WHEN GENES ARE NOT ENOUGH: GENOMICS AND EPIGENOMICS

GIULIO PAVESI (*)

Nota presentata dal s.c. Martino Bolognesi
(Adunanza del 2 maggio 2013)

SUNTO. – Le tecnologie di sequenziamento, che permettono di determinare l'esatta sequenza delle basi che compongono una molecola di DNA, hanno costituito un passo in avanti di portata storica per tutte le scienze della vita. La loro applicazione al sequenziamento di interi genomi, primo tra tutti quello umano, ha tuttavia rivelato alcuni aspetti inattesi dell'organizzazione dei geni delle cellule viventi, e della relativa regolazione dell'espressione. Questo articolo riassume brevemente come il passaggio dalla genetica alla genomica, ovvero allo studio dell'intero genoma e non delle singole unità che lo compongono, ha portato alla revisione e aggiornamento di alcuni concetti tradizionali della genetica classica, nonché alla scoperta del ruolo fondamentale della regolazione epigenetica dell'espressione genica.

***

ABSTRACT. – Sequencing technologies, that permit to determine the exact sequence of base pairs forming a DNA molecule, have been an historical breakthrough for all life sciences. Their large scale application to the sequencing of entire genomes, like the human genome, has on the other hand revealed some unexpected features of the organization of genes, as well as of the regulation of their expression. This article briefly summarizes how the post-genomic era, in which the whole genome could be studied, instead of the single parts composing it, has brought to the revision and the update of some of the most basic assumptions of classic genetics, as well as the discovery of the fundamental role of the epigenetic regulation of gene expression.

(*) Department of Biosciences University of Milan, Italy.
E-mail: giulio.pavesi@unimi.it

Sequencing and genomes

The introduction of DNA sequencing technologies, that given a DNA molecule permit to determine the exact arrangement of base pairs forming its sequence, has been one of the milestones of research in every aspect of life sciences. Their large scale application to the sequencing of the whole genome of different organisms has in turn produced some of the most recent milestones of science in general, the most important being the sequencing of the human genome at the turn of our century.

Once the sequence of the human genome has been available, the next logical step was to annotate the genes it contained. According to the canonical definition of a gene at the molecular level, that can be found also today in any genetics textbook, it meant looking for DNA regions that are transcribed into a messenger RNA, that, in turn, is to be translated (according to the genetic code) into a protein. Hence, sequencing techniques applied this time to RNA sequences (instead of DNA – although they had to be transformed first in cDNA molecules), and the following assignment of each sequence to the corresponding genomic locus permitted to identify the position and the sequence of genes, as well as the protein they encoded. This annotation process involved thousands of laboratories and research groups worldwide, working together or independently, and from the very beginning it produced quite a surprising result: human genes were few, an constituted just a little fraction of the overall DNA sequence of the genome. All in all, the final count (valid also as we know today) was little more than 20,000 protein coding genes – much less, at least by one order of magnitude – than initial estimates. So, a human being, according to the canonical definition of gene, was build by a surprisingly low number of proteins. Even if it was known that through mechanisms like alternative splicing a single gene could produce alternative transcripts and hence more than one protein, this latter phenomenon was considered an exception, rather than the rule. Furthermore, the regions of DNA actually encoding for proteins, that is, forming the protein coding portion of mRNAs, constituted no more than the 3-4% of the overall genomic sequence.

More surprising results kept arriving while the sequencing machines available kept churning out the sequence of the whole genome of species other than human, first of all of the "model" organisms used for studies in genetics and molecular biology, from mouse, to

rat, to the fruit fly, to the zebrafish, to plant models like Arabidopsis thaliana. Not only human genes were "few", but also less than the number of other species that, if the complexity of an organism had to be correlated with the number of protein coding genes, should have had much less genes.

And, furthermore, comparative genomic studies, that as the word says compared genes and genomes of different species to one another, revealed striking similarities and conservation both in gene number and in sequence. For example, the number of genes is virtually the same in human and mouse, and any other mammal. And, at the sequence level, each gene has remained conserved throughout evolution to the point that the protein it encodes can be assumed to have the same structure, and hence function, in either species (homologous genes). Hence, the conclusion that "men and mice had the same genes" (and all the other mammals, for that matter). And, human (and mammals) had less protein coding genes than, for example, fishes and plants; but also less than much "simpler" species like C.elegans, a nematode worm barely visible by the naked eye.

So, the general conclusion was that genes "weren't enough", at least to understand and explain different levels of complexity and evolution in the different species.

Gene expression and its regulation

Once whole genome annotations were available, other experimental techniques like oligonucleotide microarrays were introduced for measuring the level of expression of genes. Given a uniform cell population, these methodologies permitted to first of all to identify which genes were active and produced transcripts, but also to quantify their transcript level, or, better, its variation across different samples or conditions, by capturing the corresponding RNAs. It should be kept in mind that while DNA can be considered static, that is, contained in the same sequence by all the cells of an organism, RNA is dynamic, in other words, not all genes are actively transcribed and expressed by all the cells. Their expression levels depend on the type of cell and its status, and then expression changes according to cell type, developmental stage, external stimuli, and so on. The results of thousands of experiments of this kind, in which different comparisons were made (*e.g.*

developmental stages; different adult tissues; normal vs cancer cells; etc.) revealed that gene expression, or better, transcription is a very finely modulated phenomenon, in every species. Groups of hundreds of genes sharing similar functions are activated, blocked, or change their expression simultaneously in a similar fashion. And, also, disease could often be associated to the over- (or under-) expression of groups of key genes. Hence, there had to be a very precise regulatory system, in living cells, able to orchestrate the activation or repression of the transcription of genes, according to some precise rules. Some of the key players of this system were already known to be transcription factors, protein encoded by the genome itself that, binding DNA at the right positions (usually next to the genes, and however outside the genes themselves), are able to recruit the transcriptional apparatus to the right point (the transcription start site of genes), at the right time, and with the right frequency. Transcription factors bind DNA in a sequence-specific way, that is, bind DNA when they find a precise arrangement of nucleotides on the sequence. For example, factor TBP (TATA-binding protein) was named from the fact that it binds DNA when it finds four nucleotides forming the sequence TATA; likewise for the factors of the family GATA; and so on.

At this point, if the mechanisms of regulation of transcription were solely dependent on the action of transcription factors, we should be able to find on DNA a precise "regulatory code", with binding sites for different combinations of transcription factors associated to each gene, and on the other hand genes with similar expression patterns having similar "codes". But, apart for a few anecdotal cases, there is little or no evidence of such a code, or at least not enough to be able to reconstruct exactly the mechanisms behind the changes of expression observed. That is, binding sites for transcription factors are found at the "right" positions with respect to genes when we know they are transcribed. But, if we reverse the approach and look for DNA for the binding sites trying to infer where is the corresponding gene and their effect on regulation we get little or no result. There are hundreds of thousands of "TATA" or "GATA" nucleotides spread along the three billions base pair of the human genome. But, only a small fraction of these sites are actually bound by the corresponding factor, not always at the same time, and by looking at the DNA sequence alone there is no way to determine which are actually the functional ones. So, once again, there was some piece of information lacking: if genes were not enough,

then the genome was not enough as well, to explain the complex mechanisms of expression regulation by sequence alone.

Next-generation sequencing and next-generation genes

The two previous sections shortly recapitulate what was the situation more or less ten years ago. The "post-genomic" era, in which organisms could be studied from the point of view of their entire genome had given some answers, but had also raised new questions. However, at the time another major breakthrough took place, with the introduction of "next-generation" sequencing technologies. Without delving into details, a single new generation sequencing machine can be seen as encompassing millions and millions of "first generation" sequencers, able thus to sequence in parallel millions of DNA molecules. More importantly, at a fraction of the original costs: while the first draft of the sequence of the human genome has been a worldwide cooperative effort, lasting more than ten years, involving hundreds of researchers and technicians, and with a final price tag of hundreds of millions of dollars, a human genome can be sequenced today by a small lab, in a week, for a few thousands dollars.

These new technologies thus constituted a paradigm shift, where from the genome of a species research could move to the individual genome, like in the "1000 genomes project" aimed at pinpointing variation between different human individuals, or genome wide association studies looking for mutations or any other type of variation that could be associated with disease.

But, more importantly, the sheer number of sequences that can be produced nowadays permit to observe more in depth genomes, their genes, and their products, that is, RNAs. One of the most relevant discoveries has been that, with respect to the original definition of gene that assigned one RNA to each gene, with some exceptions, a single eukaryotic gene instead produces several different RNAs, through alternative splicing of the same pre-mRNAs. Latest estimates assign 7-8 transcripts per human gene. Hence, the "few" human genes have the potential of producing a much larger repertoire of proteins, through "creative" usage of their RNAs. And the "one-to-one" correspondence, for example, between human and mouse genes is no longer kept at the level of their alternative transcripts. So, while genes are the same, the

two species use them in a slightly different way. A striking example is that in a human-mouse comparison we can observe that genes active in muscle cells have very similar splicings and alternative transcripts in the two species. But, on the average, genes expressed in the human brain seem to be more complex and to produce more alternative transcripts than their mouse counterparts. Hence, the complexity of a species, not related to the overall count of its protein coding genes, seems instead to correlate with the complexity of the genes themselves and their usage.

The application of next-generation sequencing, however, yielded more revelations, the most important being that a gene, or a region transcribed into a RNA, does not necessarily have to encode for a protein. This was known for a handful of RNAs (tRNAs, rRNAs) that are used by the cell to translate messenger RNAs into protein. But, deep sequencing revealed the presence of thousands of non protein coding RNAs in eukaryotic genomes, some very short (microRNAs), some very long and similar to normal messenger RNAs, but anyway non coding. The function of these RNAs is in several cases not completely understood yet, the most likely being in turn the regulation of the expression of protein coding genes. In any case they constitute another layer of complexity and information of the genome, being non coding RNA "genes" present in the human genome at least in the same number of protein coding genes. All in all, these recent discoveries point to the fact that the classic definition of gene as DNA unit encoding for a protein has to be revised an extended. A protein coding gene can produce several different proteins, as a rule and not as an exception as previously believed. And, a transcribed region, hence a "gene", non necessarily produces a mRNA encoding for a protein as a rule.

Next-generation sequencing and epigenomes

Studying DNA and genomes at the sequence level has permitted to reach new milestones of modern science. It should be kept in mind, however, that inside nuclei of living cells DNA is a molecule, with a precise structural organization, and not a linear sequence. The double strand of DNA is bound to protein complexes forming chromatin. The nucleosome is the fundamental subunit of chromatin. Each nucleosome is composed of a little less than two turns of DNA wrapped around a set of eight proteins called histones, which are known as a histone octa-

mer. The nucleosome core particle consists of approximately 147 base pairs of DNA wrapped in 1.67 left-handed superhelical turns around a histone octamer consisting of 2 copies each of the core histones H2A, H2B, H3, and H4 (see *Fig. 1*). Nucleosomes are folded through a series of successively higher order structures to eventually form a chromosome. The key point is that the structures formed by nucleosomes can be different at different positions of a chromosome, with the main effect being the DNA sequence more or less accessible. Hence, nucleosomes provide the first level of regulation of gene expression: if a chromatin region is "closed", DNA is not accessible, and the genes located in the region cannot be transcribed simply because the transcriptional machinery cannot contact DNA.

The main factor determining the structure of nucleosomes and chromatin, and hence the accessibility of DNA, are biochemical modifications brought by specific factors on DNA itself, and more importantly the histones around which DNA is wrapped. That is, according to

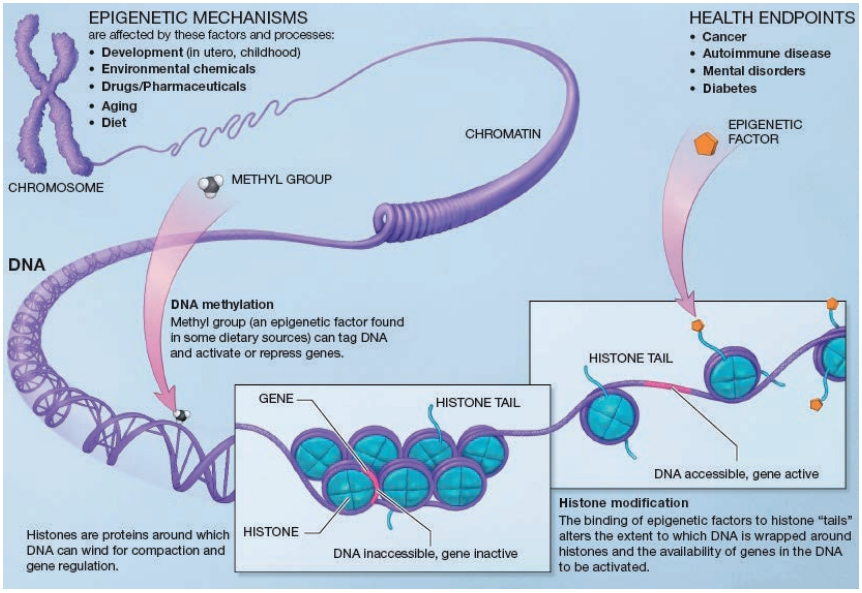

Fig. 1 – The structure of chromosomes, nucleosomes, and the main epigenetic factors: DNA methylation and histone modifications.

the biochemical state of the histones, chromatin can be closed, or be partially or completely open, exposing DNA. All these phenomena were already known before the genomic era, and studied by a branch of genetics called "epigenetics", that, in general, studied anything other than DNA sequence that influenced the development of an organism.

Quite curiously, a major breakthrough in this field, not interested in the DNA sequence, came from the very introduction of next-generation sequencing technologies. The key point, in this case, was that they are not applied to the whole genome, but to selected parts of it isolated through experiments like chromatin immunoprecipitation (ChIP). Since histones can carry several different chemical modifications, ChIP experiments permit, given a modification of interest, to isolate only the DNA regions that are wrapped around nucleosomes whose histones carry the modification itself. In order to identify which are these regions, we can simply sequence them with a next-generation sequencing platform. In this way, the complete map on the whole genome of the localization of each modification can be built, and its effect on DNA structure and accessibility studied more in depth. Before the introduction of these experiments some pieces of information were already available, where some histone modifications seemed to correlate with gene transcription, and other with silencing. But the large scale application of next-generation experiments to several different histone modifications in different cell lines (as for example in the ENCODE or Roadmap Epigenomics Project) permitted to unveil a more complex picture, showing a very precise "histone code" regulating the transcription of genes. The different histone modifications, and the different ways in which they can be combined on a nucleosome, can be seen as "signals" on DNA, marking for the transcriptional machinery if and where the transcription of a gene should start, and where it should end. Other modifications have the effect of "enhancing" the frequency of transcription, while others block the transcriptional machinery like "do not enter" signs. Bioinformatic experiments, where the position of several histone modifications on the genome was crossed with the level of transcription of genes revealed that indeed this code is able to explain the patterns of expression we observe, since it is possible to predict with high accuracy not only if a gene is transcribed or not, but also its transcript level by looking at the conformation of the histones around it.

While the genome is static, the epigenome, that is the map of the

position of histone modifications, or other factors like DNA methylation, is highly dynamic, as much as the expression of genes it regulates. Hence, each cell has a different epigenome, according to its developmental stage, or tissue, and so on. For example, cell differentiation from totipotent stem cells to pluripotent and finally unipotent adult cells is characterized by a beautiful and precise evolution of some epigenetic factors. And, more importantly, cells can change their epigenome according to external stimuli. In other words, the epigenome is where the first reaction and the adaptation of an organism to environment takes place. We can clearly observe change in the epigenome, for example, of starved animals or plants exposed to cold or drought. And, much more importantly, epigenetic features like DNA methylation and histone modifications are inherited. That is, a cell passes its epigenetic state to its daughter cells, and hence how it was adapted and the expression of its genes in response to the environment. So, if our nature is written in our DNA and our genes, which are protected from modifications by very sophisticated repair mechanisms, then our nurture is maybe written in our epigenome, which instead changes and saves in our cell the effect the environment had on us.

I started teaching Bioinformatics and Genomics at the University of Milan right at the beginning of the post-genomic era, and I remember telling my students that we were, at the time, like Galileo, who had just built a telescope for observing the sky. More than ten years later I still think that this metaphor holds true, and day after day we keep looking at our cells with more and more powerful telescopes, discovering something new at each observation: but we are still far from having a complete picture of the wonderful universe that is inside every single living cell.