Clinical implications of the ENCODE project

L. Sastre

Received: 2 October 2012 / Accepted: 3 October 2012 / Published online: 11 October 2012
© Federación de Sociedades Españolas de Oncología (FESEO) 2012

The first results of the ENCODE (Encyclopedia of DNA Elements) project have been recently presented in 30 articles published in *Nature*, *Genome Research* and *Genome Biology* (see for example the article published in *Nature* 489, pp 57–74, by the ENCODE Project Consortium, related articles in the same issue of *Nature* and references therein). These articles present the results obtained by the collaborative work of 32 research groups and more than 440 scientists during the last 5 years. The ENCODE project arose as a consequence of the human genome sequence published 11 years ago. Analyses of the human DNA genome showed that only 1.2 % of the nucleotides coded for about 20,000 proteins present in our organism and opened the intriguing question of what is the function of the remaining 98.8 % of the DNA. The ENCODE project was aimed to answer this fundamental question and now reports the biochemical function for 80 % of the genome. It is remarkable that most of the regions characterized participate in the regulation of gene expression that is central to embryonic development, cell differentiation, cellular homeostasis and physiopathological processes.

One of the objectives of the ENCODE project was to identify all the RNA molecules transcribed from the DNA and to determine all the transcription initiation sites. Surprisingly, about 60 % of the genome is transcribed into RNA, including intron and intergenic regions. Many DNA regions are transcribed from both strands, originating different RNA molecules. These studies have allowed the discovery of new classes of RNA molecules, which mainly

participate in the regulation of gene expression. The project has also analysed the structure of the chromatin and the DNA-binding site of a large number of proteins involved in the regulation of chromatin structure and gene expression, such as different histone isoforms, modified histones (acetylated or methylated), components of the RNA polymerase complex and transcription factors. These studies were based on the use of genome-wide chromatin immunoprecipitation methods. The use of massive sequencing techniques has also been crucial for the identification of different RNA molecules, chromatin structures and protein-binding sites. All these different techniques were initially applied on a region of 1 % of the genome, in a pilot project, and scaled up to the rest of the genome from 2007. Gene regulation is specific to each of the thousands of cell types present in our organism so that these studies were made in 147 cell types to cope with some of this diversity. There is also considerable diversity in the number of chromatin modifications and transcription factors that operate in the different cell types. Therefore, 13 chromatin-associated variations, 14 RNA polymerase and basal-transcription factors and 87 sequence-specific transcription factors were analysed, in addition to DNA methylation and DNase I hypersensitive sites. In total, 1,640 genome-wide data sets were prepared and analysed in the project.

The results obtained in the ENCODE project offer a new perspective on the structure and function of the genome. It becomes clear that a very significant part of the genome is transcribed and that most of this transcription is related to gene expression regulation. The project has also provided an estimate number of the variants generated for each protein-coding gene, on average 6.3 alternatively spliced transcripts per locus. Mapping of transcription start sites has identified 62,403 sites, about 44 % of them are close to the 5' end of messenger RNAs but the rest might

L. Sastre (✉)
Instituto de Investigaciones Biomédicas, CSIC/UAM, Centro de
Investigación en Red de Enfermedades Raras (CIBERER),
Madrid, Spain
e-mail: lsastre@iib.uam.es

correspond to the start site of novel types of RNA. The data obtained on DNA methylation, DNase I sensitivity and protein-binding studies have generated a huge amount of information on the mechanisms of gene expression regulation. Thousands of regulatory regions, including enhancer regions, have been identified. Comparison of the determined protein-binding and chromatin-structure patterns allowed the identification of molecular signatures characteristic for different regulatory and structural elements. Integration of the multiple data has been used for the development of predictive models of gene expression. Actually, this accumulation of genome-wide information is changing our view of gene transcription regulation.

A very important implication of the ENCODE project is that now it is possible to assign a function to many regions of the genome that were previously completely unknown. The relevance of this change is shown by the fact that many nucleotide polymorphisms or small insertions and deletions related to diseases or disease predisposition map to previously uncharacterized intergenic regions. This has been an important constrain in the many genome-wide association studies (GWAS) carried out in the last years, trying to link DNA sequences variations with specific traits and diseases. The information provided by the ENCODE project makes now possible to assign a function to many of the

polymorphisms identified in non protein-coding regions of the genome which will surely have a large impact on the prognosis and diagnosis of these diseases and might even open the way to new therapeutic strategies.

The new genome landscape offered by the project can also affect the design of new studies on the genetics bases of hereditary diseases or cancer. Because of technical and economic reasons, many of the studies presently made are focused on the sequencing of the protein-coding regions of the genome, the exome or the transcriptome. However, GWAS studies have shown that almost 90 % of the associated variants fall outside protein-coding genes and would not be detected in many of the present studies. The availability of cheaper deep-sequencing methods, together with the functional annotation of many genome regions provided by the ENCODE project, will probably contribute to a significant increase of genome-wide studies in cancer and hereditary traits and diseases that can provide an extensive view of all the possible nucleotide polymorphisms involved in these pathologic processes.

The data generated by the ENCODE project are public and can be explored on line through a visualization tool designed by the *Nature scientific journal* (<http://www.nature.com/ENCODE>).