CONTRIBUTIONS to SCIENCE, **8** (2): 155–159 (2012)
Institut d'Estudis Catalans, Barcelona
DOI: 10.2436/20.7010.01.149 ISSN: 1575-6343
www.cat-science.cat

Genetic and Socio-Cultural Risk Contributions to Disease

What is our level of knowledge about the genome today?*

Miguel Beato
Centre for Genomic Regulation, Barcelona

Resum. L'elucidació de la seqüència del genoma humà va marcar el començament d'una nova fase de la comprensió de la biologia humana. El pas del genotip al fenotip, per mitjà de la qual la informació continguda en el DNA es transcriu en forma de RNA missatger (mRNA) i després es tradueix en proteïnes, és complicada i no s'esdevé de forma directa. A més, la major part dels mecanismes que guien aquest complex procés encara no s'han resolt. Des del concepte de gen fins als diferents tipus d'informació que estan codificats en el DNA genòmic del nostre nucli i l'epigenètica, aquest article resumeix l'estat dels nostres coneixements, però també pren nota del que encara no sabem, així com de les perspectives de futur en els estudis del genoma.

Paraules clau: flux d'informació genètica · expressió gènica · regulació del DNA · dominis d'associació topològica · xarxes genètiques · epigenètica

Summary. Elucidating the sequence of the human genome marked the beginning of a new phase of understanding of human biology. The pathway from genotype to phenotype, through which the information contained in the DNA is transcribed in the form of messenger RNA (mRNA), and then translated into proteins, is a complicated one, and does not happen in a straight line. Furthermore, most of the mechanisms guiding this complexity have yet to be unravelled. From the concept of 'gene,' to the different types of information that are encoded in the genomic DNA in our nucleus, to epigenetics, this article summarises our current level of knowledge, but also takes note of what we do not yet know as well as the future perspectives in genome studies.

Keywords: flow of genetic information · gene expression · DNA regulation · topological association domains · genetic networks · epigenetics

Carved into the Temple of Apollo, there is an inscription that reads one of the Delphic maxims, "Know thyself." Humans have never followed this advice as literally as when they decided to elucidate the sequence of their own genome. However, despite all the talk about the human genome in the past decade, we still know very little about it and there are many things that are still completely unknown. Here I summarise our current level of knowledge, but also take notes of what we do not yet know as well as the future perspectives in genome studies.

Elucidating the sequence of the human genome marked the beginning of a new phase of understanding of human biology. But the genome is more or less the same in every organism, with the same combination of nucleobases (A, G, T, C) producing thousands of different life forms, and we still do not know how this happens. The challenge of modern research in genomics is to understand the relationship between genotype and phenotype, how the former establishes the latter. Genotype is the genetic makeup of a cell, or organism. While it has been traditionally ascribed to the DNA, the current view is that RNA also plays an important role, in that there may be a pool of RNA associated with sperm and oocytes that is transmitted from generation to generation. Phenotype refers to form, structure, and function. It is the composite of an organism's observable characteristics or traits, which we refer to when we speak of morphology, development, biochemical and physiological properties, phenology, and behaviour. It is the result of proteins, RNA and other macromolecules and it is established de novo in each generation.

An organism's genotype comprises a set of inherited instructions that are carried within its genetic code. The phenotype is the result of genotype expression but it is also influenced by environmental factors and interactions between the two. Francis Crick provided an explanation for the flow of genetic information within a biological system in 1958, and restated it in paper published 1970,

* Based on the lecture given by the author at the Parliament of Catalonia, Barcelona, on 23 October 2012 for the annual conference of the EPTA network, 'From genes to jeans: challenges on the road to personalised medicine.'

Correspondence: M. Beato, Centre de Regulació Genòmica, Edifici PRBB, C/ Dr. Aiguader 88, E-08003, Barcelona, Catalonia, EU. Tel. +34-933160100. Fax +34-933160099. E-mail: miguel.beato@crg.es

> "The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back to protein to either protein or to nucleic acid." [2]

Or as 1968 Nobel Prize Winner Marshall Nirenberg said, "DNA makes RNA makes protein."

But how does this work? In the very complicated pathway from genotype to phenotype, the information contained in the DNA is transcribed in the form of messenger RNA (mRNA), which is then translated into proteins. But this pathway is not a direct one; there are many intermediaries. During the synthesis of proteins from DNA via mRNA, proteins interact with mRNA, with DNA, with other proteins, etc. Development is certainly not a straight line, nor are we merely a homunculus that simply grows in size. Rather, in virtually all species the embryo is completely different from the adult. A further mystery is how, after the complicated but similar processes of gene expression and the many other transformative processes that constitute development, we can end up with organisms ranging from a fly to a human.

The DNA code is commonly described as a succession of 'letters' (representing the four nucleobases) whose sequence confers meaning on the code, analogous to the way in which the sequence of alphabet letters forms a word. But, in fact, this is a very misleading analogy, because DNA is not a binary code. It is a complicated three-dimensional chemical structure with its own dynamics. It has no defined structure derived from its sequence, but can change shape and conformation. Thus, although the genotype is encoded in the DNA, DNA is also part of the phenotype, it actively interacts with proteins and RNA to convert its own information into the final phenotype.

How this comes about is one of the main questions so far left unanswered.

Types of genomic information

1. Information encoding mRNA and proteins

If take a step by step look at the types of information that are encoded in the genomic DNA in the nucleus of our cells, it is very clear that it all begins with our genetic code. There are specific genes that are transcribed to yield mRNA, which together with tRNA guides the translation of proteins. The code itself is very clear: three nucleotides define an amino acid, with the sequence of amino acids progressively producing a protein. But while this process seems quite straightforward, it is actually very complex. Previously, it was accepted that one gene gives rise to one mRNA, which in turn gives rise to one protein. But we now know that one gene can generate twenty, thirty, or even a thousand different mRNAs, through the re-combination of the different exons of a gene, and thus to hundreds of different proteins. Again, the mechanisms guiding this complexity have yet to be unravelled. Furthermore, the coding part of the genome represents < 2 % of our genome. In other words, the part of our genome that is translated into proteins is merely a tiny fraction of all the information that we inherit from our parents.

Likewise, the classic concept of a gene as the molecular unit of heredity in a living organism is obsolete. The more we learn about the genome, the more difficult it becomes to define what a gene is. I recently attended a lecture by a scientist from The ENCODE Project: the Encyclopedia of DNA Elements (http://www.genome.gov/10005107), an international collaboration of research groups that aims to build a comprehensive list of functional elements in the human genome. Even this scientist could not come up with a definition of what a gene is. There are areas of the genome that encode proteins that carry out certain functions, but overlapping with this information may be information that affects genes located in another region of the genome. The nature of the communication between genes, the genome, and the chromosome is another mystery.

2. Information regulating gene expression and DNA replication

Although only 2 % of our DNA codes for proteins, most of the human genome is still transcribed into RNA. This RNA carries information; it is not junk as was thought until just a few years ago. Nonetheless what this information is has yet to be determined. I think we have to expand our view of the genome as not only a source of information that we can exploit but also as a source of other types of information. For example, it clearly contains regulatory sequences to control the expression of protein-encoding genes. These sequences make use of a different alphabetic system: rather than three nucleotides per one amino acid, it is the base pairs themselves that are important, with short sequences of base pairs recognised by DNA binding proteins. However, these regulatory sequences also constitute just a tiny fraction of the whole genome, approximately 1 %.

How do proteins read information? Figure 1 shows the double helix of DNA. As the strands are not symmetrically located with respect to each other, there are unequally sized spaces, or grooves, between them. Proteins read regulatory information from the DNA by contacting base pairs through the major groove. By the interactions between the amino acids of regulatory proteins and reactive chemical groups of the DNA, the DNA sequence is read. Depending on that sequence and on the nature of the factors that read it, a gene is kept silent, activated, expressed in coordination with several other genes, etc., because the process is controlled by the same regulatory protein(s). This process underlies, for example, embryogenesis. Yet, even in this case, we are still talking about only a tiny fraction of the DNA contained in a genome.

3. Topological information

Other information that can be considered important in our genome is that which is conserved through evolution and involved in processes such as mRNA stability. This information is poorly

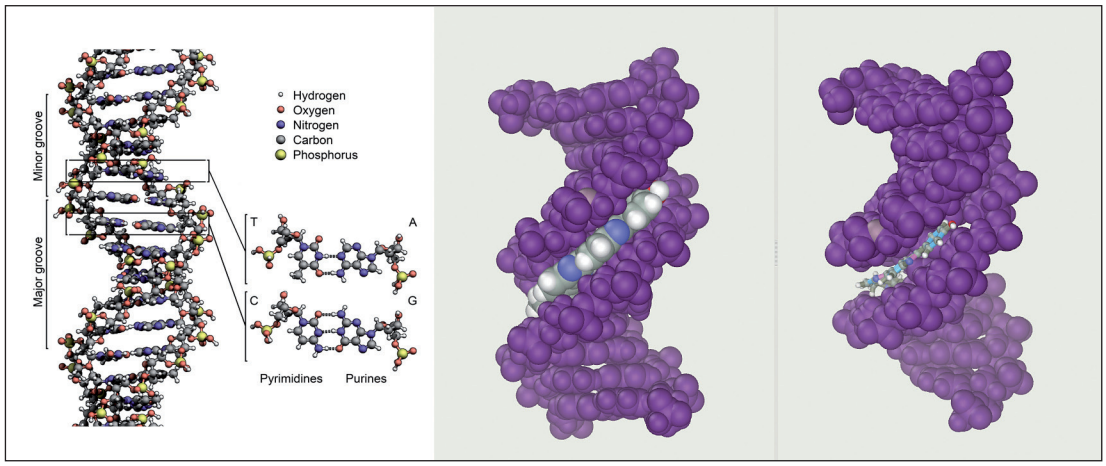

Fig. 1. Left: The structure of DNA showing with detail the structure of the four bases (adenine, cytosine, guanine, and thymine) and the location of the major and minor grooves. Right: Major and minor grooves of DNA. Source: Richard Wheeler (Wikimedia Commons).

understood whereas there is much that we know about the topological information provided by DNA: conformational properties of the sequence, organisation in nucleosomes and organisation in chromosomes. A cell nucleus contains 2 m of DNA that is roughly 10 μm in diameter. The efficient packaging of this DNA is quite dramatic, and the way it is folded depends on the DNA sequence itself. Specifically, the way DNA is wrapped around the histone cylinder and how the chromatin fibre is compacted depends on the tendency of its sequence to bend in particular directions, and this property is encoded in the sequence of dinucleotides, i.e., the sequence of consecutive nucleotides. In other words, there are three nucleotides for each amino acid, there are base pairs containing regulatory information, and there are dinucleotides that guide DNA topology. In nature, DNA is found in at least three conformations: A-DNA, B-DNA, and Z-DNA, with B-DNA as the predominant form and the one described by James Watson and Francis Crick (Fig. 2).

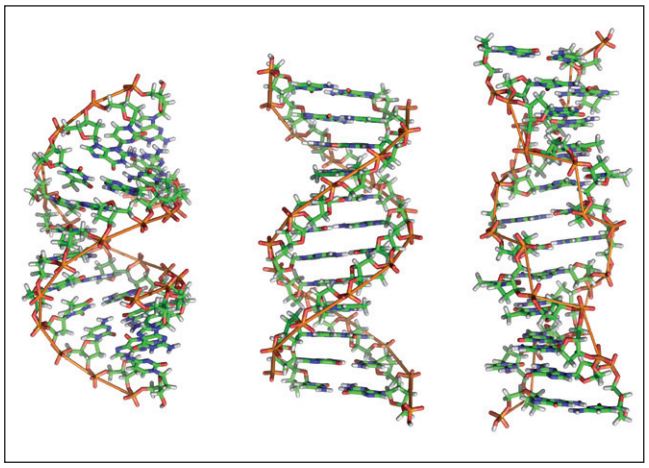

Fig. 2. DNA conformations. From left to right, A-, B-, and Z-DNA. The structure of a DNA molecule depends on its environment. In aqueous environments, including the majority of DNA in a cell, B-DNA is the most common structure. The A-DNA structure dominates in dehydrated samples and is similar to the double stranded RNA and DNA/RNA hybrids. Z-DNA is a rarer structure found in DNA bound to certain proteins. Source: Richard Wheeler (Wikimedia Commons).

A free double helix of DNA is packed around a cylinder of eight histone proteins to form a nucleosome, the basic unit of DNA packing in the cell (Fig. 3). Thus, in our cells, DNA never comes alone, it is always wrapped in proteins. As noted above, the ability of DNA to bend is a function of the ways in which the dinucleotides can be deformed—information that is encoded in their sequence. Bending, in turn, determines which part of the sequence can or cannot be read, depending on whether it is on the inside or the outside of the helix. However, there are also other types of information that are less well understood, such as genetic networks, chromatin domains, etc.

4. Other types of information

Other types of genomic information include genetic networks, chromatin domains and morphogenetic and metabolic programs. Figure 4 provides an example of a genetic network that determines the first 24 h of sea urchin development, during formation of the mesoderm and endoderm, two embryonic lay-

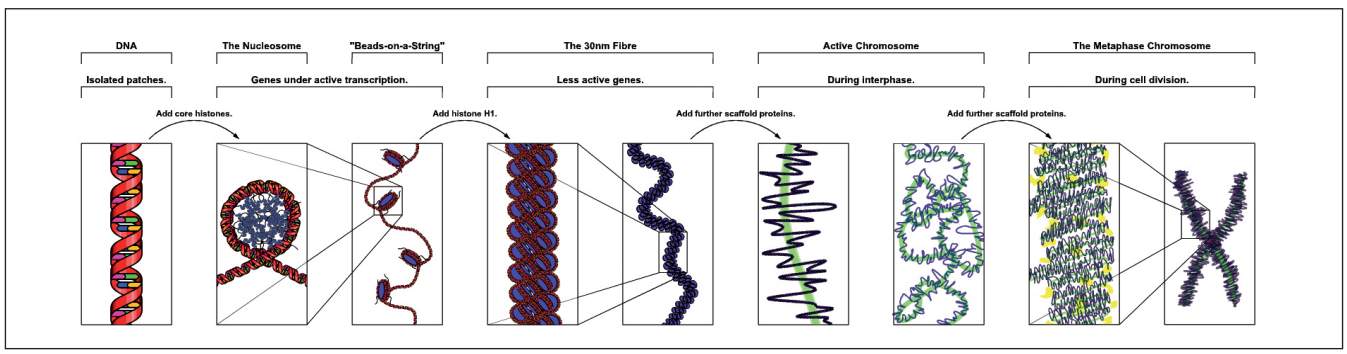

Fig. 3. The current chromatin compaction model. Source: Richard Wheeler (Wikimedia Commons).

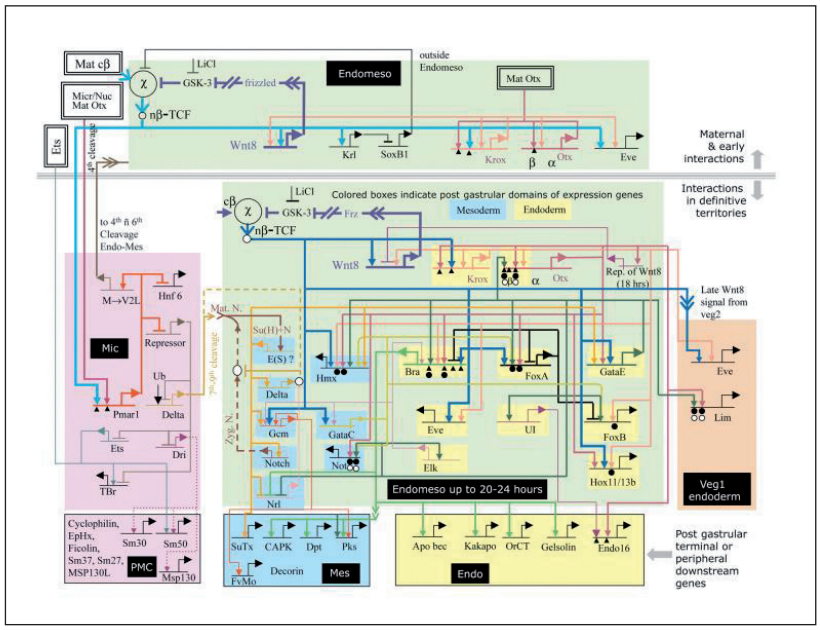

Fig. 4. Regulatory gene network for endomesoderm specification in the purple sea urchin (*Stronglycentrotus purpuratus*): the view from the genome [3].

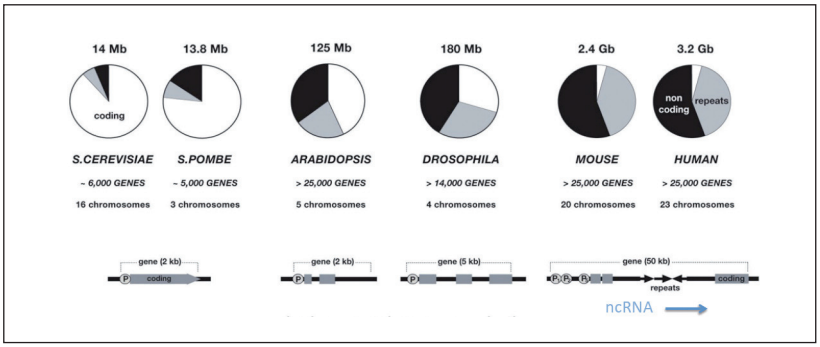

Fig. 5. Simple vs. complex gene organisation.

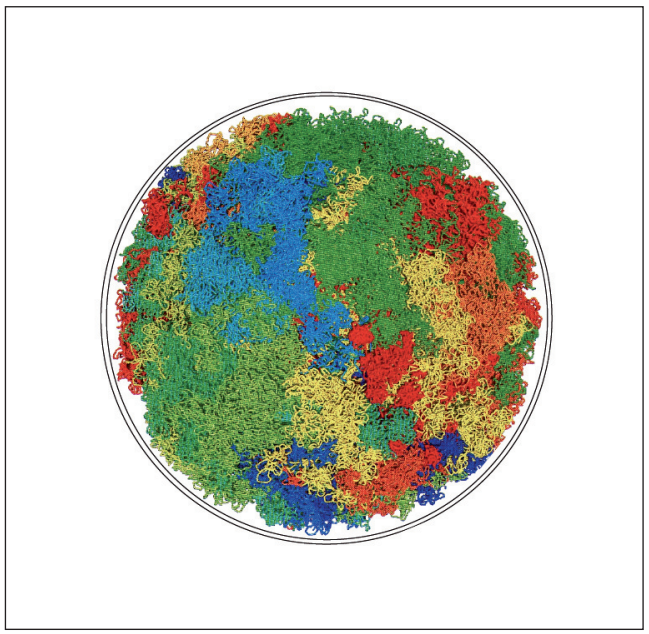

Fig. 6. Colour-coded chromosomes in the cell nucleus.

ers. Sea urchins have long provided developmental biologists with a model organism to study the processes of embryogenesis, gastrulation, and tissue differentiation. The network of genes involved in endomesoderm formation, as depicted in Fig. 4, resembles a computer circuit, including backwards and forwards loops with numerous interactions. Yet, although the various steps in this very complex system have been well elucidated, we still do not understand the nature of genetic programs and how they are organised.

An important consideration is that the number of genes in the genome does not account for the complexity of an organism. If we compare yeast (*S. cerevisiae, S. tombe*), a very simple plant (*Arabidopsis*), a fruit fly (*Drosophila*), a mouse, and a human, we find that there is little difference in the size of the coding proportion of the genome (Fig. 5). The plant has 25,000 genes, just like humans, although we are much more complex. What is different is the non-coding part of the DNA. And again, genes are not just fixed units, with a beginning and an end, but may have several possible starts and overlap with non-coding DNA, etc.

The topology of DNA in the nucleus is one key component of the genetic information contained in the nucleus. Figure 6 shows a cell nucleus with its chromosomes, with the chromosomal DNA shown in different colours. We see that the chromosomes are not intermingled: they occupy specific positions in the nucleus and have specific neighbours with which they interact. Space is therefore an important aspect of the genome, and not merely 3-dimensional space, but 4-dimensional space at that.

In our laboratory, we study the interaction of chromosomes in breast-cancer cells. Figure 7 shows how different regions

Fig. 7. Chromosome correlation matrixes for human ductal breast epithelial tumour cell line, T47D.

of the chromosome interact with each other. In this matrix of interactions, we can identify neighbouring structures in the genome. This information allows us to construct a map and a model of the chromosomes that can be confirmed by high-resolution microscopy. We have found that chromosomal structure is such that certain regions are closer to each other than others, which may have functional implications.

This spatial view of the genome dramatically changes our way of thinking about it. Among other things, it means that the genome's 3D configuration must be taken into account in mechanistic considerations. Spatial relationships add another level of complexity to the genome because elements that are close to each other do not necessarily interact equally or more efficiently. Instead, the genome contains topological association domains, regions that may include five or ten genes behaving as a unit. This is an interesting approach to understanding how the genome works: it emphasises the importance not of the gene but of the topological domain. In fact, if we look at a genomic region in greater detail, we see that the way that it folds correlates with the properties of that particular genomic compartment. These compartments can be modelled in order to predict what happens when certain genes are activated. For example, when just 10 nmol of a steroid hormone is added to a breast cancer cell for 60 min, organizational changes occur: some regions unfold and others compress [1]. This means that the genome is interpreting the signal, not only in terms of the local interaction of genes or regulatory sequence, but also in terms of the global compaction of the sequence.

To summarise, the genome can be viewed as an information cycle in which different types of information are encoded in the DNA, but with different coded sequences giving rise to different consequences. The coding information produces proteins, some of which have a regulatory function and thus recognise regulatory information, such as when this coding information should be expressed. Other proteins, such as histones, wrap the DNA in a specific manner to topologically control access to regulatory information. And finally, there are genetic networks, chromatin domains, etc., all of which are as yet poorly understood but which are known to be higher-level hierarchies that regulate complex biological processes.

Further complicating our understanding of the genome is epigenetics, i.e., the changes in gene expression or cellular phenotype cause by mechanisms other than changes in the underlying DNA sequences. Epigenetic phenomena include DNA methylation; chemical modification by histones that change the way DNA is packaged and determine its accessibility, etc. The epigenome is driveable; it can be dynamically altered by environmental conditions. Changes in the epigenome can cause changes in both the chromatin and the structure of the genome. Furthermore, these changes can be passed down to an organism's offspring. For these reasons, epigenetics is currently of great interest to the pharmaceutical industry, especially in cancer therapeutics.

Many questions about the genome remain to be answered, some have implications for normal growth and development, others are related to disease. Clearly there is much research to be done in this field. Elucidating the sequence of the human genome was an enormous collaborative undertaking, but to fully understand the structure and function of the genome will require even greater efforts.

References

1. Baù C, Marti-Renom M (2011) Structure determination of genomic domains by satisfaction of spatial restraints. Chromosome Res 19:25-35
2. Crick F (1970) Central dogma of molecular Biology. Nature 227:561-563
3. Davidson et al. (2002) A Genomic Regulatory Network for Development. In: Livi CB. Spblimp1/krox: A transcriptional regulator with a central role in endomesoderm specification in sea urchin embryos. Dissertation (Ph.D.), California Institute of Technology [http://thesis.library.caltech.edu/2458/]