Biol Philos
DOI 10.1007/s10539-014-9441-3

Junk or functional DNA? ENCODE and the function controversy

Pierre-Luc Germain · Emanuele Ratti · Federico Boem

Received: 19 July 2013 / Accepted: 11 March 2014

Abstract In its last round of publications in September 2012, the Encyclopedia Of DNA Elements (ENCODE) assigned a biochemical function to most of the human genome, which was taken up by the media as meaning the end of 'Junk DNA'. This provoked a heated reaction from evolutionary biologists, who among other things claimed that ENCODE adopted a wrong and much too inclusive notion of function, making its dismissal of junk DNA merely rhetorical. We argue that this criticism rests on misunderstandings concerning the nature of the ENCODE project, the relevant notion of function and the claim that most of our genome is junk. We argue that evolutionary accounts of function presuppose functions as 'causal roles', and that selection is but a useful proxy for relevant functions, which might well be unsuitable to biomedical research. Taking a closer look at the discovery process in which ENCODE participates, we argue that ENCODE's strategy of biochemical signatures successfully identified activities of DNA elements with an eye towards causal roles of interest to biomedical research. We argue that ENCODE's controversial claim of functionality should be interpreted as saying that 80 % of the genome is engaging in relevant biochemical activities and is very likely to have a causal role in phenomena deemed relevant to biomedical research. Finally, we discuss ambiguities in the meaning of junk DNA and in one of the main arguments

P.-L. Germain (✉) · E. Ratti · F. Boem
Department of Experimental Oncology, European Institute of Oncology (IEO), Milan, Italy
e-mail: pierre.germain@ieo.eu

E. Ratti
e-mail: emanuele.ratti@ieo.eu

F. Boem
e-mail: federico.boem@ieo.eu

P.-L. Germain · E. Ratti · F. Boem
Dipartimento di Scienze Della Salute, Università degli Studi di Milano, Via Adamello 16, 20139 Milan, Italy

Published online: 21 March 2014

raised for its prevalence, and we evaluate the impact of ENCODE's results on the claim that most of our genome is junk.

Keywords Biological function · Causal role · Selected effect · ENCODE · Junk DNA

Introduction

The ENCODE project (the Encyclopedia Of DNA Elements) culminated in September 2012 in the release of 30 publications in major scientific journals. In the introductory paper, the Consortium claimed to have been "able to assign biochemical functions for 80 % of the [human] genome" (ENCODE 2012, p. 57). As natural selection is generally held to be the primary explanation for the presence of functions, this claim seemed at odds with estimates according to which only 3–8 % of the genome has been undergoing purifying selection. In some accidental or intentional deformation, media coverage reported that most of the DNA which allegedly used to be "dismissed as junk" is in fact "active and needed" (Kolata 2012). In the headlines of The Washington Post one could read "Junk DNA' concept debunked by new analysis of human genome." (Brown and Boytchev 2012). The ENCODE Consortium, which made little attempt to clarify their point, was accused of contributing to this misunderstanding:

> In attempting to popularize the result and the project, in press releases and interviews, ENCODE leaders fit the results to a seductive 'all the textbooks are wrong' narrative. (Eddy 2013, p. R259)

ENCODE's claims provoked heated debate among biologists, taking place both in scientific publications and in less formal contexts. According to critics, ENCODE's result have not changed the fact that the vast majority of the genome is junk. The core of the debate hinges on the related meanings of the terms 'function' and 'junk DNA', with virtually all critics agreeing that ENCODE has been using the wrong concept of function.

We do not aim to provide a wholesale vindication of the ENCODE project, but to give it a proper defence against the specific criticism that it used the wrong notion of function. As we see it, critics worry that ENCODE's notion of function will lead to a multiplication of irrelevant functions, whereas a selected-effect account of function, they argue, successfully identifies *relevant* functions. We argue that the selection-criterion is only one way (and an imperfect one) to identify relevant functions, and that ENCODE's notion of function is suited to the aims and scope of the project.

The paper is divided in two parts: the first discusses the ENCODE controversy and the notion of function, while the second addresses the topic of junk DNA. In the first section, we briefly introduce the ENCODE project and discuss some of its major contributions. We then present the main criticism to ENCODE's functionality claim (Doolittle 2013; Graur et al. 2013) and the problem of spurious/irrelevant functions. In order to clarify the debate, we then discuss the most important philosophical accounts of functions and the way they address this problem. We

point out the limits of evolutionary accounts of function when biomedical research is concerned, calling for an alternative strategy to identify relevant functions. We propose that ENCODE is a discovery tool aimed at identifying relevant activities in view of functional analysis. We illustrate this through examples (from basic mechanistic explanations to higher-level system analyses of the human regulatory network) and argue that it addresses the issue of spurious functions in a way that is appropriate given the specific aims and scope of biomedical research. In light of this analysis, we propose that ENCODE's controversial claim of functionality should be interpreted as saying that 80 % of the genome is engaging in relevant biochemical activities that are very likely to have causal roles in phenomena deemed relevant to biomedical research.

In the second part of the paper, we address the question of whether ENCODE's results are compatible with the idea that most of our DNA is junk. We show that the notion of junk DNA is itself ambiguous, and that this ambiguity can lead to a problematic argument for junk DNA. Taking into account the different meanings the term can adopt, we relate it back to notions of function and assess in what sense and to what extent our genome can be said to contain junk. We argue that ENCODE's results are not in disagreement with the conventional view of our DNA as being evolutionary 'junk'.

The ENCODE project and the function controversy

The widely advertised aim of the ENCyclopedia Of DNA Elements (ENCODE) was to complement the sequencing of the human genome with an annotation—to provide a "parts list" of the genome (Pritchard and Gilard 2012, p. 55). Like its predecessor (the Human Genome Project), ENCODE is an instance of 'big science', involving 442 members from 32 institutes and a budget around 288 million USD. In late 2012, this effort led to the simultaneous publication of 30 papers in *Nature*, *Genome Biology* and *Genome Research*. But perhaps more importantly, the ENCODE is also a massive amount of publicly available data, with a total of 1,649 high-density experiments on 147 cell types (at the time of the 2012 round of publications). The project's contributions also include technical standards (both in 'wet' protocols and computational analysis), novel tools or algorithms, and a careful assessment of the strengths and weaknesses of different technologies.

Although such contributions and technological investigations have been part of the project from its very conception (NHGRI 2002; ENCODE 2004), it is the encyclopaedia itself that grabbed the attention: the annotation of transcripts (coding or not), binding sites, enhancers, insulators and other DNA elements throughout the genome. From a practical point of view, one of the most important tools provided by the ENCODE is a segmentation of the human genome (performed in different cell types), with each segment being tagged according to its putative biochemical function. As in previous such endeavours of molecular biology (such as the early chromosome maps), this is often pictured through an explicit metaphor of mapping the genome (see Gaudillière and Rheinberger 2004): the project is said to have been

"designed to populate this terrain", especially "the vast desert regions" not coding for proteins (Maher 2012).

The vast majority of these regions have traditionally been considered junk. It has long been known that protein-coding sequences represent barely more than 1 % of the human genome—preciously little when compared with the proportion covered by the most important classes of apparently function-less elements, such as introns (±30 %, or ±10 % when excluding transposable elements) or transposable elements (at least 44 %). Since the 1970's, scientists have perceived the genome as "loaded with functionless DNA", or "junk DNA" (Ohno 1973). To be fair, the possibility that *some* of junk DNA might, after all, not be junk—that it might for instance play an important role in gene regulation—has always been clear to proponents of junk DNA (see Comings 1972). Even retroviral infections could be co-opted and become selected for in the course of evolution (e.g. Lynch et al. 2011). In the last couple of decades, many such DNA regions were shown to play a role in the regulation of various biological phenomena. Nevertheless, they are widely believed to represent only a small proportion of the genome: the fact that only a very small proportion of the genome—3 to 9 % according to most estimates—appears to have undergone purifying selection[1] has led to the widespread consensus that the vast majority of the human genome is unnecessary, useless junk.

In its main paper, the ENCODE Consortium declared that they were able "to assign biochemical functions for 80 % of the genome" (ENCODE 2012, p. 57). This claim was taken up by the media as meaning that 80 % of the genome is functional, or even "critical" and "needed" (Kolata 2012). While their press releases may have encouraged this interpretation, the introductory ENCODE publication had stated very clearly what it meant by "functional":

> The Encyclopedia of DNA Elements (ENCODE) project aims to delineate all functional elements encoded in the human genome. Operationally, we define a functional element as a discrete genome segment that encodes a defined product (for example, protein or non-coding RNA) or displays a reproducible biochemical signature (for example, protein binding, or a specific chromatin structure). (ENCODE Consortium 2012, p. 57)

The "biochemical signature strategy" is central to the ENCODE project. According to Stamatoyannopoulos (2012), one of the leading ENCODE scientists, such a strategy was motivated by the limitations of a reductionist strategy (looking only at the sequence, independently of the context, to understand how the genome works) and by "the recognition of common biochemical or biophysical events that invariably attended certain types of noncoding functional elements" (Stamatoyannopoulos 2012, p. 1602). For example, enhancers are typically characterized by a combination of histone modifications—modifications not pertaining to the DNA sequence itself, but to the proteins around which the DNA is wrapped. Likewise, transcription factors typically bind to promoters or other regulatory regions, and

[1] Because natural selection tends to remove deleterious mutations from the pool, we can expect to observe less mutations in DNA sequences important for the survival and reproduction of the organism. As there are a number of technical hurdles in the detection of such selection, estimates vary (some going up to 15 %). For a discussion, see Ponting and Hardison (2011).

therefore regions bound by transcription factors are expected to be of regulatory relevance. The strategy, then, was to look for these signatures over the entire genome, in order to identify the related functional elements. Annotating the genome on this basis means to study it not as isolated DNA molecules, but in cellular contexts.

A few months after the ENCODE publications, different criticisms of its claims about function were published in scientific journals (Eddy 2012, 2013; Doolittle 2013; Graur et al. 2013; Niu and Jiang 2013). In addition, the debate was the topic of several blog entries.[2] There are two main reasons why the debate got so heated. The first is that the ENCODE Consortium, especially in its public relations, was perceived as deliberately encouraging misinterpretations to artificially boost its success (and hence future funding).[3] The second reason is that in reaction to this alleged dishonesty, one of the critiques (Graur et al. 2013) was unusually polemic and aggressive ("vitriolic" according to Eddy 2013), escalating the debate and its visibility (somewhat ironically, given how Graur et al. complain about ENCODE's marketing strategies).

Part of the objections raised regard technical issues, most of which are justified criticisms, and few others debatable. While these are scientifically important, they will not be discussed here. Taking them in consideration would reduce the figure of 80 % (perhaps to something around 50–60 %[4]), but would not make a difference to the deeper disagreement. Instead, the core of the debate is that, from the point of view of the critics, ENCODE dismissed junk DNA merely rhetorically, by using an inappropriate and much too inclusive notion of function.

Both Doolittle (2013) and Graur et al. (2013) rightly point out that the term function can be given different meanings that should not be conflated.[5] In a first sense, a 'biological function' is a 'selected effect' (SE), referring to the evolutionary dimension (Wright 1973; Millikan 1989; Neander 1991). More specifically, this means that a trait T has a function F if and only if T's performing F is the reason why T has been selected and maintained along evolution. In a second sense, which the authors attribute to Cummins (1975), the term 'function' can refer to the 'causal role' (CR) that a trait has in performing a certain activity. Oversimplifying

[2] See for example http://genomeinformatician.blogspot.ch/2012/09/encode-my-own-thoughts.html and http://www.homolog.us/blogs/blog/2013/04/09/homolog-us-blog-calls-for-sean-eddy-be-fired-for-the-sake-of-good-science/.

[3] In fact, it must be noted that most criticisms are not directly aimed at what is written in ENCODE's scientific publications, which are careful in their formulations, but instead at its interpretation. This is not limited to the mass media, but also to the coverage the results were given in prestigious scientific journals such as Science (Pennisi 2012).

[4] On the day the embargo was lifted on the last round of ENCODE's publications (and therefore long before the publication of ENCODE's criticisms), Ewan Birney, ENCODE's lead analysis coordinator, published a post on his personal blog providing his personal perspective of the project (Birney 2012a, b). In this post, Birney acknowledges that the proportion of the genome that is "functional" depends on how stringent one is, and preempts some of the most important technical criticisms addressed at the project.

[5] Doolittle for instance writes: "Those of us who speak of excess DNA as informationally junk mean that its presence is not to be explained by past and/or current selection at the level of organisms—that it has no informational function construable historically as an SE [selected effect]. Those who say that almost the whole of the human genome is functional informationally do so on the basis of an operational diagnosis embracing a non-historical CR [causal role] definition of function." (Doolittle 2013, p. 5299).

Cummins' account, they write that in this sense, "for a trait, *Q*, to have a 'causal role' function, *G*, it is necessary and sufficient that *Q* performs *G*" (Graur et al. 2013, p. 579).

As we understand it, the critics' main worry with the latter account is that it will lead to a multiplication of irrelevant functions:

> The causal role concept of function can lead to bizarre outcomes in the biological sciences. For example, while the selected effect function of the heart can be stated unambiguously to be the pumping of blood, the heart may be assigned many additional causal role functions, such as adding 300 g to body weight, producing sounds, and preventing the pericardium from deflating onto itself. (Graur et al. 2013, p. 579)

For this reason, and following Dobzhansky's famous statement that "nothing in biology makes sense except in the light of evolution", these authors claim that the relevant use of the term function in biology is the selected effect. Doolittle urges a call to arms, writing that "we need as biologists to defend traditional understandings of function" (Doolittle 2013, p. 5294).

In what follows, we will argue that the selected-effect account of function is only one means—and an imperfect one—of avoiding spurious functions, and that ENCODE's approach is successful in identifying functions that are relevant to biology and medicine. To do so, we begin with a brief discussion of the main philosophical accounts of functions.

Biological functions

Few other notions are so widespread in biology as the notion of *function*. Despite its pervasiveness, the notion is deemed problematic because of its teleological connotation, and its naturalization has therefore triggered a deep theoretical investigation. As ENCODE's critics have pointed out, such attempts are generally grouped into two theoretical approaches: causal-role approaches (Cummins 1975), and aetiological theories (Wright 1973).

Cummins' (1975) proposed his *causal role* -account (CR) of functions against what he claimed to be a problematic and unjustified assumption of aetiological theories: namely that the point of functional ascription is to explain the presence of the function-bearer. Indeed, in the context of biology such an explanation would amount to an evolutionary explanation, making the concept of function redundant. Instead, Cummins argued that the point of functional ascription is to explain a system's behaviour by appealing to the contribution of the function-bearer.[6] Complex phenomena are explained by decomposing them into component activities. In this context, "[t]o ascribe a function to something is to ascribe a capacity to it which is singled out by its role in an analysis of some capacity of a containing system." (Cummins 1975, p. 765). Nothing has a function by itself: a

[6] In a similar way, Weber (2005) has argued that "elucidating the evolutionary history of some system or subsystem is supplementary to analyzing its function; it is not part of it." (Weber 2005, p. 40) This is easily shown by the fact that expatiations (features which start being used in a way for which they have not been selected) are generally considered functional.

function is always a role *in* something, a contribution *to* something. Causal roles therefore situate an item/activity with respect to the organization of a system, for instance a cell or an organism. For example, for a DNA sequence X binding a transcription factor to have a causal role, the binding must contribute to some capacity or behaviour of the containing cell or organism: for instance, it might promote the transcription of a nearby gene Y. In this context, we would say that X's binding the transcription factor contributes to the regulation of Y's expression, or in other words, that the function of X in the regulation of Y's expression is to bind the transcription factor.

The problem of spurious or irrelevant functions stems from this reference to a capacity or behaviour of the containing system, for any effect can be said to contribute to something: as Cummins noted, the heart's beating contributes to our bodily noises. Several authors therefore argued that functional analysis in biology has to be framed, and an evolutionary perspective was proposed as an objective way to do so (Millikan 1989). Such approaches can be divided into two families. The first, which is often referred to as the *selected-effect* (SE) account of function, is backward-looking (T has function F iff T *was* selected because it did F in the organism's ancestors), while the second is forward-looking: T has function F iff T is currently/will be selected because it does F (or to put it differently, T has function F iff T's doing F increases the fitness of its bearer). This distinction has been emphasized several times in the literature (Tinbergen 1963; Bigelow and Pargetter 1987; Griffiths 2009); here, we will follow Wouters (2003) and others in referring to the first account as 'selected effect', and to the second as 'biological advantage'. As Griffiths (2009) has shown, the notion of biological advantage is logically prior to that of selected effect, because in order to know what a trait was selected for, one has to know how the trait benefited its bearers; i.e. the trait's causal role in its bearers' fitness. If this is possible for ancestors, then in principle it must also be possible for current organisms, hence functional analysis of current biological systems does not require a backward-looking concept of function, although Griffiths argues that it requires a notion of biological advantage (forward-looking concept).

For purposes (such as medicine) that are not strictly related to evolutionary biology, functions as selected-effect are relevant only insofar as they represent a proxy to biological advantages. Medicine does not generally care about what was advantageous to our ancestors, but rather about what contributes to the wellbeing of the patients it is presently concerned with. However, selected effects are most often also currently advantageous, and are therefore in many cases a useful proxy to relevant functions.

The notion of biological advantage, in turn, implies Cummins' causal roles because it ascribes to an item/activity a contribution in the organism's capacity to survive and/or reproduce (Griffiths 1993). The evolutionary perspective is important in that it establishes the topmost capacities of the system, thereby defining the *explanandum* (the phenomenon to be explained) of functional decomposition *à la* Cummins (see for instance Weber 2005; Griffiths 2009). Systems whose functions contribute to the topmost capacities of reproduction and survival are in turn decomposed into the functions of nested subsystems. This view therefore acknowledges that Cummins's account is the 'right' conceptual analysis of what

functions are, but complements it with the claim that in biology, there are 'right' *explananda* that should be the target of functional analysis.

The idea of 'right' *explananda* defined in an evolutionary perspective can be understood in two different ways: either natural systems have natural goals, by virtue of which they are to be functionally decomposed, or it is a fact about biology as a science that survival and reproduction represent the fundamental *explananda*.

Postulating natural goals is not without difficulties. For instance, Bunzl pointed out that it is difficult to provide reasons for singling out reproduction and survival as the relevant capacities of the system without getting into an upward regress:

> if individual survival and reproduction are claimed to be constitutive of what is proper working order for individual organisms, this needs to be justified by reference to the role that such organisms play in maintaining some larger system in proper working order. (Bunzl 1980, p. 118)

This larger system is obviously populations, or species, but the same problem arises at these levels. As evolution itself does not have a goal, there seems to be no ultimate way out of this logical impasse. Another aspect of the same problem is that DNA regions contribute to the survival and reproduction of many things, including, for instance, genes. Evolution does not, by itself, tell us which entity should determine functional ascription.[7]

The second version of the claim is far less problematic. Weber, for instance, argued that "[t]he capacity for self-reproduction is the most salient capacity that we want to understand in biological organisms" (Weber 2005, p. 41). Indeed, faced with the diversity and complexity of life, one of the most fundamental questions the biologist asks is how an organism can persist, and how an organism came to be the way it is. But are these the only meaningful biological *explananda*? Many philosophers have acknowledged that there might be contexts where one is interested in understanding other phenomena: that there are phenotypes and behaviours of biological systems which, although they are not biological advantages, are somehow of interest to us. The relevant question is when this is the case.

To sum up, we have argued in this section that the CR account is the 'right' analysis of the notion of function, and that it is presupposed by evolutionary accounts. The latter, we argued, complements it with the claim that in biology, fitness is the only meaningful target of functional explanation. However, we proposed that there might legitimate *explananda* in the life sciences that are not biological advantages, but can nonetheless be targets of functional analysis *à la* Cummins. In the next section, we will argue that this might especially be the case in biomedical research.

[7] Furthermore, the reader should be aware that different lines of research suggest that the modern synthesis is insufficient to understand evolution. See for instance the work edited by Pigliucci and Müller (2010) on the need of extending the standard evolutionary paradigm. Other directions/suggestions have been explored by Shapiro (2011), Gissis and Jablonka (2011 edited by), and Kauffman (1993, 1996).

ENCODE's aims and the limits of evolutionary accounts

The "right" notion of function is one that is able to identify relevant functions for given research purposes. In order to understand whether ENCODE's notion of function is adequate, one must therefore pay attention to its broader aims and to the context in which it was conceived. After the immense investment of the Human Genome Project and the public realization that sequencing would not by itself revolutionize medicine, the NHGRI prompted a project that would enable researchers to harness the medical potential of the genome. This aim was still very explicit in 2011 when the Consortium published a first glimpse of their results:

> The mission of the Encyclopedia of DNA Elements (ENCODE) Project is to enable the scientific and medical communities to interpret the human genome sequence and *apply it to understand human biology and improve health*." (ENCODE Consortium 2011, p.1, emphasis added)

Likewise, in the introduction of the 2012 round of papers the authors explicitly frame their enterprise as providing "an important resource for the study of human biology and disease" (ENCODE Consortium 2012, p. 57). This is particularly important because *what is relevant to contemporary medicine is not necessarily evolutionarily relevant.*[8] This becomes obvious when one pays attention to the major concerns of contemporary medicine. For example, 10 years ago a chief scientist at GlaxoSmithKline (Connor 2003) made much noise in the media by saying what was already obvious to most practitioners: "Our drugs do not work on most patients". The question is not only one of inefficiency: in the United States, adverse drug effects are estimated to seriously affect 2 million persons and are incurring billions of dollars of costs annually (Agency for Healthcare Research and Quality 2001). As the vast majority of drugs work for a minority of people, one of the cornerstones of contemporary biomedical research has been to investigate the differences in patients' reaction to disease and treatment. A growing body of evidence suggests that most of these differences are due to genetic variations within the population. Likewise, most diseases show considerable heritability attributable to a high number of genetic variations, each of which has a small impact on the risk

[8] Ernst Mayr (1961) proposed a distinction between two research projects within biology, which he labeled f*unctional biology* and *evolutionary biology*. According to Mayr, while functional biology seeks proximate causes and therefore investigates how certain phenomena occur, evolutionary biology is devoted to understand why or the evolutionary reason for the presence of the very same phenomena. Our point is that given traits can be involved in the explanations of functional biology independently of whether they have been selected for—as Cummins puts it: "Flight is a capacity that cries out for explanation in terms of anatomical functions regardless of its contribution to the capacity to maintain the species." (Cummins 1975, p. 756). We obviously do not claim that Mayr's two domains of biology are insulated from each other, but rather that they pursue two legitimate aims which share many, though not all of their means (Laland et al. 2011). Each domain is important in the investigation of the other. However, a reduction of all the functional relevance of the genome to its evolutionary dimension (i.e. function as selected effect or biological advantage) fails to give enough attention to the different research projects of the life sciences.

of developing the disease. There are a number of reasons why such variations might escape natural selection: perhaps their effect is too small or visible only in "unnatural" conditions (such as treatment or old age), or because of antagonistic pleiotropy. In any case, these genetic variations were not under strong purifying selection; otherwise they would not be widespread in human populations. If the aim of a research program is to understand, for instance, how genetic variations broadly present in human populations affect susceptibility to age-related diseases, success of treatments or adverse effects, then there is a good chance that many DNA regions would be missed by conservation studies. For whatever is required to build a breathing human will not be variable across human populations, and hence will not, for instance, be of direct interest for pharmacogenomics. In our society, making people healthier most often has very little to do with granting them more offspring.

ENCODE's critics have dismissed the causal role account of function on the ground that it allows a multiplicity of irrelevant or spurious functions, and proposed instead that selected-effects identify relevant functions. In the previous section, we argued that selected-effects are only proxies to relevant functions as causal roles, and in this section, we argued that they might be inadequate for the identification of those causal roles important to biomedical research. The ENCODE has bet on a different strategy to identify functions relevant to this research context. In order to understand this strategy, the next section looks at how ENCODE's annotation of the genome is actually employed.

ENCODE identifies activities with an eye towards causal roles

Biological understanding often comes from functional decomposition of complex systems/behaviours into contributions of simpler components—in their explanations via causal roles (Bechtel and Richardson 2010; Craver 2007). In many circumstances, although not all, the *explanandum* is related to an organism's fitness. When this is the case, evidence of past evolutionary conservation indicates a probable (although not certain) causal role relevant to the organism's fitness. But in all cases, any activities indicate a potential (although not certain) causal role into phenomena of interest. In this picture, the ENCODE project identified a specific subset of biochemical activities (transcription, transcription factor binding, and specific combinations of histone modifications, etc.) which very often contribute and make a difference to the phenomena scientists are interested in. We suggest that a research strategy particularly important in contemporary sequencing-based biology consists in first finding evidence of the biochemical activities in which given components are involved (e.g., a DNA region binds a transcription factor, or a protein has an enzymatic activity), before relying on this annotated "store" to explain different phenomena. It is in this latter step that biologists ascribe causal roles to components (e.g., the DNA region thereby enhances expression of gene x, promoting process y). There are therefore two distinct discovery steps:

1. one produces evidence that a particular item does something likely relevant (an activity of the item);

2. one looks for the contribution this activity may bring to phenomena of interest (a causal role of the item).[9]

The strength of this two-steps strategy lies in the fact that evidence of (1) gives clues to discover (2). It therefore depends on a careful choice of a relevant subset of activities, which constitutes the core of ENCODE's biochemical signatures.

ENCODE's critics have also criticized this choice

> Why make a big fuss about 74.7 % of the genome that is transcribed, and yet ignore the fact that 100 % of the genome takes part in a strikingly "reproducible bio-chemical signature"—it replicates! (Graur et al. 2013, p. 580)

However, ENCODE's criteria are far from arbitrary. We would argue that ENCODE identifies *relevant activities*—that is, activities likely to make a relevant difference[10] to some phenomena scientists are likely to care about. Indeed, while functional ascription is, in the CR account, relative to the *explanandum*, biologists are typically interested in a subset of all *explananda* (not all of which are necessarily evolutionarily relevant). ENCODE's contention is that their biochemical activities are the most likely to play a causal role in *explananda* of this kind. This is by no means arbitrary, but empirically motivated, for cases of DNA regions making a contribution to relevant *explananda* by the sheer fact that they are replicated during cell division are extremely rare, while cases of DNA regions making a difference to *explananda* by being transcribed, or bound to, are beyond count. Just like evolutionary conservation, ENCODE's biochemical activities are an imperfect but useful proxy to relevant functions.

Let us illustrate the two-steps strategy first by means of a simple example. The gene SF3B1 codes for the splicing factor 3b subunit 1, which has a number of activities (step 1): it interacts with a number of other proteins (splicing factors, RNA helicases, etc.), forms a spliceosome complex (U2 snRNP) with other splicing factors, and binds to RNA molecules (and, most likely, does many other things). It is also involved in mRNA processing (step 2): evidence suggests that it contributes to

[9] Note that the two-step strategy we propose is not in conflict with Bechtel and Richardson's (2010) strategy of decomposition and localization. According to the latter, scientists decompose a complex phenomenon into less complex subsystems or contributions, and attempt to localize these functions to physical components of the system (e.g. organelles). Our point is that in the step of localization, the physical components are not entirely uncharacterized, and the earlier characterization of their activities provides important hints as to which function localizes where. Our two-step strategy is however to be distinguished from another very common strategy in biology. Mutants identified through reverse genetics, for instance, can establish the relevance of a part in a given phenomenon before identifying any of its activities. Obviously, the strategy we describe is but one of the many general strategies available to biologists.

[10] This has to be understood as making a *relevant* difference, for any change to DNA makes a phenotypic difference at least insofar as the genome is also part of the structure of the organism. Even a transcription factor binding site in the middle of nowhere, not leading to any transcription, is having an effect on relevant gene functions, at least insofar as it sequesters the transcription factor and hence reduces the amount of the protein available for important binding sites. In the same way, non-coding RNA can have an influence on the expression of coding genes because they are bound by miRNAs which would normally regulate the coding genes (Salmena et al. 2011). However, such impacts may be so small as to be imperceptible.

intron removal by anchoring the spliceosome to pre-mRNAs. Steps 1 and 2 may be considered as two separate epistemological moments. We may say that the first step identifies biochemical activities of the splicing factor 3b, while the second step situates *some* of those activities with respect to a broader biological phenomenon we are trying to understand. While the activities are the same in both steps, step 2 ascribes a causal role to some of these activities. Importantly, knowledge of some activities of the gene product (interacting with splicing factors, binding mRNA) are highly suggestive of its potential causal roles. Note, however, that it may acquire new causal roles according to the phenomena of interest. For instance, evidence suggests that particular mutations in SF3B1 enhance intron retention within specific transcripts (Wang et al. 2011), leading to differential splicing of certain genes involved in cancer development.

This basic strategy can be applied to more complex *explananda*, as illustrated by another example. For instance, the ENCODE can be seen as an effective way to characterize higher-level activities, which will only later be used for the functional decomposition of biological phenomena. For instance, Gerstein et al. (2012) tried to 'unveil' the principles of the human transcriptional regulatory network by studying the combinatorial binding of many transcription factors (TFs), i.e. how different combinations of the binding of different TFs lead to different biological meanings. In order to accomplish this task, a map of all TFs was needed, or (more realistically) a map of a reasonable amount of them. Gerstein and colleagues analysed the genome-wide binding profiles of 119 TFs, in order to understand the organization and dynamics of the human regulatory network (their *explanandum*). This strategy is motivated by the fact that the human genome regulation is composed by an enormous number of components and, even if they display a clear activity, their contribution to the regulation of the human genome depends on the interactions they have with the other components. First, Gerstein and colleagues were able to organize the binding patterns in a hierarchy representing the system-level regulatory wiring of the network of associations among TFs. Then, they were able to highlight the combinatorial and context-specific fashion of the co-associations between TFs. Moreover, they showed that different parts of the hierarchical transcription factor network exhibit distinct properties. These findings provide insights into the complex human regulatory network (the *explanandum*) by decomposing it in the contributions of different components (the activities to be situated in the more complex *explanandum*).

Another major usage of the ENCODE as a discovery tool is in the analysis of genome-wide association studies (GWAS). GWAS are based on the comparison of hundreds of thousands single-nucleotide polymorphisms (SNPs) between different subpopulations (e.g. healthy versus diseased). Because they are simultaneously testing very large amounts of potential associations, such studies pose major statistical challenges, leading to many false positives. Moreover, such methods cannot distinguish mere statistical associations from causally-relevant genetic variations. In the last years, GWAS have successfully identified many variants *associated* to particular phenotypes, without however proving their causal relevance to these phenotypes. Given the number of such associations requiring validation, it is usually important to prioritize some of them for further study. To this end,

different groups (e.g. Chanock 2012; Schaub et al. 2012) have relied on ENCODE's annotation for the analysis of GWAS results. Here, ENCODE's sense of function is extremely relevant to filter out from this mass of candidates a smaller fraction that are within regions of likely functional relevance. Intersecting the polymorphisms with ENCODE data can be of some help in identifying a subset of SNPs that have the highest chance to be disease-relevant. With this first characterization of SNPs in mind, biologists can start to investigate whether the SNPs selected (the ones that are functional according to ENCODE) have a causal role in the phenotype of interest. In other words, ENCODE is a useful tool to discard some hypotheses, or to 'prune the hypotheses tree'.

ENCODE's critics show scepticism regarding this strategy. For instance, Graur et al. criticize ENCODE's ability to restrict the space of possible mechanisms. Since according to ENCODE 80 % of the genome engages in relevant activities, critics say that

> without ENCODE, researchers would have had to examine 3.5 billion nucleotides in search of function, with ENCODE, they would have to sift through 2.7 billion nucleotides. (Graur et al. 2013, p. 34)

This is, however, an oversimplification. The ENCODE does not only say whether given regions are active or not, but gives rich information about the activities, and therefore potential causal roles they might play. As mentioned earlier, the rationale of the two-steps strategy is that the identified activities hint at specific causal roles. Transcribed non-coding regions might have hybridization-mediated effects on gene expression, enhancers are likely to regulate expression of genes in their domain, chromatin marks at the promoter/TSS region might regulate transcription, while marks in the gene body might slow or stall polymerase, and so on and so forth. The ENCODE, therefore, does not only say "these are the parts to be considered", but proposes, for each, very specific hypotheses to be investigated. While it may not eliminate many DNA regions as non-functional, it does eliminate many possible ways in which these regions might be functional.

In all cases discussed, the basic strategy is to identify activities, which can then serve as building blocks for mechanistic explanations of phenomena of interest. To use the words of Birney, "you've got to put all the parts down on the table before putting it together" (quoted in Maher 2012). The identification of a store of possible entities and activities puts significant constraints on the space of possible mechanisms for a phenomenon (Darden 2006; Craver 2007).

It may be useful to summarize what is, on our account, the epistemological strategy in which the ENCODE is embedded with the help of a diagram (Fig. 1). At the top there is a repository of all DNA elements for which a relevant biochemical activity has been found. Most (if not all) of these elements play a causal role in one phenomenon or another, although only a subset of such phenomena is of interest to scientists. For any given *explanandum*, *some* activities of *some* elements are selected. Elements can be hypothesized to be relevant for the *explanandum* because they display activities of a certain kind. Their potential causal role in the phenomenon is then investigated. In the example on the left, the *explanandum* is the human regulatory network, for which some activities of Element1 and Element2 are

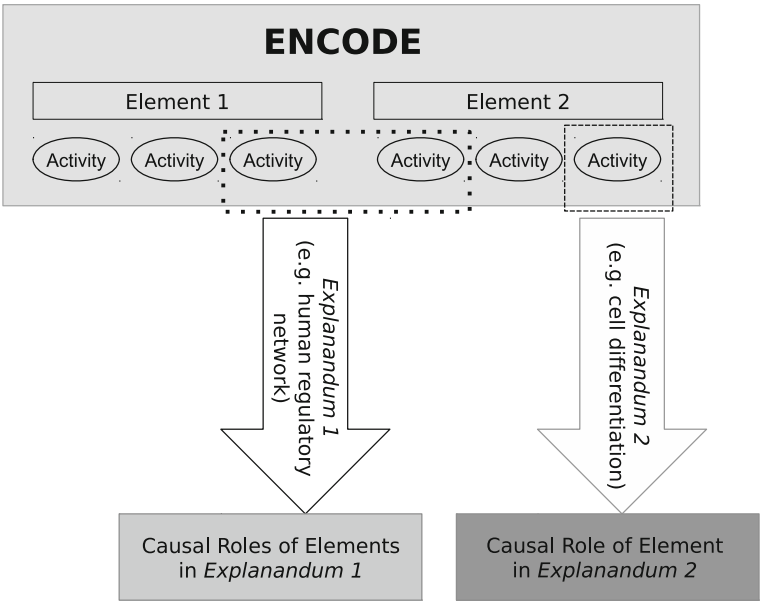

Fig. 1 Epistemological strategy in which the ENCODE is embedded

selected. However, Element2 may also be investigated according to another *explanandum*.

[In light of this epistemological strategy, we believe that the best interpretation of ENCODE's claim of function is methodological: it is the claim that *80 % of the genome is engaging in relevant biochemical activities and is very likely to have a causal role in phenomena deemed relevant to biomedical research*. In other words, while this 80 % cannot strictly speaking be called functional (according to the CR account) as ENCODE claimed, it is very likely to be. This means that, according to ENCODE, most of what has been called junk DNA cannot be ignored in biomedical research, and that 80 % of the genome is potentially relevant.

The emphasis on biochemical activities is, in our opinion, justified. As we have argued, identifying biochemical activities is the most appropriate way to get at functions that are relevant to the research context to which ENCODE aims to contribute.

Unravelling the onion: Junk DNA, essentiality and function

While some of ENCODE's critics might agree with what has been said in the previous sections, the deeper ground to reject ENCODE's claim has been its pretension to the death of junk DNA. The claim that most of our DNA is junk is relevant in two contexts. One is our broader understanding of genome structure and evolution. The other is the attack on the teaching of evolution by the religiously motivated 'Intelligent design' movement. The existence of junk DNA seems at odds

with the view that the genome is as efficient as possible, and, hence, that it is the work of an intelligent force or designer. In this section we discuss the meaning of junk DNA and whether and in what sense ENCODE's claim that most of our genome has a function conflicts with the claim that most of our genome is junk. We argue that ENCODE's results do not contradict the claim that most of our genome is evolutionary junk. Building on the previous sections, we argue that the only version of the claim to junk DNA with which ENCODE's results conflict is that most of our genome does not make a difference to biological phenomena of interest to biomedical research. Such an interpretation of 'junk DNA', however, is neither widespread among scientists nor particularly useful conceptually. Under more conventional accounts of junk DNA, most of our genome can still be considered as junk, although this does not imply that it can be dismissed for the purposes of biomedical research.

What is 'junk DNA'?

The expression "junk DNA" is generally attributed to a presentation given at the Brookhaven Symposium by Susumu Ohno in 1970, but appears to have been relatively common already in the 1960's (see Graur 2013 for a discussion). In any case, the origin of the concept of junk DNA is most likely related to the realization that genome size (the C-value) does not correlate at all with any measure or informal understanding of organism complexity. This led to the conclusion that many organisms have "an unreasonable excess of DNA" (Comings 1972, p. 313), which Ohno described as "garbage DNA" (Ohno 1970, p. 62), "junk DNA" (Ohno 1972, 1973) or "functionless DNA" (Ohno 1973). In the course of the following decades, advances in biology have provided different explanations for the existence of this excess DNA: pseudo-genes (broken down duplicates), retroviruses, transposable elements, etc.

It is important to note that biologists never held junk DNA to be *entirely* useless (Comings 1972, pp. 316–317), and there is a minimal sense of "not entirely useless" to which most biologists would probably agree: junk DNA might for instance have been useful in the past, or it might gives evolution something to tinker with.[11] Both scientific reviews (e.g. Makalowski 2003) and popular science articles[12] suggest that junk DNA was widely understood as *currently* useless DNA which however might turn out handy in the future (might be mutated into something functional), or might have been useful in the past and not yet cleared away.

[11] This is well illustrated by Brenner's (1998) distinction between "junk" and "garbage": "Some years ago I noticed that there are two kinds of rubbish in the world and that most languages have different words to distinguish them. There is the rubbish we keep, which is junk, and the rubbish we throw away, which is garbage. The excess DNA in our genomes is junk, and it is there because it is harmless, as well as being useless, and because the molecular processes generating extra DNA outpace those getting rid of it. Were the extra DNA to become disadvantageous, it would become subject to selection, just as junk that takes up too much space, or is beginning to smell, is instantly converted to garbage…" (Brenner S, *Refuge of spandrels*. Curr. Biol. 8: R669, 1998, quoted in Graur et al. 2013, p. 586).

[12] E.g. "Lean Gene Machine", *Scientific American*, accessed at http://www.scientificamerican.com/article.cfm?id=lean-gene-machine.

There is an additional sense in which it is sometimes said that junk DNA can be useful. As different authors have noted (e.g. Comings 1972; Doolittle 2013), the bulk of DNA gives shape to chromosomes, and thereby determines important biological parameters. We can call this the structural role of a DNA stretch: it is not strictly dependent on the sequence of the DNA stretch, but on its shape, size, and position in the genome. For instance, having some length of DNA between an enhancer and its target gene means that depending on the shape of chromatin (which can bring the enhancer close to the gene's TSS), the enhancer can affect the gene or not, thereby adding possibilities for additional state-specific regulation. Although they are often referred to as junk, such stretches of DNA are obviously useful to the organism (perhaps even vital), but their sequence will most likely not show the traditional hallmarks of evolutionary conservation.

As Doolittle writes, "junk advocates have to date generally considered that even DNA fulfilling bulk structural roles remains, in terms of encoded information, just junk." (Doolittle 2013, p. 5295) It is not entirely clear how to render precisely the meaning of 'encoding' in ways that would allow one to make the distinction: as Griffiths (2001) noted, all available naturalized concepts of information would include developmental factors in what encodes information, and by extension would include DNA with a structural role. Nevertheless, there is a shared and useful understanding among biologists of what it means for DNA to encode for something, distinguishing functions by virtue of sequence from functions that are largely independent from the exact sequence. On this view, if the function of a stretch of DNA is unaffected by mutating its sequence, then this function is not by virtue of information encoded in the sequence. Although not all biologists would classify such DNA stretches as junk, Doolittle's view is perhaps the most widespread.

Junk should not be confused with non-essential

Unfortunately, the notion of 'junk DNA' is easily confused with another notion, namely the notion of DNA that would not be required by a skilful engineer for building a similar organism. An ambiguity between these two notions has accompanied the expression of junk DNA since its earliest usages. Comings, for instance, uses "junk DNA" interchangeably with "nonessential DNA" (see for instance Comings 1972, p. 319), and the same ambiguity is present in the key argumentative scheme that has repeatedly been raised in support for the existence of junk DNA: the C-value paradox, or the lack of correlation between DNA content and organism complexity.

As an example of this argumentative scheme, consider the "onion test" famously coined by Gregory (2007), and quoted by ENCODE's critics:

> Whatever your proposed functions are, ask yourself this question: Why does an onion need a genome that is about five times larger than ours? (T. Ryan Gregory, quoted in Graur et al. 2013, p. 578)

Doolittle (2013) proposed a very similar, although more explicit thought experiment. Suppose that we were to use ENCODE's strategy to identify functional

elements in other organisms which greatly vary in genome size. The two possible outcomes of such a comparison would be that either

1. the number of so-called functional elements is constant (or at least correlates with complexity rather than genome size), in which case either humans have junk, or by some suspicious coincidence we happen to be pretty much the only species at some sort of optimal genome size; or
2. the number of so-called functional elements increases with genome size; since "it would be hard to convince ourselves that lungfish are 300 times more complex than *Takifugu* [the putter fish] or 40 times more complex than us" (Doolittle 2013, p. 5296), we are forced to conclude that most of the so-called functional elements are in fact junk.

> If we do not think of this additional or "excess" DNA, so manifest through comparisons between and within biological groups, as junk (irrelevant if not frankly detrimental to the survival and re-production of the organism bearing it), how then are we to think of it?" (Doolittle 2013, p. 5295)

Like Doolittle, we suspect that (2) would be the most likely outcome of such an investigation. Although complexity is notoriously hard to assess, even in principle, we grant for the sake of the argument that it is very unlikely that the lungfish is 300 times more complex than *Takifugu*. The problem is that this result does not, *per se*, lead to any conclusion regarding junk DNA. Likewise, the fallacy in Gregory's onion test is to conclude, from the claim that one could make an onion (or, more precisely, an onion-like morphology and behaviour) with 5 times less DNA, to the claim that 4/5 of the onion's genome is non-functional.

To see better why this argument is flawed, consider the well-known stereotype (whose truth or falsity is irrelevant here) according to which public companies are less efficient, and require much more administration, than private companies. Suppose this were true, and that the inefficiency were real (rather than merely a trade-off with, say, more laudable goals). The claim that a private company would do the same thing as a public company with half of the workforce, and even the claim that specific employees are unnecessary, do not imply that these public employees are not doing anything or do not have a function. Rather, they do have a function, but if the system was differently organized this function might not be needed anymore. The proof is that asking the public company to get rid of half of its workforce would surely cripple it completely[13]—it would require a major reorganization to survive such a cut.

The same distinction applies to onions: they probably have many more functional elements (as defined by ENCODE) than humans, but that is not to say that these elements would all be necessary if a particularly skilled engineer was to design an onion-like morphology and behaviour. The point is that whether the engineer would need them is irrelevant to whether they are functional in the tinkered system that the onion actually is. And with the exception of some very specific research programs

[13] If evidence is required to support this claim, the reader may consider as an example the effect of budget and workforce cuts on the Greek National Health System (Kentikelenis et al 2011).

(e.g. synthetic biology), biomedical research is about understanding the actual, tinkered systems.

Another example of this mistake can be found in Doolittle (2013):

> My computer might be 5 ft from the wall socket, but if I have only a 10-ft electrical cord all 10 ft will seem functional, because cutting the cord anywhere will turn off my machine. (Doolittle 2013, p. 5297)

In this analogy, the entire 10-ft of cable is under selective pressure, and yet Doolittle refuses the claim that it is functional in its entirety. Again, this confuses the junk/functional distinction with the even more ambiguous distinction between essential and non-essential.[14]

Note that 'essential' can be understood in different ways. Doolittle's analogy suggests the most radical understanding: essential DNA is DNA without which an ideal engineer could not design a creature resembling humans in all relevant respects. But it could also be taken to mean *required* by this specific organism (required for what?). In fact, the core of the scientific disagreement seems to be precisely about this: some biologists contend that we could remove (or mutate at will) most of the DNA in our genome—up to 90 % according to some[15]—without noticing any relevant phenotypic difference. Others, such as many members of the ENCODE Consortium, would surely disagree.

Part of the disagreement is due to a different understanding of what a "relevant phenotypic difference" is (see footnote 10). Any change to DNA makes a phenotypic difference at least insofar as the genome is also part of the structure of the organism. However, not all such differences are relevant, and whether we construe relevant differences as differences in fitness or as differences in the phenomena researchers are trying to explain, we will get different conceptions of junk.

Whether or not most of our genome is irrelevant to the phenomena of interest to biomedicine remains an open empirical question, about which biologists can legitimately disagree.[16] The C-value paradox does not, as it stands, constitute a conclusive argument regarding this question. It would require the additional premise that evolution makes all organisms in an equally economical way, which we think would be to take nature for an engineer rather than a tinkerer. However, the C-value paradox does not imply the reverse either: it may well be that most of our genome is irrelevant to the development of actual humans, and we have only argued that this is not necessarily the case.

[14] This problematic move is also present in another critique of ENCODE's claims, which contrasts the question of "How much DNA does it take to design a human?" with that of "How much DNA does it take to evolve a human?" (Eddy 2013, p. R260), relating the former to function and the latter to junk (see also the interview with Eddy in Diep 2013). Function, however, does not mean ideal design.

[15] Chris Ponting (personal communication) for instance made this claim, but also emphasized the immense difficulty of identifying the remaining 10 % scattered across the genome.

[16] Perhaps the most interesting study regarding this question is that of Nobrega et al. (2004), who deleted two megabase-long non-coding regions of the mouse genome and failed to detect any relevant phenotypic difference.

ENCODE's implications for junk DNA

Junk DNA is often taken to mean many different things. Sometimes, it is used to denote DNA that was not selected for, or was not selected for its contribution to the fitness of the organism of interest. Because ENCODE remained methodologically agnostic with respect to evolution, its results neither support nor contradict the view that most of our genome is junk in this evolutionary sense. In fact, ENCODE's scientists never explicitly claimed that their results impacted this question (although Stamatoyannopoulos 2012 certainly seems to slide in this direction), however they did little to avoid a public misunderstanding of their claim (it has in fact been argued that they fuelled it, in order to benefit from the media hype—Graur et al. 2013; Eddy 2013).

When junk is not understood in an explicitly evolutionary way, its meaning is inextricably linked to an account of function: it is taken to mean either functionless DNA, or (following Doolittle) DNA whose function (if any) is not sequence-dependent. The distinction between structural function and function by virtue of encoding does not lessen the dispute as to how ENCODE's claims bear on the junk/function debate. The reason is that any philosophically sound notion of 'encoding' would most likely consider all of ENCODE's so-called "biochemical functions" as functions by virtue of encoding: the largest part of the genome was declared 'functional' because it is transcribed, and to determine the sequence of a RNA molecule is clearly to encode for something (it is sequence-dependent). Of what remains, most regions are transcription factor binding sites or, as in the case of enhancers/insulators or hyper-accessibility sites, regions that are characterized by transcription factor binding, and most biologists would agree that binding sites are 'encoded'.

The controversy, therefore, does not so much rest on the informational criterion, but rather on some criterion of relevance for functionality, which ENCODE's critics identify with selection. If functions are limited to what confers noticeable increment in fitness, then once more ENCODE's result do not conflict with the view that most of our genome is junk. However, as we argued throughout this paper, selection is not all that matters. One can have different criteria of relevance, and depending on the criterion, one will have a different notion of non-functional and as a consequence a different notion of junk DNA.

Finally, if junk DNA is instead taken to mean DNA that does not make a significant difference to biological phenomena of interest to biomedical research, then ENCODE's claim of functionality of the genome does conflict with the view that most of our genome is junk. Given the arguments presented earlier, we believe that there is compelling but not conclusive evidence that most of the genome can have an impact on relevant phenomena, where relevance is to be understood in terms of social aims such as healthcare, rather than exclusively as evolutionarily relevant. While we do not wish to endorse the claim that most of our genome is making a difference to phenomena relevant to biomedicine, we merely wish to point out that it is not implausible—that for the time being, scientists can reasonably disagree about it. And we believe that this is the core substantial debate behind the ENCODE controversy.

Intelligent design

It is worth addressing the implications of these different notions on the doctrine of Intelligent design. The claim that most of our genome is essential, in the sense that it would be needed by an ideal engineer to construct a creature similar to humans in relevant respects, does suggest an intelligent design, for nature is not an ideal engineer. However, we have argued that the very ambiguous notion of essential DNA should be kept separated from notions of function and of junk DNA. The Onion test and in general the C-value paradox present decisive arguments against the claim that most of our genome is essential, and those arguments remain unaffected by ENCODE's results.

In a similar way, the claim that most of our genome was not selected for and is not advantageous implies that we are accidental, less-than-perfect machines, and thereby conflicts with central tenets of Intelligent design. However, the reverse is not true: the claim that most of our genome was selected for—which is obviously false and was never made by ENCODE scientists—would not *per se* argue for intelligent design. It is merely compatible with it, just as it is compatible with evolution by natural selection. Indeed, we could expect natural selection to evolve lean genomes under different circumstances, for instance if the cost of non-advantageous DNA was very strong (as it is in many microorganisms, and probably explains the bladderwort's lean genome—see Ibarra-Laclette et al. 2013).

Finally, the only claim that can potentially be attributed to the ENCODE Consortium—that most of our genome is relevant to the phenomena of interest to biomedical research—does not lend any support to the hypothesis of an intelligent design. Instead, what this claim does imply is that biological systems are messy: beyond the activities selected for during evolution, a host of other elements are engaging in relevant biochemical activities, exerting a small or perhaps very contextual influence on relevant phenomena, which are sometimes the straw that breaks the camel's back.

Conclusions

As we see it, the main criticism of ENCODE's claim that ±80 % of the genome is functional is that it relies on a wrong notion of function, making its dismissal of the prevalence of junk DNA merely rhetoric. According to the critics, ENCODE's notion of function leads to a multiplication of spurious or irrelevant functions. They argue that an evolutionary account of functions avoids this pitfall, and that it is *the* right account of function for biology.

We argued that Cummins' account of functions as causal roles in the behaviour of a containing system represents the right analysis of what functions are, at least for the purposes of functional biology, and that it is actually presupposed by evolutionary accounts. Evolutionary accounts of functions follow Cummins' analysis (Griffiths 1993; Weber 2005), but establish fitness as the only *explanandum* for functional decomposition. Instead, we have argued that the *explananda* of biologists are not limited to evolution, and that what is relevant to biomedical sciences is not necessarily relevant to evolution. As a consequence, evolutionary accounts risk to miss many of

the functions relevant to phenomena of interest in biomedicine. More precisely, many phenomena of biomedicine are not necessarily related to biological advantages but are nonetheless targets of functional analysis *à la* Cummins.

Selection is neither sufficient nor necessary for function. It is a very useful *proxy* to relevant functions, but an imperfect one and not the only one. We have argued that it might be inadequate for the context of contemporary biomedical research and that ENCODE's strategy might be more adequate. We described the discovery process to which the ENCODE contributes and how its biochemical signature strategy is relevant to identifying the relevant activities of its various parts with an eye towards causal roles of interest to biochemical research.

In light of this analysis, we argued that ENCODE's controversial claim should be interpreted as saying that 80 % of the genome engages in relevant biochemical activities and is very likely to have a causal role in phenomena deemed relevant to biomedical research. In other words, that most of what has been called junk DNA cannot be ignored in biomedical research. Even this interpretation of ENCODE's claim is far from being uncontroversial, and we believe that there can be legitimate disagreement about it, but would argue that it is a plausible claim.

We can summarize our observations by relating different notions of non-functional DNA to different accounts of function. To each of them in turn can be associated a particular interpretation of ENCODE's claim that most of our DNA is functional (Table 1).

In the last part of the paper, we discussed the meaning of junk DNA in order to evaluate the impact of ENCODE's results on the claim that most of our genome is junk. We revealed an ambiguity in the meaning of junk DNA, not only among ENCODE's critics but present since the very origin of the concept, and showed how this ambiguity may lead to problematic arguments for the prevalence of junk DNA. Finally, we examined, for the different possible meanings of junk DNA, whether and how ENCODE's claims bear on it. We argued that while our interpretation of ENCODE's claims contradicts the statement that most of DNA is not currently making a significant difference to the organism's relevant phenotype (*B* in Table 1), it neither supports nor conflict with the claim that most of our genome has not been selected for (*C*). Nor does it lend any support to the rather ambiguous claim that most of our genome is essential (*D*).

Big scientific projects such as the ENCODE or the Human Genome Project (HGP) raise many philosophical issues, of which we have discussed only a small fraction. We hope that our effort will improve the understanding of the meaning and contribution of such 'big science' projects, for part of the debate we have analysed rests on this issue. Mapping efforts such as the ENCODE or the HGP cannot be evaluated on the same basis as traditional, hypothesis-driven research projects.[17] Eddy (2013) therefore diagnoses that, in order to advertise the value of the project,

[17] According to Eddy (2013), "[t]here are three categories of big science: the big experiment, the map, and the leading wedge. A big experiment is driven by a single question or hypothesis test, but requires a large scale community investment. […] A map is a data resource—comprehensive, complete, closed ended—to be used by multiple groups, over a long time, for multiple purposes. […] A leading wedge is a massed technology development effort, in an area where we need radically better methods." (Eddy 2013, p. R261) While the success of "big experiments" is generally easy to appraise, Eddy deplores that "[w]e have been too shy to defend maps and leading wedges in biology" (Eddy 2013, p. R261).

Table 1 Different notions of junk DNA and their related interpretations of ENCODE's claim

	Notion of non-functional DNA	Version of the claim that most of our DNA is functional	Underlying account of function
A	DNA that does not participate in a subset of biochemical reactions, chosen because of their usual functional relevance to phenomena of interest	Most DNA participates in a relevant subset of biochemical activities (ENCODE)	Activity… of a special kind
B	DNA that is not currently making a difference to phenomena of interest to biomedical researchers	Most DNA makes a significant difference to phenomena of interest	CR (ahistorical), likely approximated by ENCODE's subset of activities
C	DNA that did not make a difference to the fitness of the organism's ancestors	Most DNA made a difference to the fitness of the organism's ancestors	SE (backward-looking)
D	DNA that is not currently making a significant difference to the survival and/or reproduction of the organism	Most DNA is currently making a significant difference to the organism's fitness	Advantage (forward-looking)
E	DNA that would not be strictly required for building a similar organism (in the relevant respects)	An ideal engineer building a similar organism couldn't do without most of our genome	Essential (whatever that means)

> ENCODE's publicity spun it retrospectively as a hypothesis test, but the ENCODE was not designated to test anything. ENCODE is a map: it should have been published and defended as such (Eddy 2013).

In a similar way, Graur et al. (2013) write that "ENCODE's biggest scientific sin was not being satisfied with its role as data provider; it assumed the small-science role of interpreter of the data" (Graur et al. 2013, p. 587). While we agree with Eddy that the ENCODE is a map, we disagree with Graur et al. when they claim that a mapping project such as the ENCODE is merely about producing data. Just like 'collections' (Strasser 2008, 2012), maps are different from a simple accumulation of data: they are rather ways of organizing data according to specific aims, in order to make specific contrasts emerge and to enable specific kinds of investigations. As we have argued in this paper, the mapping of the biological activity of DNA elements produced by the ENCODE project is indeed a 'collection' in that sense, that allows biologists to generate precise hypotheses about the biological role of certain DNA elements.

Like all maps, the ENCODE is a means of doing—it is meant to foster a particular kind of science, which we tried to describe. In this view, ENCODE's claim of functionality, understood as a methodological claim about the proportion of the genome relevant for biomedical research, is part of the map. And as we have argued, understood as such it is a debatable, but defensible claim. The real, empirical disagreement is precisely as to the proportion of DNA one could remove (or mutate) from the human genome without noticing a relevant difference—where relevance is ultimately determined not by evolution but by our interests.

Acknowledgments We wish to acknowledge Fridolin Groß, who was part of the many discussions at the origin of this paper and carefully commented several versions of the paper. In addition, we wish to thank all those who have read drafts of this paper: Michel Morange, Michael Weisberg, Iros Barozzi, Lorenzo Del Savio, Marcel Weber and the lgBIG group in Geneva (in which the paper was discussed), Alkistis Elliot-Graves and Vera Pendino. We are also thankful to our colleagues of the FOLSATEC programme. Finally, we wish to acknowledge the two anonymous reviewers for their help in improving the text.

References

Agency for Healthcare Research and Quality (2001) Reducing and preventing adverse drug events to decrease hospital costs: research in action, issue 1. Retrieved from http://www.ahrq.gov/research/findings/factsheets/errors-safety/aderia/index.html

Bechtel W, Richardson RC (2010) Discovering complexity—decomposition and localization as strategies in scientific research. The MIT Press, Cambridge

Bigelow J, Pargetter R (1987) Functions. J Philos 84(4):181–196

Birney E (2012a) Lesson for big-data projects. Nature 489:49–51

Birney E (2012b) ENCODE: my own thoughts. Ewan's Blog: Bioinformatician at large. Retrieved September 5, 2012, from http://genomeinformatician.blogspot.ch/2012/09/encode-my-own-thoughts.html

Brenner S (1998) Refuge of spandrels. Curr Biol 8:R669

Brown D, Boytchev H (2012) "Junk DNA" concept debunked by new analysis of human genome. The Washington Post. Retrieved September 5, 2012, from http://www.washingtonpost.com/national/health-science/junk-dna-concept-debunked-by-new-analysis-of-human-genome/2012/09/05/cf296720-f772-11e1-8398-0327ab83ab91_story.html

Bunzl M (1980) Comment on "health as a theoretical concept". Philos Sci 47:116–118

Chanock SJ (2012) Toward mapping the biology of the genome. Genome Res 22(9):1612–1615. doi:10.1101/gr.144980.112

Comings DE (1972) The structure and function of chromatin. Adv Human Genetics 3:237–431

Connor S (2003) Glaxo chief: our drugs do not work on most patients. The Independent. Retrieved December 8, 2003, from http://www.independent.co.uk/news/science/glaxo-chief-our-drugs-do-not-work-on-most-patients-575942.html

Craver C (2007) Explaining the brain: mechanisms and the mosaic unity of neuroscience. Oxford University Press, New York

Cummins R (1975) Functional analysis. J Philos 72(20):741–765

Darden L (2006) Reasoning in biological discoveries. Cambridge University Press, Cambridge

Diep F (2013) Friction over function: scientists clash on the meaning of ENCODE's genetic data. Scientific American. Retrieved April 12, 2013, from http://www.scientificamerican.com/article/friction-over-function-encode/

Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci USA 110(14):5294–5300. doi:10.1073/pnas.1221376110

Eddy SR (2012) The C-value paradox, junk DNA and ENCODE. Curr Biol 22:R898–R899. doi:10.1016/j.cub.2012.10.002

Eddy SR (2013) The ENCODE project: missteps overshadowing a success. Curr Biol 23:R259–R261. doi:10.1016/j.cub.2013.03.023

Gaudillière JP, Rheinberger H-J (2004) From molecular genetics to genomics, the mapping cultures of twentieth-century genetics. Routledge, London

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Koon-Kiu Y, Chao C et al (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489:91–100. doi:10.1038/nature11245

Gissis SB, Jablonka E (eds) (2011) Transformations of lamarckism. From subtle fluids to molecular biology. MIT Press, Cambridge

Graur D (2013) The Origin of Junk DNA: A Historical Whodunnit. Judge Starling. Retrieved October 19, 2013, from http://judgestarling.tumblr.com/post/64504735261/the-origin-of-junk-dna-a-historical-whodunnit

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E (2013) On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol 5:578–590. doi:10.1093/gbe/evt028

Gregory TR (2007) The onion test. Genomicron, April 27th 2007, retrieved from http://www.genomicron.evolverzone.com/2007/04/onion-test/

Griffiths PE (1993) Functional analysis and proper functions. Br J Philos Sci 44(3):409–422. doi:10.1093/bjps/44.3.409

Griffiths PE (2001) Genetic information: a metaphor in search of a theory. Philos Sci 68(3):394–412

Griffiths PE (2009) In what sense does "nothing make sense except in the light of evolution"? Acta Biotheor 57:11–32. doi:10.1007/s10441-008-9054-9

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang T-H, Herrera-Estrella L (2013) Architecture and evolution of a minute plant genome. Nature 498(7452):94–98. doi:10.1038/nature1213

Kauffman S (1993) The origins of order: self-organization and selection in evolution. Oxford University Press, Oxford

Kauffman S (1996) At home in the universe: the search for the laws of self-organization and complexity. Oxford University Press, Oxford

Kentikelenis A, Karanikolos M, Papanicolas I, Basu S, McKee M, Stuckler D (2011) Health effects of financial crisis: omens of a Greek tragedy. Lancet 378:1457–1458

Kolata G (2012) Bits of mystery DNA, far from "Junk," play crucial role. The New York Times. p. 5–7. Retrieved September 6, 2012, from http://www.nytimes.com/2012/09/06/science/far-from-junk-dna-dark-matter-proves-crucial-to-health.html

Laland KN, Sterelny K, Odling-Smee J, Hoppitt W, Uller T (2011) Cause and effect in biology revisited: is Mayr's proximate-ultimate dichotomy still useful? Science 334:1512–1516. doi:10.1126/science.1210879

Lynch VJ, Leclerc RD, May G, Wagner GP (2011) Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genetics 43(11):1154–1159. doi:10.1038/ng.917

Maher B (2012) The human encyclopaedia. Nature 486:46–48

Makalowski W (2003) Not junk after all. Science 300(5623):1246–1247. doi:10.1126/science.1085690

Mayr E (1961) Cause and effect in biology. Science 134(3489):1501–1506. doi:10.1126/science.134.3489.1501

Millikan RG (1989) In defense of proper functions. Philos Sci 56:288–302

Neander K (1991) Functions as selected effects. Philos Sci 58:168–184

NHGRI (2002) National Human Genome Research Institute (2002) Workshop summary: the comprehensive extraction of biological information from genomic sequence, retrieved from http://www.genome.gov/10005568

Niu D-K, Jiang L (2013) Can ENCODE tell us how much junk DNA we carry in our genome? Biochem Biophys Res Commun 430:1340–1343. doi:10.1016/j.bbrc.2012.12.074

Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM (2004) Megabase deletions of gene deserts result in viable mice. Nature 431:988–993. doi:10.1038/nature02923.1

Ohno S (1970) Evolution by gene duplication. Springer, New York

Ohno S (1972) So much "junk" DNA in our genome. Brookhaven Symp Biol 23:366–370

Ohno S (1973) Evolutionary reason for having so much junk DNA. In: Pfeiffer RA (ed) Modern aspects of cytogenetics: constitutive heterochromatin in man. F.K. Schattauer Verlag, Stuttgart

Pennisi E (2012) ENCODE project writes eulogy for junk DNA. Science 337:1159–1161

Pigliucci M, Müller GB (eds) (2010) Evolution—the extended synthesis. The MIT Press, Cambridge

Ponting CP, Hardison RC (2011) What fraction of the human genome is functional? Genome Res 21:769–1776. doi:10.1101/gr.116814.110

Pritchard JK, Gilard Y (2012) Evolution and the code. Nat (News & Views) 489:55

Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP (2011) A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language?. Cell 146(3):353–358. doi:10.1016/j.cell.2011.07.014

Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012) Linking disease associations with regulatory information in the human genome. Genome Res 22:1748–1759. doi:10.1101/gr.136127.111

Shapiro JA (2011) Evolution: a view from the 21st century. FT Press, New Jersey

Stamatoyannopoulos J (2012) What does our genome encode? Genome Res 22:1602–1611. doi:10.1101/gr.146506.112

Strasser BJ (2008) GenBank—natural history in the 21st century. Science 322(5901):537–538. doi:10.1126/science.1163399

Strasser BJ (2012) Data-driven sciences: from wonder cabinets to electronic databases. Stud Hist Philos Biol Biomed Sci 43:85–87. doi:10.1016/j.shpsc.2011.10.009

The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. Science 306:636–640. doi:10.1126/science.1105136

The ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9(4):e1001046. doi:10.1371/journal.pbio.1001046

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. doi:10.1038/nature11247

Tinbergen N (1963) On aims and methods in ethology. Zeitschrift für Tierpsychologie 20(4):410–433

Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K et al (2011) SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. N Engl J Med 365:2497–2506. doi:10.1056/NEJMoa1109016

Weber M (2005) Philosophy of experimental biology. Cambridge University Press, Cambridge

Wouters AG (2003) Four notions of biological function. Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci 34:633–668. doi:10.1016/j.shpsc.2003.09.006

Wright L (1973) Functions. Philos Rev 82(2):139–168