The UCSC Genome Bioinformatics Group:
Opening the Human Genome for Exploration

By Branwyn Wagman, UC Santa Cruz

SANTA CRUZ, CA -- Sept. 22, 2003 -- The assembly and distribution of the first draft of the sequence of the human genome marked a turning point in the history of science and medicine. All of the functions of a human cell are implicitly coded in the human genome, and we have read the molecular sequence of that genome. All people on this planet share a common heritage, forged in the nucleus of our cells over billions of years of evolution. The human genome constitutes a record of the innovations of countless individual life creation events that have come before our own births.

The Genome Bioinformatics Group at UC Santa Cruz played a pivotal role in bringing this extraordinary life script into the light of science and continues to develop improved ways of presenting the genome for the benefit of researchers worldwide. Its brainchild, the UCSC Genome Browser, provides a web-based "microscope" for exploring the human genome sequence and is used daily by thousands of biomedical researchers throughout the world.

The Genome Bioinformatics Group aids the worldwide scientific community in its challenge to assemble the vast amounts of information contained in the genome sequence, to probe it with new experimental and informatics methodologies, and ultimately to decode the genetic program of the cell, laying out the plan for the complex pathways of molecular interactions that it orchestrates.

What is the human genome?

The human genome comprises a sequence of approximately 3 billion component parts, called nucleotides, which are organized into DNA molecules—the famous double helix. The nucleotides, which serve as the alphabet for the language of life, are represented by just four letters: A, C, G, and T, corresponding to adenine, cytosine, guanine, and thymine. The nucleotide alphabet codes for the sequence of amino acids the body will use to build proteins. Combinations of three nucleotides indicate one of twenty possible amino acids (for example, CCT codes for the amino acid glycine), so sets of nucleotide triplets form the instructions that cells use to build proteins. These proteins perform the work of the cells from development
throughout life. A segment of a DNA molecule that codes for one complete protein is called a gene.

UC Santa Cruz and the international genome effort

The International Human Genome Project came to UC Santa Cruz a few days before Christmas of 1999 when Eric Lander, the director of the Whitehead sequencing center, called David Haussler to request his help in annotating the human genome. In particular, he wanted help in discovering where the genes, which make up only approximately 1.5% of the sequence, are located. Haussler had previously developed mathematical models known as Hidden Markov models that had quickly become the dominant methodology in the computer gene-finding field and had been used successfully on the recently sequenced *Drosophila melanogaster* (fruit fly) genome. Haussler brought in Jim Kent, then a graduate student in UCSC’s Molecular, Cell and Developmental Biology Department, along with systems engineer Patrick Gavin and graduate students David Kulp (who had led the gene-finding effort on the *Drosophila* genome) and Terrence Furey. This was the birth of the UCSC Genome Bioinformatics Group.

It was a crucial time for the International Human Genome Project (IHGP). The private company Celera had announced their intention to sequence the human genome well in advance of the public effort. The IHGP was then performing the sequencing 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 nematode worm) 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. Celera decided to use 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 their 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 5-15 contigs, but the relative order of the contigs was not known. Unfortunately the
average size of a contig was considerably smaller than the average size of a human gene, so it was
difficult to extract useful information out of the draft clones.

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 their 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.

In May of 2000, Kent dropped his other work to focus on the assembly problem, and in a
remarkable display of energy and talent, developed a 20,000 line program that assembled the
working draft of the human genome. The program, called the GigAssembler, finished the job on June
22, 2000, just days before Celera completed its first assembly. Kent’s assembly played a key role in
the White House ceremony on June 26, 2000 announcing the completion of the first public working
draft of the human genome as well as the Celera draft. The IHGP working draft combined
anonymous genomic information from human volunteers of diverse backgrounds, accepted on a
first-come, first-taken basis, while the Celera sequence was of a single individual.

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 from
http://genome.ucsc.edu. The genome server at UCSC broadcast one half trillion bytes of
information in the first 24 hours of free and unrestricted access to the assembled blueprint of our
human species.

The UCSC Genome Bioinformatics Group’s focus today

Today the UCSC Genome Bioinformatics Group works to make the human genome
sequence even more useful for science and medicine by identifying and annotating its key
functional elements in such a way that they are easily accessible to researchers. This process of
discovery and categorization is the first stage toward a full understanding of the workings of the
human genome, a project that will occupy science and medicine for many years.

To facilitate the annotation process, Jim Kent and the growing UCSC Genome
Bioinformatics Group constructed the UCSC Human Genome Browser, available at the website
http://genome.ucsc.edu. This web-based tool allows researchers to view the genome at any scale
from a full chromosome down to an individual nucleotide. It is available worldwide without charge, and the web site receives about 4,000 visitors per day, generating in total about 140,000 page requests as they explore the genome. The site has logged visitors from 44 countries throughout the world.

The UCSC Genome Browser integrates the work of hundreds of scientists worldwide, including work generated at UCSC, in an interactive graphical display. The browser displays both experimentally validated and computer-predicted genes along with dozens of lines of evidence that help scientists recognize the key features of genes and guess their function. Extremely fast search software allows researchers to match any DNA sequence to the human genome in seconds, thereby mapping experimental data to the reference sequence. The browser helps unravel the varied splicing patterns whereby one gene can make many different protein products. This varied splicing pattern, which occurs in the human genome but not in the genomes of many other species, may explain how a human gets by with only about twice as many genes as a roundworm.

Now that the human genome sequence is available, comparable genome browser projects have sprung up, most notably those at the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/Genomes/index.html) and at the European Bioinformatics Institute (EBI, http://www.ensembl.org). These two efforts and the UCSC effort complement each other, and they also overlap considerably, leading to friendly competition that has sharpened the efforts on all sides. At the same time, collaborations between these three groups have benefited the scientific community. For example, reciprocal links between the three genome browsers allow researchers to jump from any place in the human genome via a web link to view the same region on either of the other two browsers. Researchers who are hunting for new disease genes and exploring the evolution, developmental regulation, and functional properties of existing human genes report that the UCSC web browser provides the most effective research platform. Apart from the special genome information the UCSC group computes locally, the effectiveness of the browser stems primarily from its intuitive organization and interface and its access speed.

Comparative genomics—seeing ourselves through other species

Besides developing, supporting, and continuing to improve the genome browser, the UCSC Genome Bioinformatics Group conducts research into the functional elements of the human genome that have evolved under Darwinian selection. The UCSC group employs deeper and more accurate probabilistic models to explore the genome than those employed by other groups, allowing discovery with greater sensitivity and specificity. The group has pioneered the use of multiple mammalian species and other kinds of high-throughput experimental data in a probabilistic
framework to analyze entire genomes. This work lays the foundation for future bioinformatic analyses that will incorporate ever more powerful experimental methods to interrogate genome function and pave the way for new understanding and treatment of human disease.

In addition to the human genome browser, the UCSC group and its collaborators continually develop browsers for other species, allowing comparisons that elucidate our own origins. For example, in collaboration with scientists at the Pennsylvania State University, the UCSC Genome Bioinformatics Group has compared the human genome to that of the laboratory mouse. In the approximately 75 million years since these two species diverged from their common ancestor, their genomes have independently accumulated many changes, leading to the two different species we see today. These changes include chromosomal rearrangements, gene duplications, insertion of new genetic material by retrotransposition, deletions of blocks of genetic material, and individual nucleotide substitutions. Reconstructing these changes by computational analysis is giving us a new tool for understanding our own genome.

Just as interesting as the differences between the species is the discovery that short segments in the human genome that have been extremely well-conserved throughout many millions of years of evolution nearly always occur in important functional elements like genes. Darwinian selection has prevented changes in these segments from being passed on by inheritance. These well-conserved segments stand out as small islands amidst a sea of surrounding DNA that is of less functional importance, most of it changing by random genetic drift.

Probabilistic algorithms developed by the UCSC Genome Bioinformatics Group have identified the well conserved elements in the human genome by comparison to the mouse genome and determined that they account for about 5% of the human genome sequence. This is the first estimate of the fraction of the human genome that is actively subject to Darwinian selection. These critical elements of the genome have been made available for experimental validation via the UCSC genome browser, along with a mapping of the major genomic changes between the species. Similar work is in progress using both the mouse genome and the recently completed draft of the rat genome sequence. Most of these well-conserved elements lie outside of known genes, making them fertile ground for new discoveries.

**Processing the data—the KiloKluster**

Processing the huge quantity of data encompassed in the genomes that UCSC presents to the world requires a bank of 1024 Pentium III processors running on the GNU/Linux operating system. The processors for this system, known as the KiloKluster, are housed in 8 racks of 63 compute nodes with an intermediate server,
allowing 128 processors to fit into a single rack. This provides an exceptional amount of inexpensive computing power in a minimal space. The KiloKluster is used principally for genome assembly, analysis, and comparison. To manage the high volume of work, the KiloKluster uses a custom batch scheduling system called Parasol, which Jim Kent designed for this purpose. Once the data have been processed, they are stored in tables that allow easy access at http://genome.ucsc.edu through a system of 8 servers, each with 600 GB of disk storage and 4 GB of RAM.

Possibilities for health—genome research and diseases

As we begin to better understand the molecular mechanisms responsible for human disease, entire new avenues of treatments may open up. We are only now getting a first glimmer of the molecular functions of a healthy human cell or organ, and we are still a long way from understanding the often subtle and complex ways that these can go awry. Yet knowledge of the human genome puts us on the brink of a revolution in medicine.

Research based on genome studies will form the basis for new diagnoses and therapies for human disease that will transform the practice of medicine in this century. We will see an unprecedented individualization of medicine, with therapies tailored to be most effective given an individual’s genetic makeup. Rather than relying on the traditional trial-and-error testing and design of drugs, researchers will increasingly use their knowledge of the molecular causes of diseases to design new therapies, as was done in developing the leukemia drug, Gleevec. It may someday allow us even to repair or replace the damaged genes in bodies, re-orchestrating the molecular pathways needed for health. The UCSC Genome Bioinformatics Group has paved the way for these new breakthroughs by ensuring that the sequence of the human genome is available free and without restrictions in a form useful to scientists, doctors, and students.

The genome in the world

The work of the UCSC Genome Bioinformatics Group has reached deeply into the scientific community. Since it was first published in Nature in February 2001, the scientific paper on the first working draft of the human genome has already been cited more than 2,000 times in other scientific papers. Most of these articles report exciting new discoveries made using the data made available by the human genome project. Additional details of the group’s work are reported in the papers listed at http://genome.ucsc.edu/goldenPath/pubs.html. These include more than 25 additional scientific papers written by group members over the last three years.
The work of the group has received extensive press and popular coverage in major newspapers and in popular books. In addition, a CD made at UCSC containing the human genome sequence was presented in a ceremony to President Clinton in January 2001 and was included in the National Millennium Time Capsule located at the Smithsonian Museum. The project truly found its home in popular culture on December 15, 2000, when a question about the human genome work at UCSC appeared on Final Jeopardy.

The UCSC Genome Bioinformatics Group receives funding primarily from the National Human Genome Research Institute (NHGRI), the Howard Hughes Medical Institute (HHMI), and the California Institutes of Science and Innovation. Even though this funding makes the work possible, the UCSC Genome Bioinformatics Group is ultimately accountable to the entire international community of scientists, doctors and students.

**Ethical considerations**

Aside from the fascinating scientific possibilities, the initial driving concept in the UCSC Genome Bioinformatics Group was ethical: how to make the genome that contains our common heritage available to all, free of patenting and licensing restrictions. This dovetailed with the broader aims of the international project, which sought to involve researchers worldwide rather than focus efforts on just a few centers, based on the conviction that since the genome belongs to all humanity, the work should transcend national boundaries. Having achieved the goal of unrestricted access, the UCSC group continues to consider the additional ethical issues presented by our deepening understanding of our genome and its workings.

Knowledge of the human genome can pose risks as well as benefits for society. Among the shorter term risks are discrimination in employment and health insurance based on genetic information, the possibility that certain groups in society will be stigmatized based on higher incidence of particular genetic alleles within these groups, and the psychological effects on the individual of knowing that he or she has a currently incurable genetic disease that will manifest itself later in life.

The UCSC Genome Bioinformatics group and the agencies that fund it strongly support proposed Federal legislation intended to address the discrimination issues. Beyond legislation, the group seeks to involve as many societal groups as possible in the process of developing and disseminating our new genetic knowledge, including underrepresented minorities and people of varied ethnic heritages, so that it is clear from the beginning that this is an exploration of all of the diversity and commonality of our human heritage, carried out jointly by all people. To help achieve
this goal, the group has applied for and received a supplement to our NHGRI grant that is specially designed to engage underrepresented minorities in our research and educational mission.

Since the psychological effects of genetic knowledge are best addressed by professional genetic counselors, the UCSC group strongly supports the efforts of NHGRI to work with public health agencies to train individuals to provide such services.

One significant risk for the more distant future is that knowledge of our genome, combined with some future technology for modifying or selecting the genomes of our children (not currently available), will create social pressures on parents to try to “enhance” the genomes of their children. This brings up ethical issues that demand widespread discussion engaging all segments of human society. A knee-jerk reaction might be to seek restrictions on research into the human genome sequence. This is happening in a different arena today, where necessary and justifiable bans on reproductive cloning are being extended to prevent research on techniques of regenerative medicine that could have life-saving potential for victims of stroke, spinal injury, and several types of organ failure. When a new technology evokes important social and ethical concerns, as it has for reproductive cloning and as it may one day for genetic enhancement, then all elements of society must be engaged in educated discussion about the issues.

People must learn the appropriate scientific distinctions and understand both the potentials and pitfalls of a new technology or area of inquiry before they make decisions on how broadly or severely it should be regulated. Knowledge of the human genome has a tremendous potential to heal. This potential must be appropriately balanced against the serious issues, like genetic enhancement, that this knowledge may also bring into focus.