Microreviews in Cell and Molecular Biology 2012

The ENCODE Project Decoded

Matthew Stailey
Psychology Senior
Department of Microbiology and Molecular Genetics, Oklahoma State University, Stillwater, OK 74078, USA

Key Words:
ENCODE project, chromatin, transcription, transcription factor binding, gene regulation, long-range DNA interactions, DNA mapping

The ENCODE (Encyclopedia of DNA Elements) project, started in 2003, is a consortium of 442 scientists from around the world working together to assign a function to the DNA that does not encode genes. ENCODE used 147 different cell types and many different research techniques to achieve their goal. On September 5, 2012, ENCODE released the initial results of their study. The purpose of this review is to summarize a portion of ENCODE's results.

Introduction

The Human Genome Project, completed in 2003, gave the world a copy of the complete readout of human DNA. This feat was certainly an epic milestone in the scientific community, but the results led to many more questions. It was reported that the portion of DNA that encodes genes only accounts for about 3% of our total DNA. If humans do not have many more genes than mice or trees, what makes us so much more complex? If only 3% of DNA codes for proteins, what does the other 97% of our DNA do? Is the majority of our chemical makeup "junk DNA?"

Thus in 2003, the US National Human Genome Research Institute launched the ENCODE project. The main goal of the ENCODE project was to analyze the human genome and learn whether the noncoding regions of DNA are biologically active. The project analyzed 1640 data sets collected from a team of 442 scientists from around the world. On September 5, 2012, the results were published in 30 articles in several different journals. Not surprisingly, ENCODE caused a lot of excitement in the scientific community. This review aims to summarize some of the finding from the ENCODE project found in the journal Nature.

Recent Progress

Some of the most interesting and significant findings described[1] by ENCODE are as follows. 80.4% of the human genome is biochemically active in at least one of the cell types studied. The research method chromatin immunoprecipitation and high-throughput sequencing (ChIP-seq) was commonly used to determine whether a portion of DNA has biochemical activity. Only 1% of the genome lies further than 1.7 kilo base pairs (kb) away from at least one of these biochemical events.

ChIP-seq is used to find out whether a specific protein binds to a specific DNA sequence. ChIP-seq works by isolating a section of chromatin that is cross linked to a protein, splicing this segment out, and using an antibody specific to the DNA-protein complex of interest. Next, the DNA sequence is amplified using PCR and these samples are sequenced in order to understand which regions of DNA is the target of binding proteins.

The types of cells that were analyzed were broken up into three types of categories according to priority and practicality. The tier 1 cell types were chronic myelogenous leukemia cells (K562 erythroleukaemia cells), a B-lymphoblastoid cell line (GM12878), and a type of embryonic stem cell (H1 hESC). There were 15 types of tier 2 cells, including 5 cancerous cell lines. The tier 2 cells were derived from many different tissue types, including blood vessels, lung and brain tissue, and skeletal muscle. Three hundred thirty six cell types made up tier 3. Tier 3 cells included nearly all tissue types, many of which were cancerous.

Another interesting topic of the ENCODE papers was the accessibility of special regulatory sequences on chromatin. Inside the nucleus, DNA wraps around histone proteins which then cluster closely together in order to coil the DNA. DNA wrapped around a histone protein is called a nucleosome. There are long stretches between histones that are more accessible to enzymes. As a result, the enzyme DNase 1 cleaves stretches of DNA in the

region of DNA that lies between histones. Since these linking DNA areas are about 100 times more sensitive to DNase I than DNA wrapped around a histone, they are referred to as DNase I hypersensitive sites (DHSs). DHSs appear in front of locations of transcription initiation and promoters and are required for gene expression.

Data from ENCODE[2] has furthered our understanding of how DHSs form and transcription is regulated. Transcription factors are proteins that are very specific to certain DNA sites and determine when and if a gene is transcribed. Using ChIP-seq techniques, ENCODE was able to sequence signals from all of the transcription factors used in K562 cells. Results from the data indicate that transcription factors may be responsible for the remodeling of chromatin, thus determining the location of DHSs and where transcription takes place. The recognition sequence for some of the transcription factors is closely linked with chromatin accessibility across all cell types, indicating that accessibility is driven by transcription factor binding. A transcription factor can work alone in this function, or multiple transcription factors can work together to form a complex that can either promote or repress transcription.

Another fascinating finding from ENCODE is regarding the 3D structure of DNA and the implications it has in the regulation of genes[3]. The enhancer is most often thought to target the transcription start site closest to the base pair as written in a sequence. Data from ENCODE shows that less than half of the enhancers target the start site closest in the sequence to it. How can this be? It has to do with the 3D structure of DNA. As mentioned earlier, DNA is wrapped around histone proteins like beads on a string. These beads are then coiled closely together. Therefore, it is entirely possible that an enhancer and the gene it targets could be very far away from each other when the sequence is written out linearly, but they are actually physically quite close to each other. In fact, nearly 50% of transcription start sites have at least one long range interaction with regulatory elements.

One of the ways that these long range interactions can occur is by looping interactions. Looping interactions occur when transcriptional factors bind to different sequences of DNA on the same DHS. The interaction of the transcriptional factors can force the DNA to loop, much like shoe laces, in order to bring a particular promoter or enhancer closer to what it is meant to target. Some of these interactions occur from distances as far as several Mb (10^6 bases) away. Some of these interactions were specific to only one cell type, indicating a method for cell and tissue specialization.

Discussion

Although the initial ENCODE project has provided a huge step forward, our understanding of the genome still has far to go. The research done by ENCODE has only showed us a snapshot of what happens inside a nucleus and it will still take much more work and time before the interactions between genes, proteins, and regulatory RNA are understood. There are still many pressing questions that have to do with the complexity of humans. For example, an onion has more genes than a human, but it is certainly less complex. Why is that?

This microreview only explored a small amount of the data from ENCODE, such as chromatin accessibility and long range transcription factor interactions. Other interesting topics in the ENCODE project data have to do with epigenetic regulation of RNA and histone modification, the use of machines in learning about genetics, the impact of evolution between species and within populations, and the impact these studies will have on disease research. The ENCODE project is like a map of the genome, highlighting areas that areas that effect gene expression, and this data has been made freely available to anyone.

There is no doubt the ENCODE project has massive implications. The reported data is changing the way that scientists look at the human genome, particularly in regulation of transcription. So far, 80% of the genome has been shown to perform some sort of biochemical activity in experiments that studied less than 10% of the different types of cells. Perhaps it won't be long before 100% of the genome is assigned some sort of activity. Thanks to ENCODE, even more exciting events are on the horizon for modern researchers and scientists.

References

[1] The ENCODE Project Consortium "An integrated encyclopedia of DNA elements in the human genome" *Nature* **489**, 57-74 (2012)

[2] Thurman, R. E. *et al.* " The accessible chromatin landscape of the human genome" *Nature* **489,** 75-82 (2012)

[3] Sanyal, A., Lajoie, B., Jain, G. & Dekker, J. "The long-range interaction landscape of gene promoters" *Nature* **489** 109-113 **(**2012)

