Genome Biology and Evolution Advance Access published February 20, 2013 doi:10.1093/gbe/evt028

Genome Biology and Evolution

On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE

Dan Graur[1], Yichen Zheng[1], Nicholas Price[1], Ricardo B. R. Azevedo[1], Rebecca A. Zufall[1], and Eran Elhaik[2]

[1]Department of Biology and Biochemistry, University of Houston, Texas 77204-5001, USA

[2] Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland 21205, USA.

Corresponding author: Dan Graur (dgraur@uh.edu)

Genome Biology and Evolution

Abstract

A recent slew of ENCODE Consortium publications, specifically the article signed by all Consortium members, put forward the idea that more than 80% of the human genome is functional. This claim flies in the face of current estimates according to which the fraction of the genome that is evolutionarily conserved through purifying selection is under 10%. Thus, according to the ENCODE Consortium, a biological function can be maintained indefinitely without selection, which implies that at least 80 – 10 = 70% of the genome is perfectly invulnerable to deleterious mutations, either because no mutation can ever occur in these "functional" regions, or because no mutation in these regions can ever be deleterious. This absurd conclusion was reached through various means, chiefly (1) by employing the seldom used "causal role" definition of biological function and then applying it inconsistently to different biochemical properties, (2) by committing a logical fallacy known as "affirming the consequent," (3) by failing to appreciate the crucial difference between "junk DNA" and "garbage DNA," (4) by using analytical methods that yield biased errors and inflate estimates of functionality, (5) by favoring statistical sensitivity over specificity, and (6) by emphasizing statistical significance rather than the magnitude of the effect. Here, we detail the many logical and methodological transgressions involved in assigning functionality to almost every nucleotide in the human genome. The ENCODE results were predicted by one of its authors to necessitate the rewriting of textbooks. We agree, many textbooks

SMBE

dealing with marketing, mass-media hype, and public relations may well have to be rewritten.

"Data is not information, information is not knowledge, knowledge is not wisdom, wisdom is not truth,"

—Robert Royar (1994) paraphrasing Frank Zappa's (1979) anadiplosis

"I would be quite proud to have served on the committee that designed the *E. coli* genome. There is, however, no way that I would admit to serving on a committee that designed the human genome. Not even a university committee could botch something that badly."

—David Penny (personal communication)

"The onion test is a simple reality check for anyone who thinks they can assign a function to every nucleotide in the human genome. 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* (personal communication)

Early releases of the **ENC**yclopedia **O**f **D**NA **E**lements (ENCODE) were mainly aimed at providing a "parts list" for the human genome(The ENCODE Project Consortium 2004). The latest batch of ENCODE Consortium publications, specifically the article signed by all Consortium members (The ENCODE Project Consortium 2012), has much more ambitious interpretative aims (and a much better orchestrated public-relations

campaign). The ENCODE Consortium aims to convince its readers that almost every nucleotide in the human genome has a function and that these functions can be maintained indefinitely without selection. ENCODE accomplishes these aims mainly by playing fast and loose with the term "function," by divorcing genomic analysis from its evolutionary context and ignoring a century of population genetics theory, and by employing methods that consistently overestimate functionality, while at the same time being very careful that these estimates do not reach 100%. More generally, the ENCODE Consortium has fallen trap to the genomic equivalent of the human propensity to see meaningful patterns in random data—known as apophenia (Brugger 2001; Fyfe et al. 2008)—that have brought us other "codes" in the past (e.g., Witztum 1994; Schinner 2007).

In the following we shall dissect several logical, methodological, and statistical improprieties involved in assigning functionality to almost every nucleotide in the genome. We shall only deal with a single article (The ENCODE Project Consortium 2012) out of more than 30 that have been published since the 6 September 2012 release. We shall also refer to two commentaries, one written by a scientist and one written by a *Science* journalist (Pennisi 2012a,b), both trumpeting the death of "junk DNA."

"Selected Effect" and "Causal Role" Functions

The ENCODE Project Consortium assigns function to 80.4% of the genome (The ENCODE Project Consortium 2012). We disagree with this estimate. However, before

challenging this estimate, it is necessary to discuss the meaning of "function" and "functionality." Like many words in the English language, these terms have numerous meanings. What meaning, then, should we use? In biology, there are two main concepts of function: the "selected effect" and "causal role" concepts of function. The "selected effect" concept is historical and evolutionary (Millikan 1989; Neander 1991). Accordingly, for a trait, *T*, to have a proper biological function, *F*, it is necessary and (almost) sufficient that the following two conditions hold: (1) *T* originated as a "reproduction" (a copy or a copy of a copy) of some prior trait that performed *F* (or some function similar to *F*) in the past, and (2) *T* exists because of *F* (Millikan 1989). In other words, the "selected effect" function of a trait is the effect for which it was selected, or by which it is maintained. In contrast, the "causal role" concept is ahistorical and nonevolutionary (Cummins 1975; Admundson and Lauder 1994). That is, for a trait, *Q*, to have a "causal role" function, *G*, it is necessary and sufficient that *Q* performs *G*. For clarity, let us use the following illustration (Griffiths 2009). There are two almost identical sequences in the genome. The first, TATAAA, has been maintained by natural selection to bind a transcription factor, hence, its selected effect function is to bind this transcription factor. A second sequence has arisen by mutation and, purely by chance, it resembles the first sequence; therefore, it also binds the transcription factor. However, transcription factor binding to the second sequence does not result in transcription, i.e., it has no adaptive or maladaptive consequence. Thus, the second sequence has no selected effect function, but its causal role function is to bind a transcription factor.

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 grams to body weight, producing sounds, and preventing the pericardium from deflating onto itself. As a result, most biologists use the selected effect concept of function, following the Dobzhanskyan dictum according to which biological sense can only be derived from evolutionary context. We note that the causal role concept may sometimes be useful; mostly as an *ad hoc* devise for traits whose evolutionary history and underlying biology are obscure. This is obviously not the case with DNA sequences.

The main advantage of the selected-effect function definition is that it suggests a clear and conservative method of inference for function in DNA sequences; only sequences that can be shown to be under selection can be claimed with any degree of confidence to be functional. The selected effect definition of function has led to the discovery of many new functions, e.g., microRNAs (Lee et al. 1993) and to the rejection of putative functions, e.g., *numt*s (Hazkani-Covo et al. 2010).

From an evolutionary viewpoint, a function can be assigned to a DNA sequence if and only if it is possible to destroy it. All functional entities in the universe can be rendered nonfunctional by the ravages of time, entropy, mutation, and what have you. Unless a genomic functionality is actively protected by selection, it will accumulate deleterious mutations and will cease to be functional. The absurd alternative, which unfortunately

was adopted by ENCODE, is to assume that no deleterious mutations can ever occur in the regions they have deemed to be functional. Such an assumption is akin to claiming that a television set left on and unattended will still be in working condition after a million years because no natural events, such as rust, erosion, static electricity, and earthquakes can affect it. The convoluted rationale for the decision to discard evolutionary conservation and constraint as the arbiters of functionality put forward by a lead ENCODE author (Stamatoyannopoulos 2012) is groundless and self-serving.

Of course, it is not always easy to detect selection. Functional sequences may be under selection regimes that are difficult to detect, such as positive selection, or weak (statistically undetectable) purifying selection, or they may be recently evolved species-specific elements. We recognize these difficulties, but it would be ridiculous to assume that 70+% of the human genome consists of elements under undetectable selection, especially given other pieces of evidence, such as mutational load (Knudson 1979; Charlesworth et al. 1993). Hence the proportion of the human genome that is functional is likely to be larger to some extent than the approximately 9% for which there exists some evidence for selection (Smith et al. 2004), but the fraction is unlikely to be anything even approaching 80%. Finally, we would like to emphasize that the fact that it is sometimes difficult to identify selection should never be used as a justification to ignore selection altogether in assigning functionality to parts of the human genome.

ENCODE adopted a strong version of the causal role definition of function, according to which a functional element is a discrete genome segment that produces a protein or an

RNA or displays a reproducible biochemical signature (for example, protein binding). Oddly, ENCODE not only uses the wrong concept of functionality, it uses it wrongly and inconsistently (see below).

Using the Wrong Definition of "Functionality" Wrongly

Estimates of functionality based on conservation are likely to be, well, conservative. Thus, the aim of the ENCODE Consortium to identify functions experimentally is, in principle, a worthy one. We have already seen that ENCODE uses an evolution-free definition of "functionality." Let us for the sake of argument assume that there is nothing wrong with this practice. Do they use the concept of causal role function properly? According to ENCODE, for a DNA segment to be ascribed functionality it needs to (1) be transcribed or (2) associated with a modified histone or (3) located in an open-chromatin area or (4) to bind a transcription factors or (5) to contain a methylated CpG dinucleotide. We note that most of these properties of DNA do not describe a function; some describe a particular genomic location or a feature related to nucleotide composition. To turn these properties into causal role functions, the ENCODE authors engage in a logical fallacy known as "affirming the consequent". The ENCODE argument goes like this:

1. DNA segments that function in a particular biological process (e.g., regulating transcription) tend to display a certain property (e.g., transcription factors bind to them).
2. A DNA segment displays the same property.

3. Therefore, the DNA segment is functional.

(More succinctly: if function, then property; thus, if property, therefore function.) This kind of argument is false because a DNA segment may display a property without necessarily manifesting the putative function. For example, a random sequence may bind a transcription-factor, but that may not result in transcription. The ENCODE authors apply this flawed reasoning to all their functions.

Is 80% of the genome functional? Or is it 100%? Or 40%? No wait…

So far, we have seen that as far as functionality is concerned, ENCODE used the wrong definition wrongly. We must now address the question of consistency. Specifically, did ENCODE use the wrong definition wrongly in a consistent manner? We don't think so. For example, the ENCODE authors singled out transcription as a function, as if the passage of RNA polymerase through a DNA sequence is in some way more meaningful than other functions. But, what about DNA polymerase and DNA replication? 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 biochemical signature"—it replicates!

Actually, the ENCODE authors could have chosen any of a number of arbitrary percentages as "functional," and… they did! In their scientific publications, ENCODE promoted the idea that 80% of the human genome was functional. The scientific commentators followed, and proclaimed that at least 80% of the genome is "active and

needed" (Kolata 2012) Subsequently, one of the lead authors of ENCODE admitted that the press conference mislead people by claiming that 80% of our genome was "essential and useful." He put that number at 40% (Gregory 2012), while another lead author reduced the fraction of the genome that is devoted to function to merely 20% (Hall 2012). Interestingly, even when a lead author of ENCODE reduced the functional genomic fraction to 20%, he continued to insist that the term "junk DNA" needs "to be totally expunged from the lexicon," inventing a new arithmetic according to which 20% > 80%. In its synopsis of the year 2012, the journal *Nature* adopted the more modest estimate, and summarized the findings of ENCODE by stating that "at least 20% of the genome can influence gene expression" (Van Noorden 2012). *Science* stuck to its maximalist guns, and its summary of 2012 repeated the claim that the "functional portion" of the human genome equals 80% (Annonymous 2012). Unfortunately, neither 80% nor 20% are based on actual evidence.

The ENCODE Incongruity

Armed with the proper concept of function, one can derive expectations concerning the rates and patterns of evolution of functional and nonfunctional parts of the genome. The surest indicator of the existence of a genomic function is that losing it has some phenotypic consequence for the organism. Countless natural experiments testing the functionality of every region of the human genome through mutation have taken place over millions of years of evolution in our ancestors and close relatives. Since most mutations in functional regions are likely to impair the function, these mutations will tend

to be eliminated by natural selection. Thus, functional regions of the genome should evolve more slowly, and therefore be more conserved among species, than nonfunctional ones. The vast majority of comparative genomic studies suggest that less than 15% of the genome is functional according to the evolutionary conservation criterion (Smith et al. 2004, Meader et al. 2010, Ponting and Hardison 2011), with the most comprehensive study to date suggesting a value of approximately 5% (Lindblad-Toh et al. 2011). Ward and Kellis (2012) confirmed that ~5% of the genome is interspecifically conserved, and by using intraspecific variation, found evidence of lineage-specific constraint suggesting that an additional 4% of the human genome is under selection (i.e., functional), bringing the total fraction of the genome that is certain to be functional to approximately 9%. The journal *Science* used this value to proclaim "No More Junk DNA" Hurtley 2012), thus, in effect rounding up 9% to 100%.

In 2007, an ENCODE pilot-phase publication (Birney et al. 2007) estimated that 60% of the genome is functional. The 2012 ENCODE estimate (The ENCODE Project Consortium 2012) of 80.4% represents a significant increase over the 2007 value. Of course, neither estimate is congruent with estimates based on evolutionary conservation. With apologies to the late Robert Ludlum, we shall refer to the difference between the fraction of the genome claimed by ENCODE to be functional (> 80%) and the fraction of the genome under selection (2–15%) as the "ENCODE Incongruity." The ENCODE Incongruity implies that a biological function can be maintained without selection, which in turn implies that no deleterious mutations can occur in those genomic sequences described by ENCODE as functional. Curiously, Ward and Kellis, who estimated that

only about 9% of the genome is under selection (Smith et al. 2004), themselves embody this incongruity, as they are coauthors of the principal publication of the ENCODE Consortium (The ENCODE Project Consortium 2012).

Revisiting Five ENCODE "Functions" and the Statistical Transgressions Committed for the Purpose of Inflating their Genomic Pervasiveness

According to ENCODE, 74.7% of the genome is transcribed, 56.1% is associated with modified histones, 15.2% is found in open-chromatin areas, 8.5% binds transcription factors, and 4.6% consists of methylated CpG dinucleotides. Below we discuss the validity of each of these "functions." We decided to ignore some the ENCODE functions, especially those for which the quantitative data are difficult to obtain. For example, we do not know what proportion of the human genome is involved in chromatin interactions. All we know is that the vast majority of interacting sites (98%) are intrachromosomal, that the distance between interacting sites ranges between 10^5 and 10^7 nucleotides, and that the vast majority of interactions cannot be explained by a commonality of function (The ENCODE Project Consortium 2012).

In our evaluation of the properties deemed functional by ENCODE, we pay special attention to the means by which the genomic pervasiveness of functional DNA was inflated. We identified three main statistical infractions. ENCODE used methodologies encouraging biased errors in favor of inflating estimates of functionality, it consistently

and excessively favored sensitivity over specificity, and it paid unwarranted attention to statistical significance, rather than to the magnitude of the effect.

Transcription does not equal function

The ENCODE Project Consortium systematically catalogued every transcribed piece of DNA as functional. In real life, whether or not a transcript has a function depends on many additional factors. For example, ENCODE ignores the fact that transcription is fundamentally a stochastic process (Raj and van Oudenaarden 2008). Some studies even indicate that 90% of the transcripts generated by RNA polymerase II may represent transcriptional noise (Struhl 2007). In fact, many transcripts generated by transcriptional noise exhibit extensive association with ribosomes and some are even translated (Wilson and Masel 2011).

We note that ENCODE used almost exclusively pluripotent stem cells and cancer cells, which are known as transcriptionally permissive environments. In these cells, the components of the Pol II enzyme complex can increase up to 1000-fold, allowing for high transcription levels from non-promoter and weak promoter sequences. In other words, in these cells transcription of nonfunctional sequences, i.e., DNA sequences that lack a *bona fide* promoter, occurs at high rates (Marques et al. 2005; Babushok et al. 2007). The use of HeLa cells is particularly suspect, as these cells are not representative of human cells, and have even been defined as an independent biological species (*Helacyton gartleri*) (Van Valen and Maiorana 1991). In the following, we describe three

classes of sequences that are known to be abundantly transcribed, but are typically devoid of function: pseudogenes, introns, and mobile elements.

The human genome is rife with dead copies of protein-coding and RNA-specifying genes that have been rendered inactive by mutation. These elements are called pseudogenes (Karro et al. 2007). Pseudogenes come in many flavors (e.g., processed, duplicated, unitary) and, by definition, they are nonfunctional. The measly handful of "pseudogenes" that have so far been assigned a tentative function (e.g., Sassi et al. 2007; Chan et al. 2013) are, by definition, functional genes, merely pseudogene look-alikes. Up to a tenth of all known pseudogenes are transcribed (Pei et al. 2012); some are even translated in tumor cells (e.g., Kandouz et al. 2004). Pseudogene transcription is especially prevalent in pluripotent stem cells, testicular and germline cells, as well as cancer cells such as those used by ENCODE to ascertain transcription (e.g., Babushok et al. 2011). Comparative studies have repeatedly shown that pseudogenes, which have been so defined because they lack coding potential due to the presence of disruptive mutations, evolve very rapidly and are mostly subject to no functional constraint (Pei et al. 2012). Hence, regardless of their transcriptional or translational status, pseudogenes are nonfunctional!

Unfortunately, because "functional genomics" is a recognized discipline within molecular biology, while "non-functional genomics" is only practiced by a handful of "genomic clochards" (Makalowski 2003), pseudogenes have always been looked upon with suspicion and wished away. Gene prediction algorithms, for instance, tend to zombify

pseudogenes *in silico* by annotating many of them as functional genes. In fact, since 2001, estimates of the number of protein-coding genes in the human genome went down considerably, while the number of pseudogenes went up (see below).

When a human protein-coding gene is transcribed, its primary transcript contains not only reading frames, but also introns and exonic sequences devoid of reading frames. In fact, from the ENCODE data, one can see that only 4% of primary mRNA sequences is devoted to the coding of proteins, while the other 96% is mostly made of noncoding regions. Because introns are transcribed, the authors of ENCODE concluded that they are functional. But, are they? Some introns do indeed evolve slightly slower than pseudogenes, although this rate difference can be explained by a minute fraction of intronic sites involved in splicing and other functions. There is a long debate whether or not introns are indispensable components of eukaryotic genome. In one study (Parenteau et al. 2008), 96 introns from 87 yeast genes were knocked out. Only three of them (3%) seemed to have a negative effect on growth. Thus, in the vast majority of cases, introns evolve neutrally, while a small fraction of introns are under selective constraint (Ponjavic et al. 2007). Of course, we recognize that some human introns harbor regulatory sequences (Tishkoff et al. 2006), as well as sequences that produce small RNA molecules (Hirose 2003; Zhou et al. 2004). We note, however, that even those few introns under selection are not constrained over their entire length. Hare and Palumbi (2003) compared nine introns from three mammalian species (whale, seal and human), and found that only about a quarter of their nucleotides exhibit telltale signs of functional constraint. A study of intron 2 of the human *BRCA1* gene, revealed that only 300 bp (3% of the length of the

intron) is conserved (Wardrop et al. 2005). Thus, the practice of ENCODE of summing up all the lengths of all the introns and adding them to the pile marked "functional" is clearly excessive and unwarranted.

The human genome is populated by a very large number of transposable and mobile elements. Transposable elements, such as LINEs, SINEs, retroviruses, and DNA transposons, may, in fact, account for up to two thirds of the human genome (Deininger et al. 2003; Jordan et al. 2003; de Koning et al. 2011) and for more than 31% of the transcriptome (Faulkner et al. 2009). Both human and mouse had been shown to transcribe nonautonomous retrotransposable elements called SINEs (e.g., *Alu* sequences) (Sinnett et al. 1992; Shaikh et al. 1997; Li et al. 1999; Oler et al. 2012). The phenomenon of SINE transcription is particularly evident in carcinoma cell lines, in which multiple copies of *Alu* sequences are detected in the transcriptome (Umylny et al. 2007). Moreover, retrotransposons can initiate transcription on both strands (Denoeud et al. 2007). These transcription initiation sites are subject to almost no evolutionary constraint, casting doubt on their "functionality." Thus, while some transposons have been domesticated into functionality, one cannot assign a "universal utility for retrotransposons." (Faulkner et al. 2009) Whether transcribed or not, the vast majority of transposons in the human genome are merely parasites, parasites of parasites, and dead parasites, whose main "function" would appear to be causing frameshifts in reading frames, disabling RNA-specifying sequences, and simply littering the genome.

Let us now examine the manner in which ENCODE mapped RNA transcripts onto the genome. This will allow us to document another methodological legerdemain used by ENCODE, i.e., the consistent and excessive favoring of sensitivity over specificity. Because of the repetitive nature of the human genome, it is not easy to identify the DNA region from which an RNA is transcribed. The ENCODE authors used a probability based alignment tool to map RNA transcripts onto DNA. Their choice for the type I error, i.e., the probability of incorrect rejection of a true null hypothesis, was 10%. This choice is unusual in biology, although the more common 5%, 1%, and 0.1% are equally arbitrary. How does this choice affect estimates of "transcriptional functionality"? In ENCODE, the transcripts are divided into those that are smaller than 200 bp, and those that are larger than 200 bp. The small transcripts cover only a negligible part of the genome, and in the following they will be ignored. The total number of long RNA transcripts in the ENCODE study is approximately 109 million. The mean transcript length is 564 nucleotides. Thus, a total of 6 billion nucleotides, or two times the size of the human genome, are potentially misplaced. This value represents the maximum error allowed by ENCODE, and the actual error is, of course, much smaller. Unfortunately, ENCODE does not provide us with data on the actual error, so we cannot evaluate their claim. (Of course, nothing is straightforward with ENCODE; there are close to 47 million transcripts shorter than 200 nucleotides in the dataset purportedly composed of transcripts that are longer than 200 nucleotides.) Oddly, in another ENCODE paper, it is indirectly suggested that the 10% type I error may be too stringent, and lowering the threshold "may reveal many additional repeat loci currently missed due to the stringent quality thresholds applied to the data" (Djebali et al. 2012), indicating that increasing the number

of false positives is a worthy pursuit in the eyes of some ENCODE researchers. Of course, judiciously trading off specificity for sensitivity is a standard and sound practice in statistical data analysis, however, by increasing the number of false positives, ENCODE achieves an increase in the total number of test positives, thereby, exaggerating the fraction of "functional" elements within the genome.

At this point, we must ask ourselves, what is the aim of ENCODE: Is it to identify every possible functional element at the expense of increasing the number of elements that are falsely identified as functional? Or is it to create a list of functional elements that is as free of false positives as possible. If the former, then sensitivity should be favored over selectivity; if the latter then selectivity should be favored over sensitivity. ENCODE chose to bias its results by excessively favoring sensitivity over specificity. In fact, they could have saved millions of dollars and many thousands of research hours by ignoring selectivity altogether, and proclaiming a priori that 100% of the genome is functional. Not one functional element would have been missed by using this procedure.

Histone modification does not equal function

The DNA of eukaryotic organisms is packaged into chromatin, whose basic repeating unit is the nucleosome. A nucleosome is formed by wrapping 147 base pairs of DNA around an octamer of four core histones, H2A, H2B, H3, and H4, which are frequently subject to many types of posttranslational covalent modification. Some of these

modifications may alter the chromatin structure and/or function. A recent study looked into the effects of 38 histone modifications on gene expression (Karlić et al. 2010). Specifically, the study looked into how much of the variation in gene expression can be explained by combinations of three different histone modifications. There were 142 combinations of three histone modifications (out of 8,436 possible such combinations) that turned out to yield statistically significant results. In other words, less than 2% of the histone modifications may have something to do with function. The ENCODE study looked into 12 histone modifications, which can yield 220 possible combinations of three modifications. ENCODE does not tell us how many of its histone modifications occur singly, in doublets, or triplets. However, in light of the study by Karlić et al. (2010), it is unlikely that all of them have functional significance.

Interestingly, ENCODE, which is otherwise quite miserly in spelling out the exact function of its "functional" elements, provides putative functions for each of its 12 histone modifications. For example, according to ENCODE, the putative function of the H4K20me1 modification is "preference for 5' end of genes." This is akin to asserting that the function of the White House is to occupy the lot of land at the 1600 block of Pennsylvania Avenue in Washington, D.C.

Open chromatin does not equal function

As a part of the ENCODE project, Song et al. (2011) defined "open chromatin" as genomic regions that are detected by either DNase I or by a method called Formaldehyde

Assisted Isolation of Regulatory Elements (FAIRE). They found that these regions are not bound by histones, i.e., they are nucleosome-depleted. They also found that more than 80% of the transcription start sites were contained within open chromatin regions. In yet another breathtaking example of affirming the consequent, ENCODE makes the reverse claim, and adds all open chromatin regions to the "functional" pile, turning the mostly true statement "most transcription start sites are found within open chromatin regions" into the entirely false statement "most open chromatin regions are functional transcription start sites."

Are open chromatin regions related to transcription? Only 30% of open chromatin regions shared by all cell types are even in the neighborhood of transcription start sites, and in cell-type-specific open chromatin, the proportion is even smaller (Song et al. 2011). The ENCODE authors most probably smelled a rat and, thus, come up with the suggestion that open chromatin sites may be "insulators." However, the deletion of two out of three such "insulators" did not eliminate the insulator activity (Oler et al. 2012).

Transcription-factor binding does not equal function

The identification of transcription-factor binding sites can be accomplished through either computational, e.g., searching for motifs (Bulyk 2003; Bryne et al. 2008), or experimental methods, e.g., chromatin immunoprecipitation (Valouev et al. 2008). ENCODE relied mostly on the latter method. We note, however, that transcription-factor binding motifs are usually very short and, hence, transcription-factor binding-look-alike

sequences may arise in the genome by chance. None of the two methods above can detect such instances. Recent studies on functional transcription-factor binding sites have indicated, as expected, that a major telltale of functionality is a high degree of evolutionary conservation (Stone and Wray 2001; Vallania et al. 2009; Wang et al. 2012; Whiteld et al. 2012). Sadly, the authors of ENCODE decided to disregard evolutionary conservation as a criterion for identifying function. Thus, their estimate of 8.5% of the human genome being involved in transcription factor binding must be hugely exaggerated. For starters, any random DNA sequence of sufficient length will contain transcription-factor binding sites. What is the magnitude of the exaggeration? A study by Vallania et al. (2009) may be instructive in this respect. Vallania et al. set up to identify transcription-factor binding sites in the mouse genome by combining computational predictions based on motifs, evolutionary conservation among species, and experimental validation. They concentrated on a single transcription factor, Stat3. By scanning the whole mouse genome sequence, they found 1,355,858 putative binding sites. By considering only sites located up to 10 kb upstream of putative transcription start sites, and by including only sites that were conserved among mouse and at least 2 other vertebrate species, the number was reduced to 4,339 (0.32%). From these 4,339 sites, 14 were tested experimentally; experimental validation was only obtained for 12 (86%) of them. Assuming that Stat3 is a typical transcription factor, by extrapolation, the fraction of the genome dedicated to binding transcription factors may be as low as 0.32% × 86% = 0.28% rather than the 8.5% touted by ENCODE.

But, let us assume that there are no false positives in the ENCODE data. Even then, their estimate of about 280 million nucleotides being dedicated to transcription factor binding cannot be supported. The reason for this statement is that ENCODE identified putative transcription-factor binding sites by using a methodology that encouraged biased errors yielding inflated estimates of functionality. The ENCODE database for transcription-factor binding sites is organized by institution. The mean length of the entries from three of them, University of Chicago, SYDH (Stanford, Yale, University of Southern California, and Harvard) and Hudson Alpha Institute for Biotechnology, are 824, 457 and 535 nucleotides respectively. These mean lengths, which are highly statistically different from one another, are much larger than the actual sizes of all known transcription factor binding sites. So far, the vast majority of known transcription-factor binding sites were found to range in length from 6 to 14 nucleotides (Oliphant et al. 1989; Christy and Nathans 1989; Okkema and Fire 1994; Klemm et al. 1994; Loots and Ovcharenko 2004; Pavesi et al. 2004), which is 1–2 orders of magnitude smaller than the ENCODE estimate. (An exception to the 6-to-14-bp rule is the canonical p53 binding site, which is composed of two decamer half-sites that can be separated by up to 13 bp.) If we take 10 bp as the average length for a transcription factor binding site (Stewart and Plotkin 2012), instead of the ~600 bp used by ENCODE, the 8.5% value may turn out to be ~8.5% × 10/600 = 0.14% or lower, depending on the proportion of false positives in their data. Interestingly, the DNA coverage value obtained by SYDH is ~18%. We were unable to identify the source of the discrepancy between 18% for SYDH versus the pooled value of 8.5%. The discrepancy may be either an actual error, or the pooled analysis may have

used more stringent criteria than the SYDH institutional analysis. At present, the discrepancy must remain one of ENCODE's many unsolved mysteries.

DNA methylation does not equal function

ENCODE determined that almost all CpGs in the genome were methylated in at least one cell type or tissue. Saxonov et al. (2006) studied CpGs at promoter sites of known protein-coding genes and noticed that expression is negatively correlated with the degree of CpG methylation in promoters. Thus, the conclusion of ENCODE was that all methylated sites are "functional." We note, however, that the number of CpGs in the genome is much higher than the number of protein-coding genes. A scan over all human chromosomes (22 autosomes + X + Y) reveals that there are 150,281,981 CpG sites (4.8% of the genome), as opposed to merely 20,476 protein-coding genes in the latest ENSEMBL release (Flicek et al. 2012). The average GC content of the human genome is 41%. Thus, the randomly expected CpG frequency is 8.4%. The actual frequency of CpG dinucleotides in the genome is about half the expected frequency. There are two reasons for the scantiness of CpGs in the genome. First, methylated CpGs readily mutate to non-CpG dinucleotides. Second, by depressing gene expression, CpG dinucleotides are actively selected against from regions of importance in the genome. Thus, what ENCODE should have sought are regions devoid of CpGs rather than regions with CpGs.

According to ENCODE, 96% of all CpGs in the genome are methylated. This observation is not an indication of function, but rather an indication that all CpGs have

the ability to be methylated. This ability is merely a chemical property, not a function. Finally, it is known that CpG methylation is independent of sequence context (Meissner et al. 2008), and that the pattern of CpG methylation in cancer cells is completely different from that in normal cells (Lodygin et al. 2008; Fernandez et al. 2012), which may render the entire ENCODE edifice on the issue of methylation entirely irrelevant.

Does the frequency distribution of primate-specific derived alleles provide evidence for purifying selection?

In this section, we discuss the purported evidence for purifying selection on ENCODE elements. The ENCODE authors compared the frequency distribution of 205,395 derived ENCODE single-nucleotide polymorphisms (SNPs) and 85,155 derived non-ENCODE SNPs and found that ENCODE-annotated SNPs exhibit "depressed derived allele frequencies consistent with recent negative selection." Here we examine in detail the purported evidence for selection. Of course, it is not possible to enumerate all the methodological errors in this analysis; some errors, such as disregarding the underlying phylogeney for the 60 human genomes, and treating them as independently derived, will not be commented upon.

Given that the number of SNPs in the human population exceeds 10 million, one might wonder why so few SNPs were used in the ENCODE study. The reason lies in the choice of "primate-specific regions" and the manner in which the data were "cleaned." Based on a three-step so-called EPO alignment (Paten et al. 2008a, b) of 11 mammalian species,

the authors identified 1.3 million primate-specific regions that were at least 200 bp in length. The term "primate specific" refers to the fact that these sequences were not found in mouse, rat, rabbit, horse, pig and cow. By choosing primate specific regions only, ENCODE effectively removed everything that is of interest functionally (e.g., protein coding and RNA-specifying genes as well as evolutionarily conserved regulatory regions). What was left consisted among others of dead transposable and retrotransposable elements, such as a TcMAR-Tigger DNA transposon fossil (Smit and Riggs 1996) and a dead *Alu*Sx, both on chromosome 19.

Interestingly, out of three ethnic sample that were available to the ENCODE researchers (59 Yorubans, 60 East Asians from Beijing and Tokyo, and 60 Utah residents of Northern and Western European ancestry), only the Yoruba sample was used. Because polymorphic sites were defined by using all three human samples, the removal of two samples had the unfortunate effect of turning some polymorphic sites into monomorphic ones. As a consequence, the ENCODE data includes 2,136 alleles each with a frequency of exactly 0. In a miraculous feat of "next generation" science, the ENCODE authors were able to determine the frequencies of nonexistent derived alleles.

The primate-specific regions were then masked by excluding repeats, CpG islands, CG dinucleotide and any other regions not included in the EPO alignment block or in the human genome. After the masking, the actual primate specific segments that were left for analysis were extremely small. Eighty-two percent of the segments were smaller than 100 bp and the median was 15 bp. Thus, the ENCODE authors would like us to believe that

inferences that based in part on ~85,000 alignment blocks of size 1 and ~76,000 alignment blocks of size 2 bp are reliable!

The primate-specific segments were, then, divided into segments containing ENCODE-annotated sequences and "controls." There are three interesting facts that were not commented upon by ENCODE: (1) the ENCODE-annotated sequences are much shorter than the controls, (2) some segments contain both ENCODE and non-ENCODE elements, and (3) 15% of all SNPs in the ENCODE-annotated sequences and 17% of the SNPs in the control segments are located within regions defined as short repeats, or repetitive elements, or nested repeat elements. Be that as it may, the ENCODE-containing sample had on average a frequency that was lower by 0.20% than that of the derived alleles in the control region. Of course, with such huge sample sizes, the difference turned out to be highly significant statistically (Kolmogorov–Smirnov test, $p = 4 \times 10^{-37}$).

Is this statistically significant difference important from a biological point of view? First, with very large numbers of loci, one can easily obtain statistically significant differences. Second, the statistical tests employed by ENCODE let us believe that the possibility of linkage disequilibrium may not have been taken into account. That is, it is not clear to us whether the test took into account the fact that many of the allele frequencies are not independent because the alleles occupy loci in very close proximity to one another. Finally, the extensive overlap between the two distributions (approximately 99.958% by

our calculations) indicates that the difference between the two distributions is likely too small to be biologically meaningful.

Can the shape of the derived allele frequency distribution be used as a test for selection? That is, is the excess of extremely rare alleles (private alleles) necessarily indicative of selection? Actually, such a distribution may also be due to bottlenecks (Nei et al. 1975), demographic effects, especially rapid population growth (Slatkin and Hudson 1991; Williamson et al. 2005), background selection (Charlesworth et al. 1993; Kaiser and Charlesworth 2009), or sequencing errors (MacArthur et al. 2012; Achaz 2008; Knudsen and Miyamoto 2009). Before claiming evidence for selection, ENCODE needs to refute these causes.

"Junk DNA is Dead! Long Live Junk DNA"

If there was a single succinct take-home message of the ENCODE consortium, it was the battle cry "Junk DNA is Dead!" Actually a surprisingly large number of scientists have had their knickers in a twist over "junk DNA" ever since the term was coined by Susumu Ohno (1972). The dislike for the term became more evident following the "disappointing" finding that protein-coding genes occupy only a minuscule fraction of the human genome. Before the unveiling of the sequence of the human genome in 2001, learned estimates of human protein-coding gene number ranged from 50,000 to more than 140,000 (Roest Crollius et al. 2000), while estimates in GeneSweep, an informal betting contest started at Cold Spring Harbor Laboratory in 2000, in which scientists

attempted to guess how many protein-coding sequences it takes to make a human (Pennisi 2003), reached values as high as 212,278 genes. The number of protein-coding genes went down considerably with the publication of the two draft human genomes. In one publication it was stated that there are "26,588 protein-encoding transcripts for which there was strong corroborating evidence and an additional ~12,000 computationally derived genes with mouse matches or other weak supporting evidence" (Venter et al. 2001). In the other publication, it was stated that there "appear to be about 30,000–40,000 protein-coding genes in the human genome" (International Human Genome Sequencing Consortium 2001). When the "finished" euchromatic sequence of the human genome was published, the number of protein coding-genes went down even further to "20,000–25,000 protein-coding genes" (International Human Genome Sequencing Consortium 2004). This "diminishing" tendency continues to this day, with the number of protein-coding genes in Ensembl being 21,065 and 20,848 on May 2012 and January 2013, respectively.

The paucity of protein-coding genes in the human genome in conjunction with the fact that the "lowly" nematode *Caenorhabditis elegans* turned out to have 20,517 protein-coding genes, resulted in a puzzling situation similar to the C-value paradox (Thomas 1971), whereby the number of genes did not sit well with the perceived complexity of the human organism. Thus, "junk DNA" had to go.

The ENCODE results purported to revolutionize our understanding of the genome by "proving" that DNA hitherto labeled "junk" is in fact functional. The ENCODE position

concerning the nonexistence of "junk DNA" was mainly based on several logical misconceptions and, possibly, a degree of linguistic prudery.

Let us first dispense with the semantic baggage of the term "junk." Some biologists find the term "junk DNA" "derogatory," "disrespectful" even (e.g., Brosius and Gould 1992). In addition, the fact that "junk" is used euphemistically in off-color contexts does not endear it to many biologists.

In dissecting common objections to "junk DNA," we identified several misconceptions, chief among them (1) a lack of knowledge of the original and correct sense of the term, (2) the belief that evolution can always get rid of nonfunctional DNA, and (3) the belief that "future potential" constitutes "a function."

First, we note that Susumu Ohno's original definition of "junk DNA" referred to a genomic segment on which selection does not operate (Ohno 1972). The correct usage implies a genomic segment that has no immediate use, but that might occasionally acquire a useful function in the future. This sense of the word is very similar to the colloquial meaning of "junk," such as when a person mentions a "garage full of junk," in which the implication is that the space is full of useless objects, but that in the future some of them may be useful. Of course, as in the case of the garage full of junk, the vast majority of junk DNA will never acquire a function. This sense of the term "junk DNA" was used by François Jacob in his famous paper "Evolution and Tinkering" (Jacob 1977): "[N]atural selection does not work as an engineer… It works like a tinkerer—a tinkerer

who does not know exactly what he is going to produce but uses whatever he finds around him whether it be pieces of string, fragments of wood, or old cardboards… The tinkerer… manages with odds and ends. What he ultimately produces is generally related to no special project, and it results from a series of contingent events, of all the opportunities he had to enrich his stock with leftovers."

Second, there exists a misconception among functional genomicists that the evolutionary process can produce a genome that is mostly functional. Actually, evolution can only produce a genome devoid of "junk" if and only if the effective population size is huge and the deleterious effects of increasing genome size are considerable (Lynch 2007). In the vast majority of known bacterial species, these two conditions are met; selection against excess genome is extremely efficient due to enormous effective population sizes, and the fact that replication time and, hence, generation time are correlated with genome size. In humans, there seems to be no selection against excess genomic baggage. Our effective population size is pitiful and DNA replication does not correlate with genome size.

Third, numerous researchers use teleological reasoning according to which the function of a stretch of DNA lies in its future potential. Such researchers (e.g., Makalowski 2003; Wen et al. 2012) use the term "junk DNA" to denote a piece of DNA that can never, under any evolutionary circumstance, be useful. Since any piece of DNA may become functional, many are eager to get rid of the term "junk DNA" altogether. This type of

reasoning is false. Of course, pieces of junk DNA may be coopted into function, but that does not mean that they presently are functional.

To deal with the confusion in the literature, we propose to refresh the memory of those objecting to "junk DNA" by repeating a 15-year old terminological distinction made by Sydney Brenner, who astutely differentiated between "junk DNA," one the one hand, and "garbage DNA," on the other: "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 1998).

It has been pointed to us that junk DNA, garbage DNA, and functional DNA may not add up to 100% because some parts of the genome may be functional but not under constraint with respect to nucleotide composition. We tentatively call such genomic segments "indifferent DNA." Indifferent DNA refers to DNA sites that are functional, but show no evidence of selection against point mutations. Deletion of these sites, however, is deleterious, and is subject to purifying selection. Examples of indifferent DNA are spacers and flanking elements whose presence is required but whose sequence is not

important. Another such case is the third position of four-fold redundant codons, which needs to be present to avoid a downstream frameshift.

Large genomes belonging to species with small effective population sizes should contain considerable amounts of junk DNA and possibly even some garbage DNA. The amount of indifferent DNA is not known. Junk DNA and indifferent DNA can persist in the genome for very long periods of evolutionary time; garbage is transient.

We urge biologists not be afraid of junk DNA. The only people that should be afraid are those claiming that natural processes are insufficient to explain life and that evolutionary theory should be supplemented or supplanted by an intelligent designer (e.g., Dembski 1998; Wells 2004). ENCODE's take-home message that everything has a function implies purpose, and purpose is the only thing that evolution cannot provide. Needless to say, in light of our investigation of the ENCODE publication, it is safe to state that the news concerning the death of "junk DNA" have been greatly exaggerated.

"Big Science," "small science," and ENCODE

The Editor-in-Chief of *Science*, Bruce Alberts, has recently expressed concern about the future of "small science," given that ENCODE-style Big Science grabs the headlines that decision makers so dearly love (Alberts 2012). Actually, the main function of Big Science is to generate massive amounts of reliable and easily accessible data. The road from data to wisdom is quite long and convoluted (Royar 1994). Insight, understanding,

and scientific progress are generally achieved by "small science." The Human Genome Project is a marvelous example of "big science," as are the Sloan Digital Sky Survey (Abazajian et al. 2009) and the Tree of Life Web Project (Maddison et al. 2007).

Did ENCODE generate massive amounts of reliable and easily accessible data? Judging by the computer memory it takes to store the data, ENCODE certainly delivered quantitatively. Unfortunately, the ENCODE data are neither easily accessible nor very useful—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. 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, thereby performing a kind of textual hermeneutics on a 3.5-billion-long DNA text. Unfortunately, ENCODE disregarded the rules of scientific interpretation and adopted a position common to many types of theological hermeneutics, whereby every letter in a text is assumed a priori to have a meaning.

So, what have we learned from the efforts of 442 researchers consuming 288 million dollars? According to Eric Lander, a Human Genome Project luminary, ENCODE is the "Google Maps of the human genome" (Durbin et al. 2010). We beg to differ, ENCODE is considerably worse than even Apple Maps. Evolutionary conservation may be frustratingly silent on the nature of the functions it highlights, but progress in understanding the functional significance of DNA sequences can only be achieved by not ignoring evolutionary principles.

High-throughput genomics and the centralization of science funding have enabled Big Science to generate "high-impact false positives" by the truckload (The PLoS Medicine Editors 2005; Platt et al. 2010; Anonymous 2012; MacArthur 2012; Moyer 2012). Those involved in Big Science will do well to remember the depressingly true popular maxim: "If it is too good to be true, it is too good to be true."

We conclude that the ENCODE Consortium has, so far, failed to provide a compelling reason to abandon the prevailing understanding among evolutionary biologists according to which most of the human genome is devoid of function. The ENCODE results were predicted by one of its lead authors to necessitate the rewriting of textbooks (Pennisi 2012). We agree, many textbooks dealing with marketing, mass-media hype, and public relations may well have to be rewritten.

Acknowledgements

We wish to thank seven anonymous and unanonymous reviewers, as well as Benny Chor, Kiyoshi Ezawa, Yuriy Fofanov, T. Ryan Gregory, Wenfu Li, David Penny, and Betsy Salazar for their help. Drs. Lucas Ward, Ewan Birney, and Ian Dunham helped us navigate the numerous formats, files and URLs in ENCODE. Although we have failed to recruit 436 additional coauthors, we would like to express the hope that Jeremy Renner will accept the title role in the cinematic version of the *ENCODE Incongruity*. DG wishes to dedicate this article to his friend and teacher, Prof. David Wool, on the occasion of his

80th birthday. David Wool instilled in all his students a keen eye for distinguishing between statistical and substantive significance.

References

Abazajian KN, et al. 2009. The Seventh Data Release of the Sloan Digital Sky Survey. ApJS 182: 543-558.

Achaz G. 2008. Frequency spectrum neutrality tests: one for all and all for one. Genetics 179:1409-1424.

Alberts B 2012. The end of “small science”? Science 337: 1583.

Anonymous 2012. Error prone. Nature 487: 406.

Anonymous 2012. Genomics beyond genes. Science 338: 1528-1528.

Babushok DV, et al. 2007. A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. Genome Res. 17: 1129-1138.

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

Brosius J, Gould SJ 1992. On "genomenclature": A comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA." Proc. Natl. Acad. Sci. USA 89: 10706-10710.

Brugger P 2001. From haunted brain to haunted science: a cognitive neuroscience view of paranormal and pseudoscientific thought. In: *Hauntings and Poltergeists: Multidisciplinary Perspectives*, edited by J. Houran and R. Lange (North Carolina: McFarland & Company, Inc. Publishers).

Bryne JC, et al. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. 36: D102-D106.

Bulyk ML 2003. Computational prediction of transcription-factor binding site locations. Genome Biol. 5: 201.

Chan WL, et al. 2013. Transcribed pseudogene ψPPM1K generates endogenous siRNA to suppress oncogenic cell growth in hepatocellular carcinoma. Nucleic Acids Res. [Feb 1, Epub ahead of print]

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134, 1289-1303.

Christy B, Nathans D 1989. DNA binding site of the growth factor-inducible protein Zif268. Proc. Natl. Acad. Sci. USA 86: 8737-8741.

Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr 2003. Mobile elements and mammalian genome evolution. Curr. Opin. Genet. Dev. 13, 651-658.

Cummins R 1975. Functional analysis. J. Philos. 72: 741-765.

Dembski WA 1998. Reinstating design within science. Rhetoric Pub. Affairs 1: 503-518.

Denoeud F, et al. 2007. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res. 17: 746-759.

Djebali S, et al. 2012. Landscape of transcription in human cells. Nature 489: 101-108.

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

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

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

Faulkner GJ, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. Nature Genet. 41: 563-71.

Fernandez AF, Huidobro C, Fraga MF 2012. De novo DNA methyltransferases: oncogenes, tumor suppressors, or both? Trends Genet. 28: 474-478.

Flicek P, et al. 2012. Ensembl 2012. Nucleic Acids Res. 40: D84-D90.

Fyfe S, Williams C, Mason OJ, Pickup GJ 2008. Apophenia, theory of mind and schizotypy: Perceiving meaning and intentionality in randomness. Cortex 44: 1316-1325.

Gregory TR. 2012. ENCODE spokesperson: 40%, not 80%. http://www.genomicron.evolverzone.com/2012/09/encode-spokesperson-40-not-80/

Griffiths PE 2009. In what sense does "nothing in biology make sense except in the light of evolution"? Acta Biotheoretica 57: 11-32.

Hall SS 2012. Journey to the genetic interior. Sci. Am. 307(4): 80-84.

Hare MP, Palumbi SR 2003. High Intron Sequence Conservation Across Three Mammalian Orders Suggests Functional Constraints. Mol. Biol. Evol. 20: 969-978.

Hazkani-Covo E, Zeller RM, Martin W. 2010. Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. 6:e1000834.

Hirose T, Shu MD, Steitz JA 2003. Splicing-Dependent and -Independent Modes of Assembly for Intron-Encoded Box C/D snoRNPs in Mammalian Cells. Mol. Cell 12: 113-123.

Hurtley S 2012. No more junk DNA. Science 337: 1581.

International HapMap 3 Consortium 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. Nature 409: 860-921.

International Human Genome Sequencing Consortium 2004. Finishing the euchromatic sequence of the human genome. Nature 431: 931-945.

Jacob F 1977. Evolution and tinkering. Science 196: 1161-1166.

Jordan IK, Rogozin IB, Glazko GV, Koonin EV 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 19: 68-72.

Kaiser VB, Charlesworth B. 2009. The effects of deleterious mutations on evolution in non-recombining genomes. Trends Genet. 25:9–12.

Kandouz M, Bier A, Carystinos GD, Alaoui-Jamali MA, Batist G 2004. Connexin43 pseudogene is expressed in tumor cells and inhibits growth. Oncogene. 23: 4763-4770.

Karlić R, Chung HR, Lasserre J, Vlahovicek K, Vingron M 2010. Histone modification levels are predictive for gene expression. Proc. Natl. Acad. Sci. USA 107: 2926-2931.

Karro JE, et al. 2007. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res. 35: D55-60.

Klemm JD, Rould MA, Aurora R, Herr W, Pabo CO 1994. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA binding modules. Cell 77: 21-32.

Knudsen B, Miyamoto MM. 2009. Accurate and fast methods to estimate the population mutation rate from error prone sequences. BMC Bioinformatics 10:247.

Knudson AG. 1979. Our load of mutations and its burden of disease. Am. J. Hum. Genet. 31:401-413.

Kolata G. 2012. Bits of Mystery DNA, Far From 'Junk,' Play Crucial Role. http://www.nytimes.com/2012/09/06/science/far-from-junk-dna-dark-matter-proves-crucial-to-health.html?pagewanted=all&_r=0

de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7: e1002384.

Lee RC, Feinbaum RL, Ambros V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854.

Li T, Spearow J, Rubin CM, Schmid CW 1999. Physiological stresses increase mouse short interspersed element (SINE) RNA expression in vivo. Gene 239: 367-372.

Lindblad-Toh K, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476-482.

Lodygin D, et al. 2008. Inactivation of miR-34a by aberrant CpG methylation in multiple types of cancer. Cell Cycle 7: 2591-2600.

Loots GG, Ovcharenko I 2004. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Res. 32: W217-W221.

Lynch M. 2007. The Origins of Genome Architecture. Sinauer Associates, Sunderland.

MacArthur DG, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335: 823-828.

Maddison DR, Schulz K-S, Maddison WP 2007. The Tree of Life web project. Zootaxa 1668:19–40.

Makalowski W 2003. Not junk after all. Science 300: 1246-1247.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3: e357.

Meader S, Ponting CP, Lunter G 2010. Massive turnover of functional sequence in human and other mammalian genomes. Genome Res. 20: 1335-1343.

Meissner A, et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454: 766-770.

Millikan RG 1989. In defense of proper functions. Philos. Sci. 56: 288-302.

Moyer AM 2012. Handling false positives in the genomic era. Clin. Chem. 58: 1605-1606.

Neander K 1991. Functions as selected effects: the conceptual analyst's defense. Philos. Sci. 58: 168-184.

Nei M, Maruyama T, Chakraborty R 1975. The bottleneck effect and genetic variability in populations. Evolution 29: 1-10.

Ohno S 1972. So much "junk" DNA in our genome. Brookhaven Symp Biol. 23: 366-70.

Okkema PG, Fire A 1994. The Caenorhabditis elegans NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. Development 120: 2175-2186.

Oler AJ, et al. 2012. Alu expression in human cell lines and their retrotranspositional potential. Mobile DNA 3: 11.

Oliphant AR, Brandl CJ, Struhl K 1989. Defining the Sequence Specificity of DNA-Binding Proteins by Selecting Binding Sites from Random-Sequence Oligonucleotides: Analysis of Yeast GCN4 Protein. Mol. Cel. Biol. 9: 2944-2949.

Parenteau J, et al. 2008. Deletion of many yeast introns reveals a minority of genes that require splicing for function. Mol. Biol. Cell 19: 1932-1941.

Paten B, Herrero J, Beal K, Fitzgerald S, Birney E 2008a. Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 18: 1814-1828.

Paten B, et al. 2008b. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. 18: 1829-1843.

Pavesi G, Mereghetti P, Mauri G, Pesole G 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 32: W199-W203.

Pei B, et al. 2012. The GENCODE pseudogene resource. Genome Biol. 13: R51.

Pennisi E 2003. A low number wins the GeneSweep pool. Science 300: 1484.

Pennisi E 2012. Genomics' big talker. Science 337: 1167-1169.

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

Platt A, Vilhjálmsson BJ, Nordborg M 2010. Conditions under which genome-wide association studies will be positively misleading. Genetics 186: 1045-1052.

The PLoS Medicine Editors 2005. Why bigger is not yet better: the problems with huge datasets. PLoS Med. 2(2): e55.

Ponjavic J, Ponting CP, Lunter G 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 17: 556-565.

Ponting CP, Hardison RC 2011. What fraction of the human genome is functional? Genome Res. 21: 1769-1776.

Raj A, van Oudenaarden A 2008. Stochastic gene expression and its consequences. Cell 135: 216-226.

Royar R 1994. New horizons, clouded vistas. Comput. Compos. 11: 93-105.

Sassi SO, Braun EL, Benner SA. 2007. The evolution of seminal ribonuclease: pseudogene reactivation or multiple gene inactivation events? Mol. Biol. Evol. 24:1012–1024.

Saxonov S, Berg P, Brutlag DL 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. USA 103: 1412-1417.

Schinner A. 2007. The Voynich manuscript: evidence of the hoax hypothesis. Cryptologia 31: 95-107.

Shaikh TH, Roy AM, Kim J, Batzer MA, Deininger PL 1997. cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts. J. Mol. Biol. 271: 222-234.

Sinnett D, Richer C, Deragon JM, Labuda D 1992. Alu RNA transcripts in human embryonal carcinoma cells: Model of post-transcriptional selection of master sequences. J. Mol. Biol. 226: 689-706.

Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562.

Smit AF, Riggs AD 1996. Tiggers and DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA 93: 1443-1448.

Smith NG, Brandström M, Ellegren H 2004. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. Genomics 84, 806-813.

Song L, et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 21: 1757-1767.

Stamatoyannopoulos JA 2012. What does our genome encode? Genome Res. 22: 1602-11.

Stewart AJ, Plotkin JB. 2012. Why transcription factor binding sites are ten nucleotides long. Genetics 192: 973–985.

Stone JR, Wray GA 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Mol. Biol. Evol. 18: 1764-1770.

Struhl K 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nature Struct. Mol. Biol. 14,103-105.

Thomas CA. 1971. The genetic organization of chromosomes. Annu. Rev. Genet. 5:237-256.

Tishkoff SA, et al. 2006. Convergent adaptation of human lactase persistence in Africa and Europe. Nature Genetics 39: 31-40.

Umylny B, Presting G, Ward WS 2007. Evidence of Alu and B1 expression in dbEST. Systems Biology in Reproductive Medicine 53: 207-218.

Vallania F, et al. 2009. Genome-wide discovery of functional transcription factor binding sites by comparative genomics: The case of Stat3. Proc. Natl. Acad. Sci. USA 106: 5117-5122.

Valouev A, et al. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nature Methods 5: 829-834.

Van Noorden 2012. 2012 in review. Science 492: 324-327.

Van Valen LM, Maiorana VC 1991. Hela, a new microbial species. Evol. Theory 10: 71-74.

Venter JC, et al. 2001. The sequence of the human genome. Science 291: 1304-1351.

Wang J, et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 22: 1798-1812.

Ward LD, Kellis M 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337: 1675-1678.

Wardrop SL, Brown MA; kConFab Investigators 2005. Identification of two evolutionarily conserved and functional regulatory elements in intron 2 of the human BRCA1 gene. Genomics 86: 316-328.

Wells J 2004. Using intelligent design theory to guide scientific research. Prog. Complex. Inf. Design 3.1.2

Wen Y-Z, Zheng L-L, Qu L-H, Ayala FJ, Lun Z-R 2012. Pseudogenes are not pseudo any more. RNA Biol. 9: 27-32.

Whitfield TW, et al. 2012. Functional analysis of transcription factor binding sites in human promoters. Genome Biol. 13: R50.

Williamson SH, et al. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA 102:7882–7887.

Wilson BA, Masel J 2011. Putatively noncoding transcripts show extensive association with ribiosomes. Genome Biol. Evol. 3: 1245-1252.

Witztum D, Rips E, Rosenberg Y 1994. Equidistant letter sequences in the Book of Genesis. Stat. Sci. 9: 429-438.

Zappa F 1979. “Packard Goose,” Joe's Garage: Acts II & III, FZ Records.

Zhou H, et al. 2004. Identification of a novel box C/D snoRNA from mouse nucleolar cDNA library. Gene 327: 99-105.