

UCSC GENOME RESEARCH PRIMER

All of the functions of a human cell are implicitly coded in the human genome. Now that the molecular sequence of the human genome is known, researchers have begun to mine it for clues as to how the body works in health and in disease.

Besides being the blueprint for life, the human genome constitutes a record of the innovations of countless individual life creation events that have come before our own births. All people on this planet share a common heritage, forged in the nucleus of our cells over billions of years of evolution. Research comparing the human genome with those of other species is already yielding surprising discoveries and confirming long-held ideas.

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

The Genome Bioinformatics Group aids the worldwide scientific community in its challenge to understand 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.

DISCOVERY AND ANNOTATION

While the sequence of the genome is now available, our ability to decode that sequence and tap into the wealth of information it holds is still quite limited. 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 a critical step toward fully understanding the workings of the human genome, a project that will occupy science and medicine for many years.

Genome sequences are difficult to read, because they consist of letter strings with no breaks or punctuation. The example below contains 7 different letters (genomes contain only 4). Can you understand what it is saying?

THATTHATISISTHATTHATISNOTISNO
TISTHATITITIS

With word breaks and punctuation, it starts to make sense:

THAT THAT IS, IS. THAT THAT IS
NOT, IS NOT. IS THAT IT? IT IS!

To facilitate the annotation process, Jim Kent and the growing UCSC Genome Bioinformatics

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 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, contributing to both our physical attributes and many of our less tangible features, such as behavior, learning, and predisposition to disease. A segment of a DNA molecule that codes for one complete protein is called a gene. The human genome is carried on 23 different chromosomes—or DNA molecules.

Genomes of other species contain more or fewer nucleotides and chromosomes but follow the same basic organizational scheme as the human genome.

Group constructed the UCSC Genome Browser. This web-based tool serves as a multi-powered microscope to view all 23 chromosomes of the human genome. The coarse-level view shows early chromosome maps as determined by electron microscopy, then the browser drills down to levels of increasing detail, focusing first on chromosome bands, then on gene clusters (showing known genes—mostly those linked to diseases), then single genes, then components of genes, and finally on the nucleotides—the As, Cs, Gs, and Ts that make up the genome alphabet.

Not only does the browser show the genome sequence, it delineates known areas of the genome and offers supplementary information about the genes—in effect, providing the word breaks and punctuation. The UCSC browser

brings the genome sequence to life by aligning relevant areas with experimental and computational data and images generated in the last decades by scientists from around the world. It also links to databases throughout the world, giving researchers instant access to deeper information about the genome.

The UCSC Genome Browser is available worldwide without charge, and the web site receives about 4,000 visitors per day. In a usual week, the visitors generate 1 million page requests as they explore the genome. The site has logged visitors from 44 countries throughout the world.

Since the first assembly of the human genome, the UCSC group has added a growing number of species to the UCSC Genome Browser, including roundworm, puffer fish, chicken, mouse, and chimpanzee. Along with researchers worldwide, UCSC participates in projects to compare genomes of various species as a way to better understand gene function and the process of evolution.

COMPARATIVE GENOMICS

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 natural selection. The UCSC Genome Browser allows rapid comparisons between species, which can lead to many different types of new discoveries:

- Searching the human genome for sequences that match those with known functions in other organisms can lead to new human gene discoveries.
- The molecular genetics behind disease development and progression in model organisms can be leveraged to discover potential disease-related genes in humans, moving us closer to diagnostic advances and targeted treatments.
- We can reconstruct the evolutionary history of the human genome by identifying the origins of interspecies differences and of

short segments in the human genome that have been extremely well-conserved over millions of years of evolution.

- By searching for the highly conserved segments in the human genome—those that are unchanged from like segments in the genomes of other organisms, we can begin to understand the essential elements of the blueprint for life. Researchers suspect that these highly conserved elements must be essential to function. Genes make up only a small percentage of the unchanged elements, suggesting that other unknown regulatory elements in the genome are also important for function.

POSSIBILITIES FOR HEALTH

As we begin to better understand the molecular mechanisms responsible for human disease, entire new avenues of treatments will be possible. 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.

Rather than relying on trial and error to design and test new drugs, researchers will increasingly use their knowledge of the molecular causes of diseases to design new, targeted therapies. Research based on genome studies will also form the basis for new diagnoses and therapies for human disease that will transform the practice of medicine in this century. The practice of medicine will become much more individualized, with therapies tailored to be most effective given an individual's genetic makeup. Medical tests are already available to identify individual genetic variations that affect a patient's response to commonly used medications. These tests can allow doctors to avoid adverse reactions and choose medications appropriate for specific individuals. Someday we may even be able to repair or replace the disease-causing genes, re-orchestrating the molecular pathways needed for health.