Genome as a multipurpose structure built by evolution

Michel Morange

Genome as a multipurpose structure built by evolution

Michel Morange, Centre Cavaillès, République des savoirs
USR 3608, Ecole normale supérieure, 29 rue d'Ulm, 75230 Paris Cedex05, France
Email: morange@biologie.ens.fr

Abstract

The publication in 2012 of the results of the ENCODE program generated an acrimonious debate about the role of junk DNA. This debate is a symptom of the difficulties of dovetailing functional and evolutionary descriptions of genomes.

My argument is that extant genomes are the result of a progressive evolutionary construction. To the basic function – to give rise to RNA and proteins – have been successively added other functions – regulation by microRNA and epigenetic marks, for instance. This process of complexification was not a regular one, but the result of a complex evolutionary history, different for the different genomes.

Better knowledge of the evolutionary history of genomes would help to understand these structures and their functions.

Historians and philosophers of biology have amply discussed the difficulty, even the impossibility, of providing a precise definition of a gene (Beurton et al. 2000). The genome no longer appears as a collection of similar genes, but as a multitasking structure formed of different types of genetic elements (Gingeras 2006). Participants in these debates nonetheless often agree on one issue, that the genome is (and has to be) something well defined: a collection of genes, a

reservoir of junk DNA, or a structure filled with regulatory sequences.

My point of view will be different: the genome is what evolution has progressively made of it. This vision has three important consequences. The first is that one must not expect the genome to be precisely organized, with well-defined functions for its different subparts: in particular, the action of natural selection is counteracted by the accumulation of neutral or even slightly deleterious mutations that have slipped through its sieve. The second is that these different structures and functions are not quite comparable: they do not have the same seniority, and have not been moulded by the same evolutionary mechanisms. The third is that every genome has a different evolutionary history. One has to be very cautious when extrapolating the observations made on one genome to all other genomes.

I will briefly reiterate the main steps that led to the discovery of the complexity of the genome. Then, I will analyze the controversy surrounding the publication of the results of the ENCODE program in 2012 as an example of the difficulties of dovetailing functional and evolutionary visions. And finally, I will discuss recent results supporting an evolutionary vision of the genome, and what we know, or more often don't, of this evolutionary history.

The progressive discovery of the complexity of the genome

There was a golden age of simplicity for the gene extending from the early attempts by August Weismann, Hugo De Vries and others to design a material and corpuscular model of inheritance, through the "classical genetics" that dominated during the first half of the XXth century, up to the demonstration that genes were made of DNA, and the discovery of DNA structure.

In the Pilgrim Trust Lecture that he delivered in November 1945, Hermann Muller gave different but similar definitions of genes: “guides, some elements that are themselves relatively invariable and that serve as a frame of reference in relation to which the passing phases of other features are adjusted” and “relatively stable controlling structure, to which the rest is attached and about which it in a sense revolves” (Muller 1947, 1).
Each gene may have a different role, but the types of functions that are fulfilled by genes are similar. It was this unified vision of genes that was included in the Modern Evolutionary Synthesis. Genes can be defined by their variations, and the way they affect the characteristics of organisms. All genes are somehow equivalent since their variations lead in one way or another to a modification of organisms. The discovery that all genes have the same chemical nature strengthened this unified conception. Genomes were simply the collection of individual genes.

But the molecularization of the gene was also the origin of the difficulties that progressively appeared at the end of the 1950s and beginning of the 1960s. Seymour Benzer was the first to separate different characteristics that had been so far associated with the gene: to be a unit of function, a unit of recombination, and a unit of mutation. He introduced three different words for the objects bearing these different characteristics that have not been retained. This unified conception of the gene was also challenged by the distinction made by François Jacob and Jacques Monod between structural and regulatory genes (Jacob and Monod 1961). In their model, not all genes have the same type of function, which also means that the evolutionary consequences of their variations will be different. Another aspect of the operon

model was also puzzling: the product of the regulatory gene interacted with a DNA sequence called an operator, positioned upstream of the structural genes that it controlled. In their first publication, Monod and Jacob called this sequence the "operator gene" (Jacob and Monod 1961, 344), before subsequently dropping the name "gene". In the same years, it appeared also that genes could have different functions: encode the structure of proteins or permit the synthesis of functional RNA such as ribosomal and transfer RNA.
Another difficulty became obvious at the end of the 1960s: the size of genomes was not related to the complexity of organisms, what Charles Thomas called in 1971 the C-value paradox (Thomas 1971). More puzzling still was the fact that evolutionarily close organisms could have highly different amounts of DNA in their genomes.
The emphasis put on this C-value paradox paralleled the discovery by molecular hybridization that the genome was full of repeated sequences (Britten and Kohne 1968). These repeated sequences were rapidly suggested to have a regulatory role (Britten and Davidson 1969 and 1971). But the progressive characterization of genome sequences with the newly elaborated tools of genetic engineering at the end of the 1970s showed that many of these repeated sequences were transposons and retrotransposons. It was discovered in 1977 that eukaryotic genes are split into exons, whose sequences are present in mRNA, and introns separating them, and apparently devoid of any obvious function.
Evidence was also found for the existence of inactive copies of genes, or pseudogenes. The first pseudogene described in *Xenopus* in 1977 was a nonfunctional copy of the gene responsible for the synthesis of 5SRNA (Jacq et al. 1977). In 1979 and 1980, pseudogenes corresponding to the genes encoding globin proteins were discovered (Hardison et al.

1979; Lauer et al. 1980), and these early observations were rapidly extended to other genes and gene families.

As early as 1972, Susumu Ohno had suggested that genomes are replete with nonfunctional DNA that he called “junk DNA” (Ohno 1972). In 1980, two articles published in *Nature* (Doolittle and Sapienza 1980; Orgel and Crick 1980) popularized the view, based on Dawkins’s ideas, that most parts of the genome are formed of parasitic selfish DNA elements that are not eliminated by natural selection. The nonadaptive value of the large size of the eukaryotic genome is still widely accepted by population geneticists (Lynch 2007): the strongest argument is that “minute” genomes present in some species are fully functional, despite the elimination of most of this junk DNA (Ibarra-Laclette et al. 2013).

Regulatory regions distant from the genes that they control (enhancers) were described at the beginning of the 1980s. Epigenetic modifications of the genome have recently attracted most attention from biologists. Epigenetic marks were discovered at the beginning of the 1960s (histone modifications) and mid-1970s (for DNA methylation), but the two types of observations only converged at the end of the 1990s (Morange 2013). Simultaneously, and not fully independently, noncoding RNAs were described. The first were small interfering RNAs (siRNAs) involved in the defence against invading DNA sequences (Morange 2012). siRNAs are also involved in the control of epigenetic marks. MicroRNAs, whose mechanism of action has much in common with that of siRNAs, had been described earlier, but their important regulatory function in development only became obvious at the beginning of the 2000s. siRNAs are the results of the degradation of exogenous (viral) RNAs, whereas microRNAs are naturally produced by transcription of the

genome. More recently it has been shown that there is an abundance of long noncoding RNAs, most of which are transcribed from intergenic sequences (lincRNAs), but their function remains the topic of numerous discussions.
Beyond the complexity of these recently described phenomena, and the difficulty of attributing to some of them an obvious functional significance, these observations have raised two additional difficulties. The first is the impossibility of attributing in some (many) cases one specific function to a DNA sequence: it can be an enhancer, and simultaneously transcribed into noncoding RNAs. The same sequence can correspond to an intron, encode a microRNA, and contain binding sites for transcription factors. The second difficulty is that comparison of DNA sequences from different organisms demonstrates that their functions are far from being stable during evolution: some of the DNA fragments have recently acquired or lost functions. The most striking phenomenon is the apparently not so infrequent recruitment of noncoding sequences to encode new functional proteins (Neme and Tautz 2013).

The recent controversy about the ENCODE program

The very active controversy that emerged at the end of 2012 and beginning of 2013 on the interpretation of the results produced by the ENCODE consortium is closely related to the previous issue.
Encyclopedia of DNA elements (ENCODE) was a program launched in 2003 by the National Human Genome Research Institute (The ENCODE Project Consortium 2004) to complement the results of the human genome sequencing program: it aimed in particular to attribute a function to the 98% of the genome that does not encode proteins. The answer was looked for in the combination of different technical

approaches: characterization of the histone (and chromatin) modifications at any position in the genome; characterization of the regulatory regions by their sensitivity to a DNase I treatment; positioning on the genome of the main transcription factors; characterization of all the transcripts, including the noncoding RNA; the search for long-range chromatin interactions; etc. This was done on 147 different human cell types. Preliminary results were published in 2007 on 1% of the genome (The ENCODE Project Consortium 2007) and the full results in 30 articles at the end of 2012 (The ENCODE Project Consortium 2012). More recently, a similar program for model organisms, the nematode *C. elegans* and *Drosophila*, has been launched, and some researchers are pushing to do the same for the zebrafish (Sivasubbu *et al.* 2013).

The abstract of the 2012 presentation reported that it had been possible to assign a biochemical function to 80% of the genome (The ENCODE Project Consortium 2012, 57). This result made a huge splash in scientific and general journals. The same week, S*cience* announced the results of ENCODE under the title "ENCODE Project writes eulogy for junk DNA" (Pennisi 2012), and the article stated that 80% of the human genome is functional – not exactly what was said in the ENCODE consortium presentation.

Many critical articles – some angry in tone – were published in the following months, arguing that the results of the ENCODE program did not demonstrate that 80% of the human genome was functional and have not sounded the death knell of junk DNA (Eddy 2012; Niu and Jang 2013; Graur et al. 2013; Doolittle 2013; Eddy 2013) – see below.

The controversy was rooted in part in an opposition to these huge time-consuming and costly programs, whose participants have to justify their utility and in so doing were accused of

having overstated the significance of the results. Also, journalists tended to focus on these “sensational” results and neglect other, less problematic but still important findings. Sean Eddy raised the interesting issue that these huge programs do not have the same ambitions as traditional scientific work, and therefore should not be evaluated using the same criteria (Eddy 2013). They can be developed to solve an important scientific question: this situation is frequent in physics – consider, for instance, the efforts deployed to find evidence for the Higgs boson – but absent in biology. These huge programs may also be designed to build a map that will be further used by other researchers, or to develop new technologies when current ones represent a bottleneck for the growth of scientific knowledge. The ENCODE program had this dual ambition, which means that its results cannot be evaluated immediately, but will be revealed through the work that they will ultimately permit. The data produced by the ENCODE consortium are a foundation on which researchers interested in a specific issue can build their own projects. Beyond the debate on the utility of these huge programmes, there was a more focused debate on the attribution of a function to 80% of the genome. The authors of the different ENCODE publications were accused of having surreptitiously mixed three different meanings of the word “function”. The attribution of a function to a DNA sequence can be deduced from the effect that a modification of this sequence has on the fitness of the organism. A biochemical function can also be attributed to a DNA sequence if, for instance, this sequence binds a transcription factor or is transcribed into RNA. A DNA sequence can also be ascribed a function if this sequence bears the putative marks of a function, for instance if it possesses the sequences required for the binding of a regulatory protein or for the initiation of transcription – in the

absence of any evidence that the protein binds or that the sequence is transcribed. The only significant functional sequences would be those that fulfil the first criterion, whereas the authors of ENCODE articles considered as sufficient the second and even the third criterion.
Observations can be explained by the fact that biological systems are noisy: transcription factors can interact at many nonfunctional sites, and transcription initiation takes place at different positions corresponding to sequences similar to promoter sequences, simply because biological systems are not tightly controlled (Struhl 2007). In addition, the biochemical functions that are detected are not all for the benefit of the organism: some of the transposons and retrotransposons present in the genome have retained their activities – transcription, production of DNA-binding proteins. The biochemical activities are those of the transposons, and nothing demonstrates that these activities have any impact on the organism.
More precise critiques targeted the way the work has been conducted, in the choice of the cell lines, for instance. The meaning of the "80%" was also discussed. A value of 80% meant that if the genome was cut into fragments of a certain size, 80% of these fragments would have a biochemical function – whatever it may be. If fragments of a different size had been selected, the result would have been different – revealing the arbitrary nature of this value.

With the flow of results coming from ENCODE, a recurrent debate re-emerged early about the possibility to elaborate a new definition of the gene, the previous simple definitions having been invalidated. It is not obvious that the new definition that was proposed – "a gene is a union of genomic sequences encoding a coherent set of potentially overlapping

functional products" (Gerstein et al. 2007) – will captivate all biologists!
Another issue was raised by Thomas Gingeras concerning the level at which the next definition of the gene had to be given (Gingeras 2007). The situation is simpler at the level of the transcripts: a messenger RNA can be clearly distinguished from a microRNA or from a long noncoding RNA, although some long RNAs can be cleaved into fragments with different functions (Tuck and Tollervey 2011). For Gingeras, "transcripts could be used to define the operational units of a genome" (Gingeras 2007). This would be a return to the hypothetical RNA world, when the genetic material was RNA, although many of the functions now present were probably not associated with these early RNA genomes.

An interesting discussion also took place about the word "junk". It means that, presently, a fragment of DNA has no function likely to be screened by natural selection. "Junk" is not synonymous with "garbage", which would imply that this DNA fragment is of a different nature, preventing it from ever being likely to have any function. A fragment of junk DNA may have had a function in the past, and may be recruited in the future to play a functional role.
This debate emphasized the difficulties that functional and evolutionary biologists have in interacting. For instance, the idea often supported by molecular biologists that junk DNA is a reservoir for future evolutionary transformations has no sense for evolutionary biologists: natural selection only operates in the present, not on future outcomes. The difficulties corresponding to the superposition of different functions on the same DNA fragment are, for an evolutionary biologist, the normal result of the game of evolution, permanently adding or suppressing functions.

The evolutionary history (and prehistory) of genomes

The existence of different functions within genomes and in the transcripts means that there are different "causal roles" that have been progressively implemented for a single structure, the genome. The motors of this process are contingency and natural selection.

The evolutionary history of genomes has yet to be written. The most informative case, because the historical path is often recent and has not yet been erased, is that of microRNAs. Although their existence is probably ancient, they increased dramatically in bilaterian animals, and there was an apparently rapid divergence in mRNA-target relations. It has now been firmly established that they have different structural origins (Berezikov 2011). Some are derived from transposons (Piriyapongsa et al. 2007). Depending upon this origin, their mechanism of action as well as stability and specificity can vary.

Genomes have different evolutionary histories: the most distant two genomes are, the most different their evolutionary histories have been. The same is true, and even more obvious, for epigenomes. Epigenetic marks and their functions differ between animals and plants. And in mammals, the same phenomenon, the inactivation in females of one X-chromosome by epigenetic marks, results from different mechanisms with distinct evolutionary histories (Escamilla-Del-Arenal et al. 2011). These studies reveal different facets of the tinkering action of evolution (Jacob 1977). But the notion of tinkering has to be complemented: it is necessary to explain why the system can be tinkered with. For instance, the existence of molecular noise in transcription and of alternative splicing were favourable to the formation of new transcripts, the production of which could ultimately be stabilized by

natural selection if they were likely to generate interesting new regulatory properties. Much work has already been done (see, for instance, Lev-Maor et al. 2003) on the way the functioning of the genome can facilitate the introduction of new functions with a limited alteration of the pre-existing ones. Tinkering must not be seen as a principle excluding any novelty from evolution: the production of new proteins from noncoding DNA that has been revealed by recent studies is a good example of the creative power of evolution.

It is obviously too early, and maybe in the end it will be impossible, to propose undisputed evolutionary scenarios for the construction of extant genomes. Current scenarios remain purely hypothetical. It seems reasonable to imagine that present genomes with one or a limited number of chromosomes were preceded by genomes formed of an ensemble of independent genetic units – somehow similar to plasmids, or to self-replicating RNAs in the hypothesis of a RNA living world having preceded the extant protein-DNA world. Regulatory signals were probably present in these independent units. These signals became more and more important with the development of gene networks. The invention of chromosomes permitted the addition of new regulatory signals, distant from the genes and possibly common to different ones: this emergence of chromosomes was not a precondition for the existence of regulatory signals but it offered additional opportunities. These various transformations of the genome interacted, although their occurrence might have been independent. For instance, the development of epigenetic marks could have permitted genomes to control the invasion by foreign sequences such as transposons (Fedoroff 2012), and therefore the accumulation of these transposons in the genome. This junk DNA might

subsequently have generated microRNAs or long RNA transcripts.
The evolutionary history of the genomes was probably as complex and as diverse as the evolutionary history of organisms. Let us hope that it will be possible in the future to have access, at least partial, to this evolutionary history. This might cast new light on the functions of genes and genomes, and probably help to discard some useless controversies such as those that emerged with the results of the ENCODE program.

Acknowledgments: The author is indebted to David Marsh for critical reading of the manuscript, and to the two reviewers for helpful suggestions.

References

Berezikov, E. 2011. Evolution of microRNA diversity and regulation in animals. *Nature Reviews/genetics* 12:846-860
Beurton, P., R. Falk, and H.-J. Rheinberger eds. 2000. *The concept of the gene in development and evolution*. Cambridge: Cambridge University Press
Britten, R. J., and D. E. Kohne. 1968. Repeated sequences in DNA. *Science* 161:529-540
Britten, R. J., and E. H. Davidson. 1969. Gene regulation for higher cells: a theory. *Science* 165:349-357
Britten, R. J., and E. H. Davidson. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol* 46:111-133
Doolittle W. F. 2013. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci USA* 110:5294-5300
Doolittle, W. F., and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284:601-603

Eddy, S. R. 2012. The C-value paradox, junk DNA and ENCODE. *Current Biol* 22(21):R898-R899
Eddy, S. R. 2013. The ENCODE project: missteps overshadowing a success. *Current Biol* 23(7):R259-R261
Escamilla-Del-Arenal, M., S. Teixeira da Rocha, and E. Heard. 2011. Evolutionary diversity and developmental regulation of X-chromosome inactivation. *Hum Genet* 130:307-327
Fedoroff, N. V. 2012. Transposable elements, epigenetics, and genome evolution. *Science* 338:758-767
Gerstein, M. B., et al. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Res* 17(6):669-681
Gingeras, T. R. 2006. The multitasking genome. *Nature Genetics* 38(6):608-609
Gingeras, T. R. 2007. Origin of phenotypes: genes and transcripts. *Genome Research* 17(6): 682-690
Graur, D., et al. 2013. On the immortality of television sets: « function » in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* 5(3):578-590
Hardison, R. C., et al. 1979. The structure and transcription of four linked rabbit β-like globin genes. *Cell* 18:1285-1297
Ibarra-Laclette, E., et al. 2013. Architecture and evolution of a minute plant genome. *Nature* 498:94-98
Jacob, F. 1977. Evolution and tinkering. *Science* 196:1161-1166
Jacob, F., and J. Monod. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3:318-356
Jacq, C., J. R. Miller, and G. G. Brownlee. 1977. A pseudogene structure in 5S DNA of *Xenopus laevis*. Cell 12:109-120
Lauer, J., C.-K. J. Shen, and T. Maniatis. 1980. The chromosomal arrangement of human α-like globin genes:

sequence homology and α-globin gene deletions. *Cell* 20:119-130
Lev-Maor, G., et al. 2003. The birth of an alternatively spliced exon: 3’ splice-site selection in *Alu* exons. *Science* 300:1288-1291
Lynch, M. 2007. *The origins of genome architecture*. Sunderland MA: Sinauer Associates
Morange, M. 2012. Transfers from plant biology: from cross protection to RNA interference and DNA vaccination. *J Biosci* 37 (6):949-952
Morange, M. 2013. The long and tortuous history of epigenetic marks. *J Biosci* 38:1-4
Muller, H. J.. 1947. The gene. *Proc Roy Soc B* 134:1-37
Neme, R., and D. Tautz. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution. *BMC Genomics* 14:117-129
Niu D.-K., and Li Jiang. 2013. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem Biophys Res Comm* 430:1340-1343
Ohno, S. 1972. So much « junk » DNA in our genome. *Brookhaven Symp Biol* 23:366-370
Orgel, L. E., and F. H. C. Crick. 1980. Selfish DNA: the ultimate parasite. *Nature* 284:604-607
Pennisi, E. 2012. ENCODE project writes eulogy for junk DNA. *Science* 337:1159-1161
Piriyapongsa, J., L. Marino-Ramirez, and I. King Jordan. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* 176:1323-1337
Sivasubbu, S., C. Sachidanandan, and V. Scaria. 2013. Time for the zebrafish ENCODE. *J Genet* 92(3):695-701
Struhl, K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Struct Mol Biol* 14:103-105

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* 306:636-640

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799-816

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57-74

Thomas, C. A. Jr. 1971. The genetic organization of chromosomes. *Annu Rev Genet* 5:237-256

Tuck, A. C., and D. Tollervey. 2011. RNA in pieces. T*rends Genet* 27:422-432