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

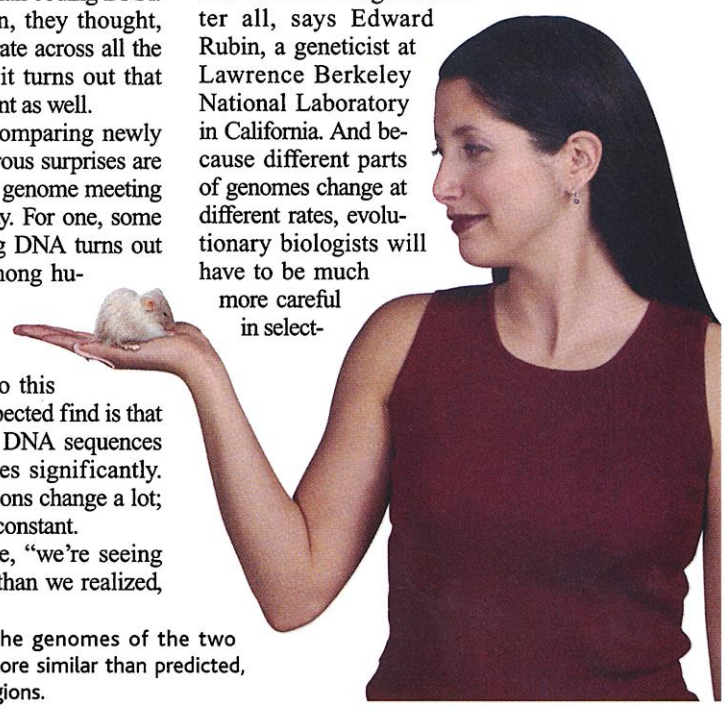
Comparisons of the sequences of the mouse and human genomes have turned up unexpected features

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

ing the DNA they use to evaluate the phylogenetic relationships among organisms.

Deserts and jungles

Some of the new findings emerged once researchers were able to survey the entire landscapes of the mouse and human genomes. Rubin and his colleagues have determined how gene density can vary dramatically in both species. Some regions—"gene jungles"—have a high density of genes, whereas "gene deserts" have very few, if any, over long stretches of DNA. At the meeting, Rubin's Lawrence Berkeley colleague Inna Dubchak reported that the human genome contains 234 gene-poor sections ranging in length from 620,000 to 4 million bases. Together, these deserts comprise about 9% of the human genome, Dubchak said.

The researchers expected little similarity between the gene deserts of the mouse and human, but when they matched the human sequence up with the draft sequence of the mouse, they found that the two species had 178 deserts in common. Jane Rogers, a sequencer at the Sanger Institute in Hinxton, U.K., describes this observation as "the most surprising thing" that's so far come out of comparisons of the two genomes.

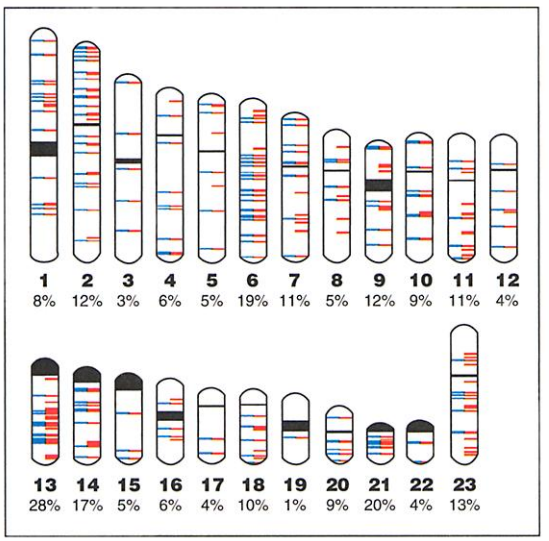
The results have driven home how narrow-minded genome explorers have been. "We've had an extremely gene-centric view," says Rubin. What's more, as Francis Collins, director of the National Human Genome Research Institute in Bethesda, Maryland, points out, the work "implies some sort of chromosome organization that has not really been appreciated." Biologists are also scrambling to identify a function for the noncoding regions. Otherwise, Rubin asks, "what has preserved them over millions of years of evolution?"

He suggests that they might assist in the pairing of chromosomes that takes place prior to cell division. Or they might help keep chromosomes organized. Either job would be sufficiently critical to help keep those sequences relatively constant through the millennia, says Rubin. To find out, he, Dubchak, and their colleagues are in the midst of breeding mice that lack particular gene deserts to study the resulting effects of these genome deficiencies.

However those experiments come out, the close similarities between deserts and genes in the two species threaten to make genome analyses more difficult. Because the typical gene has characteristic DNA at its beginning and end, spotting coding regions is easy—but not trivial.

Researchers had hoped that, with two genomes in hand, they would be able to go beyond just picking out genes to finding the DNA involved in regulating gene expression—DNA thought to be critical in the evolution of new species and in human disease. But such regulatory sequences have been harder to find than expected, Collins explains, because for them "we do not have good signatures."

One problem is that these DNA sequences can lie far away from the genes they regulate. Genomicists thought they would be able to discern some of these regions by comparing mouse and human sequences, as many key regulatory regions should be the same or similar in both. But the presence of so much conserved sequence that's not regulatory DNA, particularly in the deserts, complicates that search.



Shared topography. In the human genome, a subset of gene deserts (red lines) also exist in mice (blue lines).

SOURCE: E. RUBIN/LAWRENCE BERKELEY NATIONAL LABORATORY

Evolutionary tempos

Given the comparisons of mouse and human sequences, researchers must also rethink notions about how genomes evolve. At one time, researchers thought the rate of change tended to bottom out in the genes, making them, from an evolutionary point of view, canyons in the genomic landscape. The DNA in between are the plateaus, with elevated, but supposedly consistent, rates of change through time.

To the contrary, the "plateaus" are pockmarked with accelerations and decelerations in sequence mutations, Robert Waterston, director of the Genome Sequencing Center at Washington University in St. Louis, Missouri, reported at the meeting. The same is true of a few bases within genes—those that don't help define a particular amino acid.

Some preliminary evidence existed for an irregular rate of evolution along noncoding DNA. But those analyses involved small regions of DNA. Only now that researchers can examine whole genomes have they been able to confirm this idea, says Haussler, whose group did much of the analysis that Waterston described.

For this work, Haussler and his colleagues focused on two types of DNA, neither of which are thought to be subjected to any evolutionary forces but instead undergo what is called neutral evolution, changing randomly over time. The first type of DNA consisted of repeats in ancient mobile elements, called transposons, that took up permanent residence in the genome of the common ancestor of the human and mouse many millions of years ago. This provided "an enormous data set, more than 50 million bases," says Haussler. The second type consisted of the last base in the three-base codons for each of eight amino acids. For all of these codons, Haussler says, "the third base is completely free to change" without affecting the identity of the amino acid specified. That means any changes would reflect neutral evolution, even though they occur within genes.

Haussler and his colleagues sought out the two types of DNA along 5-million-base windows of each human chromosome. They then calculated the rate of evolution in each window based on differences in the two species. The researchers found that the rate went up and down along the chromosomes and that these fluctuations were largely consistent whether they were looking at ancient transposons or third bases. "We're seeing strong and distinctive variation," says Haussler. Thus far, no one has found any characteristics in the genome that could explain why this is happening. "It's an absolutely intriguing puzzle."

It is also, Waterston noted, "a complication we didn't anticipate." For one, it will complicate the work of evolutionary biologists, who often attempt to date when new species emerged using so-called molecular clocks. These clocks depend on the relative number of mutations in a species and are based on the premise that they "tick" at a constant rate throughout the history of the DNA. A number of skeptics have questioned the reliability of these clocks (*Science*, 5 March 1999, p. 1435), and the new findings could provide them with new ammunition. For instance, if the genome changes at different rates in different parts of the chromosomes—rates that might also slow down or speed up through time—then molecular clocks might provide erroneous results unless researchers chose the regions they are comparing very carefully.

Researchers have just embarked on their exploration of the full mouse and human genomes, and already long-held notions are being overturned. Just think, says Rubin, of what lies ahead.

–ELIZABETH PENNISI