

COMPARATIVE GENOMICS AT UCSC

The UCSC Genome Bioinformatics Group has pioneered a computer-based, probabilistic process for elucidating the genome of one mammalian species by comparing it to the genomes of one or more other species. This works because many functional elements of mammalian genomes have been conserved through and evolved under natural selection. Analyzing the genomes of multiple species along with high-throughput experimental data using a probabilistic model can thus yield predictions that have unprecedented sensitivity and specificity. Such analyses have now shown that it is possible to reconstruct the genome of an ancestral species using predictions based on the genomes of current species.

FACILITATED BY THE UCSC GENOME BROWSER

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

- Searching the human genome for genes that are known in similar organisms, researchers can begin to determine their functions in humans. This research leverages the long history of generational experiments in model organisms such as mice.
 - Comparing genomes and disease progression in similar species may lead to the discovery of new strategies for the treatment and prevention of disease. For example, almost all human genes known to be associated with diseases have counterparts in the rat genome. Biomedical researchers can mine this information for clues that could lead to improvements in the diagnosis, prevention, and cure of disease. It is often possible to study organisms as simple as yeast or fruit flies to determine a gene's function.
 - We can reconstruct the evolutionary history of the human genome—both the origins of interspecies differences and the presence of short segments in the human genome that have been extremely well-conserved throughout many millions of years of evolution. These highly conserved regions are thought to contain the most functionally important elements of the genome. They point to areas where intensified study will lead to a better understanding of how the genome works. Natural 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 appears to be of less functional importance, most of it changing by random genetic drift.
- Inter-species genomic comparisons may help zoologists distinguish subtle differences between species, and may even lead to some reorganization of the evolutionary tree. For example recent genomic research conducted at Wayne State University¹ has led to the suggestion that the two chimpanzee species (*Pan troglodytes* and *Pan paniscus*) should be re-categorized into the same genus as humans (*Homo*).

COMPARATIVE GENOMICS WORK AT UCSC

Besides continually adding species and functions to the UCSC Genome Browser, researchers at UC Santa Cruz are exploring these aspects of comparative genomics:

- The UCSC group uses data from the comparative genomics projects at the NIH Intramural Sequencing Center (NISC) to make improved mathematical models of vertebrate molecular evolution. The project focuses on 50 regions of the human genome that are being studied by the ENCODE project, finding orthologous regions in other vertebrate species. The resulting models can be applied to reconstruct the evolutionary history of each base in the human genome. This work aims to discover both coding and non-coding functional elements.
- Exploration of the function of ultra-conserved regions in the genome and the mechanism for the conservation.
- A critical next step in the development of the human genome as a foundation for biomedical research is the completion of a high quality set of full-length mRNAs with identified coding regions.

With this goal in mind, the UCSC group works on a National Cancer Institute project to develop the Mammalian Gene Collection.

- Participation in genome sequencing consortia, both through assembling and annotating the genomes and through analyzing and comparing them to elucidate the path of evolution.

DISCOVERIES AT UCSC

The completion of the first three mammalian genome sequences, the human, the mouse, and the rat, on the UCSC Genome Browsers facilitates genome-wide comparisons that were not possible before. Researchers in the UCSC Genome Bioinformatics group, in collaboration with scientists elsewhere, compared the 2.5 billion base pairs of the mouse genome and the 2.7 billion base pairs of the rat genome with the 2.9 billion base pairs of the human genome to identify regions of these three genomes that appear to have been conserved through the natural selection process. Through the application of probabilistic algorithms developed at UCSC, these regions have been estimated to account for about 5-6% of the mammalian genome sequence.^{2,3} As genomes for other species are analyzed and compared, more accurate estimates will emerge. Interestingly, most of the well-conserved elements identified lie outside of known genes.

By looking at the conserved 5-6% of the genome, UCSC researchers working in collaboration with researchers from the University of Queensland, Brisbane, Australia, have identified 481 “ultra-conserved” regions of 200 or more DNA bases that are completely identical in the genomes of the three species.⁴ The probability of finding even one such element in the 2.9 billion bases of the human genome is almost nonexistent under a standard model of neutral evolution, where every base is equally likely to undergo independent change. All of the unchanged regions were also found in the dog and chicken genomes, and two-thirds of them were found in the fish genome. But they cannot be traced beyond the fish to sea squirt, fly, or worm. These 481 ultra-conserved regions most often either overlap genes that are involved in RNA processing or reside in the non-coding portions of genes or near genes that are involved in regulating gene transcription or development. This finding has focused attention on determining both the mechanism for this meticulous conservation and the function of these ultra-conserved regions.

RECONSTRUCTING AN ANCESTRAL GENOME

UCSC researchers, in collaboration with researchers at Pennsylvania State University, have shown that it is possible for computers to reconstruct the DNA of the ancestor common to all placental mammals, including humans.⁵ They reproduced part of the genome of this small, furry nocturnal animal that lived more than 75 million years ago, at the time of the dinosaurs. This genome reconstruction promises a better understanding of the evolution of humans and other mammals at the molecular level.



Earliest eutherian mammal reconstruction art by Mark A. Klingler/CMNH

They developed a massive software program to simulate all the known processes that modify DNA as it evolves. They then focused on a small region of the genome called the CFTR locus, which includes the gene involved in cystic fibrosis. This region--encompassing ten genes and adjacent stretches of DNA, for a total of more than one million base pairs of genetic code--had been completely sequenced in many different mammals.

Comparing the simulated evolution process with actual genomes of 19 existing species showed that the process works, yielding 98 percent agreement. Additional genome sequencing will be needed to do a complete reconstruction of the ancestral mammalian genome.

FINDING RAPIDLY-EVOLVING GENES

Computational analysis comparing the genomes of humans, chimpanzees, and other vertebrates have led UCSC researchers to identify elements of the human genome that have undergone accelerated evolutionary changes from one species to the next.

Through this process, they uncovered a gene that has undergone accelerated evolutionary change in humans and is active during a critical stage in brain development (6). Although researchers have yet to determine the precise function of the gene, the

evidence suggests that it may play a role in the development of the cerebral cortex and may even help explain the dramatic expansion of this part of the brain during human evolution.

REFERENCES CITED

1. Wildman DE, Uddin M, Liu G, Grossman LI, Goodman M. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus *Homo*. *Proc Natl Acad Sci U S A*. 2003 Jun 10;100(12):7181-8.
2. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002 Dec 5; 420(6915):520-62.
3. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*. 2004 April 1; 428:493-521.
4. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. Ultraconserved Elements in the Human Genome. *Science*. 28 May 2004; 304(5675):1321-1325, 28 May 2004. First published online 6 May 2004. *Science Express*.
5. Blanchette M, Green ED, Miller W, Haussler D. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*. 2004 Dec;14(12):2412-23.
6. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, Kern AD, Dehay C, Igel H, Ares M Jr, Vanderhaeghen P, Haussler D. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*. 2006 Sep 14;443(7108):167-72. Epub 2006 Aug 16.