MONOGRAPH
MÈTODE Science Studies Journal, 6 (2016): 135–141. University of Valencia.
DOI: 10.7203/metode.6.4083
ISSN: 2174-3487.
Article received: 11/09/2014, accepted: 02/03/2015.

GENES, GENOMES, AND CODES

REVISITING SOME KEY TERMS WITH MULTIPLE MEANINGS

EVELYN FOX KELLER

Is a genome the full complement of an organism's genes or of its DNA? Is genetics the study of genes or of heredity? Is the genetic code the mechanism for translating nucleotide sequence to amino acid sequence or to phenotype? Does «genetic information» refer to the sequences coding for proteins or to all DNA sequences? Each of these questions stems from an elision between one, concrete, meaning, and another, open-ended and ambiguous. Such elision invites the illusion that the ambiguity of the open-ended term has been resolved, and by implication, that the gap between actual achievement and promise has been closed. Yet, despite the phenomenal progress molecular biology has made, we remain without an adequate account of the organization of proteins into an organism.

Keywords: code, code-script, central dogma, genetic information, genes and genomes.

In a recent commentary celebrating the current state of (or, as he emphasizes, current gaps in) our understanding of DNA, Philip Ball, a former editor of *Nature*, wrote:

> This week's diamond jubilee of the discovery of DNA's molecular structure rightly celebrates how Francis Crick, James Watson and their collaborators launched the «genomic age» by revealing how hereditary information is encoded in the double helix. Yet the conventional narrative [...] is as misleading as the popular narrative of gene function itself in which the DNA sequence is translated into proteins and ultimately into an organism's observable characteristics, or phenotype.
>
> (Ball, 2013, p. 419)

A bit later, he added:

> A student referring to textbook discussions of genetics and evolution could be forgiven for thinking that the «central dogma» devised by Crick and others in the 1960s – in which information flows in a linear, traceable fashion from DNA sequence to messenger RNA to protein, to manifest finally as phenotype – remains the solid foundation of the genomic revolution. In fact, it is beginning to look more like a casualty of it.
>
> (Ball, 2013, p. 419)

«DESPITE THE PHENOMENAL PROGRESS MOLECULAR BIOLOGY HAS MADE, WE REMAIN WITHOUT AN ADEQUATE ACCOUNT OF THE ORGANIZATION OF PROTEINS INTO AN ORGANISM»

In other words, we celebrate the Watson-Crick revelation of «how hereditary information is encoded in the double helix» while at the same time admitting the utterly misleading nature of the «conventional narrative» of their discovery – a narrative «as misleading as the popular narrative of gene function itself».

But what exactly is it that is so misleading? Ball actually gives us two narratives – one he refers to as the conventional, the other as the popular narrative; one is a claim about hereditary information, the other a claim about the genetic code. These are not the same. Then how is it that such «misleading» narratives are so routinely perpetuated in the teaching of molecular biology?

Part of the answer to this question is to be found in the replication of this ambiguity throughout the discourse of molecular biology (preceded by a parallel set of ambiguities in the discourse of classical genetics) that has worked for sixty years to simultaneously sustain and obscure what Ball now sees as misleading. I begin with the two narratives that Ball invokes: a) DNA sequence codes for proteins which ultimately form an organism's observable characteristics, or

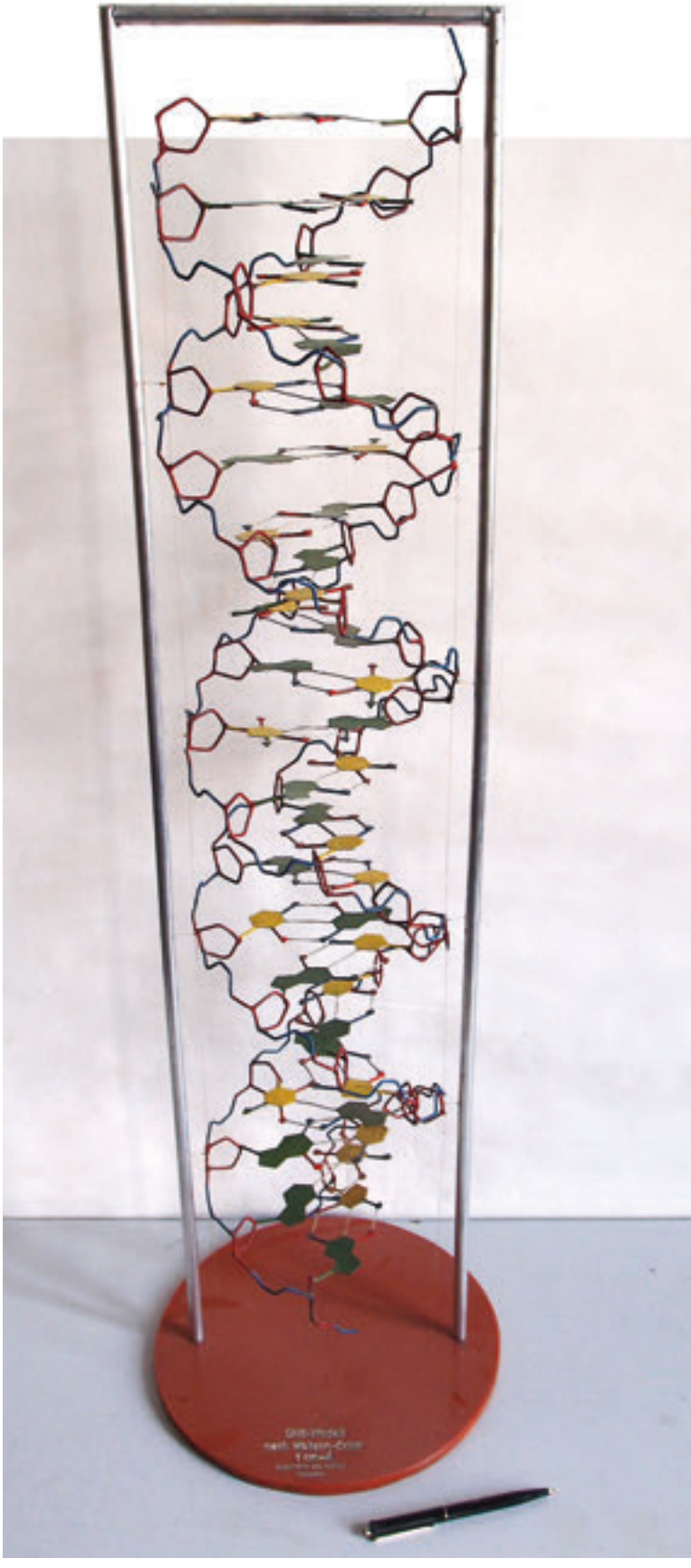
phenotype. In popular lingo, DNA makes RNA, RNA makes proteins, and proteins make us; b) hereditary information is encoded in the double helix.

The concept of code figures crucially in both. In the first, the meaning of *code* (or *encode*) is quite clear. It derives from telegraphy and cryptography and is in fact the first definition given by the dictionary: to encode is «to translate into cipher or code; to express information by means of a *code*. Colloquially, "to code"» (Oxford English Dictionary, n. d.). As in the Morse code. Indeed, Crick was explicit about this being the sense in which he used the term code in his sequence hypothesis. «Genetic code» referred to the process of translation from a text written in nucleotide sequences to one written in amino acid sequences. Incidentally, he was also careful to distinguish the sequence hypothesis from what he called the central dogma: the hypothesis that «Once information has got into a protein it can't get out again»[1].

The deciphering of the genetic code was a tremendous achievement in the history of biology and it well deserves to be celebrated. Perhaps more surprisingly, so too was the central dogma, at least as Crick understood it. Moreover, there is nothing to suggest anything misleading in either claim. What is it then that *has* the reader led astray?

The difficulty to which Philip Ball refers arises when people speak of the hereditary information encoded in the double helix, for, in this formulation, another quite different sense of *encode* is commonly invoked, namely the information required not for a set of proteins, but for an organism. More in the sense of Schrödinger's *code-script*, a notion that preceded Crick's concept of code by fifteen years; i.e., in the sense, as Schrödinger (1944) himself wrote, that:

> Every complete set of chromosomes contains the full code; so there are, as a rule, two copies of the latter in the fertilized egg cell, which forms the earliest stage of the future individual. In calling the structure of the chromosome fibres a code-script we mean that the all-penetrating mind, once conceived by Laplace, to which every causal connection lay immediately open, could tell from their structure whether the egg would develop, under suitable conditions, into a black cock or into a speckled hen, into a fly or a maize plant, a rhododendron, a beetle, a mouse or a woman.
>
> (Schrödinger, 1944, p. 21)



David Ludwig

In the early days of genetics, there would have been no obvious reason to question the understanding of the study of genetics as the study of genes. The picture shows a DNA model from the collection of the Greifswald Botanic Garden (Germany).

[1] *Ideas on protein synthesis* (1956), collected in the Francis Compton Crick Paper at the archive of the Wellcome Library for the History and Understanding of Medicine. Retrieved from http://profiles.nlm.nih.gov/ps/retrieve/ResourceMetadata/SCBBFT

But, as he also acknowledged:

> […] the term code-script is, of course, too narrow. The chromosome structures are at the same time instrumental in bringing about the development they foreshadow. They are law-code and executive power – or, to use another simile, they are architect's plan and builder's craft – in one.
>
> (Schrödinger, 1944, p. 22)

In other words, unlike that of the Morse code, Schrödinger could not say what a code-script is. Its meaning, perforce, had to remain open-ended, characterized not by what it is but by what it is expected to do, by the answer it is hoped to provide. The question that arises is this: are these two different narratives, one referring to hereditary information and the other to sequence information, one to the hereditary code-script and other to the genetic code, or are they two different versions of the same narrative, linked by a common vocabulary? I suggest that we have here two distinct narratives, both of which have played crucial roles in molecular biology, and that what Ball characterizes as misleading is the collapse of these two narratives into one.

«IS A GENOME THE FULL COMPLEMENT OF AN ORGANISM'S GENES OR OF ITS DNA? IS GENETICS THE STUDY OF GENES OR OF HEREDITY?»

But this duality is not limited to the meanings of code. Indeed, it recurs in virtually all the new terms imported into biology with the molecular revolution. Not surprisingly, the same elisions also recur in the links commonly assumed between genes, information, codes and code-script. Take, for example, Ball's own reference to Crick's «central dogma» as the thesis that «information flows in a linear, traceable fashion from DNA sequence to messenger RNA to protein, to manifest finally as phenotype», conflating Crick's own version of the central dogma as the thesis that «Once information has got into a protein it can't get out again». The difference, namely the link between a list of proteins and an organism's phenotype, is crucial, and the locus of much of the most critical slippage in the discourse of Molecular Genetics. For example, is a genome the full complement of an organism's genes or of its DNA? Is Genetics the study of genes or the study of heredity? Is the genetic code the mechanism for translating nucleotide sequence to amino acid sequence or to phenotype? Does the central dogma refer to the information in proteins or in phenotype? Similarly, does «genetic information» refer to the sequences coding for proteins or to all DNA sequences?

Each of these questions stems from a collapse of meanings, from an elision between one, concrete, meaning, and another open-ended and ambiguous. Such elision invites the illusion that the ambiguity of the open-ended term has been resolved, and by implication, that the gap between actual achievement and promise has been closed. The fact remains however that, despite the phenomenal progress molecular biology has made, we remain to this day without an adequate account of the organization of proteins into an organism.

Two points seem worth noting. First is that these elisions are not casual but systematic. Second is their simultaneous transparency and opacity. Once identified, they seem crystal clear, plain for anyone to see; recognition depends neither on special expertise, nor on new findings. Yet this style or habit of chronic slippage from one set of meanings to the other has prevailed for over fifty years; it has become so deeply ensconced as to have become effectively invisible to most readers of biological literature.

For me, the questions of primary interest are: how have these elisions affected the research trajectory of molecular biology and what makes it possible now for Ball to write that «the conventional narrative [...] is as misleading as the popular narrative of gene function itself»? To address these questions, I focus on the tacit equation that underlies the very definition of genetics, namely, the equation of the totality of an organism's genes with its genomes, its chromosomes, and its genetics – an equation, inherited directly from the earlier language of classical genetics, that has been a staple of the discourse of molecular biology since its beginning.

GENES, GENOMES, AND JUNK DNA[2]

In the early days of genetics, there would have been no obvious reason to question the understanding of the study of genetics as the study of genes, and indeed, the frequent elision between genes and mutations in that literature suggests the expectation that there was no other chromosomal locus in which heritable mutations could arise. But from the 1970s on, especially as the focus of molecular genetics shifted to the study of

[2] For a more complete account of this history, see Keller (2014).

eukaryotic organisms, and as the study of regulation assumed increasing centrality to that science, the relation between genes and genetics has become far less straightforward. To the extent that regulation is a property of DNA, it is surely genetic, but is it always attributable to genes? Clearly, the answer depends on what is meant by the word *gene*, but taking the word in its most commonly invoked sense, the question becomes: is regulation always attributable to protein-coding sequences?

A related challenge to the equivalence between genes and genetic material came from a series of discoveries of substantial expanses of nonprotein-coding («non-genic» or «extra») DNA sequences in eukaryotic genomes. Of particular importance were the discoveries of (1) large amounts of repetitive DNA, and later, of transposable elements; (2) the wildly varying relationship between the amount of DNA in an organism and its complexity; and (3) split genes (protein-coding sequences interrupted by non-coding «introns»). However, the challenge was soon blunted by the designation of such DNA as «junk» (Ohno, 1972). After 1980, with the appearance of two extremely influential papers published back-to-back in *Nature* (Doolittle & Sapienza, 1980; Orgel & Crick, 1980), the idea of «junk DNA» seemed to become entrenched.

Borrowing Richard Dawkins' notion of selfish DNA, Orgel and Crick were explicit about their use of that term:

> […] in a wider sense, so that it can refer not only to obviously repetitive DNA but also to certain other DNA sequences which appear to have little or no function, such as much of the DNA in the introns of genes and parts of the DNA sequences between genes [...] The conviction has been growing that much of this extra DNA is «junk», in other words, that it has little specificity and conveys little or no selective advantage to the organism...
>
> (Orgel & Crick, 1980, p. 604)

Until the early 1990s, the assumption that the large amounts of non-coding DNA found in eukaryotic organisms had «little or no function», that it contributed nothing to their phenotype and could therefore be ignored, remained relatively uncontested. For all practical purposes, genomes (or at least the interesting parts of genomes) could still be thought of as collections of genes. Indeed, when the Human Genome Project (HGP) first announced its intention to sequence the entire human genome, much of the opposition to that proposal was premised on this assumption. Thus, e.g., Bernard Davis complained that:

Böhringer

Humans have especially sophisticated perceptual capacities, enabling them to respond to a wide range of complex visual, auditory, linguistic, and behavioral/emotional signals in their extended environment. Recent research has begun to show that responses to such fundamentally social signals also extend way down to the level of gene expression.

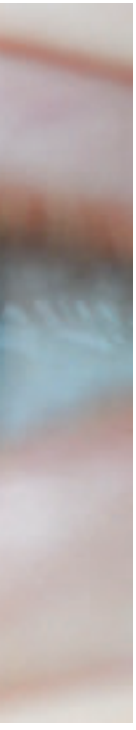
> […] blind sequencing of the genome can also lead to the discovery of new genes […] but this would not be an efficient process. On average, it would be necessary to plow through 1 to 2 million «junk» bases before encountering an interesting sequence.
>
> (Davis, 1990, p. 343)

And in a similar vein, Robert Weinberg argued:

> The sticky issue arises at the next stage of the project, its second goal, which will involve determining the entire DNA sequence of our own genome and those of several others. Here one might indeed raise questions, as it is not obvious how useful most of this information will be to anyone. This issue arises because upwards of 95 % of our genome contains sequence blocks that seem to carry little if any biological information. [...] In large part, this vast genetic desert holds little promise of yielding many gems.
>
> (Weinberg, 1991, p. 78)

Weinberg's assumptions were widely shared in the molecular biology community, and inevitably, they had consequences. Take, for example, work in medical genetics. For decades, it has been

John Goode



commonplace for medical geneticists to regard the significance of mutations in non-coding DNA exclusively in terms of their value in identifying the main actors of interest, i.e., the genes responsible for disease. Thus, for example, official descriptions of the goal of the International HapMap Project (launched in 2003) systematically confound «DNA sequence variation» (wherever it occurs) with disease «genes», promising «to develop a public resource that will help researchers find genes that are associated with human health and disease» (The International HapMap Consortium, 2004, p. 468).

A similar story can also be told about the neglect of non-coding (or non-genic) DNA in molecular evolution. For reasons partly technical and partly conceptual, work of molecular evolutionary biologists has traditionally focused on changes in the protein coding sequences of DNA, with conclusions based on the assumption that such sequences can be taken as a stand in for the entire genome. But such stories of neglect – in medical genetics, of the medical implications of non-genic DNA, in evolutionary biology – can be told now only because the assumptions on which they were based have now begun to be noticed, and accordingly, to be challenged. So what happened that made this possible (that made Ball's article possible)?

The launching of the HGP in 1990 may well have been the most significant moment in the history of our understanding of the relations between genes and *genomes*. With the rise of genomics our view of the genome as simply a collection of genes has all but collapsed. Of particular shock value were the discoveries of how few genes the human genome contained, and of how small a portion of the genome's structure is devoted to protein coding sequences. In a review article published in 2004, John Mattick displayed the ratio of non-coding to total genomic DNA as an increasing function of developmental complexity, estimating that ratio as 98.5 % for humans. The obvious question is, what is all that non-coding DNA for? Can it possibly all be junk?

«THE DECIPHERING OF THE GENETIC CODE WAS A TREMENDOUS ACHIEVEMENT IN THE HISTORY OF BIOLOGY AND IT WELL DESERVES TO BE CELEBRATED»

The notion of junk DNA handily accommodates the classical view of genomes as collections of genes. But the rise of genomic data has brought that accommodation to a breaking point. In 2003, a new metaphor came into use, one that has by now largely replaced the older one (Gibbs, 2003). Instead of «junk», non-genic DNA has become «the dark matter of the genome».

This was also the year in which the research consortium ENCODE (Encyclopedia Of DNA Elements) was formed, charged with the task of identifying all the functional elements in the human genome. Early results were reported in *Nature* in 2007, and they effectively put the kibosh on the hypothesis that non-coding DNA lacked organismic function. They confirmed that the human genome is «pervasively transcribed» even where non-coding; that the resulting transcripts are involved in forms and levels of genetic regulation heretofore unsuspected; that regulatory sequences of the resulting ncRNA may overlap protein coding sequences, or that they may be far removed from coding sequences; and that non-coding sequences are often strongly conserved under evolution.

The take-home message is clear. Genetics is not just about genes and what they code for. It is also about how the DNA sequences that give rise to proteins are transcribed, spliced, and translated into amino acid sequences, in the appropriate amounts at

With the rise of genomics, our view of the genome as simply a collection of genes has all but collapsed. In the picture, a DNA sequence at the London Science Museum.

the appropriate time and place; about how these, once assembled into proteins, navigate or are transported to the sites where, and when, they are needed, etc. All of this requires coordination of an order of complexity only now beginning to be appreciated. NcRNA transcripts of the remaining 98-99 % of the genome turn out to be crucial to the regulation of transcription, alternative splicing, chromosome dynamics, epigenetic memory, and more. They are even implicated in the editing of other RNA transcripts, and of modulating the configuration of the regulatory networks these transcripts form. In short, they provide the means by which gene expression can respond to both immediate and longer range environmental context and adapt appropriately.

Adaptation does not require direct alteration of DNA sequences: environmental signals trigger a wide range of signal transduction cascades that routinely lead to short-term adaptation. Moreover, by lending to such adaptations the possibility of intergenerational transmission, epigenetic memory works to extend short-term to long-term adaptation. As Mattick explains:

> The ability to edit RNA […] suggests that not only proteins but also – and perhaps more importantly – regulatory sequences can be modulated in response to external signals and that this information may feed back via RNA-directed chromatin modifications into epigenetic memory.
>
> (Mattick, 2010, p. 551)

Finally, environmental signals are not restricted to the simple physical and chemical molecular biology stimuli that directly impinge: organisms with central nervous systems have receptors for forms of perception that are both more complex and longer range. Humans have especially sophisticated perceptual capacities, enabling them to respond to a wide range of complex visual, auditory, linguistic, and behavioral/emotional signals in their extended environment. Recent research has begun to show that responses to such fundamentally social signals also extend way down to the level of gene expression.

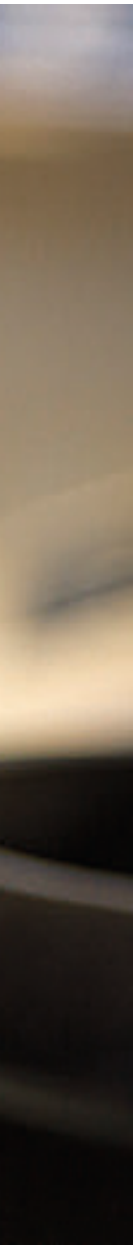
CONCLUSION: WHY IT MATTERS

The gap between a collection of protein-coding sequences and the full complement of genetic material (or DNA) of an organism is as important as it is large. There is of course debate about just how important: in particular, ENCODE's attribution of functionality to virtually *all* transcribed sequences has been hotly disputed. But scarcely anyone today would claim that non-coding DNA is without function. The questions under debate are of how much and

National Cancer Institute / Daniel Sone

In addition to providing information required for building an organism, the genome also provides a vast amount of information enabling it to adapt and respond to the environment in which it finds itself. Above, a researcher works with a DNA sample in the laboratory.

«BUT CURRENT RESEARCH DEMANDS A MORE RADICAL REFORMULATION AND, IN GOOD PART, THAT IT DOES SO BY FOCUSING ATTENTION ON FEATURES THAT HAVE BEEN MISSING FROM OUR CONCEPTUAL FRAMEWORK»



what kinds of function can be tied to ncDNA, and about the implications of such attribution. A relatively conservative response is simply to rename all transcribed sequences of the DNA as genes, attempting thereby to hold onto the view of these entities (and hence of genomes) as effectively autonomous formal agents, containing within themselves the blueprint for an organism's life – i.e., all of the biological information needed to build a living organism.

But current research demands a more radical reformulation and, in good part, that it does so by focusing attention on features that have been missing from our conceptual framework. In addition to providing information required for building an organism, the genome also provides a vast amount of information enabling it to adapt and respond to the environment in which it finds itself. Fortunately so, for without such capacity, how could organisms develop and maintain themselves in the face of environmental vicissitudes?

Rather than a set of genes initiating causal chains leading to the formation of traits, today's genome might better be described as an exquisitely sensitive reactive system – a device for regulating the production of specific proteins in response to the constantly changing signals it receives from its environment. The signals it detects come most immediately from its intra-cellular environment, but these, in turn, reflect input from the external environments of the cell and of the organism. Humans are especially reactive systems, and they are so on every level at which they are capable of interacting: cultural, interpersonal, cellular, and even genetic. We have long understood that organisms interact with their environments; that interactions between genetics and environment, between biology and culture, are crucial to making us what we are. What research in genomics seems to show is that, at every level, biology itself is constituted by those interactions – even at the level of genetics. If much of what the genome «does» is to respond to signals from its environment, then the bifurcation of developmental influences into the categories of genetic and environmental makes no sense.

Such a reformulation does, however, leave us with an obvious question: if the genome is so responsive to its environment, how is it that the developmental process is as reliable as it is? This is a question of major importance in biology, and it is rapidly becoming evident that the answer must be sought not only in the structural (sequence) stability of the genome, but also in both the relative constancy of environmental inputs and the dynamic stability of the system as a whole (Keller, 2000). Genomes are responsive, but far from infinitely so; the range of possible responses is severely constrained, both by the organizational dynamics of the system in which they are embedded and by their own structure.

«IF MUCH OF WHAT THE GENOME "DOES" IS TO RESPOND TO SIGNALS FROM ITS ENVIRONMENT, THEN THE BIFURCATION OF DEVELOPMENTAL INFLUENCES INTO THE CATEGORIES OF GENETIC AND ENVIRONMENTAL MAKES NO SENSE»

REFERENCES

Ball, P. (2013). DNA: Celebrate the unknowns. *Nature, 496*, 419–420. doi: 10.1038/496419a

Davis, B. D. (1990). The human genome and other initiatives. *Science, 249*: 342–343.

Doolittle, W. F., & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature, 284*, 601–603. doi: 10.1038/284601a0

Gibbs, W. W. (2003). The unseen genome: Gems among the junk. *Scientific American, 289*, 46–53. doi: 10.1038/scientificamerican1103-46

Keller, E. F. (2000). *The century of the gene*. Cambridge, MA: Harvard University Press.

Keller, E. F. (2014). From gene action to reactive genomes. *The Journal of Physiology*, *592*(11), 2423-2429. doi: 10.1113/jphysiol.2014.290991

Mattick, J. S. (2004). RNA regulation: A new genetics? *Nature Reviews Genetics, 5*, 316–323. doi: 10.1038/nrg1321

Mattick, J. S. (2010). RNA as the substrate for epigenome-environment interactions: RNA guidance of epigenetic processes and the expansion of RNA editing in animals underpins development, phenotypic plasticity, learning, and cognition. *Bioessay, 32*, 548–552. doi: 10.1002/bies.201000028

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven Symposium in Biology*, *23*, 336–370.

Orgel, L. E., & Crick, F. H. (1980). Selfish DNA: The ultimate parasite. *Nature, 284*(5757), 604–607. doi: 10.1038/284604a0

Oxford English Dictionary. (n.d.). Code. In F. McPherson (Ed.) *Oxford English dictionary* online. Retrieved from http://www.oed.com/view/Entry/35578?rskey=ppAXGm&result=1#eid

Schrödinger, E. (1944). *What is life?* Cambridge: Cambridge University Press.

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. doi: 10.1038/nature05874

The International HapMap Consortium. (2003). The International HapMap Project. *Nature*, *426*, 789-796.

The International HapMap Consortium. (2004). Integrating ethics and science in the International HapMap Project. *Nature Reviews Genetics, 5*, 467–475. doi:10.1038/nrg1351

Weinberg, R. A. (1991). The human genome initiative. There are two large questions. *The FASEB Journal, 5*, 78.

Evelyn Fox Keller. Professor Emerita of History and Philosophy of Science at the Massachusetts Institute of Technology (USA). She is the author of several books, such as *The century of the gene* (Harvard University Press, 2000), *Making sense of life: Explaining biological development with models, metaphors and machines* (Harvard University Press, 2002) and *The mirage of a space between nature and nurture* (Duke University Press, 2010).