International J. of Healthcare & Biomedical Research, Volume: 1, Issue: 1, October 2012 P: 1-5

Editorial Review: Encyclopedia of DNA Elements (ENCODE) Project: A Major Scientific Milestone.

Dr. Prathamesh Kamble*, Dr. Motilal Tayade**, Dr. Shital Maske#,
Dr. Kirankumar Jadhav*, Dr. Sunil Bhamare*.

Assistant Professor, B. J. Govt Medical college, Pune, ** Assistant Professor, Rural Medical College, PIMS , Loni , # Assistant Professor, RCSM Govt Medical college, Kolhapur. Corresponding author: Dr.Prathamesh Kambale .Contact no: 09423295943

ABSTRACT: On 5th September 2012, one of the greatest milestones in the field of science and ultimately the mankind was passed. On this day **ENCODE** project, that is, **Encyclopaedia of DNA Elements**, project have released their initial but path breaking results. ENCODE project is one of the most critical projects launched by US-NHGRI. These results have forced the scientists all over the world to rethink the majority of views about genetics. Some of these views are what is a gene? What is the definition of gene? Is there any 'Junk DNA' in human genome? Also it made the scientist to volte-face about the gene and disease relationship. These results published recently are from the initial phase of ENCODE project and this project is still far from complete. But, these results seem to be promising and pivotal in understanding of human genetics. This ENCODE project recently have gained lot of attention of scientist and genetic researchers. This present review is an attempt to highlight the goal, features, findings and the future scope of this project.

KEYWORDS: ENCODE, Encyclopedia of DNA elements.

Introduction: The Encyclopaedia of DNA Elements (ENCODE) is a public research consortium. [1] That means it is an association and joint efforts of many scientists, scientific groups, laboratories all over the world. This ambitious project was launched in September 2003 by the US National Human Genome Research Institute (NHGRI) [2, 3, 4] and it is the second most critical projects launched by NHGRI after successful completion of Human Genome Project. ENCODE project was designed to pick up where Human Genome project left off. The goal of ENCODE project is to search all functional elements which make up human genome and to make a blueprint of human biology. To understand this first we will discuss about the Human Genome Project.

Human Genome Project (HGP) was an international research project began in October 1990 with the ultimate objective to sequence the human DNA. That means to determine the sequence of nucleotide base pairs (A, T, G and C) which make up human genome. In May 2006, this project was declared to be complete and approximately 20000 to 25000 human genes were identified, mapped and published. [5] HGP provided us the read out of our DNA; the underlying code for human life. Sequencing human genome created lots of raw data but it has given us very little information about how this genetic blue print is used in the body, how it is organized and most importantly how it is controlled? To address this issue NHGRI launched further continuation of HGP in the form of ENCODE project.

Human Genome Project was like listing different ingredients required to prepare a tasty and mouth-watering dish. But simple list of ingredients will not be sufficient to prepare a gourmet meal but the real challenging task is to blend the ingredients in precise amount and controlled manner. The primary objective of ENCODE project was similar to a recipe of human genome. It was aimed to enlist the functional elements in human genome and then to study its organisation, use and control system in the human body.

Introduction and Project overview of ENCODE :

In the present century, human genome data interpretation is one of the leading challenges of scientists all over the world. To take up this challenge, in 2003, NHGRI started this most ambitious project - The Encyclopaedia of DNA Elements. They organised a consortium with involvement of more than 450 consortium members, 32 institutions and hundreds of scientists around the world. [6] This project has completed its Pilot phase and Production phase is on board.

Pilot phase:

The pilot phase was started to study different DNA techniques for use in later stages with budget of $12 million. All the existing DNA techniques were used to analyse about 1 % portion of the human genome that is approximately 30 million base-pairs. This phase was spanned from 2003- 2007. Under this study phase, 50% of the sample area was selected manually whilst remaining 50% was selected at random. [7] Methods used for evaluation were mainly chromatin immunoprecipitation (ChIP) and quantitative PCR. The results of these analyses were evaluated based on their ability to identify regions of DNA which were known or suspected to contain functional elements. The pilot ENCODE project released all the data into public databases rapidly. [8] In 2007, pilot phase was successfully finished. The results were published in a special issue of Genome Research [9] and in Nature [3].

Production phase:

In 2007, the ENCODE project was expanded to study complete human genome. [10] New advanced technologies were added to gain higher accuracy, time and cost effectiveness. Some of these assay methods are ChIP-seq, DNase I Hypersensitivity, RNA-seq and assays of DNA methylation. For this production phase, NHGRI reorganised this project as open consortium and awarded grant more than $ 80 million. Also they established a system for sharing functional genomic data. [11]

On 5th September 2012, ENCODE project published 30 papers in several journals and released their extensive set of results. The publication included Nature journal 6 papers, genomic biology 18 papers and genomic research 6 papers. [12] These results are based on 1640 genome wide data prepared from 147 cell types.

Most striking new findings of ENCODE:

- **Junk DNA:** Until recently, it was thought that 80% of human DNA is a 'Junk DNA' that is it has no functions. These genes are probably the one which were maintained during the process of evolution. But recent ENCODE consortium reports could assign functions to more than 80% of human genome. The previously non-functional or overlooked regions in the genes are now known to be filled with regulatory proteins like promoters, enhancers and RNA transcripts. They have important role in regulation of genetic expression. [13]
- **Redefining gene:** Gene is a stretch of DNA that is transcribed to make a protein; this was a simplistic view of gene until recently. But now according to ENCODE consortium results, each gene can be transcribed in different ways. There are lot of transcripts for expression of a protein and there is high overlapping between the transcripts. Some transcripts are connected to previously unconnected genes. This finding has fundamentally changed the concept of gene. Thus these findings forced us to rethink and change the definition of gene. [14]
- **DNase I hypersensitivity assay** is a marker of regulatory DNA. Studies based on this assay have identified a comprehensive map of DNase hypersensitive sites and regulatory DNAs. This map identifies nearly 3 million binding sites of transcription

factors. They have doubled the number of known recognition sequence for DNA binding protein in human. [15, 16]

- **Hierarchies of transcription factors:** ENCODE consortium have presented hierarchies of transcription factors and intertwining network of these transcription factors. This has added a lot to our present understanding of principle behind the wiring of transcription factor networks. [17]
- **Long range signals:** Until recently the concept of regulatory genome was that it lies in close proximity of gene to be expressed. But now ENCODE consortium results proves that it is just the oversimplification. They could map more than 1000 long range signals in each cell type. [18]
- **Genome wide association studies:** For last few decades, researchers are attempting to understand the relation between the disease and genetics. Over these days, they have collected a long list of SNPs that correlate with different diseases. [19] The ENCODE consortium have mapped all these SNPs and many more in their data depository. Ahead of that they have found that only 12 % of these SNPs lie within coding region of gene. 60% or more of disease associated SNPs lie within the region of gene which was previously thought to be 'Junk DNA'. This region is noncoding region but function containing promoters and enhancers. This has provided fresh leads to our understanding of gene and the disease. [20]

Even though these results from the 2nd phase are already published by ENCODE consortium in various journals in 30 publications. But still the ENCODE data is vast. It is not possible to contain all the data in research papers and then to retrieve and use it easily. To tackle this problem ENCODE team have developed a new method on ENCODE portal site. On this site one can choose among the 13 topics of his interest. Then reader just needs to follow the online thread further to get all the information linked with it. This system has made the retrieval of required data really easy.

ENCODE project: Future challenges

- The most challenging feature of ENCODE project is achieving complete coverage of all functional elements in human genome. Human body contains hundreds of distinct cell types. Each cell type has a different set of genome which shows uniqueness in their expression.
- Another major challenge is to study dynamic or real time aspect of regulation of gene expression. Present day study assay methods provide information about gene regulation at a point of time. But for comprehensive understanding of genetic expression; there is need for real time assay methods.
- There is need to develop new technologies which can provide us information about dynamic phase and simultaneous capture of multiple sites of gene regulation network.

Further challenge is to study the different life processes in our life starting from birth to ageing and finally death. At each stage of life genetic expression and its regulation is different in each cell type in the body. To study all these processes poses the greatest challenge.

As this list of future challenges, range of experiments and scope of this project are continuously expanding and it will further keep on expanding in the future. There are high chances that this project could unfold endlessly. Some scientists have already raised the concern that this project could go on forever. Therefore we should answer the question, where to stop?

CONCLUSION:

The overall impact and importance of ENCODE project consortium can be assessed only after complete assembling of human genomic data. However, already the initial results of this project have dramatically transformed and enhanced our concept and understanding of genomics and its regulation. The reference data sets are already been used by many scientist worldwide. Therefore ENCODE project is indeed a major scientific milestone!

REFERENCES

1. Maher B. ENCODE: The human encyclopaedia. Nature. 2012. 489 :7414; 46–48.
2. Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, Cline MS, et al. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Res. 2012;40(Database issue):D912-7.
3. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447 (7146): 799–816.
4. Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, et al. EGASP: The human ENCODE Genome Annotation Assessment Project. Genome Biology. 2006; 7: S2.
5. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001; 409 (6822): 860–921.
6. Birney E. The making of ENCODE: Lessons for big-data projects. Nature. 2012; 489(7414):49-51.
7. ENCODE Pilot Project: Target Selection". The ENCODE Project: ENCyclopedia Of DNA Elements. United States National Human Genome Research Institute. 2011-08-01. Retrieved 2011-08-05.
8. ENCODE Project at UCSC". University of California at Santa Cruz. Retrieved 2011-08-05.
9. Weinstock GM. ENCODE: more genomic empowerment. Genome Res. 2007; 17(6):667-8.
10. Genome.gov/ENCODE and modENCODE Projects (http://www.genome.gov/ 10005107) . The ENCODE Project: ENCyclopedia Of DNA Elements. United States National Human Genome Research Institute. 2011-08-01. http://www.genome.gov/10005107. Retrieved 2011-08-05.
11. National Human Genome Research Institute-Organization http://www.nih.gov/about/almanac /organization/NHGRI.htm).The NIH Almanac.United States National Institutes of Health http://www.nih.gov/about/almanac/organization/NHGRI.htm . Retrieved 2011-08-05
12. Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, Segal E. Genomics: ENCODE explained. Nature. 2012; 489(7414):52-5.
13. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57-74.
14. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, et al. Landscape of transcription in human cells. Nature. 2012; 489(7414):101-8.
15. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489(7414):75-82.
16. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012; 489(7414):83-90.
17. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, et al. Architecture of the human regulatory network derived from ENCODE

18. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012; 489(7414):109-13.
19. Mardis ER. A decade's perspective on DNA sequencing technology. Nature. 2011; 470(7333):198-203.
20. Hardison RC. Genome-wide Epigenetic Data Facilitate Understanding of Disease Susceptibility Association Studies. J Biol Chem. 2012; 287(37):30932-40.

Manuscript submission: 12 September 2012
Peer review approval: 04 October 2012
Final Proof approval: 12 October 2012
Date of Publication: 16 October 2012