



**URL of this page:** <http://www.genome.gov/25521554>



**National Human Genome Research Institute**  
[www.genome.gov](http://www.genome.gov)

## **New Findings Challenge Established Views on Human Genome**

### ***ENCODE Research Consortium Uncovers Surprises Related to Organization and Function of Human Genetic Blueprint***



Cover art: Darryl Leja, NHGRI

**BETHESDA, Md.**, Wed., June 13, 2007 - An international research consortium today published a set of papers that promise to reshape our understanding of how the human genome functions. The findings challenge the traditional view of our genetic blueprint as a tidy collection of independent genes, pointing instead to a complex network in which genes, along with regulatory elements and other types of DNA sequences that do not code for proteins, interact in overlapping ways not yet fully understood.

In a group paper published in the June 14 issue of *Nature* and in 28 companion papers published in the June issue of *Genome Research*, the ENCYclopedia Of DNA Elements (ENCODE) consortium, which is organized by the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health (NIH), reported results of its exhaustive, four-year effort to build a parts list of all biologically functional elements in 1 percent of the human genome. Carried out by 35 groups from 80 organizations around the world, the research served as a pilot to test the feasibility of a full-scale initiative to produce a comprehensive catalog of all components of the human genome crucial for biological function.

"This impressive effort has uncovered many exciting surprises and blazed the way for future efforts to explore the functional landscape of the entire human genome," said NHGRI Director Francis S. Collins, M.D., Ph.D. "Because of the hard work and keen insights of the ENCODE consortium, the scientific community will need to rethink some long-held views about what genes are and what they do, as well as how the genome's functional elements have evolved. This could have significant implications for efforts to identify the DNA sequences involved in many human diseases."

The completion of the Human Genome Project in April 2003 was a major achievement, but the sequencing of the genome marked just the first step toward the goal of using such information to

diagnose, treat and prevent disease. Having the human genome sequence is similar to having all the pages of an instruction manual needed to make the human body. Researchers still must learn how to read the manual's language so they can identify every part and understand how the parts work together to contribute to health and disease.

In recent years, researchers have made major strides in using DNA sequence data to identify genes, which are traditionally defined as the parts of the genome that code for proteins. The protein-coding component of these genes makes up just a small fraction of the human genome - 1.5 percent to 2 percent. Evidence exists that other parts of the genome also have important functions.

However, until now, most studies have concentrated on functional elements associated with specific genes and have not provided insights about functional elements throughout the genome. The ENCODE project represents the first systematic effort to determine where all types of functional elements are located and how they are organized.

In the pilot phase, ENCODE researchers devised and tested high-throughput approaches for identifying functional elements in the genome. Those elements included genes that code for proteins; genes that do not code for proteins; regulatory elements that control the transcription of genes; and elements that maintain the structure of chromosomes and mediate the dynamics of their replication.

The collaborative study focused on 44 targets, which together cover about 1 percent of the human genome sequence, or about 30 million DNA base pairs. The targets were strategically selected to provide a representative cross section of the entire human genome. All told, the ENCODE consortium generated more than 200 datasets and analyzed more than 600 million data points.

"Our results reveal important principles about the organization of functional elements in the human genome, providing new perspectives on everything from DNA transcription to mammalian evolution. In particular, we gained significant insight into DNA sequences that do not encode proteins, which we knew very little about before," said Ewan Birney, Ph.D., head of genome annotation at the European Molecular Biology Laboratory's European Bioinformatics Institute (EBI) in Hinxton, England, who led ENCODE's massive data integration and analysis effort.

The ENCODE consortium's major findings include the discovery that the majority of DNA in the human genome is transcribed into functional molecules, called RNA, and that these transcripts extensively overlap one another. This broad pattern of transcription challenges the long-standing view that the human genome consists of a relatively small set of discrete genes, along with a vast amount of so-called junk DNA that is not biologically active.

The new data indicate the genome contains very little unused sequences and, in fact, is a complex, interwoven network. In this network, genes are just one of many types of DNA sequences that have a functional impact. "Our perspective of transcription and genes may have to evolve," the researchers state in their *Nature* paper, noting the network model of the genome "poses some interesting mechanistic questions" that have yet to be answered.

Other surprises in the ENCODE data have major implications for our understanding of the evolution of genomes, particularly mammalian genomes. Until recently, researchers had thought that most of the DNA sequences important for biological function would be in areas of the genome most subject to evolutionary constraint - that is, most likely to be conserved as species evolve. However, the ENCODE effort found about half of functional elements in the human genome do not appear to have been obviously constrained during evolution, at least when examined by current methods used by computational biologists.

According to ENCODE researchers, this lack of evolutionary constraint may indicate that many species' genomes contain a pool of functional elements, including RNA transcripts, that provide no specific benefits in terms of survival or reproduction. As this pool turns over during evolutionary time,

researchers speculate it may serve as a "warehouse for natural selection" by acting as a source of functional elements unique to each species and of elements that perform the similar functions among species despite having sequences that appear dissimilar.

### Other highlights of the ENCODE work include:

- Identification of numerous previously unrecognized start sites for DNA transcription.
- Evidence that, contrary to traditional views, regulatory sequences are just as likely to be located downstream of a transcription start site on a DNA strand as upstream.
- Identification of specific signatures of change in histones, which are the proteins that organize DNA, and correlation of these signatures with different genomic functions.
- Deeper understanding of how DNA replication is coordinated by modifications in histones.

"Teamwork was essential to the success of this effort. No single experimental approach can be used to identify all functional elements in the genome. So, it was necessary to conduct multiple, diverse experiments and then analyze them using multiple computational methods," said Elise A. Feingold, Ph.D., program director for ENCODE in NHGRI's Division of Extramural Research, which provided most of the funding for the pilot project.

Authors of the ENCODE papers include researchers from academic, governmental and industry organizations located in Australia, Austria, Canada, Germany, Japan, Singapore, Spain, Sweden, Switzerland, the United Kingdom and the United States. The ENCODE project has been open to all interested researchers who agree to abide by the consortium's guidelines.

"Following the Human Genome Project's model of free and rapid data access, we have designated ENCODE as a community resource project. This designation means all ENCODE data were deposited in public databases as soon as they were experimentally verified," said Peter Good, Ph.D., program director for genome informatics in NHGRI's Division of Extramural Research.

The main portal for ENCODE data is the University of California, Santa Cruz's ENCODE Genome Browser (<http://genome.ucsc.edu/ENCODE>); the analysis effort is coordinated from Ensembl, a joint project of EBI and the Wellcome Trust Sanger Institute, at ([http://www.ensembl.org/Homo\\_sapiens/encode.html](http://www.ensembl.org/Homo_sapiens/encode.html)). Much of the primary data have been deposited in databases at the NIH's National Center for Biotechnology Information at (<http://www.ncbi.nlm.nih.gov/projects/geo/info/ENCODE.html>) and EBI at (<http://www.ebi.ac.uk/arrayexpress/>). For more detailed information on the ENCODE project, including the consortium's data release and accessibility policies and a list of NHGRI-funded participants, go to: [www.genome.gov/ENCODE](http://www.genome.gov/ENCODE).

"It would have been impossible to conduct a scientific exploration of this magnitude without the skills and talents of groups representing many different disciplines. Thanks to the ENCODE collaboration, individual researchers around the world now have access to a wealth of new data that they can use to inform and shape research related to the human genome," said Eric D. Green, M.D., Ph.D., director of NHGRI's Division of Intramural Research, which has multiple investigators participating in the ENCODE research consortium.

In addition to contributing to the *Nature* paper, NHGRI intramural researchers authored two of the ENCODE papers in *Genome Research*. The first study, led by Elliot H. Margulies, Ph.D., an investigator in the Genome Technology Branch, analyzed the genomes of 23 mammalian species for all ENCODE targets. This paper details how Dr. Margulies and his colleagues explored the correlation - and, in some cases, lack of correlation - between DNA sequences that are constrained across mammalian evolution and DNA sequences that act as functional elements. In the second study, a bioinformatics team led by NHGRI's Deputy Scientific Director Andreas D. Baxeavanis, Ph.D., along with Laura L. Elnitski, Ph.D., an investigator in NHGRI's Genome Technology Branch, and Tyra G.

Wolfsberg, Ph.D., an associate investigator in the same branch, describes how they built a Web portal that provides simplified access to data from the ENCODE consortium. That portal, called ENCODEdb, is freely accessible to the research community at <http://research.nhgri.nih.gov/ENCODEdb>.

In a related development, NHGRI last month launched a companion project to ENCODE that will identify all functional elements in the genomes of the fruit fly (*Drosophila melanogaster*) and the round worm (*Caenorhabditis elegans*). That four-year effort, dubbed model organism ENCODE (modENCODE), will examine the functional landscape of the smaller, and therefore more manageable, genomes of the two key model organisms, which should aid efforts to tackle such questions in humans. The scientific community relies heavily on the fruit fly and round worm to identify common genes, regulatory sequences and processes that underlie human conditions.

The National Human Genome Research Institute is part of the National Institutes of Health. For more about NHGRI, visit [www.genome.gov](http://www.genome.gov).

The National Institutes of Health - "The Nation's Medical Research Agency" - includes 27 institutes and centers, and is a component of the U.S. Department of Health and Human Services. It is the primary federal agency for conducting and supporting basic, clinical and translational medical research, and it investigates the causes, treatments and cures for both common and rare diseases. For more, visit [www.nih.gov](http://www.nih.gov).

**Contact:**

**Geoff Spencer** NHGRI  
(301) 402-0911  
[spencerg@mail.nih.gov](mailto:spencerg@mail.nih.gov)

**Anna-Lynn Wegener**  
European Molecular Biology Laboratory  
00 49 6221 387452  
[Wegener@embl.de](mailto:Wegener@embl.de)

 [Top of page](#)

---

*Posted: June 13, 2007*