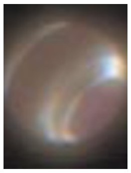
Universities
Canada is to create 2,000 research chairs, says prime minister
p731

Space
NASA and European Space Agency plan science missions
p734

Nuclear tests
US Senators ignore scientific advice on test ban treaty
p735



Neptune
Earth-based cameras put planet in sharp focus
p737

# Venter's *Drosophila* 'success' set to boost human genome efforts

**Paris**

Maverick, dreamer or visionary? The jury is still out on Craig Venter, head of Celera Genomics Systems, whose ambition is to sequence and piece together the entire human genome by 2001 using a controversial 'whole-genome shotgun' strategy.

But Venter, who thrives on his image as the *enfant terrible* of genomics, seems poised to score a victory over sceptics who predict that shotgun sequencing will not work with large, complex genomes.

Last month, Celera finished sequencing the 180-million-base-pair (Mbp) genome of *Drosophila melanogaster*. The company intends to publish the complete genome at the beginning of 2000, in collaboration with the publicly funded Drosophila Genome Project, headed by Gerry Rubin at the University of California at Berkeley.

Does this mean that Venter will repeat the feat in humans? Succeeding with *Drosophila* would be a convincing proof of the feasibility of shotgunning large genomes, but the ultimate test will be the 3,500-Mbp human genome. But Celera is unlikely to sequence the human genome using only a shotgun strategy, as it will almost certainly be quicker to combine this data with conventional maps generated by the publicly funded international Human Genome Project (HGP).

Indeed, if there is one lesson from the *Drosophila* project it is that the collaboration between Celera and Berkeley has speeded up the assembly of the genome and greatly improved the quality of the data obtained.

The Berkeley team have produced 26.5 Mbp of sequence and a low-resolution physical map that will help Celera align its shotgun sequences along the genome. Celera has provided a '10×' sequence of the entire genome, which means that ten bases of sequence have been generated for every base of genome. The completeness of a genome increases with the number of times it is sequenced. The norm is 5–6×, and 10× is the current gold standard. One scientist who has seen Celera's unpublished 10× *Drosophila* data says it is "terrific".

Ironically, in the race to complete the human genome, Francis Collins, director of the US human genome effort, and Michael Morgan, head of genomics at the Wellcome Trust in the United Kingdom, recently revised their goal from a 10× map by 2003 to a 5× draft by spring 2000.

The *Drosophila* sequence is quite an achievement. "The *Drosophila* community is very grateful that this speed has been possible, and the input of Celera made a big difference," says Allan Spradling, an embryologist at the the Carnegie Institution of Washington and the Howard Hughes Medical Institute.

Shattering a genome into fragments for sequencing is easy. Arranging the fragments in the order in which they occur on the chromosome is more difficult. Celera's big achievement is that it seems to be making solid progress in piecing together the 3.2 million sequence fragments of the fruitfly genome.

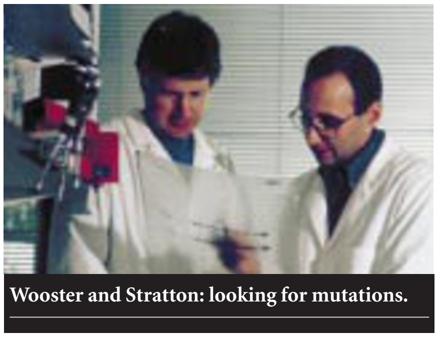
In the conventional 'clone-by-clone' approach being taken by the HGP, each fragment is cloned and propagated in a library by inserting it in the genome of a bacterial artificial chromosome (BAC). The jumbled

▶

# Wellcome funds cancer database

**London**

Britain's Wellcome Trust is to provide up to £10 million ($16.5 million) over the next five years to detect the genes that are mutated in human cancers and release the information to a public database.



Wooster and Stratton: looking for mutations.

ICR

The project, to be based at the Sanger Centre near Cambridge, will use sequence information from the international Human Genome Project — of which the trust is providing one third of the cost — as the basis for screening genomes for the mutations underlying oncogenesis.

The work is "one of the first 'next step', post-genome projects," says Mike Dexter, director of the trust. "Although it will focus on cancer-related genes, it will also act as a 'proof of concept' for ideas that can be applied to most, if not all, polygenic diseases."

The idea of screening cancer cell genomes for mutations was proposed by Mike Stratton and Richard Wooster of the Institute for Cancer Research (ICR) in Sutton. They led the team that located and then identified the *BRCA2* breast-cancer susceptibility gene, based on a sequence from the Sanger Centre.

"The completion of the Human Genome Project will enable us to develop technology that we can use to compare theDNA sequence of normal and cancerous tissue from the same individual and hence identify the mutated genes implicated in tumour development," says Stratton.

Providing substantial support for cancer research is a new move for the Wellcome Trust. But Dexter argues that the basic nature of the work, its direct link to the sequencing project, and its relevance to other polygenic diseases make it suitable for the trust's support. Additional funding (£1 million initially) is being provided from the ICR, and further funding is being sought.

Data will be made accessible to all researchers, in line with trust policy. "Initially we will be focusing on the common cancers, but ultimately the aim is to provide as much data as possible about the broad spectrum of human neoplastic disease," says Stratton, who will head the project at the Sanger Centre. **David Dickson** set of chunks, or 'contigs', are then rearranged into a physical map, typically by looking for overlapping fragments sharing short sequences of DNA.

Once BAC contigs spanning the genome have been constructed, the sequences of the 40- to 400-kilobase-pair (kbp) BACs are compiled, yielding the sequence of each chromosome and eventually the entire genome. The HGP requires building a library of some 30,000 BACs.

Venter is taking a different tack, processing the millions of DNA fragments, or 'reads', directly and feeding them into a supercomputer. Gene Myers, head of Celera's computational biology group, is developing software to assemble the jigsaw by matching their sequences.

Until 1995, the upper limit for such whole-genome shotgunning was 40 kbp. Then Graham Sutton at Venter's Institute for Genomic Research in Rockville, Maryland used the technique to sequence the 1,800-kbp genome of the bacterium *Haemophilus influenzae*. Celera now appear to have pushed the upper limit through the ceiling.

Myers presented data at the eleventh International Genome Sequencing and Analysis Conference in Miami, Florida last month. He claimed that, in a test run on a 6 × 28-Mbp stretch of *Drosophila* sequence prepared at Berkeley using a clone-by-clone approach, a scaffold of contigs generated by his computer assembler gave a perfect match. The only discrepancy was traced back to an erroneous chimaeric clone in the Berkeley data set.

The demonstration has won over some critics of whole-genome shotgunning. "The tests they did were stringent. It is hard to argue with a match like that," says Spradling. "It is impressive." Michael Ashburner, from the European Bioinformatics Institute in Cambridge, says: "I'm now fairly convinced the technique works."

## Difference in complexity

But others remain unconvinced. Philip Green, a biocomputing expert at the University of Washington, says: "If they get it right in *Drosophila* I'll be impressed, but it will not persuade me that they will succeed in humans." This view is shared by both John Sulston, head of the Sanger Centre, in Cambridge, and Francis Collins, director of the US National Human Genome Research Institute.

They point out that the human genome is larger and more complex than that of *Drosophila*. Green says the human genome contains millions of repeat sequences. He is sceptical that Myers will be able to position these, arguing that doing so is easier using a clone-by-clone approach, where each clone will have fewer copies of any given repeat.

But Myers insists that the *Drosophila* demonstration "proves that we are not being confounded by the repeats. We are able to identify all the unique stretches of the genome, assemble them and order them without any mistakes."

Myers: claims his software gives a perfect match, spots repeats and creates a virtual genome.

Myers claims that his software is able to spot repeats in the genome. He also says he has created virtual genomes on screen, incorporating all sorts of nasty repeats, and that his assembler is able to deal with them.

Myers admits, however, that his software — which runs to 150,000 lines of code — is incomplete. He is confident that he can finish the software within weeks.

Much of his confidence stems from the decision by Celera to use a variant of classical shotgunning, called double-barrelled shotgunning. Most sequencing projects collect data from only one end of a DNA fragment but Celera is using sequence from both ends.

If you take a 2,500-bp insert and read 500 bp off each end, you know that you have a 'mated' pair of reads which, in the final assembly, should be roughly 1,500 bp apart and pointing at each other. This is a powerful tool, which in principle allows assembly across repeats.

Myer says: "Normally if I shotgun *Drosophila* I get 3.2 million reads. Period. You know nothing about them. What we get using mates is 3.2 million reads where for a great number they come in pairs which are at a certain distance apart in the genome. That is a significant additional clue about how to do the assembly."

Mates are commonly used to check genome assemblies, but have not been widely used as core assembly strategies because they require a quality of sequence data that has been difficult to achieve using conventional slab gel machines, which typically give ten per cent false pairing of mates. Celera has invested in some 250 capillary electrophoresis systems —Perkin Elmer's ABI Prism 310 Genetic Analyzer — which cut the error rate to between 1 and 0.1 per cent.

But Green reckons that Myers will not be able to solve the problem of large, low-frequency repeats. Ironically, the easiest way for Celera to resolve large repeats will be to use long-range BAC data generated by the HGP. Myers admits that he will probably need the 600,000 BAC end sequences from the HGP to detect these longer repeats. Indeed, to speed things up, Venter will not rely solely on whole-genome shotgunning, but will also use the mapping data flowing from the HGP. Green predicts that Venter will map the HGP reads to the rough BAC contig map.

## Collaboration

Many genome scientists believe that the HGP should follow Rubin's lead and strike a deal with Celera. But when Myers and James Weber, director of medical genetics at the Marshfield Medical Research Foundation in Wisconsin, first proposed a whole-genome shotgun of the human genome in 1997, it was perceived as an alternative, and a threat, to the public clone-by-clone approach.

This has left smouldering resistance to Venter, says Sulston, who argues that the two approaches are complementary. "Myers is doing the human genome in a larger and bolder way and it is excellent that this experiment is being done; but we need both approaches."

Sulston says: "By continuing the robust HGP in the public domain using a clone-by-clone approach we ensure that the thing is produced in a timely and complete way. It is an insurance policy against failure at Celera, and to ensure that the data is made public.

"If Celera combined their data publicly with us that would be the best possible option." But attempts at formal collaboration have stalled, mainly because of the HGP's demand that Venter abide by its rule to release data within 24 hours (see *Nature* **397,** 93; 1999). "We would like full collaboration. We cannot because the sequence must be fully and openly released," says Sulston.

Paul Gilman, head of policy planning at Celera says that "Celera continues to believe that through cooperation we could accelerate the completion of the genome, not a draft, to even earlier than Celera's 2001 completion date."

Venter shipped the *Drosophila* sequence to commercial subscribers to his database on 4 October. The deal with Rubin provides for the data to be submitted to Genbank, perhaps as soon as this month. The annotation of the *Drosophila* genome will also begin next month, when 20 Celera scientists and 30 scientists from the *Drosophila* community meet for an "annotation jamboree".

Rubin says that the annotation data will now also be submitted to the Genbank database on publication of the genome in early 2000. "The entire annotation record will be provided to FlyBase for public release without restriction."

**Declan Butler**