

BOOK REVIEW

Open Access



A deeper confusion

Georgi K. Marinov* 

Book details

The Deeper Genome: Why there is more to the human genome than meets the eye, edited by John Parrington, (Oxford, United Kingdom: Oxford University Press), 2015. pp. xx + 272. ISBN:978-0-19-968873-9. H/c \$22.61.

Junk DNA: A Journey Through the Dark Matter of the Genome, edited by Nessa Carey, (New York, United States: Columbia University Press), 2015. pp. xx + 360 + 61 b&w illustrations. ISBN:978-0-23-117084-0. H/c \$29.95.

Keywords: Junk DNA, Epigenetics, Genomics, Evolutionary theory, Population genetics

There is an old 1967 Bulgarian movie titled *Whale*, largely unknown in the West, but a cult classic in its country of origin in part due to the censorship it was subjected to by the authorities at the time. Its storyline is both simple and absurd, but it was still a bit too much of an accurate satire of day-to-day reality for the censors, and as a result for many years the movie received only very limited screening. It starts on board of a small fishing vessel in the Black Sea, which has for many days not been able to catch even a single fish. Finally, they “succeed” by finding one tiny sprat in their nets. Having a plan to fulfill, they report using the radio that they have caught 30kg of mackerel, hoping that they will have more luck with their catch on their way back to the shore. However, long before their arrival, the news has traveled further up the bureaucratic hierarchy, and in the process changed quite significantly. Their immediate superior, who also has a quota to meet and is very far from reaching it, reports to his boss that the catch is in fact 300kg of an even larger species. At the next level, the exaggeration continues, and the catch is now a dolphin, and finally the dolphin ends up being reported as a whale (even though no whales have ever been observed in the Black Sea in historic times).

Reading *The Deeper Genome: Why there is more to the human genome than meets the eye*, by John Parrington, and *Junk DNA: A Journey Through the Dark Matter of the Genome*, by Nessa Carey, I could not help but be reminded of that movie. Readers well versed in the

details of the controversy over the extent of functionality of the human genome might also find many parallels between its story and the way what we truly know about genome biology has traveled (and transformed along the way) from experimental tubes, FASTQ files, and white boards, filled with equations, to the pages of the popular books on the subject that are now coming out, similarities which extend to the institutional and social factors that drive these distortions.

These are the first books to attempt filling the niche of communicating the findings of modern functional genomics to a wider audience. They differ greatly in the technical level and quality of the exposition, but the main message is the same—the idea that most of the human genome consists of “junk DNA” has been overturned by recent discoveries, with large-scale functional genomics efforts such as the ENCODE Project Consortium sealing the case (the ENCODE Project was set up with the goal of cataloging all candidate functional elements in the human genome, using a combination of functional genomics assays measuring transcript production, regulatory protein occupancy and chromatin structure; it identified reproducible biochemical activity over the majority of the genome, discussed in detail below, which was taken by many to mean that junk DNA theory has been overturned). However, that is a conclusion that is not supported by the classical theory on the subject, by the biological discoveries of recent decades (many of which are in fact largely irrelevant to that question), and by the functional genomics data that has generated so much excitement in the last few years. If taken

*Correspondence: GKM359@gmail.com
Department of Biology, Indiana University, 47405 Bloomington, IN, USA

uncritically, these texts can be expected to generate even more confusion in a field that already has a serious problem when it comes to communicating the best understanding of the science to the public.

They will also certainly provide ammunition for intelligent design proponents and other creationists. The debunking of junk DNA and the quest to find function for the whole of the human genome have constituted major focus points for such groups in their crusade against evolution (Wells 2011; Tompkins 2012; Wells 2013)—it is assumed (justifiably or not) that a creator would not design genomes full of “junk”, therefore any scientific result that seems to show that more of the genome is functional than previously thought is warmly embraced by them as evidence against junk DNA theory as a whole.

The deeper genome: why there is more to the human genome than meets the eye

John Parrington is an associate professor in the Department of Pharmacology at the University of Oxford. However, as he details in the introduction section of the book, he writes *The Deeper Genome* from the perspective of a science journalist rather than someone who is intimately familiar with the field, and his interest in the subject was in large part sparked by the announcement of the results from the second phase of the ENCODE Project on September 5th, 2012, when he was spending time at *The Times* in London on a British Science Association Media Fellowship. The introduction also states the main message of the book:

So while the original Human Genome Project provided the sequence of letters that make up the DNA code, ENCODE appeared to have gone substantially further and told us what all these different letters actually do. Perhaps most exciting was its claim to have solved one of the biggest conundrums in biology: this is the fact that our genes, which supposedly define us as a species, but also distinguish you or I or anyone else on the planet from each other, make up only 2 % of our DNA. The other 98 per cent had been written off as “junk”; however, this raised the question of why our cells should spend vital energy replicating and storing something with no function. [...] By scanning through the whole genome rather than just the genes, and using multiple, cutting-edge approaches to measure biochemical activity, ENCODE had come to the startling conclusion that, far from being junk, as much as 80 per cent of these disregarded parts of the genome had an important function (pp. 2–3)

This view is defended not just on the basis of the results from the ENCODE Project (to be discussed below). The

author also brings together discoveries from many other hot research areas, which are presented as supporting the idea that the whole genome is functional.

The first two chapters of *The Deeper Genome* take the reader on a brief walk through the history of genetics and genome biology in the last 200 years. Chapter 1, “The Inheritors”, covers the problem of the mechanisms of inheritance that early evolutionary theory faced and that was never resolved by Darwin in his lifetime, through Mendel’s insights into that question and their rediscovery at the turn of the 20th century, and the subsequent establishment of the chromosomal theory of inheritance and modern genetics. Chapter 2, “Life as a Code”, goes over the discovery of DNA, the identification of DNA as the carrier of genetic information, the deciphering of its double helix structure, the molecular biology revolution of the 1950s and 1960s and the establishment of the Central Dogma (Crick 1958).

Chapter 3, “Switches and Signals”, presents the early history of regulatory biology, focusing on the pioneering work of Jacob and Monod, and ending with a brief discussion of chromatin and epigenetic marks. Chapter 4, “The Spacious Genome”, continues with the 1970s discoveries of enhancers, introns and splicing, before going into a discussion of the C-value (the discrepancy between genome size and the perceived complexity of organisms; Thomas 1971) and g-value (the discrepancy between the number of genes and organismal complexity; Hahn and Wray 2002) paradoxes, junk DNA and the development of some of the classic explanations for their existence.

Up to the junk DNA part, these first chapters are an enjoyable read and will be useful to general readers and possibly even working biologists, given that, as with most scientific disciplines, the teaching of biology up to and including the graduate level rarely includes the development of good understanding of the intellectual history of the field, a gap that students are left to fill on their own if they are so inclined. However, the junk DNA section marks the beginning of the problems that plague the rest of the book, specifically the imprecise presentation of facts and the omission of important and powerful arguments to the contrary of the position defended in the text. For example, Parrington states that:

In this sense, what more perfect demonstration is there that nature is “an excellent tinkerer, not a divine artificer”, than the fact that 98 per cent of our own genome is useless? [...] This is a powerful argument, and one that I have much sympathy with, guided as I am by the principle that both life and the universe can be explained by purely materialist principles. However, using the uselessness of so much of the genome for such a purpose is also

risky, for what if the so-called junk turns out to have an important function, but one that hasn't yet been identified? Whether such important functions exist within non-coding DNA has been one of the most hotly debated topics in genetics over the last few years (p. 72).

However, no knowledgeable person has ever defended the position that 98 % of the human genome is useless. The 98 % figure corresponds to the fraction of it that lies outside of protein coding genes, but the existence of distal regulatory elements, as nicely narrated by the author himself, has been at this point in time known for four decades, and there have been numerous comparative genomics studies pointing to a several-fold larger than 2 % fraction of the genome that is under selective constraint (Siepel et al. 2005; Lindblad-Toh et al. 2011; Davydov et al. 2010; Meader et al. 2010), largely lying in noncoding areas. Thus there is (and there has been) no real debate regarding whether noncoding DNA can have important functions—it absolutely does, this is well known, and it is misleading to state otherwise, let alone later use that as an argument in favor of the functionality of the whole genome.

Chapter 5, “RNA Out of the Shadows”, explores the wide variety of roles that noncoding RNA plays in cells, from ribozymes (and the RNA world hypothesis), to small RNAs and RNA interference, and finally, lincRNAs (long intergenic noncoding RNAs). While one of the purposes of the chapter is to use the multitude of noncoding RNAs to support the functionality of most of the genome, it actually underestimates their diversity; it also gets some of its facts wrong:

Currently, there are four known classes of non-coding RNAs, although each class almost certainly include many subclasses. First, there are the siRNAs, which, as we've just discussed, regulate gene expression by destroying their target mRNAs. The second class are known as microRNAs, or miRNAs for short. [...] Third, there are the piRNAs [...] The fourth class are the long non-coding RNAs, or lncRNAs [sic]. These are defined mainly by length, all being over two hundred bases long, in contrast to the other three classes which are typically much smaller, at around twenty bases (pp. 83–84).

Of course, in reality there are many more functional types of noncoding RNAs than just these four—aside for the fundamental for gene expression tRNAs and rRNAs, the snRNAs (small nuclear RNA, components of the spliceosome) and snoRNAs (small nucleolar RNAs that guide the chemical modifications of other RNAs) are also large classes of noncoding RNAs that have been known for

decades. We can then mention the RNA component of the telomerase, the 7SK RNA, the SRP RNAs, Y RNAs, Vault RNAs, RNase P, and numerous others, and this is just within eukaryotes phylogenetically close to humans; prokaryotes have a number of unique to them noncoding RNAs, as do various eukaryote clades. Importantly, many of these RNAs have been known for nearly three decades or more (Walter and Blobel 1982; Brown et al. 1991; Borsani et al. 1991; Greider and Blackburn 1987; Reddy et al. 1984; Lerner et al. 1981; Kedersha and Rome 1986; Blum et al. 1990; Ray and Apirion 1979; Guerrier-Takada et al. 1983), and they occupy only a small fraction of the genome (Kellis et al. 2014)—for example, according to version 19 of the GENCODE annotation of the human genome (Harrow et al. 2012), the exons of lincRNAs cover only 0.2 % of the human genome and miRNAs comprise a minuscule 0.013 % (see Table 1 below), i.e., their existence is hardly grounds for rejecting the notion of junk DNA.

Chapter 6, “It’s a Jungle in There!” is the centerpiece of the book, focusing on the ENCODE Project and its results. Unfortunately, the author derives his information mainly from press releases and interviews and not from the primary literature, which leads him down a path towards some erroneous and not supported by the data conclusions as a result of the compound overhyping of the data at each step of reporting. All of the content of the chapter is based on the 2012 main integration paper of the ENCODE Consortium (ENCODE Project Consortium 2012), and even that primarily comes from writings about the paper rather than the paper itself, while the probably more important with respect to the question of how much of the genome is functional later ENCODE publication (Kellis et al. 2014) is ignored.

Instead of providing an accurate summary of the current understanding of the issue, the book just repeats the claim that ENCODE has found “important function” for basically the whole genome. But this is not really what the ENCODE paper read on its own claims. Here are the key quotes from it (emphasis mine):

Table 1 Fraction of the human genome occupied by the exons of annotated transcripts

Class of transcripts	Fraction of genome (exons) (%)
protein coding (%)	2.95
pseudogenes (%)	0.38
lincRNAs (%)	0.28
miRNAs (%)	0.013
snoRNAs (%)	0.0089
snRNAs (%)	0.0069

These data enabled us to assign biochemical functions for 80 % of the genome, in particular outside of the well-studied protein-coding regions

...

Operationally, we define a functional element as a discrete genome segment that encodes a defined product (for example, protein or non-coding RNA) or displays a reproducible biochemical signature (for example, protein binding, or a specific chromatin structure)

...

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

Given these definitions, and given the limitations imposed by the resolution of the assays used, that 80 % of the genome (which is indeed equivalent to close to 100 % as between 15 and 25 % of it is not uniquely mappable with short sequencing reads and is thus “invisible” in these analyses) is “functional” is indeed correct. But this is only under these particular definitions of function and following the biochemical criterion for functionality, which is not on its own proof of function, much less that it is an “important” one. Here is a quote from Kellis et al. (2014) (the ENCODE publication explicitly dedicated to the question of assessing functionality):

However, biochemical signatures are often a consequence of function, rather than causal. They are also not always deterministic evidence of function, but can occur stochastically. For example, GATA1, whose binding at some erythroid-specific enhancers is critical for function, occupies many other genomic sites that lack detectable enhancer activity or other evidence of biological function (70). Likewise, although enhancers are strongly associated with characteristic histone modifications, the functional significance of such modifications remains unclear, and the mere presence of an enhancer-like signature does not necessarily indicate that a sequence serves a specific function (71, 72). In short, although biochemical signatures are valuable for identifying candidate regulatory elements in the biological context of the cell type examined, they cannot be interpreted as definitive proof of function on their own.

Reminiscent of the ways a mackerel can transform into a cetacean, somewhere along the chain of transmission of information “biochemical function”, operationally defined, transformed into “important function” in the sense in which the term is traditionally understood.

Parrington lists four types of evidence that ENCODE used to “assess function” (the correct term would be “identify candidate functional elements”). The first one is mentioned as “identifying all the places in the genome where transcription factors bind to the DNA”, which presumably refers to transcription factor ChIP-seq (Chromatin Immunoprecipitation coupled with high-throughput sequencing; Johnson et al. 2007). The second involves the mapping of open chromatin (DNase-seq and digital genomic footprinting, or DGF; Thurman et al. 2012; Neph et al. 2012). The third approach mentioned by him is the mapping of DNA methylation. Finally, the transcriptome maps generated using RNA-seq are listed.

The inclusion of DNA methylation in this list is quite puzzling. The main ENCODE integration paper indeed lists DNA methylation as one of the assays used; however, first, it was not applied as a proxy for functionality but for other scientific purposes, second, the particular technique used was Reduced Representation Bisulfite Sequencing (Meissner et al. 2005), which does not give a truly genome-wide measurement of DNA methylation, and third, DNA methylation can hardly be used as a criterion for functionality, because most of the CpG sites in somatic mammalian genomes are usually methylated (Lister et al. 2009), with some important exceptions in regulatory elements and elsewhere (Jones 2012), and because DNA methylation is one of the mechanisms used to silence one of the classic examples of junk DNA, transposable elements (Yoder et al. 1997).

This is not the only problem, as the presentation of the other methods, what they can and cannot, and what they in fact do tell us about how much of the genome is functional, is very incomplete. As a first example, the genome is indeed pervasively transcribed; however, on its own this is an oversimplification that is very far from telling the complete story. Here are some more quotes from Kellis et al. (2014):

In agreement with prior findings of pervasive transcription (85, 86), ENCODE maps of polyadenylated and total RNA cover in total more than 75 % of the genome. These already large fractions may be underestimates, as only a subset of cell states have been assayed. However, for multiple reasons discussed below, it remains unclear what proportion of these biochemically annotated regions serve specific functions

...

For example, RNA transcripts of some kind can be detected from ~75 % of the genome, but a significant portion of these are of low abundance [...]. For

polyadenylated RNA, where it is possible to estimate abundance levels, 70 % of the documented coverage is below approximately one transcript per cell (100–103). The abundance of complex nonpolyadenylated RNAs and RNAs from subcellular fractions, which account for half of the total RNA coverage of the genome, is likely to be even lower, although their absolute quantification is not yet achieved

That a large fraction of the genome is transcribed is not surprising—after all, while annotated exons might occupy only 2 % of it, the introns of those same genes cover a much larger fraction of the genome (Table 2).

This is DNA that is transcribed in order to produce mRNAs, and many of the products of its transcription are present in the various subcellular fractions assayed by ENCODE (in addition to polyadenylated RNA, which is the mature state of mRNAs, ENCODE also analyzed polyA+ and non-polyA transcripts from total cell, cytosolic, nuclear, nucleoplasmic, nucleolar, and chromatin cellular subfractions). But we cannot expect complete absence of transcription outside of annotated genes either. Another quote from Kellis et al. (2014):

At present, we cannot distinguish which low-abundance transcripts are functional, especially for RNAs that lack the defining characteristics of known protein coding, structural, or regulatory RNAs. A priori, we should not expect the transcriptome to consist exclusively of functional RNAs. Zero tolerance for errant transcripts would come at high cost in the proofreading machinery needed to perfectly gate RNA polymerase and splicing activities, or to instantly eliminate spurious transcripts

No serious attention is given in the book to the fact that much of the observed transcription is at low levels, or that, as shown in Kellis et al., the strength of the biochemical signal correlates quite well with evolutionary conservation, i.e. regions of the genome expressed at high levels or more strongly occupied by transcription factors are more likely to be conserved than those with low levels

Table 2 Fraction of the human genome occupied by annotated transcripts (exons + introns)

Class of transcripts	Fraction of genome (exons + introns) (%)
protein coding (%)	40.4
lincRNAs (%)	6.21
pseudogenes (%)	1.59
miRNAs (%)	0.067735089821

of signal, and what all this means for the question of the extent of functionality of the genome (Kellis et al. 2014):

Thus, one should have high confidence that the subset of the genome with large signals for RNA or chromatin signatures coupled with strong conservation is functional and will be supported by appropriate genetic tests. In contrast, the larger proportion of genome with reproducible but low biochemical signal strength and less evolutionary conservation is challenging to parse between specific functions and biological noise.

Another issue that is ignored is the resolution of the assays used and how they contribute to the 80 % number. The biggest contribution to it comes from the transcriptome, but the fraction of the genome occupied by ChIP-seq peaks is also quite large. However (Kellis et al. 2014):

Biochemical methods, such as ChIP or DNase hypersensitivity assays, capture extended regions of several hundred bases, whereas the underlying transcription factor binding elements are typically only 6–15 bp in length

The upward bias on biochemical functionality estimates imposed by technical limitations is even more of an issue with histone marks, where even the resolution of the assay is not that much of a problem as is the fact that a single enhancer or promoter with a limited number of functionally constrained bases pairs can induce changes in the chromatin state of several neighboring nucleosomes.

The best available assay for accurately constraining the size of the whole regulatory lexicon is DGF (digital genome footprinting, which provides “footprints” of the occupancy of transcription factors and other regulatory proteins on DNA thanks to the protection of DNA against DNase digestion that they provide), even if the footprints derived from it are often also slightly extended relative to the actual occupied site. Indeed, a very large number of footprints are identified; however, they only occupy ~10 % of the genome, and the transcription factor binding motifs residing in them cover ~5 %, i.e. a number much smaller than the whole genome or even the majority of it.

Parrington notes that a large fraction of the identified putative regulatory elements show little conservation between human and mouse, as revealed by the parallel mouse ENCODE project (Yue et al. 2014; Cheng et al. 2014; Stergachis et al. 2014). This is indeed a fascinating and very important observation, but its real significance is not that it highlights the uniqueness of humans, as interpreted in the book, but that it actually supports the view that mammalian genomes are shaped in large part

by neutral evolutionary forces (Villar et al. 2014; Marinov 2014).

This is the final major issue with the chapter—the results of the ENCODE Project are presented as rejecting the junk DNA theory, without much real discussion of what that theory is based on and why so many scientists hold it to be true. A brief overview of its main components is in order here:

1. Based on early estimates of the mutation rate in humans and the size of the human genome, simple calculations done decades ago showed that only a small fraction of it could be constrained at the sequence level, otherwise there would be too many deleterious mutations in every generation for the species to survive (Ohno 1972). The estimates of the mutation rate have been revised somewhat since then, and empirical estimates on constraint within the human population have become available too (Ward and Kellis 2012), but this has not resulted in raising the estimate on the fraction of the genome that could be selectively constrained to anything remotely close to the majority of it.
2. The C-value paradox revealed wide disparities between genome sizes in different organisms that are difficult to explain through other means than most of the large genomes being junk. More recently, the “onion test” has been formulated (Gregory 2007), as a means of testing alternative theories for explaining these paradoxes (it consists of asking such proposals to explain why onion needs much more DNA than humans for regulation, structural maintenance, or protection against mutagens, and why some species of onion need 5 times more DNA than other members in the *Allium* genus for the same purposes).
3. The discovery of selfish DNA elements and that they occupy a large fraction of big genomes (Orgel and Crick 1980; Doolittle and Sapienza 1980).
4. The understanding of the limits on the power of natural selection imposed by the population genetic environment of a species, and in particular the role of the effective population size (N_e), which determines the relative influence of selection and drift in the evolution of a lineage. So it happens (partly for obvious ecological reasons having to do with the physical size of organisms) that large-bodied multicellular organisms are among the lineages with the lowest N_e , meaning that the power of natural selection is weakest within their populations, which readily explains many of the nonadaptive or even maladaptive features of their genomes (Lynch and Conery 2003; Lynch 2007a, b), such as the presence of large amounts of junk DNA.

Regrettably, there is no engagement with these arguments in the book, when in fact ENCODE data fits comfortably within that framework, while in the same time providing a much richer understanding of the process of genome evolution. For example, the complexity of the regulatory apparatus in metazoans, and its rapid evolution, as evidenced by its divergence between rodents and primates, revealed by ENCODE and mouse ENCODE and by other studies (Villar et al. 2014) can be understood as a consequence of the low- N_e population genetic environment of these organisms, which facilitates the evolution of new regulatory elements and the complexification of regulatory networks, as many of the intermediates in the process are either effectively neutral or maladaptive (Lynch 2007c) and are not as readily tolerated in organisms with large N_e . These lines of reasoning, and the more general concept of constructive neutral evolution (Stoltzfus 1999, 2012; Lukes et al. 2011), are entirely absent from the book.

Chapter 7, “The Genome in 3D”, moves onto recent work towards characterizing the three-dimensional organization of the genome and its role in gene regulation. This is a very interesting topic and Parrington does a reasonably good job at presenting some of the basics, but it has no real relevance to the question whether there is junk DNA or not—first, if genes and regulatory elements are separated by large physical distances in linear space, it becomes in a way a necessity to have complex gene regulation happening in 3D, and second, that the genome is folded in complex and regulated manner does not in any way imply that all of it is functional—all of that can be accomplished with a small fraction of it serving as anchor points for chromatin looping, nuclear matrix and nuclear lamina attachment.

Chapter 8, “The Jumping Genes”, is dedicated to transposable elements (TEs). The story of their discovery by Barbara McClintock and a brief exposition on the main classes of TEs are followed by examples of both the negative effects they often have on their host and of TEs being exapted into various functions in the cell. Of course, these examples do not mean that all TEs are functional, and fortunately, Parrington does not make that claim.

The chapter on TEs also serves as prelude to Chapter 9, “The Marks of Lamarck”, the main idea of which is that much evidence has accumulated in recent years for Lamarckian evolution being a real phenomenon. This is indeed true (Koonin and Wolf 2009); however, the chapter does not talk about the bona fide examples of Lamarckian evolution such as the CRISPR systems of prokaryotes, but focuses almost entirely on epigenetic inheritance in metazoans, a phenomenon that is fascinating, but still very poorly characterized, and very far from being proven to play a significant role in vertebrate evolution.

Chapter 10, “Code, Non-Code, Garbage, and Junk”, revisits the subject of junk DNA, again presenting numerous appeals for much of it being functional and playing a regulatory role. Not much needs to be said about it, except that it also features a bizarre argument for the functionality of pseudogenes. One would have thought that the ceRNA (competitive endogenous RNA) hypothesis (the idea that noncoding RNAs can regulate the expression of other RNAs by acting as “sponges” for miRNAs; Salmena et al. 2011) would have been used for that purpose even though powerful arguments have been presented against it (Denzler et al. 2014). Instead, Parrrington brings up pseudoenzymes, which are proteins that clearly belong to a family of functional enzymes but lack catalytic activity. However, pseudoenzymes are not at all pseudogenes, as they are normal products of functional genes!

Chapter 11, “Genes and Disease”, is dedicated to the genetic basis of diseases. Somewhat surprisingly, it devotes very little space to discussing the exciting area of research emerging at the intersection of the results from the ENCODE Project and GWAS studies, given that this has been one of the major accomplishments of the former, which is also the inspiration for the book.

Chapter 12, “What Makes Us Human?”, talks about human evolution, in particular in the light of paleogenomics, while Chapter 13, “The Genome That Became Conscious”, discusses the cellular foundations of human brain function, with emphasis on epigenetics and gene regulation.

The concluding chapter, “The Case for Complexity” reiterates how much more complex genome biology is than previously thought. This is indeed the case, and it is also true that the last decade has seen a technological revolution that has allowed us to dig deeper than ever before into it. Popular books accurately conveying that complexity in an accessible manner are much needed. However, *The Deeper Genome* misses the opportunity of being the book that fills that gap, by making the unwarranted conclusion that all of the genome is functional its core message, and overhyping the importance of the findings it discusses. In short, the argument boils down to the following:

1. Junk DNA theory predicts that most of the genome would be completely biochemically inert.
2. Biochemical activity can be equated with function in the traditional sense of the word.
3. Human genome biology is extremely complex and much of the genome shows at least a trace of biochemical activity.
4. Therefore junk DNA theory has been falsified.

However, the premises do not hold—junk DNA theory predicts no such thing, biochemical activity can only identify candidate functional elements, and the complexity of genome biology is not mutually exclusive with most of the genome being junk.

Junk DNA: a journey through the dark matter of the genome

Still, despite a few unfortunate mistakes, *The Deeper Genome* is well written and gets many of its facts right, even if they are not interpreted properly.

This is in stark contrast with Nessa Carey’s *Junk DNA: A Journey Through the Dark Matter of the Genome*. Nessa Carey has a PhD in virology and has in the past been a Senior Lecturer in Molecular Biology at Imperial College, London. However, *Junk DNA* is a book not written at an academic level but instead intended for very broad audience, with all the consequences that the danger of dumbing it down for such a purpose entails. We are hit with the first (and biggest) such problem, at the very beginning, in a brief “Notes on Nomenclature”:

There’s a bit of a linguistic difficulty in writing a book on junk DNA, because it is a constantly shifting term. This is partly because new data change our perception all the time. Consequently, as soon as a piece of junk DNA is shown to have a function, some scientists will say (logically enough) that it’s not junk. But that approach runs the risk of losing perspective on how radically our understanding of the genome has changed in recent years.

Rather than spend time trying to knit a sweater with this ball of fog, I have adopted the most hard-line approach. Anything that doesn’t code for protein will be described as junk, as it originally was in the old days (second half of the twentieth century). Purists will scream, and that’s OK. Ask three different scientists what they mean by the term “junk”, and we would probably get four different answers. So there’s merit in starting with something straightforward. (p. xi)

Purists will indeed scream, and with a good reason. As mentioned above, no knowledgeable scientist has ever thought of all noncoding RNAs or of regulatory and other noncoding elements in the genome as junk, and to dismiss the concept of junk DNA based on such misunderstanding is an egregious fallacy. Much of the book derives from this foundational error, which is reproduced repeatedly throughout it, together with an even more extreme version of the failure to properly acknowledge the theoretical and empirical basis of junk DNA theory

that also plagues *The Deeper Genome*. In the “Introduction” chapter we read that:

For years, scientists had no explanation for why so much of our DNA doesn't code for proteins. These non-coding parts were dismissed with the term “junk DNA”. But gradually this position has begun to look less tenable, for a whole host of reasons. (p. 2)

Of course, scientists have had a very good idea why so much of our DNA does not code for proteins, and they have had that understanding for decades, as outlined above. Only by completely ignoring all that knowledge could it have been possible to produce many of the chapters in the book. The following are referred to as junk DNA by Carey, with whole chapters dedicated to each of them (Table 3).

The inclusion of tRNAs and rRNAs in the list of “previously thought to be junk” DNA is particularly baffling given that they have featured prominently as critical components of the protein synthesis machinery in all sorts of basic high school biology textbooks for decades, not to mention the role that rRNAs and some of the other non-coding RNAs on that list play in many “RNA world” scenarios for the origin of life. How could something that has so often been postulated to predate the origin of DNA as the carrier of genetic information (Jeffares et al. 1998; Fox 2010) and that must have been of critical importance both before and after that be referred to as “junk”?

Chapter 14 is dedicated to the ENCODE Project and the ground for it is prepared in Chapter 3, which discusses the C value and gvalue paradoxes. Unlike Parrrington, who is often very careful to warn against the dangers of phylogenetic chauvinism and the popular assumption that humans are the pinnacle of all of creation, Carey has no such qualms:

Table 3 Well known functional elements considered “junk” in Nessa Carey’s “Junk DNA”

Genomic feature	Chapter
telomeres	5
centromeres	6
The Xist lincRNA	7
lincRNAs	8 and 9
imprinting control elements	10
rRNAs and tRNAs	11
enhancers	12 and 15
insulators	13
5'UTRs and 3'UTRs in mRNAs	16
snRNAs and the spliceosome	17
small RNAs	18

In 2001, amidst all the hoopla, scientists were poring over the data from the human genome sequence and pondering a simple question: where on earth were all the genes? Where were all the sequences to code for the proteins that carry out the functions of cells and individuals? No other species is as complex as humans. No other species builds cities, creates art, grows crops or plays ping-pong. (p. 28)

...

Although startling, these data had been foreshadowed by indirect analyses in the previous decade by scientists trying to understand why humans are so complicated. This was the problem by which so many people had been puzzled ever since the completion of the human genome sequence failed to find a larger number of protein-coding genes in humans than in much simpler organisms. (p. 188)

As is the case with the corresponding parts of *The Deeper Genome*, there is no real discussion of the subtleties of functional genomic data (with the possible exception of a brief section on how to think about the low expression levels of noncoding transcripts), and there is no engagement with actual evolutionary theory, population genetics, and the preexisting body of arguments in favor of most of the genome being junk (in the proper sense of the term).

Instead, the main message is that junk DNA theory has been overturned and that the parts of the genome that used to be thought of as junk are responsible for the marvelous complexity and evolutionary superiority of the human species (never mind all the examples of much “simpler” organisms with more bloated genomes or more complex aspects of their genome biology than ours). All this is presented in overtly simplistic terms, and with a quite heavy-handed emphasis on making the book relevant to the wide public by focusing on the role of real or imaginary “junk DNA” in human disease.

The book also contains several elementary errors that distort the data presented—for example, Fig. 3.1 shows the relative sizes of the human, worm and yeast genomes, and also the numbers of protein coding genes, all of which are represented as circles; however the actual numbers (not shown in the figure) clearly correspond to the radius of the circles and not to their area, something that will be deeply misleading to anyone who has not memorized the true values in the course of working with these genomes for many years.

Given the accessibility of its exposition, *Junk DNA* is a book that will appeal to a wider audience. Much of it, however, is likely to be of intelligent design and creationist bent, and will accordingly use the book as ammunition

against evolutionary theory, rather than as a useful source of information (which by and large it is not anyway) while ignoring the important historical and factual details that the book gets wrong, a most unfortunate and regrettable outcome.

Conclusions

In conclusion, the complexity of genome and regulatory biology is indeed immense, and year after year we are gaining a deeper understanding of it. However, the cases in which these new discoveries have truly radically transformed our view of the genome or have overturned real and imagined “dogmas” of the old are actually very few. In the case of junk DNA, to the extent that function is a binary rather than a continuously distributed property, and that the genome can be neatly divided into “functional” and “nonfunctional” portions, genomics has indeed raised the estimates of how much of the human genome falls into the former category. But it has not raised that estimate to the majority of the sequence, nor has it rejected junk DNA.

The reason why scientific results become so distorted on their way from scientists to the public can only be understood in the socioeconomic context in which science is done today. As almost everyone knows at this point, science has existed in a state of insufficient funding and ever increasing competition for limited resources (positions, funding, and the small number of publishing slots in top scientific journals) for a long time now. The best way to win that Darwinian race is to make a big, paradigm shifting finding. But such discoveries are hard to come by, and in many areas might actually never happen again—nothing guarantees that the fundamental discoveries in a given area have not already been made. It is a situation very reminiscent of the trap in which both the sailors of the fishing ship and their superiors at each level of bureaucratic chain of command above them were finding themselves in *Whale*, and to the real-life conditions it satirized. This naturally leads to a publishing environment that pretty much mandates that findings are framed in the most favorable and exciting way, with important caveats and limitations hidden between the lines or missing completely. The author is too young to have directly experienced those times, but has read quite a few papers in top journals from the 1970s and earlier, and has been repeatedly struck by the difference between the open discussion one can find in many of those old articles and the currently dominant practices.

But that same problem is not limited to science itself, it seems to be now prevalent at all steps in the chain of transmission of findings, from the primary literature, through PR departments and press releases, and finally, in the hands of the science journalists and writers who

report directly to the lay audience, and who operate under similar pressures to produce eye-catching headlines that can grab the fleeting attention of readers with ever decreasing ability to concentrate on complex and subtle issues. This leads to compound overhyping of results, of which *The Deeper Genome* is representative, and to truly surreal distortion of the science, such as what one finds in Nessa Carey's *Junk DNA*.

The field of functional genomics is especially vulnerable to these trends, as it exists in the hard-to-navigate context of very rapid technological changes, a potential for the generation of truly revolutionary medical technologies, and an often difficult interaction with evolutionary biology, a controversial for a significant portion of society topic. It is not a simple subject to understand and communicate given all these complexities while in the same time the potential and incentives to mislead and misinterpret are great, and the consequences of doing so dire. Failure to properly communicate genomic science can lead to a failure to support and develop the medical breakthroughs it promises to deliver, or what might be even worse, to implement them in such a way that some of the dystopian futures imagined by sci-fi authors become reality. In addition, lending support to anti-evolutionary forces in society by distorting the science in a way that makes it appear to undermine evolutionary theory has profound consequences that given the fundamental importance of evolution for the proper understanding of humanity's place in nature go far beyond making life even more difficult for teachers and educators of even the general destruction of science education. Writing on these issues should exercise the needed care and make sure that facts and their best interpretations are accurately reported. Instead, books such as *The Deeper Genome* and *Junk DNA* are prime examples of the negative trends outlined above, and are guaranteed to only generate even deeper confusion.

Acknowledgements

The author would like to thank Glenn Branch for critical reading of the manuscript and helpful suggestions on improving it.

Competing interests

The author declares that he has no competing interests.

Received: 22 October 2015 Accepted: 26 October 2015

Published online: 20 December 2015

References

- Blum B, Bakalara N, Simpson L. A model for RNA editing in kinetoplastid mitochondria: “guide” RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell*. 1990;60:189–98.
- Borsani G, Tonlorenzi R, Simmler MC, Dandolo L, Arnaud D, Capra V, et al. Characterization of a murine gene expressed from the inactive X chromosome. *Nature*. 1991;351:325–9.

- Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, et al. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*. 1991;349:38–44.
- Cheng Y, Ma Z, Kim BH, Wu W, Cayting P, Boyle AP, et al. Principles of regulatory information conservation between mouse and human. *Nature*. 2014;515:371–5.
- Crick FH. On protein synthesis. *Symp Soc Exp Biol*. 1958;12:138–63.
- Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*. 2010;6:e1001025.
- Denzler R, Agarwal V, Stefano J, Bartel DP, Stoffel M. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Mol Cell*. 2014;54:766–76.
- Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. *Nature*. 1980;284:601–3.
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
- Fox GE. Origin and evolution of the ribosome. *Cold Spring Harb Perspect Biol*. 2010;2:a003483.
- Gregory TR. The onion test. 2007. <http://www.genomicron.evolverzone.com/2007/04/onion-test/>.
- Greider CW, Blackburn EH. The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell*. 1987;51:887–98.
- Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*. 1983;35:849–57.
- Hahn MW, Wray GA. The g-value paradox. *Evol Dev*. 2002;4:73–5.
- Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22:1760–74.
- Jeffares DC, Poole AM, Penny D. Relics from the RNA world. *J Mol Evol*. 1998;46:18–36.
- Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science*. 2007;316:1497–502.
- Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*. 2012;13:484–92.
- Kedersha NL, Rome LH. Isolation and characterization of a novel ribonucleoprotein particle: large structures contain a single species of small RNA. *J Cell Biol*. 1986;103:699–709.
- Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci USA*. 2014;111:6131–8.
- King JL, Jukes TH. Non-Darwinian evolution. *Science*. 1969;164:788–97.
- Koonin EV, Wolf YI. Is evolution Darwinian or/and Lamarckian? *Biol Direct*. 2009;4:42.
- Lerner MR, Boyle JA, Hardin JA, Steitz JA. Two novel classes of small ribonucleoproteins detected by antibodies associated with lupus erythematosus. *Science*. 1981;211:400–2.
- Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*. 2011;478:476–82.
- Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009;462:315–22.
- Lukes J, Archibald JM, Keeling PJ, Doolittle WF, Gray MW. How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life*. 2011;63:528–37.
- Lynch M, Conery JS. The origins of genome complexity. *Science*. 2003;302:1401–4.
- Lynch M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci USA*. 2007a;104:8597–604.
- Lynch M. The origins of genome architecture. Sunderland: Sinauer Associates; 2007b.
- Lynch M. The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet*. 2007c;8:803–13.
- Marinov GK. Functional genomic studies of the structure and regulation of eukaryotic transcriptomes. *Dissertation (Ph.D.), California Institute of Technology*; 2014.
- Meador S, Ponting CP, Lunter G. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res*. 2010;20:1335–43.
- Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res*. 2005;33:5868–77.
- Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*. 2012;489:83–90.
- Ohno S. So much “junk” DNA in our genome. *Brookhaven Symp Biol*. 1972;23:366–70.
- Orgel LE, Crick FH. Selfish DNA: the ultimate parasite. *Nature*. 1980;284:604–7.
- Ray BK, Apirion D. Characterization of 10S RNA: a new stable rna molecule from *Escherichia coli*. *Mol Gen Genet*. 1979;174:25–32.
- Reddy R, Henning D, Subrahmanyam CS, Busch H. Primary and secondary structure of 7–3 (K) RNA of Novikoff hepatoma. *J Biol Chem*. 1984;259:12265–70.
- Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*. 2011;146:353–8.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005;15:1034–50.
- Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*. 2014;515:365–70.
- Stoltzfus A. On the possibility of constructive neutral evolution. *J Mol Evol*. 1999;49:69–181.
- Stoltzfus A. Constructive neutral evolution: exploring evolutionary theory’s curious disconnect. *Biol Direct*. 2012;7:35.
- Thomas CA Jr. The genetic organization of chromosomes. *Ann Rev Genetics*. 1971;5:237–56.
- Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. *Nature*. 2012;489:75–82.
- Tompkins JP. Junk DNA myth continues its demise. *Acts Facts*. 2012;41:11–3.
- Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat Rev Genet*. 2014;15:221–33.
- Walter P, Blobel G. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature*. 1982;299:691–8.
- Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*. 2012;337:1675–8.
- Wells J. *The Myth of Junk DNA*. Seattle: Discovery Institute Press; 2011.
- Wells J. Not junk after all: non-protein-coding DNA carries extensive biological information. *Biol Inf N Perspect World*. 2013;210–31.
- Yoder JA, Walsh CP, Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*. 1997;13:335–40.
- Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*. 2014;515:355–64.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com
