from the 10X it originally promised in 1998, although the company says its sequence covers 99% of the genome. On the basis of the data presented publicly, it is impossible to verify whether Celera's assembly is correctly orientated and ordered throughout the genome. But Celera has also produced a second map by incorporating data from the public project — which will increase its depth of coverage and allow it to check its shotgun assembly.

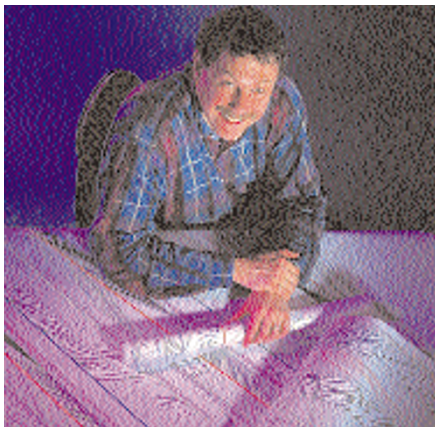
Despite the preliminary nature of both sets of data, the White House has been encouraging the two projects to bury their differences and declare their drafts complete. US politicians were appalled at the media portrayal of the sequencing project as a battle between Celera and the HGP. Clinton's aides hope that a joint announcement will end the rancour and lead to greater public recognition of the achievements of both projects.

The rapprochement between Collins and Venter was brokered by senior figures including Eric Lander, director of the Whitehead Institute at the Massachusetts Institute of Technology, one of the main sequencing centres for the HGP, and Ari Patrinos, head of biological and environmental research at the Department of Energy, who hosted meetings over beer and pizza at his home in Rockville. The agreement has four parts: Monday's choreographed joint announcement; a pledge to publish the two draft sequences, simultaneously but separately, later in the year; a loosely defined plan to hold a joint meeting of the two research teams after publication; and a promise to keep open lines of communication between the HGP and Celera.

At the White House event, Collins struck a spiritual note: "It is humbling for me and awe-inspiring to realize that we have caught the first glimpse of our own instruction book, previously known only to God." Venter was philosophical: "The complexities and wonder of how the inanimate chemicals that are our genetic code give rise to the imponderables of the human spirit should keep poets and philosophers inspired for the millenniums."

For scientists working on the HGP, the main hope is that the task of finishing the genome sequence does not get subordinated to other activities. Parallels drawn with the Apollo lunar programme — which soon fizzled out after the space race was won — provide a warning. "What we most want to avoid is the fate of achieving this heroic goal at such great cost and to the neglect of the long-term goals," says Maynard Olson, a geneticist at the University of Washington in Seattle. ■

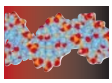
Additional reporting from David Dickson in London.



SAM OGDEN

Lander: helped to broker a genomic ceasefire.

Draft data leave geneticists with a mountain still to climb



Declan Butler and Paul Smaglik

Now the race to obtain a draft sequence of the human genome has been declared an honourable draw, attention will switch to the task of finishing the sequence and 'annotating' the entire genome — characterizing all its genes and working out their functions. The annotation is so formidable that it may need the largest Internet 'collaboratory' yet attempted.

Given that Celera has now stopped sequencing, the task of finishing the genome — in which, to ensure accuracy, each base has been sequenced 10 times over (10X coverage) — will fall to the public Human Genome Project (HGP). In that regard, says Tim Hubbard of the Sanger Centre at Hinxton, near Cambridge, the HGP got a pleasant surprise last weekend, when its data were subjected to a "brute force" computer analysis. Hubbard had expected to find that the HGP had sequenced the genome to an average depth of 5X, but instead, a figure of 7X emerged. This, and the fact that the draft seems to contain fewer gaps than expected, bodes well for finishing the genome ahead of the stated 2003 deadline, says Hubbard.

But annotation poses a much bigger challenge. The first step is to identify all of the protein-coding regions, which will give a good idea of how many genes there are. Most geneticists think the figure lies somewhere between 35,000 and 150,000. Beyond that will come detailed studies of the structure of individual genes, including their regulatory elements, and attempts to assign functions to them.

David Lipman, director of the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, believes that the draft sequence will allow researchers to use computational tools to pinpoint the position of many of the gene fragments catalogued in cDNA libraries of expressed genes. In many cases, it will then be possible to extract an entire gene from the draft sequence — and by comparison with other genes, begin to establish its function. But many biologists are unconvinced. "The current perception is that annotating finished sequence is much less difficult than annotating 'sequence in progress,'" says Richard Gibbs of Baylor College of Medicine in Houston. "And no matter how

biotech stocks represented a conflict of interest. Francis Collins of the University of Michigan appointed as his replacement.

June 1992 Venter leaves NIH to set up The Institute for Genomic Research (TIGR) in Rockville, Maryland. SmithKline Beecham provides $125 million to finance TIGR and to develop its findings commercially through a company called Human Genome Sciences.

July 1992 Britain's Wellcome Trust emerges as a major player in genomics by announcing funding of £50 million for projects including sequencing of the nematode worm *Caenorhabditis elegans*. HGP is by this time a global endeavour, involving government- and charity-funded scientists from many developed nations.

October 1993 NIH and DoE publish revised plan for 1993–1998. Goal set of 80 megabases of DNA sequence by end of 1998. Full completion of human genome sequence set for 2005.

October 1993 Wellcome Trust and UK Medical Research Council open Sanger Centre at Hinxton Hall, south of Cambridge, to sequence the human genome and those of model organisms.

September 1994 French and American researchers publish a complete genetic linkage map of the human genome, one year ahead of schedule.

December 1995 Another collaboration, again led by scientists from the United States and France, publishes a physical map of the human genome containing 15,000 marker sequences.

February 1996 At a meeting in Bermuda, international HGP partners agree to release sequence data into public databases within 24 hours.

April 1996 International consortium announces complete genome sequence of the yeast *Saccharomyces cerevisiae*.

January 1997 NCHGR renamed as National Human Genome Research Institute.

June 1997 TIGR breaks links with Human Genome Sciences following tensions over publication policy.

May 1998 Venter announces formation of a company — later named as Celera — to sequence human genome "within three years". Venter says he will use an ambitious 'whole-genome shotgun' method, but Celera's data access policy will not follow the Bermuda declaration.

BOB BOSTON/WASHINGTON UNI., ST LOUIS

Carry on sequencing: genome scientists, here preparing DNA for analysis, will get no respite.

you cut it, the draft is sequence in progress."

Even with the finished sequence in hand, experience with the two human chromosomes for which this has been achieved — numbers 21 and 22 — indicates that annotating the genome will be a mammoth task. "With 21 and 22 it was not possible to reliably identify and delineate all of the genes," says Philip Green, a biocomputing expert at the University of Washington in Seattle.

In the case of the genome of the fruitfly *Drosophila*, annotation was kickstarted by a two-week 'jamboree' held at Celera. This brought together over 40 academic fly geneticists and 50 Celera scientists, and compared the outcome of dozens of different annotation techniques. This experience should serve Celera well. "We basically trained their annotation team to annotate the human genome," observes Martin Reese, formerly of the *Drosophila* Genome Center at the Lawrence Berkeley National Laboratory in California and now with ValiGen, a company near Paris.

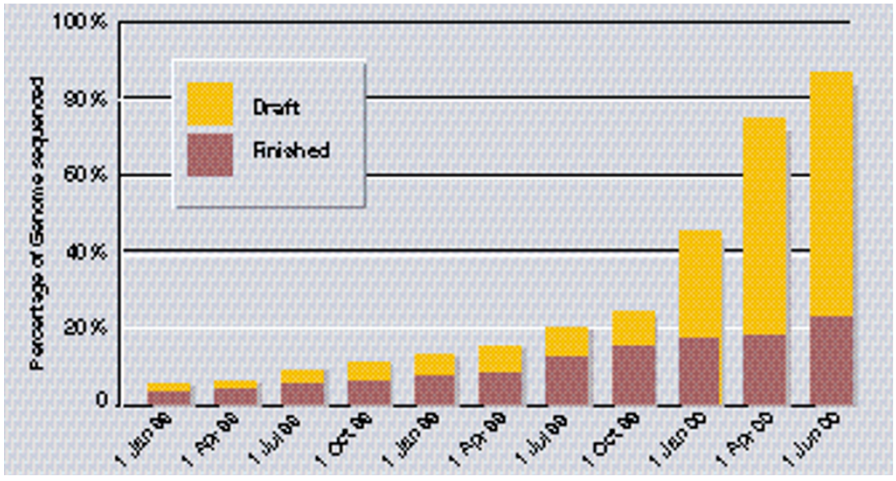
The news that Celera and HGP researchers will hold a joint scientific meeting after publishing simultaneous papers of their draft sequences (see lead story) initially raised hopes of a similar human jamboree. However, as HGP head Francis Collins pointed out to *Nature*, Celera cannot really share its annotation, as it will be its core product for sale to its subscribers. Rather, the meeting is expected to look at discrepancies between the public and private sequences with the goal of 'cleaning up' one another's data.

Celera has said little publicly about its annotation capacity, but it uses specialized software to combine the output of multiple gene finding tools — mostly those available to the public sector. But while Celera's annotation team is at the cutting edge, many experts argue that no single team is currently in a position to annotate the entire genome. "No one really knows how to do it completely," says John Quackenbush of The Institute for Genomic Research in Rockville, Maryland.

On the public side, annotating the genome might mean a rethink on how the HGP's data are organized. Lipman acknowledges that the main sequence database, NCBI's Genbank, has its limitations. "It does not represent what we know of biology at any given time," he says. "It only represents what the author put in." Indeed, while scientists deposit data in Genbank because many journals make this a condition for publication, some do not bother to correct and update it.

"With annotation we will need much more active curation," says Lipman. Many experts believe this may require a 'collaboratory' approach, using the Internet to leverage the talent of biologists worldwide. The NCBI intends to set up a system in which named biologists around the world will 'adopt' a gene or gene family, becoming the curators responsible for gathering information from the wider research community. But Lipman remains against the idea of a free-for-all in which any biologist can annotate the genome — the problem, he says, is that most do not fully understand database syntax, and so tend to make errors when they input data. "What we really want is their knowledge," says Lipman.

The Ensembl annotation project, run by the Sanger Centre and the European Bioinformatics Institute, is plotting a genuinely distributed effort. Hubbard foresees a system where a geneticist in Germany could annotate a gene online, and have his or her interpretation challenged almost in real time by a biologist in Boston. Ensembl's vision has been inspired by a radical suggestion, made by Tom Slezak of the Lawrence Livermore National Laboratory in California and Lincoln Stein of the Cold Spring Harbor Laboratory on Long Island, to use 'Napster' technology for genome annotation. This allows computer users worldwide to share MP3 music files, and could, in theory, let biologists share and annotate genome data (see *Nature* **404**, 694; 2000). If these ideas catch on, the genome project's future could be one of annotation by anarchy. ■



SOURCE: EBI/NCBI

Work in progress: data have accumulated rapidly but the public sequence is far from finished.

May 1998 Wellcome Trust responds by announcing that it will double its support for HGP, taking on responsibility for one third of the sequencing.

October 1998 NIH and DoE publish goals for 1998–2003: one third (1 gigabase) of human sequence and 'working draft' of the remainder of the genome by end of 2003, full sequence by end of 2003.

December 1998 Researchers in Britain and the United States announce genome sequence of *C. elegans*.

December 1998 NIH and Wellcome Trust block proposed collaboration between Celera and DoE, arguing that the terms would conflict with HGP's open data access policy.

March 1999 NIH brings forward planned date for working draft to spring 2000. NIH and Wellcome Trust announce end of 'pilot phase' and start of full sequencing. Costs reduced to 20–30 cents per base.

September 1999 NIH launches project to sequence mouse genome.

November 1999 HGP celebrates sequencing of one-billionth base of human DNA.

December 1999 First complete human chromosome sequence — for chromosome number 22 — published. Celera and HGP discuss possible collaboration.

January 2000 Celera announces compilation of DNA sequence covering 90% of human genome.

March 2000 Celera and academic collaborators release "substantially complete" sequence of the fruitfly *Drosophila melanogaster*, achieved using whole-genome shotgun method.

March 2000 Plans for HGP and Celera to collaborate founder amid considerable acrimony. Data access policy is again the stumbling block.

March 2000 HGP announces successful sequencing of two billion bases of human genome.

April 2000 Celera announces completion of 'raw sequencing stage' of human genome from one individual.

May 2000 HGP consortium led by German and Japanese researchers publish complete sequence of chromosome 21.

June 2000 HGP and Celera jointly announce working draft of human genome sequence. **D. D.**