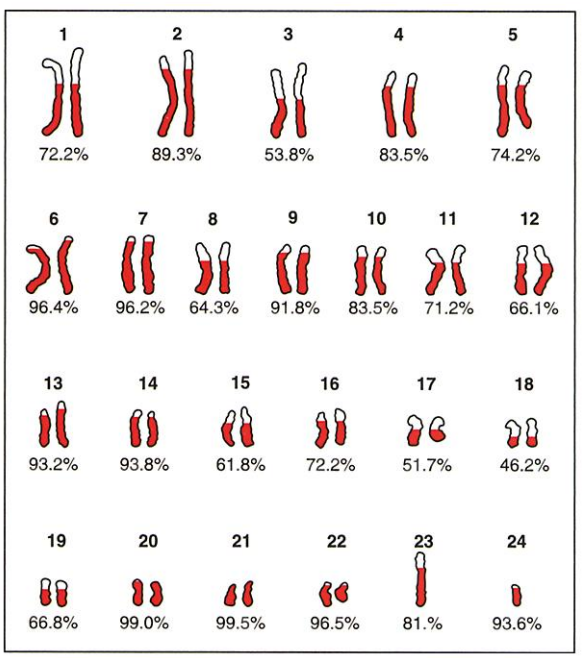
Biologists have been waiting for this prize for more than a decade. Although the draft sequence gave researchers a leg up on finding genes and a preview of what was to come, the finished product promises much more. It's "going to provide a kickoff for a whole lot of interesting science for a very broad set of scientists," says Sanger's Richard Durbin. Not only will the flood of data make it easier for geneticists to track down genes, but it will also help cell biologists understand chromosome structure and developmental biologists unravel how organisms assume their final form.

But the last 5% is always the toughest, as Francis Collins, director of the National Human Genome Research Institute (NHGRI), and others warned when the project began a decade or so ago. Parts of the chromosomes still stubbornly resist efforts to break their DNA code. Other stretches of the sequence are so alike that they are confounding efforts to piece the genome together correctly.

Yet the sequencers are undaunted. "When [the spring 2003 goal] was first mentioned 4 years ago, I was skeptical," admits Rogers. But now she pronounces it "doable."

Four chromosomes are virtually finished, Rogers points out: 20, 21, 22, and Y. Chromosomes 6, 7, 13, and 14 "are in the final stages," she notes. Nine, 10, and some others are about 85% done. And "we've got a year" to do the rest, Rogers adds confidently.

Some attribute this progress to the huge sums of money NHGRI and the Wellcome Trust poured into the big sequencing centers as the race heated up. But Collins also credits his nontraditional management style. Typically, researchers get their grants and proceed with little input or interference from the granting agency. But not the U.S. genome centers. "For a while I was riding herd pretty rough on the centers," Collins admits.



Race to finish. The percentages shown reflect how complete the sequence of each chromosome is.

SOURCE: NCBI

Each of the 16 participants in the international consortium had to agree in writing to complete its share of finished sequence, and Collins has been tracking their progress by monitoring the data deposited in GenBank. Any dawdlers must get another center to take up the slack, but Collins doesn't expect that to happen, "because no one wants to be the weak link." If the centers keep churning out bases at the current rate, he predicts that all the sequencing could be finished by the end of the year. Then all they have to add are the "final touches."

Unfortunately, these "final touches"—which involve assembling all the DNA in the right order without gaps and without mistakes—make sequencing the entire genome look easy. Finishing, as this phase is known, cannot be automated to the same extent as sequencing is automated. What's more, it leaves to humans all the problems that the computers couldn't solve when they tried to assemble the genome.

DNA hiccups

One of the bigger challenges that has confounded the sequencers is repetitive DNA: sections along the chromosomes where the same base or base sequences are repeated once, a dozen, or more times. These reside mostly at the telomeres, or the ends of the chromosomes, and the centromeres, positioned near the middle of each chromosome. For reasons not well understood, it is sometimes difficult or even impossible to copy, or clone, pieces of the DNA from these two regions in bacterial artificial chromosomes, a key first step to sequencing. Those pieces that are sequenced are often misassembled, notes Evan Eichler of Case Western Reserve University in Cleveland, Ohio.

The teams that published the four finished human chromosomes opted for expedience: They simply skipped these regions. But for the finished genome, that won't do. "We've already pushed hard into the unclonable regions," says Collins. "We are really stretching the technology." And a few brave souls have refocused their efforts on new strategies to make the job feasible.

Harold Riethman, a molecular biologist at the Wistar Institute in Philadelphia, Pennsylvania, is tackling the telomeres. Researchers worked out the repeated sequence of the telomere DNA several years ago but have not been able to decipher the DNA nearby: the so-called subtelomeric regions. So Riethman has turned to a special type of yeast artificial chromosome called a half-YAC, which contains just a small amount of yeast DNA relative to the inserted human DNA. Into these half-YACS, Riethman has inserted pieces of

human DNA that butt up against the telomere's signature repetitive sequence.

The half-YACS duplicate these chunks of DNA, providing fodder for sequencers ready to take on these regions. The DNA bits are still difficult to decipher, and the sequence hard to reassemble. The biggest problem is that these subtelomeric regions often contain double, triple, or even more copies of segments from that and various other human chromosomes: Deciding exactly where one fits is tough when several seem to match. Nonetheless, 37 of these subtelomeric regions have been partially sequenced and joined to the appropriate chromosome by collaborators in the sequencing centers, Riethman reported at the meeting. "He has done an outstanding job," says David Haussler, a bioinformaticist at the University of California, Santa Cruz.

The centromeres are proving even more vexing, says the geneticist who has taken them on: Case Western's Eichler. "In general, the pericentric regions are the hardest thing left to do," Haussler says. The centromeres, too, are plagued by blocks of DNA up to 10 million bases long that are a mishmash of duplicated segments from elsewhere in the genome. Despite the challenge, these duplicated regions need to be sequenced, says Eichler: "They can be hotspots for rapidly evolving genes," and they are implicated in some two dozen diseases, including Prader-Willi syndrome and DiGeorge syndrome.

Beyond the centromere

Eichler and his colleagues have identified even more duplicated regions outside the centromeres. These segments might also cause big headaches to sequencers trying to assemble the complete human genome, Eichler's postdoc Vicky Choi reported at the meeting. These duplicated regions, which can be more than 200,000 bases long, make up at least 5% of the genome, and any two duplications can be as much as 99% similar.

To get the finished sequence right, each chunk must be in correct chromosomal location, but the computer can't easily place them; in fact, it often doesn't know they exist. A computer might, for example, superimpose a chunk from one chromosome over a sequence that looks like it on another, or discard it altogether because it looks like something that's already been incorporated into the genome. And if the duplicated segments are located close together on the same chromosome, the computer is likely to "collapse" them: treat them as one and ignore the sequences in between. "This is a very, very important issue in finishing the genome," says Haussler.

Help is on the way from Choi and Eichler. After writing a computer program to ferret out these duplicated regions, Choi compared human sequence data from the Human Genome Project and its rival, Celera Genomics in Rockville, Maryland. She found some 24,000 fragments in which the assembly might have been confused by duplicated DNA and 89 places where two copies have been merged and intervening sequence lost.

Impressed by Choi's and Eichler's talks, Collins wasted no time asking for their help at the meeting. Knowing where the duplications are is the first step to dealing with them correctly, says Collins. "For a long time, assemblers weren't paying attention [to these regions]," says Piu-Yan Kwok of the University of California, San Francisco. "Now there is a simple test to help them detect [duplicated regions]."

These and other efforts should help ensure that the May 2003 version of the human genome meets the agreed-upon criteria for "finished": all the bases in the right order with the sequence running from telomere to centromere to telomere for each chromosome. The only gaps allowed are those that could not be filled after an exhaustive effort was made with "the set of techniques that are currently available," says Rogers. Collins predicts that each chromosome might have a dozen gaps. Although nitpickers might argue that the genome is not finished, says Eddy Rubin, a geneticist at Lawrence Berkeley National Laboratory in Berkeley, California, in reality, biologists will at last have all they need.

–ELIZABETH PENNISI

GENOMICS
ARCHITECTURE

Charting a Genome's Hills and Valleys

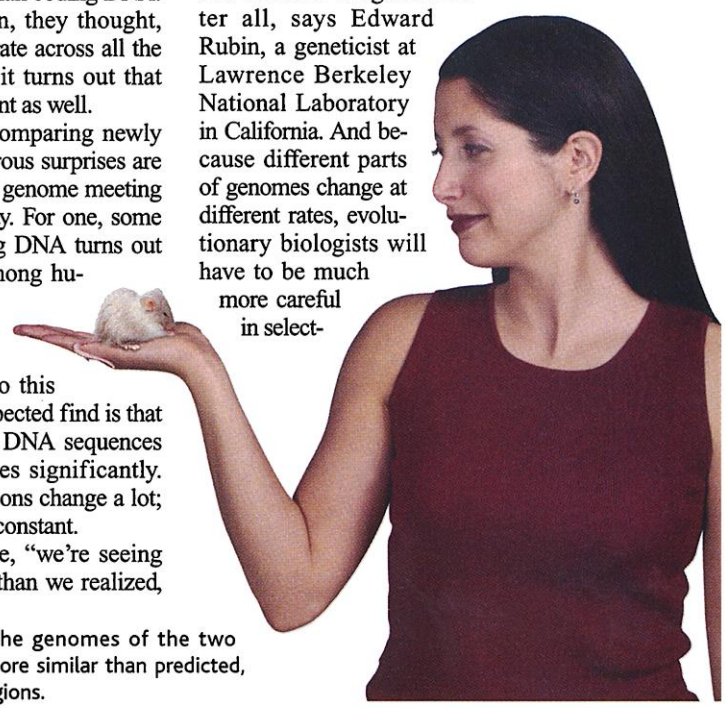
Comparisons of the sequences of the mouse and human genomes have turned up unexpected features

COLD SPRING HARBOR, NEW YORK—Time was when geneticists thought the human genome was quite uniform—consisting simply of genes strung together one after another. Then in the late 1970s, they realized that long stretches of seemingly useless DNA are sandwiched between—and even within—genes. Researchers thought that this intervening DNA constituted a second type of DNA, one that is less essential to an organism's survival and thus likely to accumulate more mutations over time than coding DNA. This accelerated evolution, they thought, would occur at a constant rate across all the noncoding regions. Now, it turns out that they were wrong on that front as well.

As researchers begin comparing newly sequenced genomes, numerous surprises are emerging, as described at a genome meeting here held from 7 to 11 May. For one, some of that "useless" noncoding DNA turns out to be highly conserved among humans and mice (see also Research Article on p. 1661 and Perspective on p. 1617). In addition to this conservation, another unexpected find is that the rate at which different DNA sequences change through time varies significantly. Some noncoding DNA regions change a lot; many others remain nearly constant.

With each new genome, "we're seeing there's more to the story" than we realized, says David Haussler, a bioinformaticist at the University of California (UC), Santa Cruz. Adds Pui-Yan Kwok, a geneticist at UC San Francisco, "Genomes are evolving in a completely nonuniform way."

As a result, biologists are rethinking their views of how genomes operate—and shedding some of their "gene-centric" views in the process. In particular, the high degree of conservation of some noncoding DNA is helping convince them that these sequences are somehow useful to the genome after all, says Edward Rubin, a geneticist at Lawrence Berkeley National Laboratory in California. And because different parts of genomes change at different rates, evolutionary biologists will have to be much more careful in select-



Of mice and (wo)men. The genomes of the two species are proving to be more similar than predicted, particularly in noncoding regions.

CREDITS: G. K. AND VIKKI HART/PHOTODISC; AMOS MORGAN/PHOTODISC