

## THE HUMAN GENOME PROJECT RACE

At the time UC Santa Cruz entered the International Human Genome Project (IHGP), the IHGP was assembling the sequence one piece (or, in the jargon of molecular biology, one “clone”) at a time and intending to string the pieces together based on a precisely constructed clone map. This approach had been shown to work very well with *Caenorhabditis elegans* (a roundworm) and with human chromosome 22. But the process of making sure every last part of the sequence is read and put together properly is quite labor intensive.

The private company Celera Genomics was using an alternative approach, a so-called whole genome “shotgun,” where small bits of sequence are read at random from the genome, and then a computer program assembles these bits into an approximation of the genome as a whole. By using this approach, Celera’s assembly would still have numerous gaps and ambiguities, but the entire project from start to finish could be done in less than half the time the IHGP planned for their effort.

At least partly in response to competition from Celera, the IHGP changed its focus from producing finished clones to producing draft clones. To sequence a clone, the IHGP adopted a shotgun approach in miniature. Bits of a clone were read at random, and the bits were stitched together by a computer program into pieces called “contigs.” After the shotgun phase, a clone was typically in 5-50 contigs, but the relative order of the contigs was not known. This was the state of the genome when David Haussler from UCSC first attempted to locate the genes computationally, and he quickly discovered that computational gene-finding was nearly impossible, since the average size of a contig was considerably smaller than the average size of a human gene.

A number of groups within the IHGP were working on a second stage of assembly that would merge the approximately 400,000 contigs into larger pieces and order them along the human chromosomes, so that the UCSC Genome Bioinformatics group, along with other groups, could find the human genes. This was necessary if the IHGP’s draft sequence was to have similar utility to Celera’s sequence, and in particular to prevent Celera and its clients from locking up significant portions of the human genome under patents. However the second stage assembly turned out to be like an extremely difficult jigsaw puzzle, with many layers of conflicting evidence of contig proximity

and overlap. This slowed the progress of the other teams considerably.

### **PUSH TO THE FINISH LINE**

In May of 2000, UCSC team member Jim Kent dropped his other work to focus on the assembly problem. In a remarkable display of energy and talent, Kent developed in just 4 weeks a 10,000-line computer program that assembled the working draft of the human genome. The program, called GigAssembler, constructed the first working draft of the human genome on June 22, 2000, just days before Celera completed its first assembly. Since the public consortium finished the genome ahead of the private company, the genome and the information it contains are available free to researchers worldwide.

Kent’s assembly was celebrated at a White House ceremony on June 26, 2000 announcing the completion of the first drafts of the human genome by the IHGP and Celera.

On July 7, 2000, after further examination by the principal scientists of the public genome project, the UCSC Genome Bioinformatics Group released this first working draft on the web at <http://genome.ucsc.edu>. The scientific community downloaded one-half trillion bytes of information from the UCSC genome server in the first 24 hours of free and unrestricted access to the assembled blueprint of our human species.

*The IHGP working draft combined anonymous genomic information from human volunteers of diverse backgrounds, accepted on a first-come-first taken basis. The Celera sequence was of a single individual.*