Historical Review

A Decade of Human Genome Project Conclusion: Scientific Diffusion About Our Genome Knowledge

Fernanda Moraes
Andréa Góes*

From the Rio de Janeiro State University, Science and Biology Teaching Department—Biology Institute, Rio de Janeiro, Rio de Janeiro, Brazil

Abstract

The Human Genome Project (HGP) was initiated in 1990 and completed in 2003. It aimed to sequence the whole human genome. Although it represented an advance in understanding the human genome and its complexity, many questions remained unanswered. Other projects were launched in order to unravel the mysteries of our genome, including the ENCyclopedia of DNA Elements (ENCODE). This review aims to analyze the evolution of scientific knowledge related to both the HGP and ENCODE projects. Data were retrieved from scientific articles published in 1990–2014, a period comprising the development and the 10 years following the HGP completion. The fact that only 20,000 genes are protein and RNA-coding is one of the most striking HGP results. A new concept about the organization of genome arose. The ENCODE project was initiated in 2003 and targeted to map the functional elements of the human genome. This project revealed that the human genome is pervasively transcribed. Therefore, it was determined that a large part of the non-protein coding regions are functional. Finally, a more sophisticated view of chromatin structure emerged. The mechanistic functioning of the genome has been redrafted, revealing a much more complex picture. Besides, a gene-centric conception of the organism has to be reviewed. A number of criticisms have emerged against the ENCODE project approaches, raising the question of whether non-conserved but biochemically active regions are truly functional. Thus, HGP and ENCODE projects accomplished a great map of the human genome, but the data generated still requires further in depth analysis. © 2016 by The International Union of Biochemistry and Molecular Biology, 44:215–223, 2016.

Keywords: *human genome project; ENCODE project; scientific popularization; science teaching*

Introduction

The term "molecular biology" was coined by Warren Weaver, in 1938. By that time, it was used to describe biological phenomena related to the structure of the molecules and their interactions. As this area of biology expanded, the DNA (deoxyribonucleic acid) turned its major focus [1, 2]. The DNA became prominent in the scientific field in the 1930s, although it had been discovered in 1869 by the physician Johann Friedrich Miescher. At the beginning of the 20th century, studies accomplished by Albrecht Kossel and Phoebus Levene established that DNA is composed by nucleotides which are formed by a deoxyribose (a sugar molecule), a phosphate group and one of the four nitrogen bases (adenine, thymine, guanine and cytosine) [2]. But only in the beginning of 1950s, researches were committed to discovering the structure of the DNA molecule. Watson and Crick unraveled the mystery, with the good images provided by Rosalind Franklin, and confirmed that the DNA structure consisted of a double helix made up of two antiparallel strands. The article with these results was published by *Nature* in 1953.

In the second half of the 1970s, the main molecular biology technologies appeared. Undoubtedly, the Sanger sequencing technique, named chain termination method, was a milestone [3]. Such technique was based on the interruption of DNA synthesis by the incorporation of modified nitrogen bases, analogues of the natural nucleotides. The Sanger sequencing technique became automated in the 1990s and it was used to sequence the hole human

Volume 44, Number 3, May/June 2016, Pages 215–223
**Address for correspondence to: Rio de Janeiro State University, Science and Biology Teaching Department—Biology Institute, Rio de Janeiro, Rio de Janeiro, Brazil. E-mail: acgoes@uerj.br*
Received 10 June 2015; Revised 9 October 2015; Accepted 29 November 2015
DOI 10.1002/bmb.20952
Published online 7 March 2016 in Wiley Online Library (wileyonlinelibrary.com)

genome during the Human Genome Project (1990–2003). The years between 2000 and 2010 were then marked by the detailed description of the human genome. To predict functional DNA elements, the data generated by the Human Genome Project started to be processed through scientific experiments as well as by *in silico* analyses.

The aim of this review is to analyze the evolution of scientific knowledge during and following the Human Genome Project completion, which opened the way to the ENCyclopedia Of DNA Elements (ENCODE) project. We analyzed scientific articles, journals and magazines published during the period 1990–2014. This period comprises the development of Human Genome Project and the 10 years following the release of the first draft of the complete human genome sequence. We hope this manuscript will contribute to draw an overview about the construction of scientific knowledge related to molecular biology events.

Results and Discussion

The Human Genome Project

In 1984 scientists from United States Energy Department met to discuss a project that would devise a technique to sequence the human genome. The aim was to launch studies to detect mutations in DNAs from Second World War survivors of the atomic bomb in Japan. Researchers from the National Institute of Health in the United States quickly joined the group and James Watson was designated to head the Human Genome Research Institute, which became National Human Genome Research Institute (NHGRI) in 1989. Later, several countries joined the effort, particularly the United Kingdom, France, Japan, Canada, Germany and China and it became an international public consortium coordinated by the Human Genome Organization (HuGO) [4].

The Human Genome Project (HGP) began in 1990 and it was expected to finish its task in 2005. The initial goals of the human genome sequencing were soon expanded. The great expectations included reaching the "holy mystery of biology" and responding to the demands imposed by genetic diseases and aging. Such exaggerated ideas were widely spread in the media [5].

James Watson remained just for a few years at the head of NHGRI (from 1989 to 1992) due to conflicts of interest. Watson's purpose was to define the DNA sequence, to understand the logic behind the genes localization within the DNA molecule and how this would influence the organism's biology. His substitute, Francis Collins, expected much more than him and believed the answers to diseases cure would be gotten by analyzing the human genome sequencing. With his approach, Collins delighted the media and the United States Congress. Until 1995, the project focused on the creation of maps of the genome. However, at the same time, Craig Venter surprised the scientific community with the publication of an article containing the complete sequencing of *Haemophilus influenzae Rd*. bacteria genome, the first living organism to be sequenced. With this publication, Venter revealed a quicker and cheaper way of large scale DNA sequencing. His method was named whole-genome shotgun and consisted in sequencing random DNA fragments digested by restriction enzymes. Venter counted on a strong bioinformatics team to undertake the task of overlapping the randomly sequenced fragments [4].

In 1998, Venter started a quarrel against the HGP with the creation of the Celera Genomics Corporation private company. He announced he would finish sequencing the human genome, employing the whole-genome shotgun method, before the year 2001, while the HGP had predicted to deliver it in 2005. The period between 1998 and 2001 was marked by disputes between the two groups. It is important to note that the Venter shotgun method was successfully employed in the human genome sequencing because it relied on data from publically funded project.

Finally, the public and private groups decided to consider the idea of publishing their data simultaneously, as the results of one would complement the results of the other. So, on June 26th 2000, Francis Collins and Craig Venter got together in the White House with the president of the United States, Bill Clinton and The British Prime Minister, Tony Blair. An armistice and a joint effort involving both groups was announced. On February 15, 2001, the human genome draft produced by the public consortium was published in *Nature*; on February 16th of the same year, *Science* published the draft from the private company Celera Genomics Corporation [6, 7]. The first draft generated by the public consortium is still online at the platform of the University of California (Human Genome Project Working Draft—http://genome.cse.ucsc.edu/) and in the National Center for Biotechnology Information—NCBI (http://www.ncbi.nlm.nih.gov). Two years later, in April 2003, the human genome sequence was fully released to celebrate the 50th anniversary of the DNA molecule description [8]. The intention was not naïve. A link was established between the HGP and the opening event in 1953 which gave origin to modern molecular biology.

During the 13 years of the HGP development, the science related to genetics evolved considerably. It was estimated that the genomes of nonrelated people differ by about 1 in every 1.200 to 1.500 DNA bases. The variation from person to person takes place as single nucleotide polymorphism (SNP) and in the copy number variations (CNV). In addition, it was found that more than 40% of human genome proteins are similar to the fly and wormy proteins, and 50% of the human genes present high similarity to the genomic sequences of other organisms [6]. It was concluded that the human genome is as complex and as special as any other organisms. These findings demystified the special expectations created around the human DNA.

It was observed that genes are not uniformly distributed throughout the 24 human chromosomes (22

autosomal chromosomes and X and Y sex chromosomes). This means that rich gene clusters regions are lying alongside poor gene clusters regions. These desertic regions correspond to 20% of the genome. Only about 2% of the human genome is committed to protein synthesis, namely ~20,000 genes are protein-coding genes. This fact was one of the biggest reasons for disappointment by the end of the HGP. It was predicted that the human genome would encode ~100,000 genes. The scientific community was astonished that the number of human genes is equal to that of a rather unsophisticated nematode. This finding was considered quite provocative and an amazing question was raised: where does the complexity of an organism derive from? It was realized that the complexity of the human being is based on the codification of different proteins and not on the quantity of genes. Part of these genes is used in the construction of different proteins during the splicing process of the messenger RNA [6]. With this discovery, the concept "one gene–one protein," which was so far part of the central dogma in biology, needed to be revised.

By that time, the DNA repeated sequences were considered junk DNA. Nevertheless, researchers suspected that learning about the role of these sequences would help to figure out the chromosomes structure and its dynamics. It was believed that these repetitions had reformulated the genome along the evolution, rearranging it and thus creating new genes or modifying the existing ones [6].

Despite all the data generated, many questions still remained unanswered. Neither the key to the understanding of genetic diseases had been found nor had the "secret of life" been disclosed. When the HGP ended, there was a great frustration in relation to the objectives put forward at the beginning of the project.

The ENCyclopedia of DNA Elements Project

By the end of the human genome sequencing, new challenges were proposed. It was time to understand how DNA works, which elements regulate it and how this regulation occurs. In September 2003, the ENCyclopedia of DNA Elements (ENCODE) Project was launched in order to interpret the data generated by the HGP. By using experimental and bioinformatics methods, it would be possible to analyze the DNA structure and its functional components. The project aimed at preparing a complete catalog which contained all functional elements codified in the human genome, for example, the protein-coding genes and noncoding ones, elements that regulate transcription, elements responsible for the structure of the chromosome and any other functional sequence considered relevant [9].

The ENCODE consortium defined a functional element as a discrete segment of the genome which encoded a certain product (e.g., a protein) or displayed a reproducible biochemical signature (e.g., a specific chromatin structure) [9]. Adopting a conventional view of genome organization, the transcripts were encoded by distinct loci and each transcript had its biological role (e.g., encoding specific proteins). Nowadays, this view has been redesigned as other forms of RNA molecules have been described, such as the small nuclear RNAs and micro RNAs, which are encoded from overlapping protein-coding gene regions. These small transcripts play an important role in the maintenance of chromatin and in other regulatory systems [9].

The transcriptional factors regulate the transcription process through protein interaction with specific DNA regions (promoters, enhancers, silencers and insulators and loci of the control regions). The promoter is defined as "the region containing all the binding sites capable of promoting transcription with normal efficiency and appropriate control" [10]. These promoter regions were widely analyzed by the ENCODE project.

At the beginning of the ENCODE project, the technology of large scale identification was performed to identify specific functional elements: genes, promoters, enhancers, repressors or silencing genes, exons, replication origins, termination sites of DNA replication, methylation sites, DNase I hypersensitive sites, transcription factors binding sites, chromatin modifications and conserved sequences in many species with known functions [9].

The chromatin immunoprecipitation (ChIP) and DNase I hypersensitive site assays were the two main techniques widely used to obtain the results. The ChIP technique consists of incubating chromatin (DNA and proteins) with specific antibodies to the target proteins previously bound to their cognate DNA sequence, followed by precipitation and purification of this DNA sequence. The analysis of the enrichment of the target protein in this specific DNA segment is done by either sequencing (ChIP-seq), real-time PCR, or microarray. Alternatively, the antibody target can be a specific histone modification (methylation or acetylation). For example, the antibody can be specific for H3K4me1 histone (the antibody will recognize the lysine aminoacid in the position 4 of histone 3, when it is monomethylated). This type of analysis allows the verification of the DNA transcriptional activity status [9].

The DNase I hypersensitivity assay is used to spot the DNA sites not associated to nucleosomes and therefore sensitive to DNase digestion. As these sites are free from nucleosomes they are probably accessible for interaction with regulatory elements. This means that chromatin accessibility characterized by DNase I hypersensitivity is a hallmark of regulatory DNA regions. As the ChIP technique, DNase I hypersensitive site analysis points out the DNA transcriptional activity status [10]. It was observed that several DNase I hypersensitive sites are located near or inside transcription start sites [11]. Figure 1 shows the mechanism in chromatin which generates the DNase sensitivity.

The ENCODE project was developed by 32 research groups forming a team of 440 scientists. They performed a number of experiments, including ChIP and DNase I hypersensitive site assays [12]. The ENCODE project was

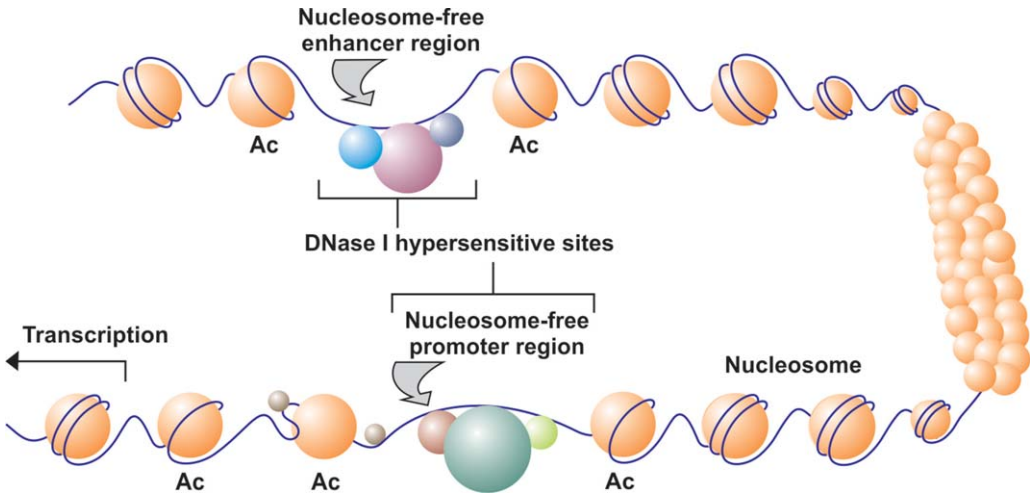

FIG 1 *Representative scheme of mechanisms in chromatin which generates the DNase sensitivity. Modified image from http://en.wikipedia.org/wiki/hypersensiblesite. Ac: acetylation - corresponds to open chromatin regions. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]*

deployed in three phases. The first one, the pilot project, lasted 4 years (from 2003 to 2007). In this phase, the data production related to the protein-coding regions was accomplished. Also, the strategies for identifying the various types of genomic elements were evaluated. The second phase began in 2007 and was concluded in 2012. The aim of this phase was to analyze the 99% of the human genome considered as nonfunctional or noncoding regions. The third and last phase of the ENCODE project started in 2012 and it is expected to be finished in 2016.

The First Phase of the ENCODE Project

The pilot ENCODE phase focused on the analysis of 30 Megabases (Mb) of DNA characterized as functional or protein-coding regions which corresponded to 44 discrete genomic regions or ~1% of the genome. From this 30 Mb, ~15 Mb reside in 14 regions for which there was already substantial biological knowledge, whereas the other 15Mb reside in 30 regions chosen randomly. Transcription analyses were made in these 44 regions in order to better understand the coding RNA repertoire molecules, as the transcripts are involved in many cellular functions, activating directly or indirectly biological molecules. The ChIP tests were performed during the pilot project for the analysis of 18 transcription factors and general components of the transcription machinery (e.g., RNA polymerase and transcription factor II-B) [11].

In this pilot phase, it was possible to obtain sequences of genome regions that are orthologous to the target ones of a set of non-humans vertebrates (mouse, rat, dog, cow, chicken, chimpanzee, monkey, frog, zebra, and fish). The data generated from these comparisons allowed the precise identification of evolutionarily preserved elements, making possible the inferring of biological functions [9].

The main conclusions obtained during the 4 years of the pilot ENCODE project are listed below [11]:

1. The transcription occurs in almost the whole genome such that most of its bases are committed with at least one primary transcript. Many transcripts link distal loci segments to protein-coding regions.
2. Various novel nonprotein coding transcripts were identified. Many of these transcripts originate from overlapping protein-coding loci and from regions previously considered transcriptionally silent.
3. Many transcription start sites were identified. Many of them present chromatin structure and protein-binding specific sequences similar to the well-known promoters.
4. The regulatory sequences that surround the transcription start sites are symmetrically distributed, with no bias towards upstream regions.
5. The accessibility to chromatin and histone modification patterns are highly predictive of both the presence and the activity of transcription start sites.
6. The DNA replication timing is related to the chromatin structure.
7. A total of 5% of the bases in the genome can be considered under evolutionary restriction in mammals. For 60% of these bases, there is evidence for function based on results of experimental tests accomplished to date.
8. A general overlapping between the genomic regions identified as functional by experimental tests and those under evolutionary restriction was not observed.

One of the most surprising conclusions from this first phase concerns the remarkable excess of experimentally identified functional elements which lack evolutionary constraint. This means that apparently many functional elements are not restricted to mammal evolution. The consortium suggested the existence of a large pool of neutral elements that

are biochemically active, but that do not provide a particular benefit to the organism. This pool may serve as a storage to natural selection, potentially acting as a source of lineage specific elements. As concluded by the consortium, this surprise suggests that we take a more "neutral" view of many of the functions conferred by the genome [11].

The Second Phase of the ENCODE Project

The second phase of the ENCODE project began in 2007 and the results were published in 2012. For this phase, the goal was to analyze the remaining 99% of the human genome. Several techniques were used in many types of cellular lineages. Between 2003 and 2012 (first and second phases of the ENCODE project), 1.640 data sets were produced and 24 types of experiments performed in 147 cellular lineages. These analyses consisted in quantifying the different RNA species from both whole cells and cellular compartments, mapping protein-coding regions, histone modifications and transcription factor binding sites by the ChIP technique, as well as mapping sites of DNA methylation [13]. It was observed that 80.4% of the genome is functional in at least one cell type [13, 14].

The ENCODE project also mapped pseudogenes, which are scattered in the genome as gene duplications, but apparently they are transcriptionally inactive [15]. The 20.687 protein-coding genes were annotated in an average of 6.3 transcripts per locus, taking alternative splicing into account. The exons found in these genes correspond to 2.94% of the genome [13]. Table I shows the quantity and types of genes annotated by the ENCODE project.

In an attempt to identify genome regulatory regions, the mapping of DNA binding sites for 119 proteins, including both transcription factors and RNA polymerase components, was carried out in 72 cellular types by the ChIP technique [13]. The transcription factors analyzed were classified into six categories according to the contribution regarding gene expression regulation [14]: transcription factors binding to specific sequences, non-specific transcription factors, chromatin structure factors, remodeling chromatin factors, histone specific methyltransferase, and RNA polymerase III associated factors.

Epigenetics is a reversible mechanism that modifies the genome and can be inherited during cell division, but it does not imply changes in the DNA sequence as a mutation does [16]. Histones are susceptible to epigenetic modifications by addition or removal of methyl and acetyl groups in the lysine amino acid located in its amino terminal region. Epigenetic mechanisms act by changing the chromatin accessibility to transcriptional regulation (Fig. 2). The ENCODE project evaluated the chromosomal locations of histone modifications in 46 types of cells. A highly variable pattern of modification across cell types was observed in parallel with changes in transcriptional activity [13]. Table II shows the main histone modifications and their putative roles described by the ENCODE project.

TABLE I ***ENCODE annotated genes***

Types of genes	*Quantity of genes*
Protein-coding genes	20.687
Nonprotein coding novel transcripts	33.977
Long RNA nonprotein coding (lncRNA) loci	9.640
Transcribed pseudogenes	863
Non-transcribed pseudogenes	10.353

Table modified from The Encode Project Consortium [21].

The ENCODE project also mapped 2.89 millions of DNase I hypersensitive sites (DHSs) in 125 kinds of cellular lineages. Approximately one third of the sites were found in only one type of cell and 3,700 sites were found in all the other types of cellular lineages, suggesting that genetic regulation in each cell type is differential. Approximately 75% of these DHS sites were found in introns or intergenic regions, indicating that introns exert functionality in gene expression regulation [14].

Methylation of cytosines located in CpG islands (groups of cytosines and guanines located generally in promoter regions) is another mechanism of epigenetic regulation. Typically, methylation of the promoter region is associated with the repression of transcription whereas methylation within gene sites is involved in the transcription activation (Fig. 3). The technique of bisulfite sequencing was used by the ENCODE consortium in order to determine the cytosines methylation profiles. The sodium bisulfite treatment of DNA converts cytosine residues to uracil, but leaves 5-methylcytosine residues unmodified (Fig. 4). Thus, bisulfite treatment introduces specific changes in the DNA sequence that depends on the methylation status of individual cytosine residues. Following the treatment, the genome is sequenced in order to retrieve this information. The DNA

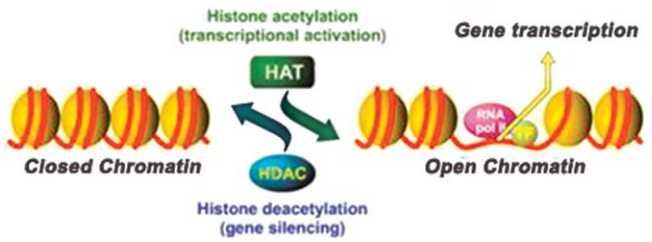

FIG 2 *Representative scheme of transcriptional regulation by histone modification. From: http://www.cyberounds.com/cmecontent/art467.html?pf=yes. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]*

TABLE II *ENCODE histone modifications*

Histone modification	*Putative functions*
H3K4me2	Mark of regulatory elements associated with enhancers and promoters
H3K4me3	Mark of regulatory elements associated with promoters/transcription starts
H3K9ac	Mark of active regulatory elements with preference for promoters
H3K9me1	Preference for 5′ end of the genes
H3K9me3	Repressive mark associated with constitutive heterochromatin and repetitive elements
H3K27ac	Mark of active regulatory elements
H3K27me3	Repressive mark associated with repressive domains and silent developmental genes

Table modified from The Encode Project Consortium [20]. H: histone; K: lysine aminoacid; me: methylation; ac: acetylation. Example: H3K27ac—lysine aminoacid in the position 27 of histone 3 which is acetylated.

methylation profile was assessed for an average of 1.2 million CpGs in 82 cell lines and tissues. It was found that 96% of these CpGs islands exhibited differential methylation in at least one cell type or tissue tested. Also, DNA methylation levels correlated with chromatin accessibility [13].

During the second phase of the ENCODE project, important aspects about the organization and function of the human genome were discovered [13] such as:

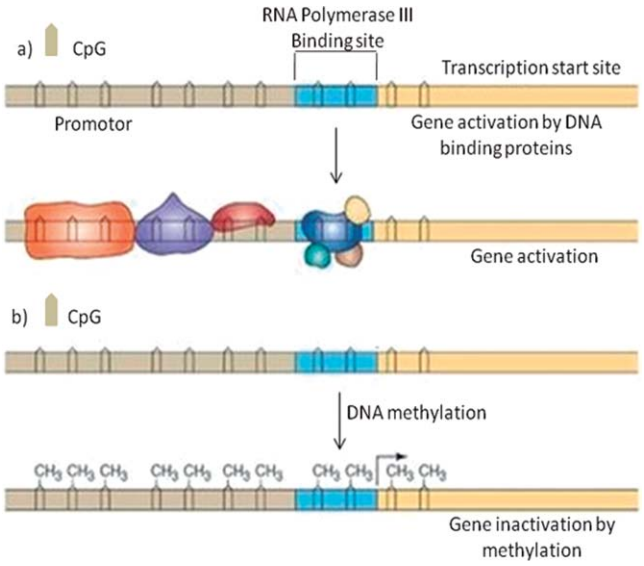

FIG 3 *Methylation of CpG islands in a gene promoter region. (a) Gene activation. (b) Gene repression. Image modified from google.com/site/biotechnology3bioq/controledaexpresãogenética 2232. CpG: cytosine and guanine rich regions. CH3: methyl group. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]*

1. Most of the human genome (80.4%) takes part in at least one biochemical RNA and/or chromatin-associated event in at least one kind of cell. A total of 99% of the known bases in the genome are within 1.7 kb of any ENCODE element, whereas 95% of bases are within 8 kb of a transcription factor binding motif.
2. The classification of the genome in seven chromatin states (signature pattern of histone modification) pointed out a set of 399.124 regions with enhancer-like features and 70.292 regions with promoter-like features as well as a lot of quiescent regions.
3. It is possible to correlate quantitatively RNA production and processing with both chromatin markers and transcription factor binding at promoters.
4. Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions. This number is at least as large as those that lie in protein-coding genes.
5. Single nucleotide polymorphisms (SNPs) associated with diseases are located mainly in non-coding functional elements.

Undoubtedly, the verification that the human genome is pervasively transcribed and almost fully active remains as one of the most important molecular biology discoveries.

The Ongoing Last Phase of the ENCODE Project: Perspectives and Controversies

The last phase of the ENCODE project began in 2012 and it is expected to be concluded in 2016. Basically, the consortium is refining the previous results and applying the knowledge to basic biological questions and disease studies through large-scale genomics studies.

However, some "philosophical" concerns have been ventilated at this moment. A great discomfort was

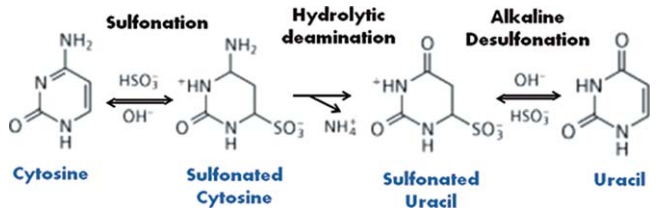

FIG 4 *Representation of the chemical reaction chain in bisulfite DNA treatment. Authors illustration. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]*

generated with the finding that the biochemically active regions cover a much larger fraction of the genome than do evolutionarily conserved regions, raising the question of whether non-conserved but biochemically active regions are truly functional. The ENCODE opponents argue that the consortium, ignoring a century of population genetics theory, did not consider that a biological function cannot be maintained indefinitely without selection and that they regard the genome as invulnerable to deleterious mutations, either because no mutation can ever occur in these "functional" regions or because no mutation in these regions can ever be deleterious. The evolutionary biologists claim that neither transcription, nor open chromatin, nor histone modification, nor transcription factor binding, nor DNA methylation equal function. They also observe that the ENCODE project message that everything has a function implies purpose, and purpose is the only thing that evolution cannot provide [17]. In fact, although a deterministic view of the organisms is a biological conception no longer acceptable, the ENCODE genomics assumption that any specific and reproducible biochemical event must correspond to a meaningful biological function prevails.

According to Kellis *et al.* [18], despite the pressing need to identify and characterize all functional elements in the human genome, it is important to recognize that there is no universal definition of what constitutes function, nor is there agreement on what sets the boundaries of an element. So a great controversy is created and the ENCODE antagonists assert the consortium adopted a wrong and much too inclusive notion of function. According to Eddys [19], ENCODE's goal was nebulous because "functional element" was ill defined and had to be operationalized. All reproducible biochemical events were claimed to be "critical" and "needed". He also points out that the ENCODE project had not shown what fraction of these activities play any substantive role in human gene regulation.

In fact, according to Germain *et al.* [20], ENCODE's strategy of biochemical signatures successfully identified activities of DNA elements with an eye towards causal roles of interest to biomedical research. And this is fully true and apprehensible if we take into account that the major biomedical Big Science projects are sponsored by the big pharmaceutical groups.

Revisiting the Flow of Information and Some Molecular Biology Dogmas

According to The Encode Project Consortium [13], the perspective of transcription and genes may have to evolve to

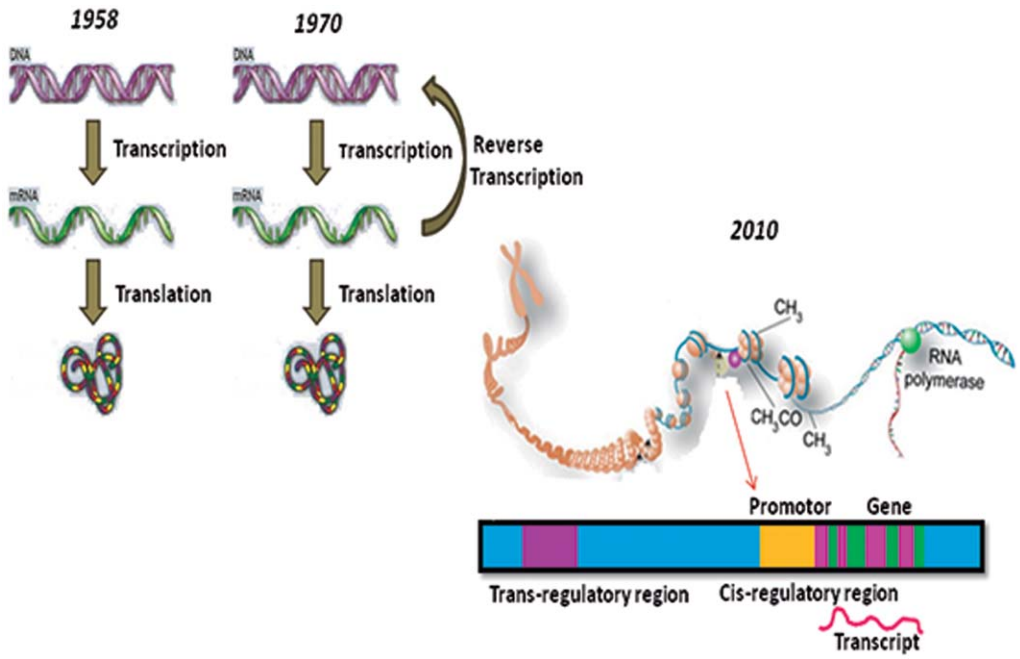

FIG 5 *Diagram illustrating the evolution of molecular biology central dogma. Images modified from http://www.academiamalhacao.com.br/nikolascte/?paged=32 and http://genome-mirror.duhs.duke.edu/ENCODE/. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]*

some surprises that challenge the current dogma on biological mechanisms. As they observed the presence of numerous intercalated transcripts spanning the majority of the genome, they asserted that a simple view of it as having a defined set of isolated loci transcribed independently does not seem to be accurate.

But let's tell the history from the beginning. In 1958, Francis Crick, the one that had discovered 5 years before the structure of the DNA molecule, postulated the central dogma of molecular biology, which explains that DNA codes for RNA, which codes for proteins. He cogitated that the flow of genetic information was transmitted unidirectionally from DNA to RNA and from this latter to the proteins, which determine the cellular and organism phenotype [21]. The information transfer from proteins to DNA, RNA or other proteins as well as the transfer from RNA to DNA, would have been considered incongruous at that time. But the conception of unidirectional information was not supported for a long time.

By the end of 1960, new forms of viruses were described in which the genetic material is composed of RNA. These viruses are capable to revert RNA in DNA during a process called reverse transcription, carried out by the reverse transcriptase enzyme [22]. Such information first contradicted the idea of unidirectional gene flow, and the central dogma was revised by Crick himself in 1970 [23] (Fig. 5). The process of RNA editing, named splicing, was then described [24]. The splicing process is achieved by the interaction of proteins with RNA. This is another reason to contradict the unidirectional flow. It was observed that proteins are present whatever the process. For example, to produce a copy of a DNA strand, the cell relies in an accurate protein repair machinery.

The disclosure of the RNA editing process opened the way to the RNA world, and it was soon verified that differential splicing is an important aspect of biological regulation and differential expression of genomic information. It was also found that some RNA molecules could undergo structural changes in the absence of proteins [25]. It means that RNA molecules can underlie catalytic processes in many ways analogous to those of proteins. So, according to Shapiro [26], the information content of RNA molecules has many potential inputs besides the sequence of the DNA template from which it was transcribed. It is well know that the proteins, as the RNAs, are not ready to act just after translation. The proteins are subject to a number of modifications (acetylation, methylation, phosphorylation, and so forth) before implementing theirs functions.

The RNA splicing process consists in the excision of intronic regions from the gene, lasting only the sequence that will be used in the protein translation (exon). By the time of the splicing elucidation, a central dogma was established in which the genome can be functionally discriminated into two regions: a significant protein-coding region and a non-significant non-coding region. In 2012, the ENCODE project clarified that the human genome is pervasively transcribed and it has been fully established that the 99% of the genome play an important role in regulatory processes. There is no doubt that the complexity of our genome is related to different regulation processes. The conservation of gene order (known as syntheny) between species reflects the need to preserve the regional regulatory structures and sequences. For example, events of chromosome translocation are rare in relation to small mutations, as insertions, deletions and inversions. Therefore, the sequences between genes can change, but the linear order of the genes within the segments is more constant [27].

Briefly, the central dogma was completely reviewed with the HGP and ENCODE projects discoveries. The mechanistic operation view of the genome has been reformulated. Nowadays, the human genome is seen as a much more complex perspective in which infinite possibilities of interactive systems regulate the cellular processes [26]. Thus, biology scientists are free from the mechanicist and reductionist view of the first steps of the molecular biology. But it is important to note that the central dogma postulated by Crick remains inviolable, taking into account that once sequencial information has passed into proteins it cannot get out again.

As a last reflection in order to close this debate, we have to mention the actual meaning of the genes. We are cured from the old postulate "a gene, a function." The gene can no longer be seen as a unitary and deterministic component, as the result of the expression of one single region of the genome. Each element of the genome has multiple components that interact directly or indirectly with many other genomic elements.

Final Considerations

The HGP and ENCODE projects contributed to the mapping of human genome and in the evolution of the central dogma of molecular biology. These studies revealed that the complexity of our genome does not rely in protein-coding genes quantity but in a great network of transcripts that allows the interactions for genome regulation. The concept of the central molecular biology dogma was reformulated. There is no unidirectional flow of information from one class of molecule to another. All the process is feedback interconnected. We deviated from a strict genetic determinism. A gene-centric conception of the organism has to be reviewed. Finally, the HGP and ENCODE projects accomplished a great map of the human genome, but the big data generated remain to be carefully analyzed. We've been endeavoring to catalog a number of phenomena in order to understand the nature language. But it is not so evident. The key for understanding the "secret of life" has not been revealed.

Acknowledgements

This work was supported by FAPERJ (Rio de Janeiro State Research Support Foundation, grant number: E-26/110.802/2011). The authors thank Tiago Caldas and Gina Arêdes for figures preparation.

References

[1] Weaver, W. (1970) Molecular biology: Origins of the term. Science 170, 581–582.

[2] Hausmann, R. (2002) To Grasp the Essence of Life: A History of Molecular Biology, Springer, Netherlands, New York.

[3] Sanger, F., Nicklen, S., and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74, 5463–5467.

[4] Venter, J. C. (2007) A Life Decoded, 1st ed., Viking, New York.

[5] Oliveira, B. V. X. and Góes, A. C. S. (2014) The human genome project: A portrait of scientific knowledge construction by the Ciência Hoje magazine. Ciência & Educação. 20, 561–577.

[6] International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409, 860–914.

[7] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001) The sequence of the human genome. Science 291, 1304–1351.

[8] International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

[9] The Encode Project Consortium (2004) The ENCODE (ENCyclopedia of DNA elements) project. Science 306, 636–640.

[10] Lewin, B. (2007) Genes IX, 9the ed., Jones & Bartlett Learning, Miami.

[11] The Encode Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 779–814.

[12] Maher, B. (2012) ENCODE: The human encyclopaedia. Nature 489, 46–48.

[13] The Encode Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

[14] Qu, H. and Fang, X. (2013) A brief review on the human encyclopedia of DNA elements (ENCODE) project. Genom. Proteom. Bioinform. 11, 135–141.

[15] Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., and Snyder, M. (2007) What is a gene, post-ENCODE? History and update definition. Genome Res. 17, 669–681.

[16] Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007) Epigenetics: A landscape takes shape. Cell 128, 635–638.

[17] Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013) On the immortality of television sets: "Function" in the human genome according to the evolution-free Gospel of ENCODE. Genome Biol. E 5, 578–590.

[18] Kellis, M., Wold, B., Snyderd, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., Dunham, I., Elnitski, L. L., Farnham, P. J., Feingold, E. A., Gerstein, M., Giddings, M. C., Gilbert, D. M., Gingeras, T. R., Green, E. D., Guigo, R., Hubbard, T., Kent, J., Lieb, J. D., Myers, R. M., Pazin, M. J., Ren, B., Stamatoyannopoulos, J. A., Weng, Z., White, K. P., and Hardison, R. C. (2014) Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. USA 111, 6131–6138.

[19] Eddy, S. R. (2013) The ENCODE project: Missteps overshadowing a success. Curr. Biol. 23, 259–261.

[20] Germain, P. L., Ratti, E., and Boem, F. (2014) Junk or functional DNA?: ENCODE and the function controversy. Biol. Philos. 29, 807–831.

[21] Crick, F. H. (1958) On protein synthesis. Symp. Soc. Exp. Biol. 12, 138–163.

[22] Temin, H. M. and Mizutani, S. (1970) RNA-dependent DNA polymerase in virions of *Rous* sarcoma virus. Nature 226, 1211–1213.

[23] Crick, F. H. (1970) Central dogma of molecular biology. Nature 227, 561–563.

[24] Chow, L. T., Gelinas, R. E., Broker, T. R., and Roberts, R. J. (1977) An amazing sequence arrangement at the 5_ ends of adenovirus 2 messenger RNA. Cell 12, 1–8.

[25] Cech, T. R. (1989) RNA as an enzyme. Biochem. Int. 18, 7–14.

[26] Shapiro, J. A. (2009) Revisiting the central dogma in the 21st Century. Nat. Genet. Eng. Nat. Genome Ed. 1178, 6–28.

[27] Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engström, P. G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., Ghislain, J., Pezeron, G., Mourrain, P., Ellingsen, S., Oates, A. C., Thisse, C., Thisse, B., Foucher, I., Adolf, B., Geling, A., Lenhard, B., and Becker, T. S. (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. 17, 545–555.