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

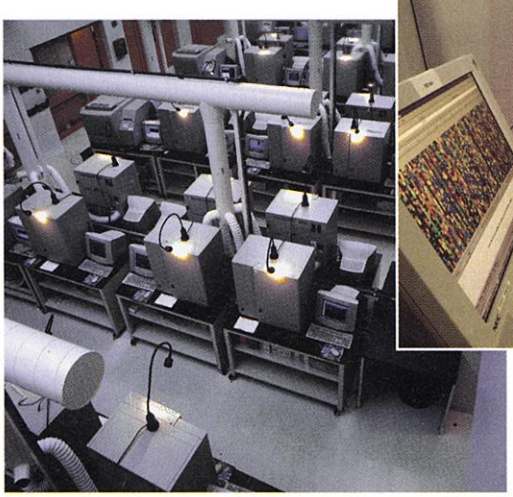
Comparison Shopping

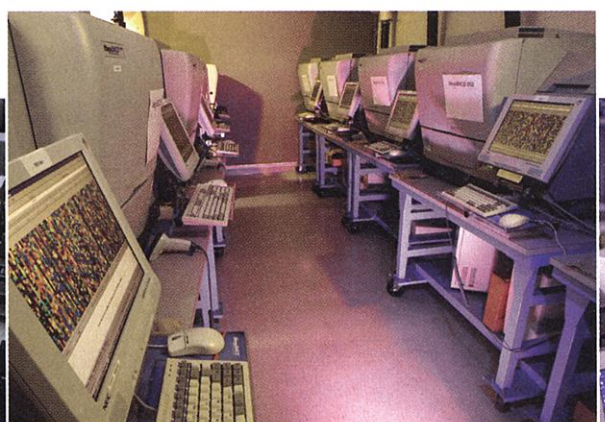
Now that the human genome has come off the production line, researchers are eager to kick the tires and take it out for a spin. They actually have two versions to test drive, one produced with private money and the other with public funds. Naturally, people are asking how the two products compare. Getting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

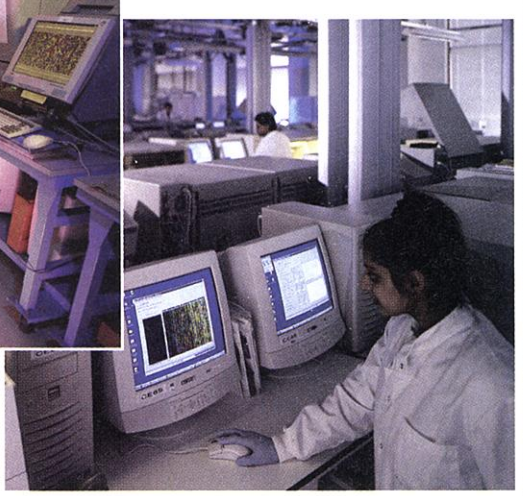
Few scientists outside the groups that produced these draft genomes have examined the results side by side. Leaders of the two sequencing groups have written up their own evaluations; not surprisingly, each one concludes that its own team has done a superior job. A few independent analysts have taken 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

DNA that isn't fully assembled, for a total of 2.9 billion base pairs. Venter calls this version "more than a draft," because he says more of the data are in order and in correct orientation than in the version assembled by the public consortium last fall. Celera is making its October version of the genome available to the public for free, on condition that the data not be used commercially or redistributed, through the company Web site (www.celera.com). The Celera team reports that more than 90% of its assembled genome is in contiguous data assemblies of 100 kilobases or more, and 25% is in assemblies of 10 megabases or more.

The publicly funded team, led by chief author Eric Lander of the Whitehead/MIT Genome Center in Cambridge, Massachusetts, reports in *Nature* this week that its version of the genome contains 2.7 billion base pairs of DNA. Like Celera's version, most of the sequence is in draft form except for chromosomes 21 and 22, which are considered "finished," or as good as they get. Indeed, fully one-third of the genome is in finished form, and Lander's group estimates that the consortium is finishing at the rate of 1 billion bases per year. Like the Celera version, this draft contains more than 100,000 gaps.

The analysis in *Nature* is based on a genome assembly completed on 7 October by bioinformatics experts David Haussler and Jim Kent of the University of California, Santa Cruz (UCSC). This version initially had a problem, though: A computational glitch caused the finished DNA sequences to be "flipped" into reverse orientation. Lander says the glitch affected "less than one-half of 1%" of the data, but he notes that some details had to be corrected in the paper, and he says an improved assembly of the genome was placed on the UCSC Web site (genome.ucsc.edu) on 9 January. The *Nature* paper reports (using an index of contiguity called N50 to describe where 50% of the nucleotides are located) that the public N50 "scaffolds" of assembled data are at least 277,000 bases long. Celera's Gene Myers says the comparable value for Celera's scaffolds is more than 3 million bases.

Although both groups have produced genomes of approximately the same size, they describe the characteristics of their sequences in different terms, which makes a quick and easy comparison difficult. It is not clear how much of the DNA in either assembly is fully contiguous, accurately positioned, or correctly oriented.

To check the congruence of the two genomes, Stanford geneticists Michael Olivier, David Cox, and colleagues used a complex genome map devised in their lab—a collection of "radiation hybrid" clones that break the genome into fragments of known dimensions. With this admittedly imprecise measure, Cox reports on page 1298 that he found that the two versions and the radiation hybrid map differed relatively little. Only 766 unique genetic markers out of a set of 20,874 were not assigned to the same chromosome.

George Church, a genome researcher at Harvard University, also attempted to compare the two genomes. But instead of using the UCSC assembly of 7 October to represent the public version, he used a different assembly made in December by the National Center for Biotechnology Information, part of the National Institutes of Health. Church notes that he was "fortunate" in doing so, because of the glitch in the 7 October data. His report, which appears this week in *Nature*, concludes that the draft assemblies are "similar in size, contain comparable numbers of unique sequences ... and exhibit similar statistics" on the number of active genes.

Researchers are eager to use these draft genomes. But the reviewers urge caution in using either one. As Lander points out, some "misassemblies" of DNA may have been "propagated into the current version of the draft genome," creating potential landmines for the unwary. **–ELIOT MARSHALL**

CREDITS: (TOP TO BOTTOM) ILLUSTRATION: CAMERON SLAYDEN; ERIC GREEN/WHITEHEAD/MIT GENOME CENTER

NEW SCIENCE:

Watching Genes Build a Body

The human genome is touted as the master plan for building an organism. But it is up to developmental biologists to decipher how that "master plan" directs construction.

Traditionally, developmental geneticists have learned how genes control development by altering a gene and observing what goes wrong in model organisms such as the fruit fly *Drosophila melanogaster*, the nematode worm, and the mouse. Complete genomes —the fly, worm, and human are now finished—have simplified the process of locating genes that cause intriguing abnormalities.

But the genomes will also have a more profound effect. Genomics "has completely revolutionized how I think about developmental biology," says Stuart Kim of Stanford University. That's because researchers can now take whole-genome snapshots of cells and tissues, instead of investigating one gene at a time. Kim and his colleagues have

completed 800 microarray experiments recording the relative activity of nearly every worm gene at different developmental stages, in different body parts, and under different conditions. The result, Kim says, is a wealth of information about each of those genes. The problem now is how to make sense of the data avalanche—the team has yet to sort through the nearly 2000 genes that are turned on during development of the genitals, for instance.

Other researchers plan to conduct similar studies on human cells. For example, the biotechnology company Geron, based in Menlo Park, California, has signed an agreement with Celera Genomics in Rockville, Maryland, to analyze which genes are switched on in human embryonic stem cells, the prized cells taken from early embryos that can develop into any cell type. Following gene activity while the cells are still undifferentiated and as they develop into certain tissue types could reveal "the essence of being a stem cell," says Kim.

–GRETCHEN VOGEL

UNSUNG HERO: LAUREN LINTON

Lauren Linton, a former biotech manager, swept into a sluggish Whitehead/MIT Genome Center in 1999 promising to boost productivity 10-fold. Instead, Whitehead rocketed it up 20-fold, becoming the top sequencer in the public consortium. Linton has now left to start her own company.