Not Junk After All: Non-Protein-Coding DNA Carries Extensive Biological Information

Jonathan Wells

Discovery Institute, 208 Columbia Street, Seattle, WA 98104, USA. jonwells2001@comcast.net

Abstract

In the 1950s Francis Crick formulated the Central Dogma of molecular biology, which states (in effect) that DNA makes RNA makes protein makes us. By 1970, however, biologists knew that the vast majority of our genome does not encode proteins, and the non-protein-coding fraction became known as "junk DNA." Yet data from recent genome projects show that most nuclear DNA is transcribed into RNAs, many of which perform important functions in cells and tissues. Like protein-coding DNA, non-protein-coding regions carry multiple overlapping codes that profoundly affect gene expression and other cellular processes. Although there are still many gaps in our understanding, new functions of non-protein-coding DNA are being reported every month. Clearly, the notion of "junk DNA" is obsolete, and the amount of biological information in the genome far exceeds the information in protein-coding regions.

Key words: Central Dogma, Sequence Hypothesis, junk DNA, non-protein-coding DNA, non-protein-coding RNA, chromatin, centromere, inverted nuclei

1. Introduction

James Watson and Francis Crick's 1953 discovery that DNA consists of two complementary strands suggested a possible copying mechanism for Mendel's genes [1,2]. In 1958, Crick argued that "the main function of the genetic material" is to control the synthesis of proteins. According to the "Sequence Hypothesis," Crick wrote that the specificity of a segment of DNA "is expressed solely by the sequence of bases," and "this sequence is a (simple) code for the amino acid sequence of a particular protein." Crick further proposed that DNA controls protein synthesis through the intermediary of RNA, arguing that "the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid, is impossible." Under some circumstances RNA might transfer sequence information to DNA, but the order of causation is normally "DNA

makes RNA makes protein." Crick called this the "Central Dogma" of molecular biology [3], and it is sometimes stated more generally as "DNA makes RNA makes protein makes us."

The Sequence Hypothesis and the Central Dogma imply that only protein-coding DNA matters to the organism. Yet by 1970 biologists already knew that much of our DNA does not code for proteins. In fact, less than 2% of human DNA is protein-coding. Although some people suggested that non-protein-coding DNA might help to regulate gene expression, the dominant view was that non-protein-coding regions had no function. In 1972, biologist Susumu Ohno published an article wondering why there is "so much 'junk' DNA in our genome" [4].

In 1976, Oxford biologist Richard Dawkins wrote: "The amount of DNA in organisms is more than is strictly necessary for building them: A large fraction of the DNA is never translated into protein. From the point of view of the individual organism this seems paradoxical. If the 'purpose' of DNA is to supervise the building of bodies, it is surprising to find a large quantity of DNA which does no such thing. Biologists are racking their brains trying to think what useful task this apparently surplus DNA is doing. But from the point of view of the selfish genes themselves, there is no paradox. The true 'purpose' of DNA is to survive, no more and no less. The simplest way to explain the surplus DNA is to suppose that it is a parasite, or at best a harmless but useless passenger, hitching a ride in the survival machines created by the other DNA" [5].

If one assumes that only protein-coding regions of DNA matter to the organism, and non-protein-coding DNA is just parasitic junk, it makes sense also to assume that only protein-coding regions would be transcribed into RNA. Why would an organism engaged in a struggle for survival waste precious internal resources on transcribing junk? Yet it turns out that organisms *do* transcribe most of their DNA into RNA — and there is growing evidence that much (perhaps even most) of this RNA performs essential functions in cells and tissues.

2. Widespread Transcription Into RNAs That Are Probably Functional

Even before the Human Genome Project was completed in 2003 [6] there had been reports of the widespread transcription of non-protein-coding DNA. In 2002, the Japanese FANTOM Consortium (for **F**unctional **AN**no**T**ation **O**f the **M**ammalian Genome) identified 11,665 non-protein-coding RNAs in mice and concluded that "non-coding RNA is a major component of the transcriptome" [7]. Other scientists reported that transcription of two human chromosomes resulted in ten times more RNA than could be attributed to protein-coding exons [8].

In 2003, the ENCODE project (for **ENC**yclopedia **O**f **D**NA **E**lements) set out to identify all the functional elements in the human genome. It soon became obvious that most of the mammalian genome is transcribed into RNA [9,10]. Preliminary data provided "convincing evidence that the genome is pervasively transcribed, such that the majority of its bases can be found in primary transcripts, including non-protein-coding transcripts" [11].

The ENCODE Project and FANTOM Consortium showed that RNAs are transcribed from *both* strands of DNA, and that antisense RNA is a major component of the mammalian transcriptome [12-15]. Not only is some RNA transcribed from the antisense strand, but RNAs can also be transcribed from several different start sites within an open reading frame. So a single open reading frame can carry multiple overlapping codes that specify both protein-coding RNAs and non-protein-coding RNAs [16-20].

Widespread transcription suggests probable function; so does sequence conservation. In 2004 and 2005, several groups of scientists identified non-coding regions of DNA that were *completely identical* in humans and mice. They called these "ultra-conserved regions (UCRs)" and noted that they were clustered around genes involved in early development. The researchers concluded that the long non-coding UCRs act as regulators of developmentally important genes [21-24].

In 2006, a team studying endothelial cells (which line the inside of human blood vessels) reported that "conserved non-coding sequences" — some within introns — were enriched in sequences that "may play a key role in the regulation of endothelial gene expression" [25]. Oxford geneticists comparing large non-protein-coding RNAs in humans, rats and mice reported conserved sequences that "possess the imprint of purifying selection, thereby indicating their functionality" [26]. And in 2009, a team of American scientists found "over a thousand highly conserved large non-coding RNAs in mammals" that are "implicated in diverse biological processes" [27].

3. Direct Evidence for Some Specific Functions of Non-Protein-Coding RNAs

There is also direct evidence for specific functions of non-protein-coding RNAs. Paraspeckles are domains inside the nuclei of mammalian cells that play a role in gene expression by retaining certain RNAs within the nucleus [28]. Several non-protein-coding RNAs are known to be essential constituents of them [29,30], binding to specific proteins to form ribonucleoproteins that stabilize the paraspeckles and enable them to persist through cell divisions even though they are not bounded by membranes [31,32].

Non-protein-coding RNAs are also involved in alternative splicing. When a eukaryotic gene is transcribed into RNA, its non-protein-coding introns are removed and the protein-coding exons are then spliced together before being translated into protein. In the great majority of cases (80-95%), the exons can be "alternatively spliced," which means that the resulting transcripts can lack some exons or contain duplicates of others [33,34]. Alternative splicing plays an essential role in the differentiation of cells and tissues at the proper times during embryo development, and many alternatively spliced RNAs occur in a developmental-stage- and tissue-specific manner [35-37].

Although introns do not code for proteins, the RNAs transcribed from them contain specific codes that regulate alternative splicing [38-40]. The mammalian thyroid hormone receptor gene produces two variant proteins with opposite effects, and the alternative splicing of those variants is regulated by an intron [41]. An intronic element plays a critical role in the alternative splicing of tissue-specific RNAs in mice [42], and regulatory elements in introns control the alternative splicing of growth factor receptors in mammalian cells [43].

In 2007, Italian biologists reported that intronic sequences regulate the alternative splicing of a gene involved in human blood clotting [44]. In 2010, a team of Canadian and British scientists studying splicing codes in mouse embryonic and adult tissues — including the central nervous system, muscles, and the digestive system — found that introns are rich in splicing-factor recognition sites. It had previously been assumed that most such sites are close to the affected exons — leaving long stretches of DNA not involved in the process of alternative splicing — but the team concluded that their results suggested "regulatory elements that are deeper into introns than previously appreciated" [45].

Introns encode other functional RNAs, as well. Short non-protein-coding RNAs are known to regulate gene expression [46], and in 2004 British scientists identified such RNAs within the introns of 90 protein-coding genes [47]. In 2007, Korean biologists reported that in humans a "majority" of short non-protein-coding RNAs originate "within intronic regions" [48]. One of these, according to American medical researchers, is involved in regulating cholesterol levels in humans [49]. Introns also encode many of the small RNAs essential for the processing of ribosomal RNAs, as well as the regulatory elements associated with such RNA-coding sequences [50,51].

Chromatin organization profoundly affects gene expression. Non-protein-coding RNAs are essential for chromatin organization [52,53], and non-protein-coding RNAs have been shown to affect gene expression by modifying chromatin structure [54,55]. A recent study of chromatin-associated RNAs in some human cells revealed that almost two-thirds of them are derived from introns [56].

Pseudogenes are transcribed into non-protein-coding RNAs that in some cases regulate the expression of the corresponding protein-coding genes. For example, pseudogenes can reduce gene expression through RNA interference. Since RNA transcribed from the antisense strand of a pseudogene is complementary to the RNA transcribed from the gene, the former binds to the latter to make double-stranded RNA that is not translated [57-59].

Pseudogenes can also increase gene expression through target mimicry. Since the non-protein-coding RNA transcribed from the sense strand of a pseudogene resembles in many respects the protein-coding RNA transcribed from the gene, the former can provide an alternative target for RNA-degrading enzymes that would normally reduce the expression of a gene by inactivating its messenger RNA [60-62].

About half of the human genome consists of non-protein-coding repetitive DNA, and about two-thirds of this is made up of **L**ong **I**nterspersed **N**uclear **E**lements (LINEs) and **S**hort **I**nterspersed **N**uclear **E**lements (SINEs). In mammals, the most common LINE has been designated L1, and in humans the most common SINEs are *Alu*s — so named because they are recognized by an enzyme from the bacterium ***A****rthrobacter* ***lu****teus.*

Human L1 sequences function by mobilizing various RNAs in the cell [63]. L1s also silence a gene that is expressed in the liver in human fetuses but not in adults [64]. In a 2008 review, an Italian biologist concluded that human L1 "regulates fundamental biological processes" [65]. LINEs also participate in the necessary inactivation of most protein-coding regions of the second X chromosome in female eutherian mammals. In 2010, British researchers reported that X chromosome inactivation depends on non-protein-coding RNAs that act more efficiently in L1-rich domains [66]. The same year, French biologists concluded that LINEs function at two different levels in X chromosome inactivation: First, LINE DNA produces a rearrangement in the chromatin that inactivates some genes; second, RNAs transcribed from LINEs coat and silence other portions of the chromosome [67].

Alu elements contain functional binding sites for transcription factors [68]. RNAs derived from *Alu* sequences repress transcription during the cellular response to elevated temperatures [69]. *Alu*-derived RNAs are also involved in the editing and alternative spicing of other RNAs and in the translation of RNAs into proteins [70-74]. In 2009, Colorado researchers studying the biological functions of *Alus* reported that they are transcribed into RNAs that help to control gene expression by controlling the transcription of messenger RNAs and by editing other RNAs. The researchers concluded: "Finding… that these SINE encoded RNAs indeed have biological functions has refuted the historical notion that SINEs are merely 'junk DNA'" [75].

4. Functions of Non-Protein-Coding DNA That Are Not Determined by Precise Nucleotide Sequences

The genome functions hierarchically, and the order of nucleotides in protein-coding and non-protein-coding DNA constitutes only the first level of that hierarchy. The length of DNA sequences (even non-protein-coding ones) is a second level; chromatin organization is a third level; and the position of chromosomes within the nucleus is a fourth [76,77]. There is evidence that DNA functions at the second, third, and fourth levels in ways that are independent of the precise nucleotide sequence.

4.1 *The Length of DNA Sequences*

In 1986, British biologist David Gubb suggested that the time needed to transcribe eukaryotic genes is a factor in regulating the quantity of protein they produce. He proposed that the sheer length of introns in some genes "would affect both the spatial and temporal pattern of expression of their gene products" [78]. In 1992, American biologist Carl Thummel likewise argued that "the physical arrangement and lengths of transcription units can play an important role in controlling their timing of expression." For example, the very long introns in certain key developmental genes could delay their transcription, "consistent with the observation that they function later in development" than genes with shorter introns [79].

In 2008, Harvard systems biologists Ian Swinburne and Pamela Silver summarized circumstantial evidence that intron length has significant effects on the timing of transcription. "Developmentally regulated gene networks," they wrote, "where timing and dynamic patterns of expression are critical, may be particularly sensitive to intron delays" [80]. So introns might have a function in gene regulation that is independent of their exact nucleotide sequence — namely, regulating the timing of transcription simply by their length.

The long stretches of non-protein-coding DNA between protein-coding regions might also affect gene expression by their length. In 1997, molecular biologist Emile Zuckerkandl suggested that DNA may function in ways that do not depend on its particular nucleotide sequence. "Along noncoding sequences," he wrote, "nucleotides tend to fill functions collectively, rather than individually." Sequences that are non-functional at the level of individual nucleotides may function at higher levels involving physical interactions [81].

Because the distance between enhancers and promoters is a factor in gene regulation, Zuckerkandl wrote in 2002, "genomic distance per se — and, therefore, the mass of intervening nucleotides — can have functional effects." He

concluded: "Given the scale dependence of nucleotide function, large amounts of 'junk DNA', contrary to common belief, must be assumed to contribute to the complexity of gene interaction systems and of organisms" [82]. In 2007, Zuckerkandl (with Giacomo Cavalli) wrote that "SINEs and LINEs, which have been considered 'junk DNA,' are among the repeat sequences that would appear liable to have teleregulatory effects on the function of a nearby promoter, through changes in their numbers and distribution" [83].

Since enhancers can be tens of thousands of nucleotides away from the genes they regulate, bringing together enhancers and promoters that are on the same chromosome requires chromosome "looping" [84,85]. The size of a chromosome loop depends on the length of the DNA. For physical reasons, a loop consisting only of DNA must be at least 500 nucleotides long, while a loop consisting of chromatin (because of its greater stiffness) must be at least 10,000 nucleotides long [86]. In such cases it may be the sheer length of the DNA that matters, not whether it encodes RNAs.

4.2 *Chromatin Organization*

Because DNA is packaged into chromatin, and because RNA polymerase must have access to the DNA to transcribe it, the structure of chromatin is all-important in gene regulation. In many cases, various proteins and RNAs mediate the attachment of RNA polymerase to the DNA by interacting with specific sequences of nucleotides, but in some cases a mere change in the three-dimensional conformation of chromatin can activate transcription by exposing the DNA to RNA polymerase [87].

In 2007, scientists in Massachusetts produced a genome-scale, high-resolution three-dimensional map of DNA and found similar conformations that were independent of the underlying nucleotide sequences. They concluded that "considerably different DNA sequences can share a common structure" due to their similar chromatin conformation, and some transcription factors may be "conformation-specific … rather than DNA sequence-specific" [88].

Two years later, scientists reported that functional non-protein-coding regions of the human genome are correlated with chromatin-related "local DNA topography" that can be independent of the underlying sequence. "Although similar sequences often adopt similar structures," they wrote, "divergent nucleotide sequences can have similar local structures," suggesting that "they may perform similar biological functions." The authors of the report concluded that "some of the functional information in the non-coding portion of the genome is conferred by DNA structure as well as by the nucleotide sequence" [89].

The clearest example of a chromatin-level function that can be independent of the exact DNA sequence is the "centromere," a special region on a eukaryotic chromosome that serves as the chromosome's point of attachment to other structures in the cell. For example, before a eukaryotic cell divides it makes a duplicate of each chromosome, and the duplicate copies of each chromosome are joined together at their centromeres until they separate and move to daughter cells.

Centromeres can form only on a foundation provided by the chromosome. Yet centromeres are built upon long stretches of repetitive DNA that some biologists have regarded as junk [90]. Although much of the DNA that underlies centromeres is now known to be transcribed into RNAs that perform a variety of functions [91-96], it turns out that centromere formation is to a great extent independent of the exact nucleotide sequence.

The DNA sequences of centromere regions vary significantly from species to species, though all centromeres function similarly [97]. If the chromosome region containing a centromere is artificially deleted and replaced by synthetic repetitive DNA, a functional centromere can form again at the same site [98]. Extra centromeres (called "neo-centromeres") can also form abnormally elsewhere on a chromosome that already has one, or on a chromosome fragment that has separated from the part bearing a centromere [99,100]. It seems that centromeres can form at many different places on a chromosome, regardless of the underlying DNA sequence.

Nevertheless, the underlying chromatin must have certain characteristics that make centromere formation possible. For example, there is evidence that some aspects of the DNA sequence are conserved [101,102]. In humans and other primates, centromere activity is normally associated with repeated blocks of 171- nucleotide subunits termed alpha-satellite DNA. (Researchers in the 1960s discovered that a fraction of DNA consisting of millions of short, repeated nucleotide sequences produced "satellite" bands when DNA was centrifuged to separate it into fractions with different densities.) Every normal human centromere is located on alpha-satellite DNA [103–105].

Human neo-centromeres form on parts of a chromosome that do *not* consist of alpha-satellite DNA, though the neo-centromere DNA still has special characteristics — most notably, an unusually high proportion of LINEs [106]. These non-protein-coding segments apparently play a role in localizing proteins that are required for the formation of the centromere and kinetochore [107,108].

In the 1980s, biologists identified several proteins associated with centromeres and called them CENPs (for **CEN**tromere **P**roteins) [109]. Subsequent research revealed that one of these, CENP-A, takes the place of some of the histones in chromatin [110]. The incorporation of CENP-A makes chromatin stiffer and provides a foundation for assembling the other components of centromeres

[111,112]. In fact, centromeres in all organisms are associated with CENP-A, which must be present for a centromere to form, though CENP-A by itself is not sufficient [113,114].

The modification of chromatin by CENP-A and other centromere-specific proteins can be passed down from generation to generation. Indeed, the location of a centromere on a particular chromosome can persist for thousands of generations. From the perspective of the Central Dogma and Sequence Hypothesis (i.e., the view that DNA sequences determine the essential features of organisms by encoding proteins), centromeres are an enigma because they show that a cell can impose an essential and heritable structure on its DNA that is independent of the precise nucleotide sequence.

4.3 *Chromosome Arrangement in the Nucleus*

Between cell divisions, chromosomes are not randomly distributed in the nucleus; instead, they occupy distinct domains [115]. Chromosome domains affect gene regulation, in part, by bringing together specific regions of chromosomes and facilitating interactions among them [116,117]. Different cell and tissue types in the same animal can have different three-dimensional patterns of chromosomes in their nuclei, which account for at least some differences in gene expression [118,119].

One notable feature of nuclear domains is their radial arrangement [120]. In 1998, biologists in New York reported that chromatin localized to the periphery of the nucleus in yeast cells tends to be "transcriptionally silent" [121]. In 2001, British biologists wrote that "most gene-rich chromosomes concentrate at the centre of the nucleus, whereas the more gene-poor chromosomes are located towards the nuclear periphery" [122]. In 2008, Dutch biologists reported that human chromosome domains associated with the periphery of the nucleus "represent a repressive chromatin environment" [123]. The same year, several teams of researchers reported independently that they could suppress the expression of specific genes by relocating them to the nuclear periphery [124–126].

These data are consistent with the observation that in most nuclei the gene-rich euchromatin is concentrated near the center while the gene-poor heterochromatin is situated more peripherally. An important exception to this radial arrangement, however, occurs in the retinas of nocturnal mammals (Fig. 1).

The retina of a vertebrate eye contains several different kinds of light-sensing cells. Cone cells detect colors and function best in bright light; rod cells are more numerous and more sensitive to low light. Nocturnal animals such as mice need to see under conditions of almost no light, so they need exceptionally sensitive rod

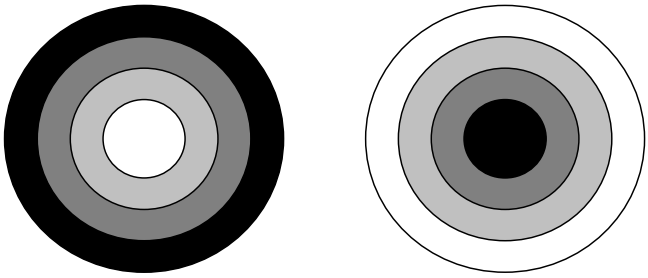

Fig. 1 Left: A simplified view of the internal arrangement of chromatin in most eukaryotic nuclei. Gene-poor heterochromatin (black) is on the periphery, and the gene content of the chromatin increases toward the center, which consists of gene-rich euchromatin (white). Right: A simplified view of the inverted chromatin arrangement found in rod cells in the retinas of nocturnal mammals. Gene-rich euchromatin is on the periphery, while gene-poor heterochromatin is in the center. The centrally located heterochromatin acts as a liquid-crystal lens that focuses the few photons available at night onto the light-sensitive outer segments of the rod cells.

cells. In 1979, medical researchers examined mouse retinas with an electron microscope and found that the heterochromatin in cone cells was located near the periphery of the nucleus, as in most other eukaryotic cells, but the heterochromatin in rod cells was concentrated in "one large, central clump" [127].

Another team of medical researchers used mice to study the genetic mutation responsible for an inherited human disease that causes nerve degeneration [128]. The team found that the mutation causes blindness in mice by altering the arrangement of the chromatin in rod cells. Instead of containing "a single, large clump of heterochromatin surrounded by a spare rim of euchromatin," the rod cells in mutant mice "showed a dramatic chromatin decondensation" and "resembled cone nuclei" [129].

Clearly, the unique localization of heterochromatin in the center of rod cells in mouse retinas is essential for normal vision in these animals. In 2009, European scientists called the unusual pattern of centrally located heterochromatin "inverted," and they reported finding an inverted pattern in the rod cell nuclei of various other mammals that are primarily nocturnal (including cats, rats, foxes, opossums, rabbits and several species of bats) but not of mammals that are primarily active in daylight (such as cows, pigs, donkeys, horses, squirrels, and chipmunks). These scientists observed that the centrally located heterochromatin had a high refractive index — a characteristic of optical lenses — and by using a two-dimensional computer simulation they showed that a main consequence of the inverted pattern was to focus light on the light-sensitive regions of rod cells [130].

In 2010, molecular biologists in France reported that the organization of the central heterochromatin in the rod nuclei of nocturnal mammals is consistent with

a "liquid crystal model" [131], and British biophysicists improved upon the 2009 study by using a new computer simulation to show that "the focusing of light by inverted nuclei" in three dimensions is "at least three times as strong" as it is in two dimensions [132].

So evidence for the functionality of non-protein-coding DNA comes from several sources: pervasive transcription of the genome, including transcription from antisense DNA and from multiple start sites within open reading frames; conservation of a substantial fraction of non-protein-coding sequences; particular sequence-dependent functions of RNAs transcribed from introns, pseudogenes, repetitive DNA (much of which is *not* conserved, but species-specific); and functions that are to a large extent independent of the exact nucleotide sequence, such as the influence of intron length on transcription timing, the role of chromatin topology in gene expression and centromere placement, and the light-focusing property of heterochromatin in inverted nuclei. Clearly, it is no longer reasonable to maintain that the vast majority of our DNA is "junk."

5. Conclusion: Multiple Levels of Biological Information

The concept of information as applied to a linear sequence — such as letters in an English sentence or nucleotides in a DNA molecule — has been extensively analyzed [133-143]. Although protein-coding DNA constitutes less than 2% of the human genome, the amount of such information in such DNA is enormous. Recent discoveries of multiple overlapping functions in non-protein-coding DNA show that the biological information in the genome far exceeds that in the protein-coding regions alone.

Yet biological information is not limited to the genome. Even at the level of gene expression — transcription and translation — the cell must access information that is not encoded in DNA. Many different RNAs can be generated from a single piece of DNA by alternative splicing, and although some splicing codes occur in intronic DNA there is no empirical justification for assuming that *all* of the information for tissue- and developmental-stage-specific alternative splicing resides in DNA. Furthermore, even after RNA has specified the amino acid sequence of a protein, additional information is needed: Protein function depends on three-dimensional shape, and the same sequence of amino acids can be folded differently to produce proteins with different three-dimensional shapes [144–147]. Conversely, proteins with different amino acid sequences can be folded to produce similar shapes and functions [148,149].

Many scientists have pointed out that the relationship between the genome and the organism — the genotype-phenotype mapping — cannot be reduced to a

genetic program encoded in DNA sequences. Atlan and Koppel wrote in 1990 that advances in artificial intelligence showed that cellular operations are not controlled by a linear sequence of instructions in DNA but by a "distributed multilayer network" [150]. According to Denton and his co-workers, protein folding appears to involve formal causes that transcend material mechanisms [151], and according to Sternberg this is even more evident at higher levels of the genotype-phenotype mapping [152].

So non-protein-coding regions of DNA that some previously regarded as "junk" turn out to encode biological information that greatly increases the known information-carrying capacity of DNA. At the same time, DNA as a whole turns out to encode only part of the biological information needed for life.

Addendum

Due to a delay in the publication of these proceedings, the material in this chapter is now (2013) over two years old. Yet it is still accurate. Indeed, the fact that most non-protein-coding DNA serves biological functions was dramatically confirmed in September 2012 by 37 papers published by the ENCODE Project in Nature, Genome Research, Genome Biology, The Journal of Biological Chemistry, *and* Science *[153-189]. The Project concluded that 80% of the genome is linked to biological functions, but Project Coordinator Ewan Birney pointed out that this conclusion was based on analyses of only 147 cell types, and "the human body has a few thousand." As more cell types are studied, Birney said, "It's likely that 80 percent will go to 100 percent." [190] A commentary accompanying the papers in* Nature *described the ENCODE results as "dispatching the widely held view that the human genome is mostly 'junk DNA'." [191] A commentary published at the same time in* Science *announced "ENCODE Project writes eulogy for junk DNA." [192]*

Acknowledgments

The author gratefully acknowledges the assistance of Richard v. Sternberg, helpful comments from reviewers, and the financial support of the Discovery Institute.

References

1. Watson JD, Crick FHC (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 171:737–738

2. Watson JD, Crick FHC (1953) Genetical implications of the structure of deoxyribonucleic acid. Nature 171:964–967
3. Crick FHC (1958) On protein synthesis. Symp Soc Exp Biol 12:138–163
4. Ohno S (1972) So much 'junk' DNA in our genome. Brookhaven Symp Biol 23: 366–70
5. Dawkins R (1976) The selfish gene. Oxford University Press, New York
6. National Human Genome Research Institute (2003) International consortium completes human genome project. http://www.genome.gov/11006929 Accessed 2013 Mar 15
7. Okazaki Y, et al (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420:563–573
8. Kapranov P, et al (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science 296:916–919
9. Carninci P, et al (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–1563
10. Pheasant M, Mattick JS (2007) Raising the estimate of functional human sequences. Genome Res 17:1245–1253
11. Birney E, et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816
12. Katayama S, et al (2005) Antisense transcription in the mammalian transcriptome. Science 309:1564–1566
13. Engström PG, et al (2006) Complex loci in human and mouse genomes. PLoS Genet 2:e47
14. He Y, et al (2008) The antisense transcriptomes of human cells. Science 322:1855–1857
15. Morris KV, et al (2008) Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. PLoS Genet 4:e1000258
16. Gustincich S, et al (2006) The complexity of the mammalian transcriptome. J Physiol 575:321–332
17. Kapranov P, Willingham AT, Gingeras TR (2007) Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8:413–423
18. Carninci P (2007) Constructing the landscape of the mammalian transcriptome. J Exp Biol 210:1497–1506
19. Wu JQ, et al (2008) Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. Genome Biol 9:R3
20. Itzkovitz S, Hodis E, Segal E (2010) Overlapping codes within protein-coding sequences. Genome Res 20:1582–1589
21. Bejerano G, et al (2004) Ultraconserved elements in the human genome. Science 304:1321–1325

22. Sandelin A, et al (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5:99
23. Woolfe A, et al (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3:e7
24. Siepel A, et al (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050
25. Bernat JA, et al (2006) Distant conserved sequences flanking endothelial-specific promoters contain tissue-specific DNase-hypersensitive sites and over-represented motifs. Hum Mol Genet 15:2098–2105
26. Ponjavic J, Ponting CP, Lunter G (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res 17:556–565
27. Guttman M, et al (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223–227
28. Bond CS, Fox AH (2009) Paraspeckles: nuclear bodies built on long noncoding RNA. J Cell Biol 186:637–644
29. Clemson CM, et al (2009) An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol Cell 33:717–726
30. Sasaki YTF, et al (2009) MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles. Proc Natl Acad Sci USA 106:2525–2530
31. Sasaki YTF, Hirose T (2009) How to build a paraspeckle. Genome Biol 10:227
32. Souquere S, et al (2010) Highly-ordered spatial organization of the structural long noncoding NEAT1 RNAs within paraspeckle nuclear bodies. Mol Biol Cell 21:4020–4027
33. Wang ET, et al (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476
34. Sultan M, et al (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956–960
35. Blencowe BJ (2006) Alternative splicing: new insights from global analyses. Cell 126:37–47
36. Prokunina-Olsson L, et al (2009) Tissue-specific alternative splicing of *TCF7L2*. Hum Mol Genet 18:3795–3804
37. Mercer TR, et al (2010) Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. Genome Res 20:1639–1650
38. Ladd AN, Cooper TA (2002) Finding signals that regulate alternative splicing in the post-genomic era. Genome Biol 3:reviews0008
39. Hui J, et al (2005) Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J 24:1988–1998
40. Nakaya HI (2007) Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. Genome Biol 8:R43

41. Hastings ML, Wilson CM, Munroe SH (2001) A purine-rich intronic element enhances alternative splicing of thyroid hormone receptor mRNA. RNA 7: 859–874
42. Nakahata S, Kawamoto S (2005) Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities. Nucleic Acids Res 33:2078–2089
43. Wagner EJ, et al (2005) Characterization of the intronic splicing silencers flanking FGFR2 exon IIIb. J Biol Chem 280:14017–14027
44. Marcucci R, Baralle FE, Romano M (2007) Complex splicing control of the human thrombopoietin gene by intronic G runs. Nucleic Acids Res 35:132–142
45. Barash Y, et al (2010) Deciphering the splicing code. Nature 465:53–59
46. Mendes Soares LM & Valcárcel J (2006) The expanding transcriptome: the genome as the 'Book of Sand', EMBO J 25:923–931
47. Rodriguez A, et al (2004) Identification of mammalian microRNA host genes and transcription units. Genome Res 14:1902–1910
48. Kim Y-K, Kim VN (2007) Processing of intronic microRNAs. EMBO J 26: 775–783.
49. Najafi-Shoushtari SH, et al (2010) MicroRNA-33 and the SREBP host genes cooperate to control cholesterol homeostasis. Science 328:1566–1569
50. Hoeppner MP, et al (2009) Evolutionarily stable association of intronic snoRNAs and microRNAs with their host genes. Genome Biol Evol 2009:420–428
51. Monteys AM, et al (2010) Structure and activity of putative intronic miRNA promoters. RNA 16:495–505
52. Lavrov SA, Kibanov MV (2007) Noncoding RNAs and chromatin structure. Biochemistry (Mosc) 72:1422–1438
53. Rodríguez-Campos A, Azorín F (2007) RNA is an integral component of chromatin that contributes to its structural organization. PLoS One 2:e1182
54. Malecová B, Morris KV (2010) Transcriptional gene silencing through epigenetic changes mediated by non-coding RNAs. Curr Opin Mol Ther 12:214–222
55. Caley DP, et al (2010) Long noncoding RNAs, chromatin, and development. ScientificWorldJournal 10:90–102
56. Mondal T, et al (2010) Characterization of the RNA content of chromatin. Genome Res 20:899–907
57. Meister G, Tuschl T (2004) Mechanisms of gene silencing by double-stranded RNA. Nature 431:343–349
58. Watanabe T, et al (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature 453:539–543
59. Tam OT, et al (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 453:534–538
60. Franco-Zorrilla JM, et al (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet 39:1033–1037

61. Piehler AP, et al (2008) The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference. BMC Genomics 9:165
62. Poliseno L, et al (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465:1033–1038
63. Garcia-Perez JL, et al (2007) Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. Genome Res 17:602–611
64. Shepard EA, et al (2007) Alternative promoters and repetitive DNA elements define the species-dependent tissue-specific expression of the FMO1 genes of human and mouse. Biochem J 406:491–499
65. Spadafora C (2008) A reverse transcriptase-dependent mechanism plays central roles in fundamental biological processes. Syst Biol Reprod Med 54:11–21
66. Tang YA, et al (2010) Efficiency of Xist-mediated silencing on autosomes is linked to chromosomal domain organization. Epigenetics Chromatin 3:10
67. Chow JC, et al (2010) LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell 141:956–969
68. Vansant G, Reynolds WF (1995) The consensus sequence of a major *Alu* subfamily contains a functional retinoic acid response element. Proc Natl Acad Sci USA 92:8229–8233
69. Mariner PD, et al (2008) Human *Alu* RNA is a modular transacting repressor of mRNA transcription during heat shock. Mol Cell 29:499–509
70. Häsler J, Strub K (2006) *Alu* RNP and *Alu* RNA regulate translation initiation in vitro. Nucleic Acids Res 34:2374–2385
71. Häsler H, Strub K (2006) *Alu* elements as regulators of gene expression. Nucleic Acids Res 34:5491–5497
72. Häsler J, Samuelsson T, Strub K (2007) Useful 'junk': Alu RNAs in the human transcriptome. Cell Mol Life Sci 64:1793–1800
73. Gu TJ, et al (2009) Alu-directed transcriptional regulation of some novel miRNAs. BMC Genomics 10:563
74. Barak M, et al (2009) Evidence for large diversity in the human transcriptome created by *Alu* RNA editing. Nucleic Acids Res 37:6905–6915
75. Walters RD, Kugel JF, Goodrich JA (2009) InvAluable junk: the cellular impact and function of *Alu* and B2 RNAs. IUBMB Life 61:831–837
76. van Driel R, Fransz PF, Verschure PJ (2003) The eukaryotic genome: a system regulated at different hierarchical levels. J Cell Sci 116:4067–4075
77. Misteli T (2007) Beyond the sequence: cellular organization of genome function. Cell 128:787–800
78. Gubb D (1986) Intron-delay and the precision of expression of homeotic gene products in *Drosophila*. Dev Genet 7:119–131
79. Thummel CS (1992) Mechanisms of Transcriptional Timing in *Drosophila*. Science 255:39–40

80. Swinburne IA, Silver PA (2008) Intron delays and transcriptional timing during development. Dev Cell 14:324–330
81. Zuckerkandl E (1997) Junk DNA and sectorial gene repression. Gene 205:323–343
82. Zuckerkandl E (2002) Why so many noncoding nucleotides? The eukaryote genome as an epigenetic machine. Genetica 115:105–129
83. Zuckerkandl E, Cavalli G (2007) Combinatorial epigenetics, 'junk DNA', and the evolution of complex organisms. Gene 390:232–242
84. Celniker SE, Drewell RA (2007) Chromatin looping mediates boundary element promoter interactions. BioEssays 29:7–10
85. Palstra RJ (2009) Close encounters of the 3C kind: long-range chromatin interactions and transcriptional regulation. Brief Funct Genomic Proteomic 8:297–309
86. Göndör A, Ohlsson R (2009) Chromosome crosstalk in three dimensions. Nature 461:212–217
87. Angermayr M, et al (2002) Transcription initiation in vivo without classical transactivators: DNA kinks flanking the core promoter of the housekeeping yeast adenylate kinase gene, AKY2, position nucleosomes and constitutively activate transcription. Nucleic Acids Res 30:4199–4207
88. Greenbaum JA, Pang B, Tullius TD (2007) Construction of a genome-scale structural map at single-nucleotide resolution. Genome Res 17:947–953
89. Parker SCJ, et al (2009) Local DNA topography correlates with functional noncoding regions of the human genome. Science 324:389–392
90. Gardiner K (1997) Clonability and gene distribution on human chromosome 21: reflections of junk DNA content? Gene 205:39–46
91. Reinhart BJ, Bartel DP (2002) Small RNAs correspond to centromere heterochromatic repeats. Science 297:1831
92. Lu J, Gilbert DM (2007) Proliferation-dependent and cell cycle regulated transcription of mouse pericentric heterochromatin. J Cell Biol 179:411–412
93. Wong LH, et al (2007) Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. Genome Res 17:1146–1160
94. O'Neill RJ, Carone DM (2009) The role of ncRNA in centromeres: a lesson from marsupials. Progr Mol Subcell Biol 48:77–101
95. Ferri F, et al (2009) Non-coding murine centromeric transcripts associate with and potentiate Aurora B kinase. Nucleic Acids Res 37:5071–5080
96. Eymery A, Callanan M, Vourc'h C (2009) The secret message of heterochromatin: new insights into the mechanisms and function of centromeric and pericentric repeat sequence transcription. Int J Dev Biol 53:259–268
97. Sullivan BA, Blower MD, Karpen GH (2001) Determining centromere identity: cyclical stories and forking paths. Nat Rev Genet 2:584–596
98. Harrington JJ, et al (1997) Formation of de novo centromeres and construction of first-generation human artificial microchromosomes. Nat Genet 15:345–355

99. du Sart D, et al (1997) A functional neo-centromere formed through activation of a latent human centromere and consisting of non-alpha-satellite DNA. Nat Genet 16:144–153
100. Warburton PE (2004) Chromosomal dynamics of human neocentromere formation. Chromosome Res 12:617–626
101. Grady DL, et al (1992) Highly conserved repetitive DNA sequences are present at human centromeres. Proc Natl Acad Sci USA 89:1695–1699
102. Jiang J, et al (1996) A conserved repetitive DNA element located in the centromeres of cereal chromosomes. Proc Natl Acad Sci USA 93:14210–14213
103. Waye JS, Willard HF (1985) Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome. Nucleic Acids Res 13:2731–2743
104. Heller R, et al (1996) Mini-chromosomes derived from the human Y chromosome by telomere directed chromosome breakage. Proc Natl Acad Sci USA 93:7125–7130
105. Murphy TD, Karpen GH (1998) Centromeres take flight: Alpha satellite and the quest for the human centromere. Cell 93:317–320
106. Lo AW, et al (2001) A 330 kb CENP-A binding domain and altered replication timing at a human neocentromere, EMBO J 20:2087–2096
107. Chueh AC, et al (2005) Variable and hierarchical size distribution of L1-retroelement-enriched CENP-A clusters within a functional human neocentromere. Hum Mol Genet 14:85–93
108. Chueh AC, et al (2009) LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. PLoS Genet 5:e1000354
109. Palmer DK, et al (1987) A 17-kD centromere protein (CENP-A) copurifies with nucleosome core particles and with histones. J Cell Biol 104:805–815
110. Yoda K, et al (2000) Human centromere protein A (CENP-A) can replace histone H3 in nucleosome reconstitution in vitro. Proc Natl Acad Sci USA 97:7266–7271
111. Black BE, et al (2007) An epigenetic mark generated by the incorporation of CENP-A into centromeric nucleosomes. Proc Natl Acad Sci USA 104:5008–5013
112. Torras-Llort M, Moreno-Moreno O, Azorín F (2009) Focus on the centre: the role of chromatin on the regulation of centromere identity and function. EMBO J 28:2337–2348
113. Van Hooser AA, et al (2001) Specification of kinetochore-forming chromatin by the histone H3 variant CENP-A. J Cell Sci 114:3529–3542
114. Vos LJ, Famulski JK, Chan GKT (2006) How to build a centromere: from centromeric and pericentromeric chromatin to kinetochore assembly. Biochem Cell Biol 84:619–639
115. Cremer T, Cremer M (2010) Chromosome territories. Cold Spring Harb Perspect Biol 2:a003889

116. Lanctôt C, et al (2007) Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nat Rev Genet 8:104–115
117. Joffe B, Leonhardt H, Solovei I (2010) Differentiation and large scale spatial organization of the genome. Curr Opin Genet Dev 20:562–569
118. Parada LA, McQueen PG, Misteli T (2004) Tissue-specific spatial organization of genomes. Genome Biol 5:R44
119. Sexton T, et al (2007) Gene regulation through nuclear organization. Nat Struct Mol Biol 14:1049–1055
120. Takizawa T, Meaburn KJ & Misteli T (2008) The meaning of gene positioning, Cell 135:9–13
121. Andrulis ED, et al (1998) Perinuclear localization of chromatin facilitates transcriptional silencing. Nature 394:592–595
122. Boyle S, et al (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Hum Mol Genet 10:211–219
123. Guelen L, et al (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453:948–951
124. Finlan LE, et al (2008) Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS Genet 4:e1000039
125. Reddy KL, et al (2008) Transcriptional repression mediated by repositioning of genes to the nuclear lamina. Nature 452:243–247
126. Ruault M, Dubarry M, Taddei A (2008) Re-positioning genes to the nuclear envelope in mammalian cells: impact on transcription. Trends Genet 24:574–581
127. Carter-Dawson LD, LaVail MM (1979) Rods and cones in the mouse retina. I. Structural analysis using light and electron microscopy. J Comp Neurol 188:245–262
128. Blackshaw S, et al., (2001) Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. Cell 107:579-589
129. Helmlinger D, et al (2006) Glutamine-expanded ataxin-7 alters TFTC/STAGA recruitment and chromatin structure leading to photoreceptor dysfunction. PLoS Biol 4:e67
130. Solovei I, et al (2009) Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. Cell 137:356–368
131. Kizilyaprak C, et al (2010) In vivo chromatin organization of mouse rod photoreceptors correlates with histone modifications. PLoS One 5:e11039
132. Kreysing M, et al (2010) Physical insight into light scattering by photoreceptor cell nuclei. Opt Lett 35:2639–2641
133. Shannon CE (1948) A mathematical theory of communication. Bell System Technical J 27:379–423,623–656
134. Brillouin L (1956) Science and information Theory, 2nd Edition. Academic Press, New York
135. Yockey HP (1992) Information theory and molecular biology. Cambridge University Press, Cambridge

136. Gitt W (1997) In the beginning was information. Christliche Literatur-Verbreitung; Bielefeld, Germany
137. Dembski WA (1998) The design inference. Cambridge University Press, Cambridge
138. Dembski WA (2002) No free lunch: Why specified complexity cannot be purchased without intelligence. Rowman & Littlefield; Lanham, MD
139. Meyer SC (2004) The origin of biological information and the higher taxonomic categories. Proc Biol Soc Wash 117:213–239
140. Sanford JC (2005) Genetic entropy & the mystery of the genome. Elim Publishing; Lima, NY
141. Dembski WA & Marks RJ II (2009) Conservation of information in search: Measuring the cost of success. IEEE Trans Syst Man Cybern A 5(5):1051–1061
142. Meyer SC (2009) Signature in the cell: DNA and the evidence for intelligent design. HarperCollins, New York
143. Ewert W, et al (2010) Efficient per query information extraction from a Hamming Oracle. Proceedings of the 42nd Meeting of the Southeastern Symposium on System Theory, IEEE, University of Texas at Tyler, March 7-9, 2010, pp. 290–297
144. Kimchi-Sarfaty C, et al (2007) A 'silent' polymorphism in the MDR1 gene changes subsrate specificity. Science 315:525–528
145. Newman SA, Bhat R (2007) Genes and proteins: dogmas in decline. J Biosci 32:1041–1043
146. Pezza JA, Serio TR (2007) Prion propagation: the role of protein dynamics. Prion 1:36–43.
147. Bryan PN, Orban J (2010) Proteins that switch folds. Curr Opin Struct Biol 20: 482–488
148. Perry AJ, et al (2006) Convergent evolution of receptors for protein import into mitochondria. Curr Biol 16:221–229
149. Nickson AA, Clarke J (2010) What lessons can be learned from studying the folding of homologous proteins? Methods 52:38-50
150. Atlan H, Koppel M (1990) The cellular computer DNA: program or data. Bull Math Biol 52:335-348
151. Denton MJ, Marshall CJ, Legge M (2002) The protein folds as Platonic forms: new support for the pre-Darwinian conception of evolution by natural law. J Theor Biol 219:325–342
152. Sternberg RV (2008) DNA codes and information: formal structures and relational causes. Acta Biotheor 56:205–232
153. Arvey A, et al (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. Genome Res 22:1723–1734
154. Bánfai B, et al (2012) Long noncoding RNAs are rarely translated in two human cell lines. Genome Res 22:1646–1657

155. Boyle AP, et al (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22:1790–1797
156. Charos AE, et al (2012) A highly integrated and complex PPARGC1A transcription factor binding network in HepG2 cells. Genome Res 22:1668–1679
157. Cheng C, et al (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res 22:1658–1667
158. Derrien T, et al (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22: 1775–1789
159. Djebali S, et al (2012) Landscape of transcription in human cells. Nature 489: 101–108
160. Dong X, et al (2012) Modeling gene expression using chromatin features in various cellular contexts. Genome Biol 13:R53
161. Dunham I, et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74
162. Euskirchen G, Auerbach RK, Snyder S (2012) SWI/SNF chromatin-remodeling factors: multiscale analyses and diverse functions. J Biol Chem 287:30897–30905
163. Frietze S, et al (2012) Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. Genome Biol 13:R52
164. Gerstein MB, et al (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489:91–100
165. Hardison RC (2012) Genome-wide epigenetic data facilitate understanding of disease susceptibility association studies. J Biol Chem 287:30932–30940
166. Harrow J, et al. (2012) GENCODE: the reference human genome annotation for the ENCODE project. Genome Res 22:1760–1774
167. Howald C, et al (2012) Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. Genome Res 22:1698–1710
168. Kundaje A, et al (2012) Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. Genome Res 22:1735–1747
169. Ladewig E, et al (2012) Discovery of hundreds of mirtrons in mouse and human small RNA data. Genome Res 22:1634–1645
170. Lan X, Farnham PJ, Jin VX (2012) Uncovering transcription factor modules using one- and three-dimensional analyses. J Biol Chem 287:30914–30921
171. Landt SG, et al (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22:1813–1831
172. Lee B-K, Iyer VR (2012) Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation. J Biol Chem 287:30906–30913
173. Maurano MT, et al (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337:1190–1195

174. Natarajan A, et al (2012) Predicting cell-type-specific gene expression from regions of open chromatin. Genome Res 22:1711–1722
175. Neph S, et al (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489:83–90
176. Park E, et al (2012) RNA editing in the human ENCODE RNA-seq data. Genome Res 22:1626–1633
177. Pei B, et al (2012) The GENCODE pseudogene resource. Genome Biol 13:R51
178. Sanyal A, et al (2012) The long-range interaction landscape of gene promoters. Nature 489:109–113
179. Schaub MA, et al (2012) Linking disease associations with regulatory information in the human genome. Genome Res 22:1748–1759
180. Sindhu C, Samavarchi-Tehrani P, Meissne A (2012) Transcription factor-mediated epigenetic reprogramming. J Biol Chem 287:30922–30931
181. Spivakov M, et al (2012) Analysis of variation at transcription factor binding sites in Drosophila and humans. Genome Biol 13:R49
182. Thurman RE, et al (2012) The accessible chromatin landscape of the human genome. Nature 489:75–82
183. Tilgner H, et al (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res 22:1616–1625
184. Vernot B, et al (2012) Personal and population genomics of human regulatory variation. Genome Res 22:1689–1697
185. Wang H, et al (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res 22:1680–1688
186. 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
187. Whiteld TW, et al (2012) Functional analysis of transcription factor binding sites in human promoters. Genome Biol 13:R50
188. Yip KY, et al (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol 13:R48
189. Zentner GE, Scacheri PC (2012) The chromatin fingerprint of gene enhancer elements. J Biol Chem 287:30888–30896
190. Quoted in Yong E (2012) ENCODE: the rough guide to the human genome. Discover Magazine Blog. Available at http://blogs.discovermagazine.com/notrocketscience/2012/09/05/encode-the-rough-guide-to-the-human-genome/. Accessed 2013 Jan 14
191. Ecker JR (2012) Serving up a genome feast. Nature 489:52–53.
192. Pennisi E (2012) ENCODE project writes eulogy for junk DNA. Science 337: 1159–1161