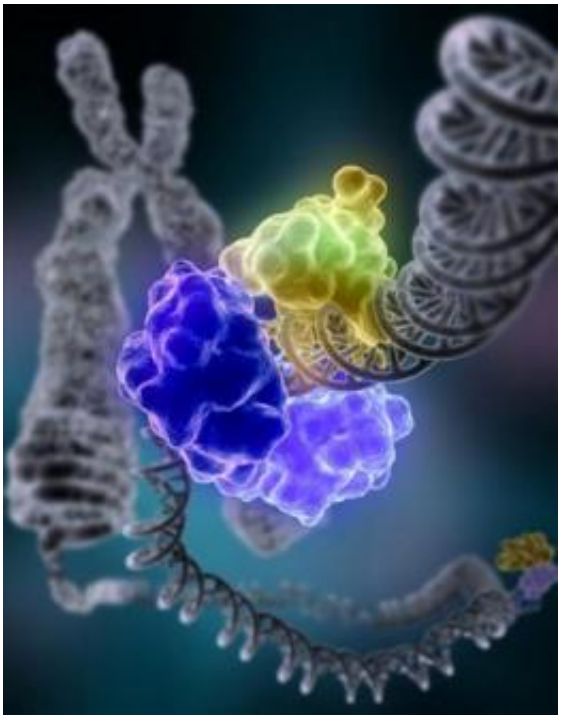
Decoding human genes is the goal of a new open-source encyclopedia

April 19 2011



ENCODE is a massive database cataloging many of the functional elements of the entire collection of human genes -- the human genome. The ENCODE data are being made available to the scientific community and to the public as an open resource, thanks to an international team of researchers. This illustration shows a group of proteins in the process of traveling along a spiraling strand of DNA, a structure comprised of genetic material. A chromosome structure, composed of tightly coiled DNA, is illustrated in the background. Credit: National Institutes of Health

A massive database cataloging the human genome's functional elements -- including genes, RNA transcripts, and other products -- is being made available as an open resource to the scientific community, classrooms, science writers, and the public, thanks to an international team of researchers. In a paper that will be published in the journal *PLoS Biology* on 19 April 2011, the project -- called ENCODE (Encyclopedia Of DNA Elements) -- provides an overview of the team's ongoing efforts to interpret the human genome sequence, as well as a guide for using the vast amounts of data and resources produced so far by the project.

Ross Hardison, the T. Ming Chu Professor of Biochemistry and Molecular Biology at Penn State University and one of the principal investigators of the ENCODE Project team, explained that the philosophy behind the project is one of scientific openness, transparency, and collaboration across sub-disciplines. ENCODE comes on the heels of the now-complete Human Genome Project -- a 13-year effort aimed at identifying all the approximately 20,000 to 25,000 genes in human DNA -- which also was based on the belief in open-source data sharing to further scientific discovery and public understanding of science. The ENCODE Project has accomplished this goal by publishing its database at genome.ucsc.edu/ENCODE, and by posting tools to facilitate data use at encodeproject.org. "ENCODE resources are already being used by scientists for discovery," Hardison said. "But what's kind of revolutionary is that they also are being used in classes to train students in all areas of biology. Our classes here at Penn State are using real data on genomic variation and function in classroom problem sets, shortly after the labs have generated them."

Hardison explained that there are about 3-billion base pairs in the human genome, making the cataloging and interpretation of the information a monumental task. "We have a very lofty goal: To identify the function of every nucleotide of the human genome," he said. "Not only are we discovering the genes that give information to cells and make proteins,

but we also want to know what determines that the proteins are made in the right cells, and at the appropriate time. Finding the DNA elements that govern this regulated expression of genes is a major goal of ENCODE." Hardison explained that ENCODE's job is to identify the human genome's functional regions, many of which are quite esoteric. "The human DNA sequence often is described as a kind of language, but without a key to interpret it, without a full understanding of the 'grammar,' it might as well be a big jumble of letters." Hardison added that the ENCODE Project supplies data such as where proteins bind to DNA and where parts of DNA are augmented by additional chemical markers. These proteins and chemical additions are keys to understanding how different cells within the human body interpret the language of DNA.

In the soon-to-be-published paper, the team shows how the ENCODE data can be immediately useful in interpreting associations between disease and DNA sequences that can vary from person to person -- single nucleotide polymorphisms (SNPs). For example, scientists know that DNA variants located upstream of a gene called MYC are associated with multiple cancers, but until recently the mechanism behind this association was a mystery. ENCODE data already have been used to confirm that the variants can change binding of certain proteins, leading to enhanced expression of the MYC gene and, therefore, to the development of cancer. ENCODE also has made similar studies possible for thousands of other DNA variants that may be associated with susceptibility to a variety of human diseases.

Another of the principal investigators of the project, Richard Myers, president and director of the HudsonAlpha Institute for Biotechnology, explained that the ENCODE Project is unique because it requires collaboration from multiple people all over the world at the cutting edge of their fields. "People are working in a coordinated manner to figure out the function of our human genome," he said. "The importance of the

project extends beyond basic knowledge of who and what we are as humans, and into an understanding of human health and disease."

Scientists with the ENCODE Project also are applying up to 20 different tests in 108 commonly used cell lines to compile important data. John Stamatoyannopoulos, an assistant professor of genome sciences and medicine at the University of Washington and another principal investigator, explained that the ENCODE Project has been responsible for producing many assays -- molecular-biology procedures for measuring the activity of biochemical agents -- that are now fundamental to biology. "Widely used computational tools for processing and interpreting large-scale functional genomic data also have been developed by the project," Stamatoyannopoulos added. "The depth, quality, and diversity of the ENCODE data are unprecedented."

Hardison said that the portion of the human genome that actually codes for protein is about 1.1 percent. "That's still a lot of data," he said. "And to complicate matters even more, most mechanisms for gene expression and regulation lie outside what we call the 'coding' region of DNA." Hardison explained that scientists have a limited number of tools with which to explore the genome, and one that has been used widely is inter-species comparison. "For example, we can compare humans and chimpanzees and glean some fascinating information," Hardison said. "But very few proteins and other DNA products differ in any fundamental way between humans and chimps. The important difference between us and our close cousins lies in gene expression -- the basic level at which genes give rise to traits such as eye color, height, and susceptibility to a particular disease. ENCODE is helping to map the very proteins involved in gene regulation and gene expression. Our paper not only explains how to find the data, but it also explains how to apply the data to interpret the human genome."

More information: www.encodeproject.org/ENCODE/

Provided by Penn State