PERSPECTIVE

Is junk DNA bunk? A critique of ENCODE

W. Ford Doolittle[1]

Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada B3H 4R2

Edited by Michael B. Eisen, Howard Hughes Medical Institute, University of California, Berkeley, CA, and accepted by the Editorial Board February 4, 2013 (received for review December 11, 2012)

Do data from the Encyclopedia Of DNA Elements (ENCODE) project render the notion of junk DNA obsolete? Here, I review older arguments for junk grounded in the C-value paradox and propose a thought experiment to challenge ENCODE's ontology. Specifically, what would we expect for the number of functional elements (as ENCODE defines them) in genomes much larger than our own genome? If the number were to stay more or less constant, it would seem sensible to consider the rest of the DNA of larger genomes to be junk or, at least, assign it a different sort of role (structural rather than informational). If, however, the number of functional elements were to rise significantly with C-value then, (*i*) organisms with genomes larger than our genome are more complex phenotypically than we are, (*ii*) ENCODE's definition of functional element identifies many sites that would not be considered functional or phenotype-determining by standard uses in biology, or (*iii*) the same phenotypic functions are often determined in a more diffuse fashion in larger-genomed organisms. Good cases can be made for propositions *ii* and *iii*. A larger theoretical framework, embracing informational and structural roles for DNA, neutral as well as adaptive causes of complexity, and selection as a multilevel phenomenon, is needed.

evolution | human genome

There is much excitement in the blogosphere, among mainstream science journalists, and within the community of practicing genome biologists about a flurry of articles and letters in the September 6th, 2012 issue of *Nature*. These papers and many others published at about the same time and since under the umbrella of the ENCODE project collectively claim function for the majority of the 3.2 Gb human genome, not just the few percent already recognized as genes (traditionally defined) or obvious gene-controlling elements. Kolata writes in *The New York Times* that "[t]he human genome is packed with at least four million gene switches that reside in bits of DNA that were once dismissed as 'junk' but that turn out to play critical roles in controlling how cells, organs and other tissues behave" (1). In a *Nature* News and View commentary, Ecker et al. (2) assert that "[o]ne of the more remarkable findings described in the consortium's entrée paper is that 80% of the genome contains elements linked to biochemical functions, dispatching the widely held view that the human genome is mostly 'junk DNA.'" The editors of *The Lancet* (3) enthuse: "Far from being 'junk,' the DNA between protein encoding genes consists of myriad elements that determine gene expression, whether by switching transcription on or off, or by regulating the degree of transcription and consequently the concentrations and function of all proteins." Succinctly, in *Science*, Pennisi (4) declares that the ENCODE publications write the "eulogy for junk DNA."

The new data—coming from high-throughput analyses of transcriptional and chromatin landscapes, transcription factor footprints, and long-range chromosomal interactions—support many current population genetic studies linking human diseases to supposedly nongenic regions, and they are truly impressive in scope and depth (5). They resonate with the current enthusiasm for assigning multiple subtle but vital regulatory roles to the still enigmatic long noncoding RNAs (lncRNAs) now known to be transcribed from much of the length of our genome (6, 7). Additionally, congruence at many sites between the many methods used (RNA sequencing, binding by one or more of 100+ DNA binding proteins, DNase I hypersensitivity, histone modification, DNA methylation, and chromosome conformation capture) leaves no doubt that many of these 4 million gene switches do represent chromosomal loci that are special in some way, in at least one cell type. However, do ENCODE's data truly require us to abandon the widespread notion that junk DNA—here specifically understood as DNA that does not encode information promoting the survival and reproduction of the organisms that bear it—is the major constituent of many eukaryotic genomes, our own genome included?

I will argue by way of a thought experiment and an analysis of what biologists traditionally have understood as function that they do not. At the very least, "junk" as it has been conceived is an apt descriptor of the bulk of many genomes larger than our own. Moreover, it almost certainly still is for much of our genome, unless we hold *Homo sapiens* to be unique among the animals in the efficiency of its chromosomal organization and not just its cultural attainments. Such genomic anthropocentrism, unacknowledged conflation of possible meanings of "function," questionable null hypotheses, and unrecognized panadaptationism are behind this most recent attempt to junk "junk."

Several of these same points have been made in brief by Eddy (8) and Niu and Jiang (9). My aim here is to remind readers of the structure of some earlier arguments in defense of the junk concept (10) that remain compelling, despite the obvious success of ENCODE in mapping the subtle and complex human genomic landscape. Also, I will suggest that we need as biologists to defend traditional understandings of function: the publicity surrounding ENCODE reveals the extent to which these understandings have been eroded. However, theoretical expansion in other directions, reconceptualizing junk, might be advisable.

Perennial Problem of C-Value

Information and Structure. The junk idea long predates genomics and since its early decades has been grounded in the "C-value paradox," the observation that DNA amounts (C-value denotes haploid nuclear DNA content) and complexities correlate very poorly

Author contributions: W.F.D. wrote the paper.

The author declares no conflict of interest.

This article is a PNAS Direct Submission. M.B.E. is a guest editor invited by the Editorial Board.

[1]E-mail: ford@dal.ca.

Downloaded by guest on January 17, 2021

with organismal complexity or evolutionary "advancement" (10–14). Humans do have a thousand times as much DNA as simple bacteria, but lungfish have at least 30 times more than humans, as do many flowering plants and some unicellular protists (14). Moreover, as is often noted, the disconnection between C-value and organismal complexity is also found within more restricted groups comprising organisms of seemingly similar lifestyle and comparable organismal or behavioral complexity. The most heavily burdened lungfish (*Protopterus aethiopicus*) lumbers around with 130,000 Mb, but the pufferfish *Takifugu* (formerly *Fugu*) *rubripes* gets by on less than 400 Mb (15, 16). A less familiar but better (because monophyletic) animal example might be amphibians, showing a 120-fold range from frogs to salamanders (17). Among angiosperms, there is a thousandfold variation (14). Additionally, even within a single genus, there can be substantial differences. Salamander species belonging to *Plethodon* boast a fourfold range, to cite a comparative study popular from the 1970s (18). Sometimes, such within-genus genome size differences reflect large-scale or whole-genome duplications and sometimes rampant selfish DNA or transposable element (TE) multiplication. Schnable et al. (19) figure that the maize genome has more than doubled in size in the last 3 million y, overwhelmingly through the replication and accumulation of TEs for example. 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 reproduction of the organism bearing it), how then are we to think of it?

Of course, DNA inevitably does have a basic structural role to play, unlinked to specific biochemical activities or the encoding of information relevant to genes and their expression. Centromeres and telomeres exemplify noncoding chromosomal components with specific functions. More generally, DNA as a macromolecule bulks up and gives shape to chromosomes and thus, as many studies show, determines important nuclear and cellular parameters such as division time and size, themselves coupled to organismal development (11–13, 17). The "selfish DNA" scenarios of 1980 (20–22), in which C-value represents only the outcome of conflicts between upward pressure from reproductively competing TEs and downward-directed energetic restraints, have thus, in subsequent decades, yielded to more nuanced understandings. Cavalier-Smith (13, 20) called DNA's structural and cell biological roles "nucleoskeletal," considering C-value to be optimized by organism-level natural selection (13, 20). Gregory, now the principal C-value theorist, embraces a more "pluralistic, hierarchical approach" to what he calls "nucleotypic" function (11, 12, 17). A balance between organism-level selection on nuclear structure and cell size, cell division times and developmental rate, selfish genome-level selection favoring replicative expansion, and (as discussed below) supraorganismal (clade-level) selective processes—as well as drift—must all be taken into account.

These forces will play out differently in different taxa. González and Petrov (23) point out, for instance, that *Drosophila* and humans are at opposite extremes in terms of the balance of processes, with the minimalist genomes of the former containing few (but mostly young and quite active) TEs, whereas at least one-half of our own much larger genome comprises the moribund remains of older TEs, principally SINEs and LINEs (short and long interspersed nuclear elements). Such difference may in part reflect population size. As Lynch notes, small population size (characteristic of our species) will have limited the effectiveness of natural selection in preventing a deleterious accumulation of TEs (24, 25).

Zuckerkandl (26) once mused that all genomic DNA must be to some degree "polite," in that it must not lethally interfere with gene expression. Indeed, some might suggest, as I will below, that true junk might better be defined as DNA not currently held to account by selection for any sort of role operating at any level of the biological hierarchy (27). However, junk advocates have to date generally considered that even DNA fulfilling bulk structural roles remains, in terms of encoded information, just junk. Cell biology may require a certain C-value, but most of the stretches of noncoding DNA that go to satisfying that requirement are junk (or worse, selfish).

In any case, structural roles or multilevel selection theorizing are not what ENCODE commentators are endorsing when they proclaim the end of junk, touting the existence of 4 million gene switches or myriad elements that determine gene expression and assigning biochemical functions for 80% of the genome. Indeed, there would be no excitement in either the press or the scientific literature if all the ENCODE team had done was acknowledge an established theory concerning DNA's structural importance. Rather, the excitement comes from interpreting ENCODE's data to mean that a much larger fraction of our DNA than until very recently thought contributes to our survival and reproduction as organisms, because it encodes information transcribed or expressed phenotypically in one tissue or another, or specifically regulates such expression.

A Thought Experiment. ENCODE (5) defines a functional element (FE) 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)." A simple thought experiment involving FEs so-defined is at the heart of my argument.

Suppose that there had been (and probably, some day, there will be) ENCODE projects aimed at enumerating, by transcriptional and chromatin mapping, factor footprinting, and so forth, all of the FEs in the genomes of *Takifugu* and a lungfish, some small and large genomed amphibians (including several species of *Plethodon*), plants, and various protists. There are, I think, two possible general outcomes of this thought experiment, neither of which would give us clear license to abandon junk.

The first outcome would be that FEs (estimated to be in the millions in our genome) turn out to be more or less constant in number, regardless of C-value—at least among similarly complex organisms. If larger C-value by itself does not imply more FEs, then there will, of course, be great differences in what we might call functional density (FEs per kilobase) (26) among species. FEs spaced by kilobases in *Arabidopsis* would be megabases apart in maize on average. Averages obscure details: the extra DNA in the larger genomes might be sequestered in a few giant silent regions rather than uniformly stretching out the space between FEs or lengthening intragenic introns. However, in either case, this DNA could be seen as a sort of polite functionless filler or diluent. At best, such DNA might have functions only of the structural or nucleoskeletal/nucleotypic sort. Indeed, even this sort of functional attribution is not necessary. There is room within an expanded, pluralistic and hierarchical theory of C-value (see below) (12, 27) for much DNA that makes no contribution whatever to survival and reproduction at the organismal level and thus is junk at that level, although it may be under selection at the sub- or supraorganismal levels (TEs and clade selection).

If the human genome is junk-free, then it must be very luckily poised at some sort of minimal size for organisms of human complexity. We may no longer think that

Downloaded by guest on January 17, 2021

mankind is at the center of the universe, but we still consider our species' genome to be unique, first among many in having made such full and efficient use of all of its millions of SINES and LINES (retrotransposable elements) and introns to encode the multitudes of lncRNAs and house the millions of enhancers necessary to make us the uniquely complex creatures that we believe ourselves to be. However, were this extraordinary coincidence the case, a corollary would be that junk would not be defunct for many other larger genomes: the term would not need to be expunged from the genomicist's lexicon more generally. As well, if, as is commonly believed, much of the functional complexity of the human genome is to be explained by evolution of our extraordinary cognitive capacities, then many other mammals of lesser acumen but similar C-value must truly have junk in their DNA.

The second likely general outcome of my thought experiment would be that FEs as defined by ENCODE increase in number with C-value, regardless of apparent organismal complexity. If they increase roughly proportionately, FE numbers will vary over a many-hundredfold range among organisms normally thought to be similarly complex. Defining or measuring complexity is, of course, problematic if not impossible. Still, it would be hard to convince ourselves that lungfish are 300 times more complex than *Takifugu* or 40 times more complex than us, whatever complexity might be. More likely, if indeed FE numbers turn out to increase with C-value, we will decide that we need to think again about what function is, how it becomes embedded in macromolecular structures, and what FEs as defined by ENCODE have to tell us about it.

Problematics of Function

Definition and Inference. What do we mean by function, informational or otherwise? Most philosophers of biology, and likely, most practicing biologists when pressed, would endorse some form of the selected effect (SE) definition of function (28–30). Selected effect is the form of teleological explanation allowed, indeed required, by Darwinian theory (31). Accordingly, the functions of a trait or feature are all and only those effects of its presence for which it was under positive natural selection in the (recent) past and for which it is under (at least) purifying selection now. They are why the trait or feature is there today and possibly why it was originally formed. Thus, we might reasonably say that the function of the *lac* operon in *Escherichia coli* is (and presumably, long has been) to allow facultative growth of bacteria on β-galactosides, because we believe that, long before *E. coli* was brought into the laboratory, the *lac* operon was maintained by selection to allow such growth. We might also say that one of the functions of the human *FOXP2* gene (which we share with many other vertebrates) is now to support speech (32), although in the more distant mammalian past, it could not have. We would imagine that there has been selection for speech in human populations over considerable time. Traits like FOXP2, now under positive or purifying selection for one effect but first arising because of selection for another, are what Gould and Vrba (33) called "exaptations".

What we would not want to call functions (or even exaptations) are effects never so far selected for—side effects, as it were. Gould and Lewontin (34) famously called these "spandrels." They comprise both undesirable but apparently unavoidable consequences, like vulnerability to phages in bacteria with pili or lower back pain in primates walking upright, and seemingly neutral ones, like the thumping noise made by the heart, to use an example beloved of philosophers. Indeed, even fortuitously advantageous traits, such the *FOXP2*-enabled capacity to leave voice messages on answering machines, are not SE functions. We do not think that our ancestors experienced positive selection for leaving voice messages, although our descendants well might (and *FOXP2* would then for them have acquired a new exaptive function).

In any case, past selection, recent or ancient, can only be inferred, and we must use indirect ways to make the inference. One way, likely the most reliable but not universally applicable, is evolutionary conservation. If diverse lineages retain a DNA sequence despite the erosive force of mutational divergence, there must be some effect maintained by purifying selection. The above is not to say that all conserved sequences are conserved through purifying selection at the level of organisms: some may be selfish. Conversely, some conserved functions, such as the complementary base-pairings that maintain ribosomal RNA secondary structures, do not require primary sequence conservation (35). Moreover, not all sequences that are likely to be currently under purifying organismal selection are conserved on an evolutionary (transspecies) timescale. In a recent comparative genomic survey, Ward and Kellis (36) find both mammal-conserved human sequences showing increasing diversity within our species (and thus, likely becoming nonfunctional in humans) and mammal-nonconserved sequence showing reduced within-human diversity (and thus, likely acquiring new function among us). Ponting and Hardison (37), using methods that they believe to take into account such turnover, "estimate that the steady-state value of α_{sel} [the proportion of all nucleotides in the human genome that are subject to purifying selection because of their biological function] lies between 10% and 15%" (37).

Another way to attribute function is through experimental ablation: whatever organism-level effect *E* does not occur after deleting or blocking the expression of a region *R* of DNA is taken to be the latter's function. This attribution is close to the everyday understanding of function, as in the function of the carburetor is to oxygenate gasoline. The approach embodies what philosophers would call a causal role (CR) definition of function and supposedly eschews evolutionary or historical justifications. Much biological research into function is done this way, but I think that most biologists consider that experimental ablation indirectly points to SE. They believe that effect *E* could, under suitable conditions, be shown to have contributed to the past fitness of organisms and most importantly, that R exists as it does because of *E*. Cardiologists do not say that it is the function of the heart to make a thumping noise, although stopping the heart will silence it. Similarly, geneticists studying Huntington disease would not say that the trinucleotide repeat in the cognate gene, reiteration of which gives rise to the disorder, has disease causation as a function—although replacing the repeat with a nonidentical set of unique triplets encoding the same amino acid sequence would eliminate the deleterious effect (38).

A third, and the least reliable, method to infer function is mere existence. The presence of a structure or the occurrence of a process or detectable interaction, especially if complex, is taken as adequate evidence for its being under selection, even when ablation is infeasible and the possibly selectable effect of presence remains unknown. Because our genomes have introns, *Alu* elements, and endogenous retroviruses, these things must be doing us some good. Because a region is transcribed, its transcript must have some fitness benefit, however remote. Because residue N of protein P is leucine in species A and isoleucine in species B, there must be some selection-based explanation. This approach enshrines "panadaptationism," which was forcefully and effectively debunked by Gould and Lewontin (34) in 1979 but still informs much of molecular

Downloaded by guest on January 17, 2021

and evolutionary genetics, including genomics. As Lynch (39) argues in his essay "*The Frailty of Adaptive Hypotheses for the Origins of Adaptive Complexity*,"

> This narrow view of evolution has become untenable in light of recent observations from genomic sequencing and population genetic theory. Numerous aspects of genomic architecture, gene structure, and developmental pathways are difficult to explain without invoking the nonadaptive forces of genetic drift and mutation. In addition, emergent biological features such as complexity, modularity, and evolvability, all of which are current targets of considerable speculation, may be nothing more than indirect by-products of processes operating at lower levels of organization.

Functional attribution under ENCODE is of this third sort (mere existence) in the main. Although FEs as defined by ENCODE might be cross-identified by several methods and even evolutionarily conserved, they could most often be the molecular equivalent of spandrels—structured elements that are the indirect consequence of selection operating on other features but are themselves selectively neutral, a form of structured noise. Demonstrations that some biochemical signatures are not neutral and may even meet SE criteria say nothing about the rest and are, of course, expected as long as opportunism and co-optation are understood to be key elements in evolution.

In taking such a liberal definitional course, ENCODE follows the lead of the Gene Ontology (GO) project, which defines molecular function in decontextualized nonhistorical terms (40):

> Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place.

Proponents of ENCODE actually are concerned with what they call "functional validation" but principally seem worried about the danger of mistaking functional specificity (recognition or activity) rather than the risk of attributing function (especially regulatory function) where none exists in the SE sense. Stamatoyannopoulos (41) writes:

> These examples illustrate a natural temptation to equate activity with patterning of epigenomic features. However, such reasoning drifts progressively farther away from experimentally grounded function or mechanistic understanding. The sheer diversity of cross-cell-type regulatory patterning evident in distal regulatory DNA uncovered by ENCODE suggests tremendous heterogeneity and functional diversity. ENCODE is thus in a unique position to promote clearer terminology that separates the identification of functional elements per se from the ascription of specific functional activities using historical experimentally defined categories, and also to dissuade the ascription of very specific functions based on a biochemical signature in place of a deeper mechanistic understanding.

Functions of Classes and Parts. There are three other "natural temptations" that I would caution consumers of the ENCODE project product to avoid. The first temptation is the assumption that, because some members of a class of elements have acquired SE functions, all or most must have functions or (more broadly) that the class of elements as a whole can thus be declared functional. Stamatoyannopoulos (41), for instance, writes:

> In marked contrast to the prevailing wisdom, ENCODE chromatin and transcription studies now suggest that a large number of transposable elements encode highly cell type-selective regulatory DNA that controls not only their own cell-selective transcription, but also those of neighboring genes. Far from an evolutionary dustbin, transposable elements appear to be active and lively members of the genomic regulatory community, deserving of the same level of scrutiny applied to other genic or regulatory features.

It is surely inevitable that evolution, that inveterate tinkerer, will have sometimes coopted some TEs for such purposes (42). However, it is an overenthusiastic extrapolation to describe TEs as a class as "active and lively members of the genomic regulatory community."

Moreover, the word "regulation" has itself degraded through use by genomicists, from designating evolved effects shown or likely to enhance fitness, presumably by efficient control of the use of resources, to more broadly denoting any measurable impact of one element or process on other elements or processes, regardless of fitness consequences. I think this broadening of definition misleads biologists such as Barroso (43) in a passage cited later in this essay. Pacemakers regulate heartbeats and that is their function: tasers and caffeine also affect cardiac rhythm, but we would not (at least in the former case) see this as regulatory function.

Regulation, defined in this loose way, is, for instance, the assumed function of many or most lncRNAs, at least for some authors (6, 7, 44, 45). However, the transcriptional machinery will inevitably make errors: accuracy is expensive, and the selective cost of discriminating against all false promoters will be too great to bear. There will be lncRNAs with promoters that have arisen through drift and exist only as noise (46). Similarly, binding to proteins and other RNAs is something that RNAs do. It is inevitable that some such interactions, initially fortuitous, will come to be genuinely regulatory, either through positive selection or the neutral process described below as constructive neutral evolution (CNE). However, there is no evolutionary force requiring that all or even most do. At another (sociology of science) level, it is inevitable that molecular biologists will search for and discover some of those possibly quite few instances in which function has evolved and argue that the function of lncRNAs as a class of elements has, at last, been discovered. The positivist, verificationist bias of contemporary science and the politics of its funding ensure this outcome.

However, what is the correct conceptual framework here? Why should either function or nonfunction for a class of elements be taken as the null hypothesis, and why should evidence for or against function, however defined, be taken as support of one or the other? Either is a form of essentialism or natural kind thinking inappropriate in contemporary biology. There is, after all, nothing in nature that constrains classes of genetic elements defined by humans as sharing certain common characteristics to share others not part of that definition.

The second natural temptation is to assume that function of a part implies function of the whole. Co-optation of the promoters of TEs or the insertion of bona fide regulatory sequences into introns is taken to impart function to the TE or intron as a whole. However, the cell does not necessarily see the rest of the TE or the intronic surround of an embedded enhancer as relevant to its activity, and only the promoter or enhancer sequence may be under selection. Even when an entire genetic element seems relevant or necessary (whole introns must be removed even if only certain sites are active in their removal), there is the possibility of excess baggage or junk-like character. Do enhancer-harboring introns really need to be so long?

Another analogy seems in order: 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. In this connection, note that much more than one-half of 80.4% of the human genome that ENCODE deems functional is so considered because it is transcribed (4), most often into an intron or lncRNA, only a tiny fraction of the length of which is likely to be involved in potentially regulatory interactions.

Downloaded by guest on January 17, 2021

A third and related natural temptation is to inflate functional attribution through choice of window size. The ENCODE Project Consortium notes (5) that:

> The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type. Much of the genome lies close to a regulatory event: 95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.

Even this last and largest percentage, assuming the average biochemical event to directly involve tens to hundreds of bases (41), assigns functions to only a minority (perhaps 10%) of the genome's base pairs. ENCODE's data are indeed unexpected and impressive, but without some principled and agreed-on metric for functional density and FE boundaries, any answer to the question "how much of the genome is functional?" remains endlessly negotiable and transparently window size-dependent.

Future Function. In her News and View editorial in the September 6, 2012 issue of *Nature*, Barroso (43) speculates as follows concerning the vast majority of human DNA, until now thought useless:

> . . .there is a good reason to keep this DNA. Results from the ENCODE project show that most of these stretches of DNA harbor regions that bind proteins and RNA molecules, bringing these into positions from which they cooperate with each other to regulate the function and level of expression of protein-coding genes. In addition, it seems that widespread transcription from non-coding DNA potentially acts as a reservoir for the creation of new functional molecules, such as regulatory RNAs.

In addition to equating regulation with having an effect, Barroso (43) revives here the notion that excess DNA is not junk because it may some day be of use, and thus it is maintained as a reservoir. Brenner (47) long ago derided this sort of reasoning as follows:

> There is a strong and widely held belief that all organisms are perfect and that everything within them is there for a function. Believers ascribe to the Darwinian natural selection process a fastidious prescience that it cannot possibly have and some go so far as to think that patently useless features of existing organisms are there as an investment for the future...

One sees similar Panglossian futuristic speculation in some of the recent literature on robustness and evolvability (critique in ref. 48). However, it cannot in general be the case that selection operating at the level of fitness of individuals within a species can favor the origin or maintenance of traits that incurs selective cost at that level, while offering only the remotest hope of future benefit to the individual and its descendants.

Other evolutionary mechanisms can. The publications by Lynch et al. (24) and Lynch (39, 49) have effectively argued that drift operating in very small populations will (by chance) encourage accumulation of DNA that can add to C-value and might, in the future, come in handy. Selection at the suborganismal (selfish DNA) level may also seem to be future-directed: TEs do sometimes later become useful through cooptation or general effects on the generation of novelty (42, 50). Indeed, Fedoroff (51) has recently proposed that TEs should not be described pejoratively as "selfish" and that the prevailing view—that the epigenetic silencing mechanisms that eukaryotes use to limit TE replication arose to do just that—puts the evolutionary cart before the horse. Rather, Fedoroff (51) suggests that these mechanisms, by also limiting recombination between repeated TEs (which makes genomes smaller), allow TEs to accumulate, growing genomes and that this was a ". . .critical step in the evolution of multicellular organisms, underpinning the ability to diversify duplicates for expression in specific cells and tissues" (51).

Fedoroff (51) concludes that

> On balance, then, the likelihood that contemporary eukaryotic genomes evolved in the context of epigenetic mechanisms seems vastly greater than the likelihood that they were invented as an afterthought to combat a plague of parasitic transposons.

However, evolutionary explanations of genome structure need not be either/or in this sense, after it is recognized that selection affecting the genome operates at all (including supraorganismal) levels of the biological hierarchy (12, 27). TEs can be selected through selfish replication at the level of DNA, while selection at the level of organisms has established silencing mechanisms to reign in TE replication, and selection at the level of clades has looked favorably on those clades with complex genomes that engender evolutionary novelty and render whole-clade extinction less likely. (That is, clades that have TE-rich dynamic genomes may have indeed because of that produced more and more interestingly diverse and evolutionarily robust and evolvable descendant species.) Evolutionary forces operate simultaneously at all levels in the same and different directions with differing strengths and results measurable in different units (such as the frequency of TEs in an individual genome, TE-enhanced individuals in a species, TE-bearing species in a genus, or classes comprising such species in a phylum). Indeed, one could reasonably argue for an eventual expansion of the SE definition of "function" to include all levels as long as we distinguish them (DNA-level function, organism-level function, clade-level function, and so forth). This definitional expansion might well lead to a reduction in the amount of DNA that we could reasonably call junk. However, I think such an expanded idea of function does not currently inform ENCODE or most genomic thinking, and in any case, the problems of inference, evidence, and appropriate null hypotheses remain.

Function as a Diffusible Quality. Quite often, we can intuit the results of a thought experiment before articulating the reasons for our intuition. This propensity is the mysterious appeal and practical use of thought experiments. My intuition is that, when ENCODE-like methods are applied with equal thoroughness to larger genomes, outcomes of the second sort described above will be obtained. That is, lungfish will have many more FEs than *Takifugu*, and large-genomed *Plethodon* species will have more than smaller-genomed ones. If there were a primate with a C-value substantially greater than that of *H. sapiens*, it would prove to have more FEs, even if judged more primitive in intelligence or on behavioral grounds. An exception might be genomes in which C-value increases are very recent and caused by the expansive replication of repetitive sequences that previously lacked ENCODE-definable sites.

Assuming these predictions are borne out, what might we make of it? Lynch (39) suggests that much of the genomic- and systems-level complexity of eukaryotes vis à vis prokaryotes is maladaptive, reflecting the inability of selection to block fixation of incrementally but mildly deleterious mutations in the smaller populations of the former. Thus, for instance, the fact that eukaryotic molecular machines comprise more interacting subunits than their prokaryotic counterparts reflects the inability of selection operating on smaller populations to enforce functional efficiency. Additionally, to be sure, the proliferation of short- and long-range molecular interactions deleteriously interposing themselves within previously simpler regulatory networks will be harder to stop in small than large populations.

Several recent publications (52–55) have revived the argument that CNE is just such an interpositional force, possibly a very

Downloaded by guest on January 17, 2021

powerful one. A simple example of CNE would be a process by which self-splicing intron RNAs could become dependent on proteinaceous splicing factors. Initially, the RNA secondary structure is sufficient to catalyze self-removal, but fortuitously bound proteins that stabilize the RNA can compensate for (presuppress) mutations that might destabilize elements of the structure necessary for independent splicing. Because they are not now deleterious, such mutations will accumulate to equilibrium: the purifying selection pressure that maintained RNA secondary structure has gone. No selection is involved at this stage. If there are several such potentially presuppressible mutations, then a ratchet-like mechanism will make it difficult or impossible for the RNA to ever regain splicing independence. Elimination of the protein will become a lethal event. Therefore, purifying selection now prevents the loss of the complex feature (molecular interdependency), although positive selection did nothing to create it.

The entire multiprotein, multi-RNA, eukaryotic spliceosome might have evolved through reiterations of this process along with very many of the intricacies of the cellular machinery (52–55). CNE would work in concert with any population-size effect. Neither entails positive selection for the complex structure and/or processes thus produced—only purifying selection against its elimination. Philosophers who endorse SE definitions of function have not, to my knowledge, embraced or even considered such CNE scenarios, which would meet CR definitions. Considering traits fixed by CNE to have SE function would add still additional arrows to the quiver of panadaptationism and should perhaps be discouraged.

A common consequence of CNE is that even structures or processes that have arisen by positive selection because they increase organismal fitness will later become more complex in terms of the number of intermolecular interactions required for their successful completion. Function diffuses. Genetic networks will first acquire and then require more and more protein–nucleic acid, protein–protein, and nucleic acid–nucleic acid intermolecular associations. Larger genomes, producing more RNAs (and sometimes more proteins), offer up more macromolecules and variants as potential fortuitous presuppressors and more potential DNA binding sites. Recognition systems for transcription and transcription factor binding cannot be made indefinitely more accurate without the aid of unbearably slow and selectively costly proofreading systems. Tradeoffs between speed, economy, and accuracy will unavoidably entail that larger genomes will produce disproportionately more noise in terms of fortuitous transcripts capable of becoming presuppressors and thus, more complex, seemingly regulatory, networks of interaction. Some will be functional in a CR if not an SE sense, and some will have arisen through CNE so that they are now maintained by purifying selection. However, many, possibly the vast majority, are just there.

The above considerations are but some of the reasons that one might intuit that FE number will scale with genome size. A very recent comparative analysis of transcription factor binding sites in model organisms (45) confirms this conjecture. Ruths and Nakhleh (45) claim that

> . . .neutral evolutionary forces alone can explain binding site accumulation, and that selection on the regulatory network does not alter this finding. If neutral forces drive the accumulation of binding sites, then, despite selective constraints, organisms with large amounts of [noncoding] DNA would evolve functional, yet 'overcomplicated' networks.

So Is Junk Bunk?

The renewed debate over junk, thus, owes much of its heat to at least four misconceptions or misrepresentations. First is the pretense that there is any definable boundary between informational and structural (genic and nongenic) functions for DNA. Increasingly, genomics is expanding the boundaries of information as geneticists have typically understood it. Minimally, gene means more than it used to mean. Djebali et al. (56) write

> . . .the determination of genic regions is currently defined by the cumulative lengths of the isoforms and their genetic association to phenotypic characteristics, the likely continued reduction in the lengths of intergenic regions will steadily lead to the overlap of most genes previously assumed to be distinct genetic loci. This supports and is consistent with earlier observations of a highly interleaved transcribed genome, but more importantly, prompts the reconsideration of the definition of a gene. As this is a consistent characteristic of annotated genomes, we would propose that the transcript be considered as the basic atomic unit of inheritance. Concomitantly, the term gene would then denote a higher-order concept intended to capture all those transcripts (eventually divorced from their genomic locations) that contribute to a given phenotypic trait.

However, regulatory loci are also informational even if not transcribed, and ENCODE has documented many long-range interactions between chromosomal regions that may be brought together physically in the nucleus, a very complex and structure-rich molecular machine, at some time during the cell cycle. Therefore, in this sense, the gross structure of the chromosome set also carries information that may be relevant to the function of genes, broadly defined. Additionally, all DNA has the job of serving as a template for its own replication—to that extent, encoding information.

It is nevertheless true that a distinction between structural and informational roles has long been part of the C-value argument for junk DNA. This line of reasoning has held that high C-value might be necessary for cellular function, but the nongenic DNA that fills the requirement is informationally junk. ENCODE's claim is that much more of the DNA is, in fact, informational (especially regulatory) than we had thought, and indeed ENCODE's focus is on sites likely to be involved directly or indirectly in transcription—on the "myriad elements that determine gene expression" to quote *The Lancet* (3). Therefore, the structure–information distinction informs the interpretation of the project's results, and without it there would be nothing novel or newsworthy in the assertion that all of the human genome has some sort of role in human biology. We have known that since the mid-1980s.

Second is the conflation of SE and CR definitions of function. 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. 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 nonhistorical CR definition of function. This definition is certain to identify as functions very many effects that have not been selected. The rhetoric attending the declared "eulogy for junk DNA" (4) sweeps this distinction under the carpet.

Third is a false natural kind ontology, essentialist in nature, that encourages (*i*) the attribution to a whole class of operationally defined genetic elements those functions known only for a few and/or (*ii*) the attribution to the whole length of such a genetic element a function that resides in only part of it. In the case of lncRNAs and intron transcripts, whose lengths together make up more than three-quarters of 80% of the genome said to be functional, this second sort of functional attribution seems especially misleading.

Fourth may be a seldom-articulated or -questioned notion that cellular complexity is adaptive, the product of positive selection

Downloaded by guest on January 17, 2021

at the organismal level. Our disappointment that humans do not have many more genes than fruit flies or nematodes has been assuaged by evidence that regulatory mechanisms that mediate those genes' phenotypic expressions are more various, subtle, and sophisticated (57), evidence of the sort that ENCODE seems to vastly augment. Yet there are nonselective mechanisms, such as CNE, that could result in the scaling of FEs as ENCODE defines them to C-value nonadaptively or might be seen as selective at some level higher or lower than the level of individual organisms. Splits within the discipline between panadaptationists/neutralists and those researchers accepting or doubting the importance of multilevel selection fuel this controversy and others in biology.

I submit that, up until now, junk has been used to denote DNA whose presence cannot reasonably be explained by natural selection at the level of the organism for encoded informational roles. There remain good reasons to believe that much of the DNA of many species fits this definition. Nevertheless, while still insisting on SE functionality, we might want to come up with new definitions of function and junk by (*i*) abandoning the distinction between informational and nucleoskeletal or nucleotypic roles for DNA, (*ii*) admitting that there may be strong selection for C-value as a determinant of many cell biological features, (*iii*) fully embracing hierarchical selection theory and acknowledging that different genomic features may have legitimate functions defined and in play at different levels, and (*iv*) expanding the SE definition of function to include traits that arise neutrally but are preserved by purifying selection (12). Much that we now call junk could then become functional. However, such a philosophically informed theoretical expansion is not what ENCODE, or at least those authors stressing the demise of junk, so far seem to have in mind (1–5).

In the end, of course, there is no experimentally ascertainable truth of these definitional matters other than the truth that many of the most heated arguments in biology are not about facts at all but rather about the words that we use to describe what we think the facts might be. However, that the debate is in the end about the meaning of words does not mean that there are not crucial differences in our understanding of the evolutionary process hidden beneath the rhetoric.

Note Added in Proof. I note that a very forcefully worded critique by Graur et al. (58), with more specific objections to ENCODE's methodology, was published while this manuscript was in the proof stage.

ACKNOWLEDGMENTS. I thank Evelyn Fox Keller, Eric Bapteste, James McInerney, Andrew Roger, and the Centre for Comparative Genomics and Evolutionary Bioinformatics (CGEB) group for comments on this manuscript and the Canadian Institutes for Health Research for support.

1 Kolata G (September 5, 2002) Bits of mystery DNA, far from 'junk', play crucial role. *The New York Times*, Section A, p. 1.
2 Ecker JR, et al. (2012) Genomics: ENCODE explained. *Nature* 489(7414):52–55.
3 Anonymous (2012) Cracking ENCODE. *Lancet* 380(9846):950.
4 Pennisi E (2012) Genomics. ENCODE project writes eulogy for junk DNA. *Science* 337(6099):1159–1161.
5 Dunham I, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
6 Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 81:145–166.
7 Barry G, Mattick JS (2012) The role of regulatory RNA in cognitive evolution. *Trends Cogn Sci* 16(10):497–503.
8 Eddy SR (2012) The C-value paradox, junk DNA and ENCODE. *Curr Biol* 22(21):R898–R899.
9 Niu DK, Jiang L (2013) Can ENCODE tell us how much junk we carry in our genome? *Biochem Biophys Res Commun* 430(4): 1340–1343.
10 Ohno S (1972) So much "junk" DNA in our genome. *Brookhaven Symp Biol* 23:366–370.
11 Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol Rev Camb Philos Soc* 76(1):65–101.
12 Gregory TR (2004) Macroevolution, hierarchy theory and the C-value enigma. *Paleobiology* 30:179–202.
13 Cavalier-Smith T (2005) Economy, speed and size matter: Evolutionary forces driving nuclear genome miniaturization and expansion. *Ann Bot (Lond)* 95(1):147–175.
14 Gregory TR (2005) Synergy between sequence and size in large-scale genomics. *Nat Rev Genet* 6(9):699–708.
15 Metcalfe CJ, Filée J, Germon I, Joss J, Casane D (2012) Evolution of the Australian lungfish (*Neoceratodus forsteri*) genome: A major role for CR1 and L2 LINE elements. *Mol Biol Evol* 29(11):3529–3539.
16 Brenner S, et al. (1993) Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* 366 (6452):265–268.
17 Gregory TR (2003) Variation across amphibian species in the size of the nuclear genome supports a pluralistic, hierarchical approach to the C-value enigma. *Biol J Linn Soc Lond* 79:329–339.
18 Mizuno S, Macgregor HC (1974) Chromosomes, DNA sequences, and evolution in salamaders of the genus *Plethodon*. *Chromosoma* 48:239096.
19 Schnable PS, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326(5956):1112–1115.
20 Cavalier-Smith T (1978) Nuclear volume control by nucleoskeletal DAN, selection for cell volme and cell growth rate, and the solution of the DNA C-value paradox. *J Cell Sci* 43:247–278.
21 Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284(5757):601–603.
22 Orgel LE, Crick FHC (1980) Selfish DNA: The ultimate parasite. *Nature* 284(5757):604–607.
23 González J, Petrov DA (2012) Evolution of genome content: Population dynamics of transposable elements in flies and humans. *Methods Mol Biol* 855:361–383.
24 Lynch M, Bobay LM, Catania F, Gout JF, Rho M (2011) The repatterning of eukaryotic genomes by random genetic drift. *Annu Rev Genomics Hum Genet* 12:347–366.
25 Fernández A, Lynch M (2011) Non-adaptive origins of interactome complexity. *Nature* 474(7352):502–505.
26 Zuckerkandl E (1986) Polite DNA: Functional density and functional compatibility in genomes. *J Mol Evol* 24(1–2):12–27.
27 Doolittle WF (1987) What introns have to tell us: Hierarchy in genome evolution. *Cold Spring Harb Symp Quant Biol* 52:907–913.
28 Garson J (2011) Selected effects and causal role functions in the brain: The case for and etiological approach to neuroscience. *Biol Philos* 26:547–565.
29 Godfrey-Smith P (1994) A modern history theory of functions. *Nous* 28:344–362.
30 Schwartz PH (1999) Proper function and recent selection. *Philos Sci* 66:S210–S222.
31 Ayala FJ (1970) Teleological explanations in evolutionary biology. *Philos Sci* 37:1–15.
32 Scharff C, Petri J (2011) Evo-devo, deep homology and FoxP2: Implications for the evolution of speech and language. *Philos Trans R Soc Lond B Biol Sci* 366(1574):2124–2140.
33 Gould SJ, Vrba E (1982) Exaptation—a missing term in the science of form. *Paleobiology* 8:4–15.
34 Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proc R Soc Lond B Biol Sci* 205(1161):581–598.
35 Noller HF, Woese CR (1981) Secondary structure of 16S ribosomal RNA. *Science* 212(4493):403–411.
36 Ward LD, Kellis M (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337(6102):1675–1678.
37 Ponting CP, Hardison RC (2011) What fraction of the human genome is functional? *Genome Res* 21(11):1769–1776.
38 La Spada AR, Taylor JP (2010) Repeat expansion disease: Progress and puzzles in disease pathogenesis. *Nat Rev Genet* 11(4):247–258.
39 Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci USA* 104(Suppl 1): 8597–8604.
40 An introduction to the gene ontology. Available at http://www.geneontology.org/GO.doc.shtml#molecular_function. Accessed December 11, 2012.
41 Stamatoyannopoulos JA (2012) What does our genome encode? *Genome Res* 22(9):1602–1611.
42 Kines KJ, Belancio VP (2012) Expressing genes do not forget their LINEs: Transposable elements and gene expression. *Front Biosci* 17:1329–1344.
43 Barroso I (2012) Non-coding but functional. *Nature* 489:54.
44 Carninci P (2010) RNA dust: Where are the genes? *DNA Res* 17(2):51–59.
45 van Leeuwen S, Mikkers H (2010) Long non-coding RNAs: Guardians of development. *Differentiation* 80(4–5):175–183.
46 Ruths T, Nakhleh L (2012) ncDNA and drift drive binding site accumulation. *BMC Evol Biol* 12:159.
47 Brenner S (1998) Refuge of spandrels. *Curr Biol* 8(19):R669.
48 Pigliucci M (2008) Is evolvability evolvable? *Nat Rev Genet* 9(1): 75–82.
49 Lynch M (2007) *The Origins of Genome Architecture* (Sinauer, Sunderland, MA).
50 Faulkner GJ, Carninci P (2009) Altruistic functions for selfish DNA. *Cell Cycle* 8(18):2895–2900.
51 Fedoroff NV (2012) Presidential address. Transposable elements, epigenetics, and genome evolution. *Science* 338(6108):758–767.
52 Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF (2010) Cell biology. Irremediable complexity? *Science* 330(6006):920–921.
53 Lukeš J, Archibald JM, Keeling PJ, Doolittle WF, Gray MW (2011) How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* 63(7):528–537.
54 Stoltzfus A (2012) Constructive neutral evolution: Exploring evolutionary theory's curious disconnect. *Biol Direct* 7:35.
55 Doolittle WF (2012) Evolutionary biology: A ratchet for protein complexity. *Nature* 481(7381):270–271.
56 Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108.
57 Frazer KA (2012) Decoding the human genome. *Genome Res* 22(9):1599–1601.
58 Graur D, et al. (2013) On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol*, 10.1093/gbe/evt028.

Downloaded by guest on January 17, 2021