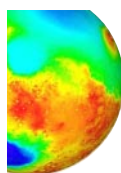




**Catch 22**  
What chromosome sequencing tells us, and which come next  
p448



**Laser inquiry**  
Mismanagement blamed for delays at US laser project  
p449



**Seas on Mars?**  
Spacecraft sends contradictory data on Martian 'ocean'  
p452



**Cabinet post**  
India's top science adviser is champion of the bomb  
p453

# 'Finishing' success marks major genome sequencing milestone ...

## Paris

Genome scientists in Britain and elsewhere have shown that they can surmount the enormous difficulties of closing the gaps between ordered chunks of sequence in the human genome — the 'finishing' step.

The evidence comes with the complete sequence of chromosome 22, published in this issue (see pages 467 and 489), and confirms the achievement as a milestone for the publicly funded international human genome consortium.

Finishing is expected to take as much time and effort as the raw sequencing of the 3.5 billion base pairs of the human genome itself, and the completed chromosome 22 gives a foretaste of what the genome project will yield by around 2003, when all the chromosomes are completed (see over).

In contrast, the draft of the human genome promised for next spring by the international Human Genome Project (HGP) will be only a rough line-up of ordered DNA stretches peppered with gaps.

"The significance [of this week's publication] is that it shows you can get very good finishing using the clone-by-clone approach," says Ian Dunham, head of the



Population of geneticists: Dunham (inset) heads the team that sequenced most of chromosome 22.

international team that sequenced chromosome 22. Dunham works at Britain's Sanger Centre, which did most of the work along with eight other laboratories.

"It is the largest contiguous sequence of DNA, and a major breakthrough," adds Michael Morgan, chief executive of the Wellcome Trust's genome campus, one of the main sponsors of the HGP.

In the consortium's clone-by-clone

sequencing strategy, each DNA fragment is cloned and propagated in a library, by inserting it into the genome of a bacterial artificial chromosome (BAC). The jumbled set of individual BAC clones is then rearranged into a 'physical map', typically by looking for overlapping fragments sharing short sequences of DNA that can be identified using the polymerase chain reaction. These are known as sequence tagged sites.

Once BAC 'contigs' spanning the entire genome have been constructed, the BACs, each about 40,000 to 400,000 base pairs long, are sequenced. The hard part is rearranging the fragments in the order in which they occur on the chromosome.

This finishing involves not just joining up raw DNA data to form contigs — groups of clones representing overlapping regions of the genome — but fixing sequence errors, and filling in gaps between contigs. The latter requires extra data and sophisticated techniques to obtain spanning clones anchored to landmarks either side of the gap.

Several scientists argue that finishing may be the Achilles' heel of Celera Genomics, the private US rival to the HGP, which is pursuing a different genome assembly approach. It is processing the millions of DNA fragments — or reads — directly, with no idea of where they might belong in the genome, and then feeding the results into a supercomputer. Craig Venter, its president,

## ... as researchers pounce on glut of data

### London

"A fantastic resource for looking at human population genetics and disease gene susceptibility in a way that has just not been really possible until now." This is how James Scott, professor of molecular medicine at Hammersmith Hospital in London, describes the availability of the complete sequence of chromosome 22 announced in this issue of *Nature*.

Scott has pioneered the field of messenger RNA editing, and has mapped the genes relevant to his research into cardiovascular disease to chromosome 22.

Benefiting from the early release of sequence data from the public Human Genome Project, Scott has already identified seven new genes at the locus, and is now working on their function and organization.

"We could not have done this work without the chromosome 22 data," says Scott, who welcomes the end to the long delays in searching for genes using positional cloning, which has been made obsolete by the availability of the complete sequence.

Single-nucleotide polymorphisms and linkage disequilibrium studies will make it

possible to study the role of genes in heart disease at the level of individuals and populations, he says. A badly working chromosome 22 is also responsible for the congenital heart disease Di George syndrome, which is caused by the loss of part of the chromosome, 22q11.

Genes linked to schizophrenia, a severe mental illness that affects around one in every 100 people worldwide, also sit somewhere on the chromosome. With the full sequence now openly available, it is unlikely to take long to track them down. **D. B.**

► hopes to use sophisticated software to assemble the enormous jigsaw by matching their sequences to work out which fragment goes where.

Venter has won over many sceptics with his recent completion of the 180-million-base-pair genome of the fruitfly *Drosophila melanogaster* (see *Nature* **401**, 729; 1999). But Philip Green, a biocomputing expert at the University of Washington, points out that the human genome is not only much larger — about 70 million fragments will need to be assembled — but also contains many more families of repeat sequences.

The main problem in sequencing the human genome is that, much more than in lower organisms, it contains many repeat sequences of DNA that look identical and are therefore difficult to place. While Dunham's team have managed to resolve all but 11 of the gaps in the 'euchromatic' region of chromosome 22, they stopped short of attempting to unravel the repeat minefield of the 'heterochromatic' regions.

These are not only difficult to sequence, but are full of repeats. This makes it difficult to resolve the sequence stretches on the short 'p' arm of chromosome 22 from similar repeated heterochromatic regions on chromosomes 21, 13, 14 and 15, for example.

Nonetheless, the success of the clone-by-clone approach in resolving repeats in the euchromatic region has galvanized its supporters. Dunham attributes much of this success to the modular design of the clone-by-clone approach. Given a problem such as a region whose depth of coverage is too low — or which contains gaps or repeats — the modular strategy allows researchers to pull the clone for that region from the fridge and do more work on it. It also allows sequencing to be customized to specific regions, such as GC-rich zones, he adds.

Green argues that the clone-by-clone approach is superior to Celera's whole-genome approach as it reduces the scale of the problem, since each clone will have less copies of any given repeat. Finishing techniques such as chromosome walking are difficult on a single clone, and sceptics predict that they would be intractable at the level of the whole genome.

Celera's approach to finishing lacks this safety net and its effectiveness in the human genome remains largely unknown. Also, Celera's approach means that it will only be able to attempt finishing late in the day, when it has the entire sequence.

"The self-sufficiency of Celera's whole-genome strategy for the human will never be put to the test," asserts David Page, a genome scientist at the Whitehead Institute at the Massachusetts Institute of Technology. "Celera's proposed sequencing of the human genome will fully exploit and be utterly dependent upon publicly available HGP mapping and sequencing data." **Declan Butler**