Received: 18 October 2023 | Accepted: 19 January 2024

DOI: 10.1002/bies.202300201

BioEssays

PROBLEMS & PARADIGMS

Are non-protein coding RNAs junk or treasure?

An attempt to explain and reconcile opposing viewpoints of whether the human genome is mostly transcribed into non-functional or functional RNAs

Nils G. Walter

Center for RNA Biomedicine, Single Molecule Analysis Group, Department of Chemistry, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Nils G. Walter, Center for RNA Biomedicine, Single Molecule Analysis Group, Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055, USA.
Email: nwalter@umich.edu

Funding information
NIH, Grant/Award Number: GM131922

Abstract

The human genome project's lasting legacies are the emerging insights into human physiology and disease, and the ascendance of biology as the dominant science of the 21st century. Sequencing revealed that >90% of the human genome is not coding for proteins, as originally thought, but rather is overwhelmingly transcribed into non-protein coding, or non-coding, RNAs (ncRNAs). This discovery initially led to the hypothesis that most genomic DNA is "junk", a term still championed by some geneticists and evolutionary biologists. In contrast, molecular biologists and biochemists studying the vast number of transcripts produced from most of this genome "junk" often surmise that these ncRNAs have biological significance. What gives? This essay contrasts the two opposing, extant viewpoints, aiming to explain their bases, which arise from distinct reference frames of the underlying scientific disciplines. Finally, it aims to reconcile these divergent mindsets in hopes of stimulating synergy between scientific fields.

KEYWORDS
bacterial riboswitches, biological function, evolutionary selection, genetics and genomics, gene expression & regulation, human genome, non-protein coding RNA

INTRODUCTION

As many stories, this essay starts with a personal experience. One fall term, teaching a special topics course on "Nucleic Acid Biochemistry" for senior undergraduates and beginning graduate students, I had asked teams of students to each develop or expand a Wikipedia entry related to our course material. One group identified "Repeated sequence (DNA)" as a rudimentary entry that needed some work. The students added—based on an exhaustive literature search—much flesh to the existing "bones" in the form of the history, types, and functions in physiology and disease of these pervasive repeat sequences. At the end of the term, the students moved the entry from their "Sandbox" to Wikipedia proper.[1] Remarkably, within 24 h an international team of retired and active geneticists and evolutionary biologists had prominently added a sentence to the entry, suggesting that these repeat sequences are likely non-functional and belong to the "junk" or "selfish" DNA of the cell. Additionally, they published an accompanying blog about their grievance that the term "junk DNA" had been removed by the students.[2]

Personally, this strong reaction surprised me as I had assumed that modern transcriptome analyses, funded in part since 2004 by the National Human Genome Research Institute (NHGRI) through the ENCODE project,[3–8] had long dispelled the early misperception that our genome harbors 90% useless "junk DNA", since most of it is transcribed into non-protein coding, or non-coding, RNAs (ncRNAs, both small and long, that is, lncRNAs; Figure 1).[9] Through a long exchange

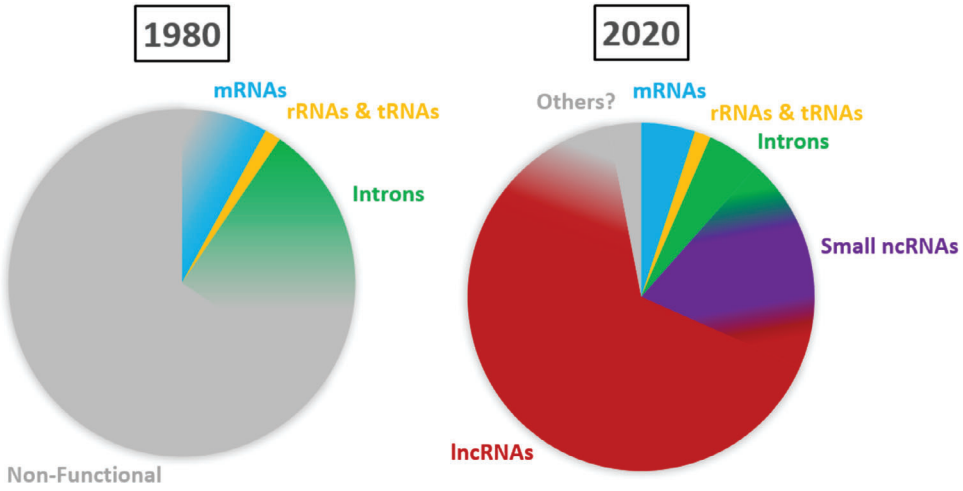

FIGURE 1 Assessment of the information content of the human genome ~20 years before (left)[110] and after (right)[111] the Human Genome Project was preliminarily completed, drawn roughly to scale.[9] This significant progress can be described per Thomas Kuhn as a "paradigm shift" flanked by extended periods of "normal science", during which investigations are designed and results interpreted within the dominant conceptual frameworks of the sub-disciplines.[9] Others have characterized this leap in assigning newly discovered ncRNAs at least a rudimentary (elemental) biochemical activity and thus function as excessively optimistic, or Panglossian, since it partially extrapolates from the known to the unknown.[75] Adapted from Ref. [9].

of thoughtful emails with several members of the community favoring the terms "junk DNA" and "junk RNA", as well as through conversations with local colleagues, I realized that much of the difference of opinion is predicated on distinctions between scientific fields. It all starts with a definition of fundamental terms such as "biological function" and "gene", as well as the perceived relevance of observing "conservation" among genomes.

To help bridge this divide, I will first define some of these fundamental terms, then aim to discuss the rationales of the two opposing points of view, and finally offer ideas for how to reconcile the geneticist's and evolutionary biologist's ways of thinking with that of the molecular biologist and biochemist.

STARTING WITH THE BASICS: DEFINING "FUNCTION" AND "GENE"

The definition of scientific terms is especially important for debates between different fields since each area of science may use a distinct frame of reference. Case in point—while the term "biological function" refers to one of the most foundational tenets of biology, it turns out to be differently defined by geneticists and evolutionary biologists versus molecular biologists and biochemists. For those biosciences studying whole organisms or ecosystems, a biological function describes the reason why an organism harbors a particular trait or behavior. In contrast, the molecular biologist and biochemist will think of biological function as a specific biochemical activity that a molecule carries out within the cells of the organism. Such a function may be as rudimentary as the recruitment of a protein by a specific segment of a nucleic acid. Accordingly, the ENCODE team defined RNA elements as functional when they bound one of the plethora of RNA-binding proteins (RBPs) found in the mammalian cell.[6] How that protein-binding RNA element fits into, for example, a specific regulatory pathway and how this pathway then leads to a particular trait or behavior of the organism as a whole is of secondary concern.

The ENCODE publications led to a heated debate in academic journals, the blogosphere and the press.[11] What the ENCODE team's results implied, at least to some, was that the idea of junk DNA was proven wrong, because almost all human DNA turned out to be functional, necessitating that textbooks be rewritten. In contrast, the defenders of the idea of junk DNA suggested that the ENCODE researchers set far too low a bar in ascribing biological function to RNA elements.[11]

Even today, a survey of human genes finds that most ncRNAs—so pervasively transcribed from the human genome—have no clear function yet.[12] While there are ever more paths toward identifying ncRNA function in service of completing the human gene catalogue, few of these tools are high in throughput, in large part because most ncRNAs participate in specific functional pathways in unique ways.[12] This tediousness leads to the fact that the definition of a broader biological function through hypothesis-driven mechanistic studies by necessity will almost always lag behind the discovery of an RNA sequence element via modern high-throughput sequencing approaches. The absence of evidence, therefore, can be argued not to be the same as evidence of absence of a function.

A related term that critically needs definition is that of a "gene". Generally, a gene is defined as a region of DNA that contains instructions for the function, growth and reproduction of an organism that is genetically inherited by the next generation of the organism. While classic, this definition has become more ambiguous as we have learned more about the complexity of multicellular organisms. First, features of the organism as a whole again are used here as central benchmarks

for defining what constitutes a gene and what does not, excluding ncRNA genes of as-yet unknown function. In addition, a broad interpretation will include the entire non-protein coding regulatory architecture, such as intronic sequences, that instructs the gene to be expressed under a given set of developmental and stress conditions.[13] In addition, genes have fuzzy boundaries since they can overlap and interweave,[14–16] so a broadly inclusive definition of a gene is as a region of DNA that expresses an RNA that may or may not be translated. Second, both geneticists and evolutionary biologists will take conservation of the DNA sequence as an important criterion for defining a gene as functional. This notion certainly holds for housekeeping genes, but runs counter to the observation that the most rapid evolution between the human and other primate genomes occurs outside of protein-coding regions, attributing evolutionary significance to regulatory sequences and RNAs.[17] In addition, single nucleotide alterations between the human genome and that of, for example, chimpanzees constitute ~1.2% of the human genome, with more extended deletions and insertions covering ~3%.[18,19] These numbers are not very far off from the estimated typical ~0.6% difference between an individual's genome and the human reference genome.[20] In fact, the human body is a genomic mosaic in that somatic genome variations accumulate during development in response to both genetic programs, which may differ from tissue to tissue, and environmental stimuli, which are often undetected and irreproducible, helping diversify our phenotypes throughout life.[21] Strikingly, modern single-cell sequencing technologies are revealing that the variation among long-living neurons in a single human results from somatic mutations that accumulate in the brain by different mechanisms and rates during development and ageing, and contribute to neuropsychiatric disorders.[22]

Taken together, these observations point to the idea that the conservation of genes may be overrated. That is, while the biological functions of ncRNAs are rooted in simple biochemical activities such as binding of a specific protein, the non-conserved sequence variations between organisms, tissues and individual cells may hold the key to understanding human evolution and complex traits such as human behavior. Just as one example, an estimated ~2300 human microRNA (miRNA) genes[23] control tissue specific gene expression by binding their Argonaute effector proteins (in humans most notably Ago2) to serve as guides towards a large fraction of human messenger RNAs (mRNAs) in such complex combinatorial fashion that the resulting regulatory network functions are challenging to fully delineate.[24,25] Despite this challenge to exactly define the function of a specific miRNA, the plethora of overlapping miRNA-mediated regulatory effects results in non-linear, emergent properties that empower the plasticity of life and shape much of a human's phenotype with respect to development, physiology, and disease.[26] More generally, regulatory sequences such as miRNAs tend to have more plasticity than those encoding proteins, enabling them to mutate and evolve rapidly under positive selection for adaptive radiation, when organisms diversify into a new environmental niche.[27]

HOW DISTINCT FRAMEWORKS ENGENDER DIVERGENT VIEWS OF ncRNAs

As a biophysical chemist who uses single molecule microscopy, in my respectful, if sometimes antagonistic, electronic discussions with the group of junk DNA champions I naturally adopted the viewpoint that function is a feature of the individual molecule. That is true even though of course only the higher-order interactions between the plethora of diverse molecules inside a cell beget the emergent properties of life.[26] To try to mitigate my own bias, I will refer to the function of an individual molecule as "elemental", in contrast to the "phenotypical" function that is the higher bar used by geneticists and evolutionary biologists for the definition of the function of, for example, a gene. Based on this difference in definition, critics of ENCODE have pointed out that merely showing the existence of a process, such as transcription of an RNA, or an elemental (or rudimentary, minimal) biochemical activity, such as the transcribed RNA binding a protein, is not sufficient to prove phenotypic (or, in their view, functional) significance.[11,28–31] Interestingly, similar controversies exist also at the level of elemental function, for example, where molecular biologists and biochemists who study RNA G-quadruplexes have long debated whether or not these RNA structures, beyond forming in vitro, also exist in the human cell and have a discernable function there.[32,33] Only recently was a specific mechanism discovered by which RNA G-quadruplexes, often found in long ncRNAs, or lncRNAs, downregulate the ability of the polycomb repressive complex 2 (PRC2) to silence chromatin by installing epigenetic marks on histones.[34] This elemental function of G-quadruplexes maintains gene loci in a transcriptionally active state. Clearly, in this case the absence of evidence did not provide evidence of the absence of function.

On some level, then, the ENCODE controversy is a philosophical disagreement on how much of an ncRNA's function has to be discovered before legitimately calling it functional, and not junk. The debate was not helped by using the derogatory term "junk". Yet, embracing the term it has been noted that, as more functions of the majority of the human genome are being discovered, ncRNAs have started to move out of the "junkyard".[35] The field indeed has made rapid strides in this direction, however, since each ncRNA tends to occupy its own niche in biology (as each new class seems to be a "one-off"[35]), mechanistic studies that hunt down each phenotypic function are slow to work through the as many as ~17 000–96 000 human lncRNAs known to date (Figure 1).[12] In addition, new classes of short ncRNAs, defined as under 200 nucleotides in length, beyond miRNAs are also still being discovered.[36] Perhaps skeptics need to cut the RNA field some slack given that we still do not know the function of a large fraction of the under 20 000 protein-coding genes[12] after over 100 years of study, taking the seminal work on enzymes by Leonor Michaelis and Maud Menten in 1913 as the starting point.[36] In fact, the ~20 years since the human genome project first succeeded in revealing all these ncRNAs seems like a comparably short time to have revealed as much as the field has (Figure 1).[12]

It is also instructive to recall the origin of the term "junk" DNA, coined over 50 years ago by Susumu Ohno based on theoretical considerations of mutational load and the large number of repetitive sequences derived from transposable elements that were assumed to be non-functional.[37,38] At the time, it was not yet clear that the phenotypic consequences of mutations in protein-coding housekeeping genes can well be catastrophic, while simultaneously the quantitative trait variations affected by mutations in the regulatory architecture that controls these genes during development and beyond can be beneficial by driving the evolution and diversification of complex organisms.[10,39,40]

Does this mean that a simple agreement to be patient until elemental function can be expanded into phenotypic function is all that is needed to resolve what remains of the ENCODE controversy?

INTO THE WEEDS: DEEP-ROOTED ARGUMENTS SUSTAIN SKEPTICISM TOWARD ncRNA FUNCTION

To better understand and address the strong reaction to the ENCODE project's conclusions we have to consider several deep-rooted, discipline-informed arguments of the skeptics. In the sections that follow, each major argument is enumerated and thoroughly discussed.

There is no phenotype!

The results of the project showed that an elemental function of many ncRNAs is the binding of specific transcription factors.[6] As the example of PRC2 indicates,[34] such binding of specific protein factors can profoundly affect transcription levels, although finding a specific phenotype for such an ncRNA will take time. In fact, genetically deleting an ncRNA may not immediately lead to an observable phenotype.[41] That fact, however, is typical of complex, critical biological functions that often involve subtle functional effects (such as that of individual miRNAs), cell-type- and stress-specific functions, or redundant pathway components to ensure a robust biology.[41,42] Furthermore, a highly transcribed ncRNA can, simply by binding a specific protein factor, deplete it from its pathways to affect a broader function; in the context of miRNAs, this is called the competing endogenous RNA (ceRNA) hypothesis.[43,44] Another potential example is found in metastasis-associated lung adenocarcinoma transcript 1 (MALAT1), which is one of the most abundant lncRNAs, highly conserved among mammals, and thought to regulate alternative pre-mRNA splicing and gene expression at both the transcriptional and post-transcriptional levels in a context-dependent manner.[45,46] While Malat1 knockout mice develop and grow normally, and do not show alterations in alternative splicing, this lncRNA has been associated with the progression of many diseases, in particular cancers,[45] supporting a profound function in the RNA metabolism of the cell, likely mediated by its many associated protein partners.[47] Other lncRNA examples include Neat1, which is involved in placental biology and cognitive function[48] and BC1, implicated in mammalian behavior and synaptic plasticity, learning and memory.[49] In fact, recent high-throughput knockout or knockdown studies of lncRNAs have started to reveal a wide range of cellular and developmental phenotypes,[50–54] with one study identifying almost 500 lncRNAs required for normal cellular proliferation, of which 89% were expressed in just a single cell type.[55] Clearly, absence of a deletion phenotype in one cell type under optimal conditions is not evidence of the absence of function in all cell types under all conditions.

Many lncRNAs are sparse!

It has been noted that fewer than 1000 lncRNAs are present at greater than one copy per cell in the typical human tissue culture cell line, suggesting that the vast majority of ncRNAs is too low in expression to be functional. However, this argument ignores the fact that RNA transcription is highly cell cycle and stress dependent,[56] opening the possibility that many lncRNAs may function primarily in single cells at critical junctures. Classic cell culture and tissue analyses, wherein millions of cells are combined to yield enough material for RNA sequencing, average out any outlier cells, and typically do not probe the vast number of stresses that cells may encounter in the human body. Notably, many stress conditions during disease are associated with a significantly changed transcriptome, including many altered ncRNAs levels.[57] Similarly, many lncRNAs are transcribed highly, but only transiently, from a plethora of enhancer sequences in the genome to control the time-limited expression of distal genes, a mechanism that guides a cell through its developmental stages.[17,58–60] RNA sequencing naturally has not yet been able to probe the transcriptional states of the trillions of cells along their developmental stages during ontogeny of the human body.

A single ncRNA molecule within a crowded cell may only make negligible functional contributions, but of course there are exceptions to this rule, such as genomic DNA molecules. Furthermore, lncRNAs may represent only 1%–4% relative to the ~500 000 mRNA molecules of a single cell, however, other ncRNAs, in particular ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs), outnumber mRNAs by a similar ~40-fold—so are mRNAs non-functional? Finally, lncRNAs often are confined to the nucleus, the root of gene expression, where they only need to bind to one or two specific genomic loci to have a major impact on the expression of nearby genes. Therefore, high expression levels are not needed, in contrast to, for example, a housekeeping mRNA that needs to saturate the cell's large number of ribosomes for efficient, competitive translation.

But the C-value paradox!

Intuitively, one may expect the human genome to be larger than that of onions or amoebae. As it turns out, however, the onion with 16 billion base pairs and the amoeba with 686 billion base pairs have ~5- and ~200-fold larger genomes, respectively, than humans.[11] In fact, genome size varies more than 200 000-fold among eukaryotes.[61] Similarly, even the same genus can show large variation in genome size,

despite similar phenotypes and karyotypes (number and shape of chromosomes). For example, among the family of buttercup plants the DNA content varies up to 80-fold,[62] whereas species of unicellular algae display a 2000-fold divergence in genome size despite similar organismal complexity.[63] This C-value paradox[64] or enigma[61] refers to the violation of the prior assumption of a constant ("C") amount ("value") of DNA per haploid set of chromosomes so that genome length is not generally correlated with the complexity of an organism.[11,28,29] However, modern sequence analyses have found that at least some species, including ancient crop plants such as corn that form allopolyploids (i.e., complete sets of chromosomes from different species) with wide hybridization between variants, provide an opportunity to unite retrotransposons in one genome following a period of divergence, which in turn leads to periodic bursts of retrotransposon and genome expansion.[65] Since the genomes of today represent a recording of sequence alterations over possibly millions of years across many ecological niches, they likely entail evolutionary imprints of ever-varying biological conditions that cannot be fully understood from just examining the end product in extant species.

Case in point—the ~21% of the human genome harboring truncated and full-length long interspersed nuclear elements (LINEs),[66] together with the ~13% of the human genome representing non-autonomous short interspersed nuclear elements (SINEs),[67] appear to be remnants of past retrotransposon expansions that propagate by target primed reverse transcription.[68] When not properly regulated, retrotransposition can result in disease as there are at least 100 examples of known pathologies caused by retroelement insertions, including some cancers and neurological disorders.[69] Notably, one such element, LINE-1 or L1, alone accounts for ~17% of the human genome and is thought to date back 56 million years, surviving evolution. Functionally, LINE-1 retrotransposition has been hypothesized to contribute to somatic mosaicism, genome diversification and genetic innovation.[70,71] Other recently discovered examples for functionality entail the LINE-1 retrotransposon Lx9, which was found to suppress a hyperinflammatory immune response and thereby support host survival upon viral infection;[72] whereas highly repetitive Alu elements appear to control enhancer-promoter looping and thus spatiotemporal gene expression.[73] Estimates of the extent of neutral evolution, or random genetic drift, of the human genome, which are often based on the assumption of the non-functionality of retrotransposon-derived sequences,[27] may have to be adjusted based on these discoveries.

The lack of ncRNA conservation is due to neutral evolution and absence of purifying selection!

Some evolutionary biologists and philosophers have suggested that sequence conservation among genomes should be the primary, or perhaps only, criterion to identify functional genetic elements.[74,75] This line of thinking is based on 50 years of success defining housekeeping and other genes (mostly coding for proteins) based on their sequence conservation. It does not, however, fully acknowledge that evolution does not actually select for sequence conservation. Instead, nature selects for the structure, dynamics and function of a gene, and its transcription and (if protein coding) translation products; as well as for the inertia of the same in pathways in which they are not involved. All that, while residing in the crowded environment of a cell far from equilibrium that is driven primarily by the relative kinetics of all possible interactions.[26] Given the complexity and time dependence of the cellular environment and its environmental exposures, it is currently impossible to fully understand the emergent properties of life based on simple cause-and-effect reasoning.[26] Consequently, the estimates of the fraction of the human genome that carries function is still being upward corrected, with the best estimate of confirmed ncRNAs now having surpassed protein-coding genes,[12] although so far only 10%–40% of these ncRNAs have been shown to have a function in, for example, cell morphology and proliferation, under at least one set of defined conditions.[50–55] Muddying the waters even more, some ncRNAs encode non-canonical micropeptides less than 100 amino acids in length with diverse and important biological roles,[76,77] which can result in dual functions in both coding and regulation.[78] Conversely, protein-coding genes can express functional lncRNAs upon alternative splicing.[79,80] Finally, the example of the lncRNA component of telomerase—the enzyme that maintains the protected chromosome ends, or telomers, indispensable for genome integrity—illustrates that widely varying RNA sizes and sequences can all be compatible with a structural topology that is conserved functionally from yeast to mammals.[81]

One example for the scope and limitations of using conservation for defining function are riboswitches, highly structured RNA motifs that are primarily found in bacteria and each bind a specific ligand of cellular importance such as an amino acid, metabolite or elemental ion.[82] Once the ligand is recognized by the so-called aptamer domain, the interlaced downstream expression platform of the riboswitch changes conformation to modulate the likelihood for transcription and/or translation (Figure 2A).[83] Over the past 20 years, Breaker and coworkers have had great success in discovering novel riboswitch classes through their aptamer's sequence conservation among bacteria of known genomic sequence (Figure 2B), leading so far to over 55 distinct riboswitch classes with at least some bioinformatic, genetic, or biochemical data to validate their functions.[84] Predictions based on power law projection suggest, however, that up to 28 000 additional, less well-represented riboswitch classes may still be hidden in the genomes of bacterial species that have fewer representatives or are more difficult to culture.[84] This projection illustrates one limitation of using conservation as a basis for discovering ncRNAs of defined function, namely the limited statistical power to capture less widely distributed sequences than housekeeping genes.

A second limitation of a sequence-based conservation analysis of function is illustrated by recent insights from the functional probing of riboswitches. RNA structure, and hence dynamics and function, is generally established co-transcriptionally, as evident from, for example, bacterial ncRNAs including riboswitches[85–90] and ribosomal RNAs,[91,92] as well as the co-transcriptional alternative splicing of eukaryotic pre-mRNAs, responsible for the important, vast diversification of the human proteome across ~200 cell types by excision of

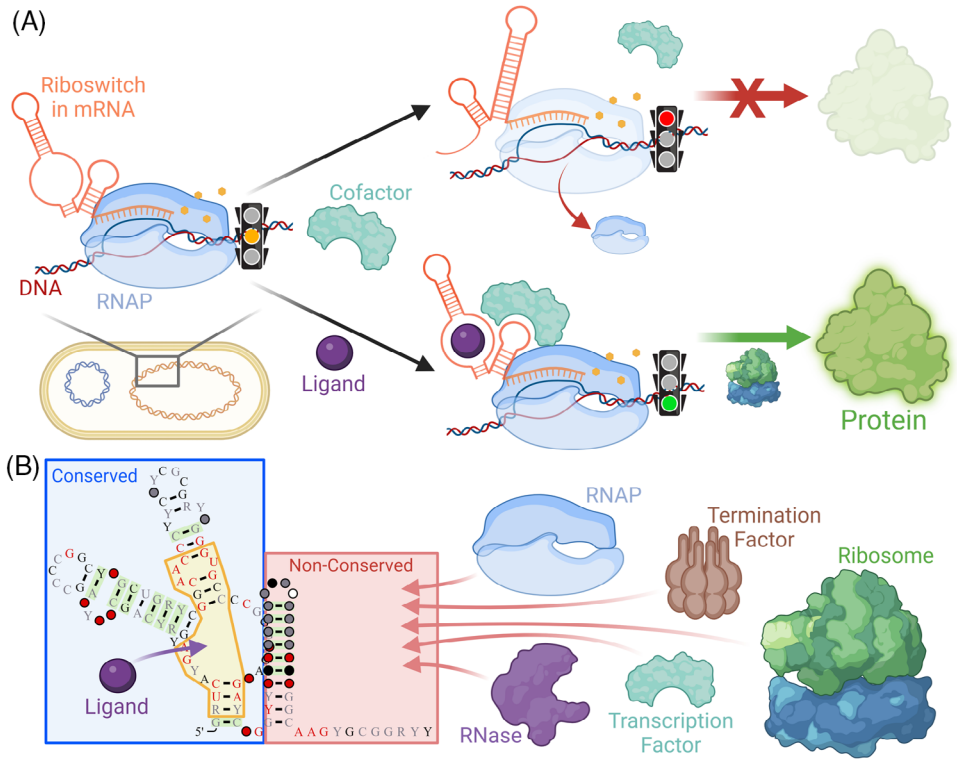

FIGURE 2 Riboswitch mechanism, function and conservation. (A) Riboswitches are highly structured RNA motifs embedded in the 5′ untranslated regions of many bacterial mRNAs, where they can enhance or suppress gene expression as they co-transcriptionally fold to bind a small-molecule or elemental ion ligand. One such mechanism involves the modulation of transcription yield by RNA polymerase (RNAP), whereas others more directly alter the likelihood of mRNA translation into protein. (B) The upstream aptamer region binds the ligand, rendering the core segment that forms the binding pocket (yellow box), as well as flanking architectural segments (blue box), highly conserved.[112,113] By comparison, the downstream expression platform shows much less conservation, most likely because it functionally interacts with many protein effectors that are idiosyncratic to the specific bacterium. Created with BioRender.com.

varying ncRNA introns.[93] In the latter case, it is becoming increasingly clear that splicing regulation involves multiple layers synergistically controlled by the splicing machinery, transcription process, and chromatin structure.[93] In the case of riboswitches, the interactions of the ncRNA with its multiple protein effectors functionally engage essentially all of its nucleotides, sequence-conserved or not, including those responsible for affecting specific distances between other functional elements.[85,89] Consequently, the expression platform—equally important for the gene regulatory function as the conserved aptamer domain—tends to be far less conserved,[83] because it interacts with the idiosyncratic gene expression machinery of the bacterium (Figure 2B). Consequently, taking a riboswitch out of this native environment into a different cell type for synthetic biology purposes has been notoriously challenging.[94] These examples of a holistic functioning of ncRNAs in their species-specific cellular context lay bare the limited power of pure sequence conservation in predicting all functionally relevant nucleotides.

It seems plausible, if not likely, that the genomes of individual species similarly work holistically. In such a model (Figure 3), non-conserved ncRNAs would simply be idiosyncratic imprints of past chromosome or gene duplications, viral infections, transposition events etc., that were accumulated over millennia and at present have been integrated into the overall functioning of the organism. In this context, it is important to recall that the generation time of humans is measured in tens of years, in contrast to the tens of minutes for bacteria, leading to vastly different timescales of evolution, which renders the detection of human genome conservation much more challenging (and the predictive power of such conservation so much weaker). Notably, however, the human body contains ~30 trillion cells, of which ~300 billion are destroyed daily since they have stopped functioning properly.[95] Combined with the observed genomic mosaicism of our body's cells, enhanced by ncRNAs, these observations imply that the sequenced genome of an individual human being itself is the outcome of relentless internal evolutionary pressures to conform to the needs of all bodily functions. Such considerations are absent from arguments invoking the inability of natural, purifying selection to weed out mildly deleterious mutations among the relatively small size (compared to the human genome) of the human population, which regards only the entire organism as the target of evolution, therefore likely inaccurately suggesting that neutral evolution and drift give rise to the many genomic ncRNAs.[75]

Furthermore, the error frequency of organismal genome replication is generally inversely proportional to their genome length, from humans to bacteria and even viruses, an observation termed Drake's

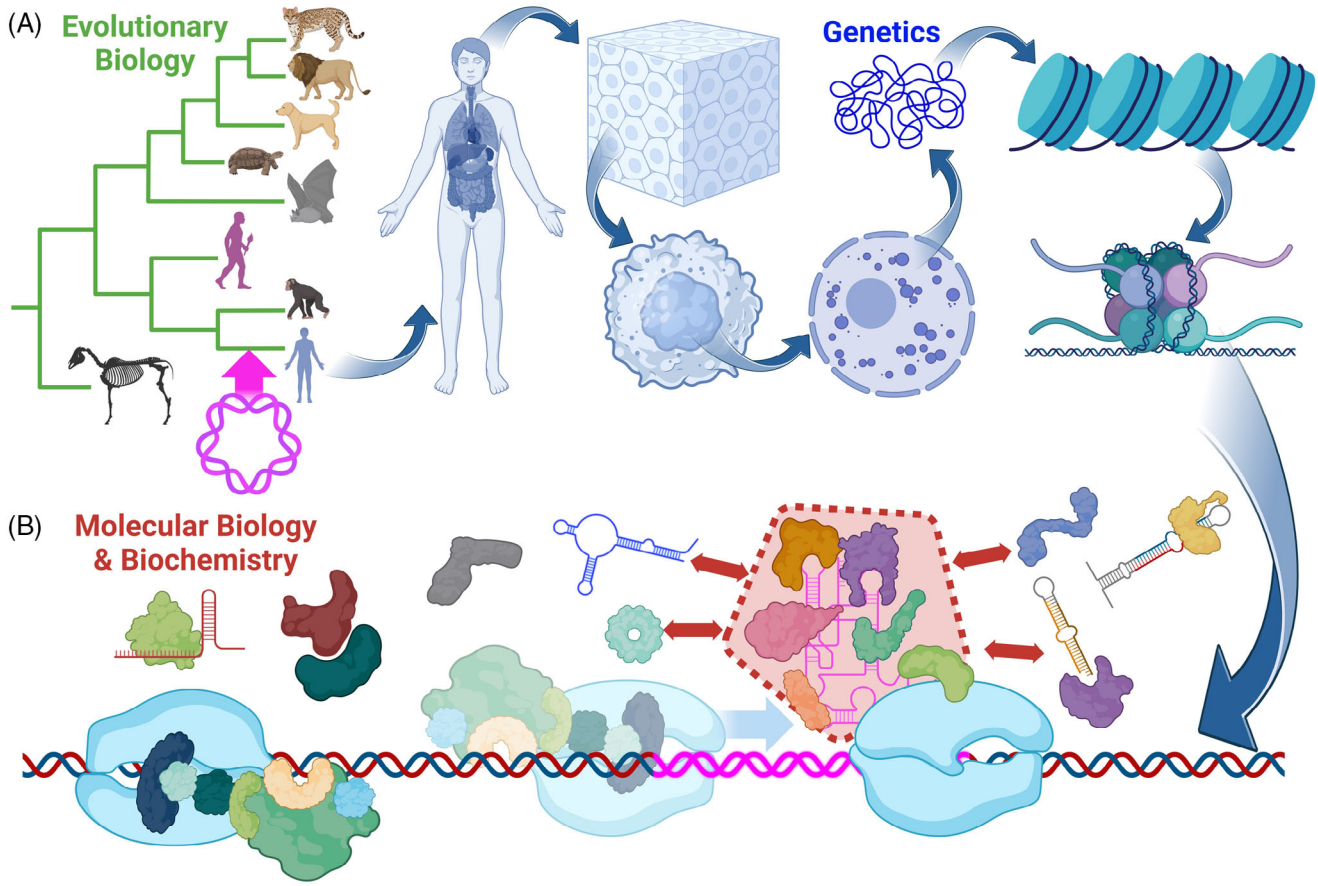

FIGURE 3 Human evolutionary biology, genetics, molecular biology and biochemistry complement each other. (A) The DNA of extant species carries numerous evolutionary imprints of past genetic events that include, for example, a parasitic DNA (pink) transpositioning into the genome and expanding it. While evolutionary biology studies the organismal level, genetics focuses on the genomic DNA tightly packaged within the nucleus of the organism's cells. (B) Reaching into the nanoscale, molecular biology and biochemistry seek to understand the molecular mechanism of RNAs and their protein partners, where a transposon (pink) may give rise to an RNA transcript that acquires specific functions by forming a complex network of interactions (red dashed box). A functional pathway is characterized by interactions that are kinetically more stable than the alternatives (double-arrows).[26] Created with BioRender.com.

rule and likely caused by genomes operating near their maximum informational storage capacity and mutation rate.[96] One insight from the COVID-19 epidemic has been the sheer speed with which viruses that produce ~1 trillion viral particles in a human body can evolve.[97] We have to assume that such evolution also occurs at the level of the human body and genome, just at a comparably glacial pace that is difficult to observe. It is therefore plausible to suggest that the human genome is selected equally on a per-nucleotide basis, just as the viral genome, even if we do not yet understand the holistic function(s) of each sequence element.

Transcription factors bind to random genome sequences!

One final argument of geneticists is the suggestion that many random pieces of DNA can promote transcription by recruiting transcription factors locally.[75] However, rarely is the transient binding of a single transcription factor sufficient to recruit an RNA polymerase molecule to a transcription site. Rather, a combinatorial cooperation between cis-regulatory sequence elements in the genome, trans-acting transcription factors and signaling molecules, and gene-distal, but cis-acting ncRNA enhancer transcripts is needed to initiate directional transcription events that govern the tissue-specific, spatiotemporally controlled expression dynamics of genes (Figure 3B).[98] Consequently, the still poorly understood constant spatial reorganization of chromosomes in the densely packed nucleus—guided by a plethora of enhancer lncRNAs (Figure 1)—is both the result of and prerequisite for correct transcriptional programs that allow for the plasticity and adaptability of the semi-autonomous gene expression observed in each individual cell of a multicellular eukaryotic organism.[99]

In fact, a majority of lncRNAs appear to arise from genetic loci that act as enhancers that dynamically and adaptably control assembly of the transcription machinery for the spatiotemporal expression of specific genes during the development of multicellular organisms.[100,101] These enhancer RNAs, or eRNAs, exhibit cell-type specific expression patterns so that they are underrepresented in many RNA sequencing datasets, and remain largely uncatalogued and of poorly defined

function.[101] Nevertheless, their large number, estimated to be in the hundreds of thousands,[100] contribution to super-enhancers that involve the rapid, transient transcription of eRNAs and their phase separation into eRNA-protein granules, fast evolution under selection for adaptive radiation, and the fact that the vast majority of single nucleotide polymorphisms in the human genome with causal links to complex diseases map to eRNA genes lend them functional relevance.[101–106] Furthermore, the observation that most eukaryotic transcription factor proteins have confusing and enigmatic functions, as well as ill-defined structures, suggests that they have been interpreted using a conceptual framework that misses cooperating eRNAs as a critical link.[40]

In this modern view of eukaryotic gene expression, only those transcription events will occur that are sufficiently robustly proofread by a sequence of kinetically controlled, reversible assembly events that have to enhance each other and outcompete a vast number of possible alternative events.[26] In the resulting holistic model (Figure 3B), the significant number of defined transcripts detected by ENCODE then become a signature of select cellular processes that are allowed to proceed among a much larger number of possible transcripts. While we still do not understand the phenotypic functions of a majority of these primarily non-protein coding RNAs, we have to assume that the likelihood is high for eventually finding many functions that evolution has preserved across the many generations of individuals. Collectively, all these organisms and their cells were exposed to a vast array of rapidly changing environmental conditions, which imprinted on their genome and were inherited by the following generations.

CONCLUSION: HOW TO RECONCILE SCIENTIFIC FIELDS

Science thrives from integrating diverse viewpoints—the more diverse the team, the better the science.[107] Previous attempts at reconciling the divergent assessments about the functional significance of the large number of ncRNAs transcribed from most of the human genome by pointing out that the scientific approaches of geneticists, evolutionary biologists and molecular biologists/biochemists provide complementary information[42] was met with further skepticism.[74] Perhaps a first step toward reconciliation, now that ncRNAs appear to increasingly leave the junkyard,[35] would be to substitute the needlessly categorical and derogative word RNA (or DNA) "junk" for the more agnostic and neutral term "ncRNA of unknown phenotypic function", or "$ncRNA_{upf}$". After all, everyone seems to agree that the controversy mostly stems from divergent definitions of the term "function",[42,74] which each scientific field necessarily defines based on its own need for understanding the molecular and mechanistic details of a system (Figure 3). In addition, "of unknown phenotypic function" honors the null hypothesis that no function manifesting in a phenotype is currently known, but may still be discovered. It also allows for the possibility that, in the end, some transcribed ncRNAs may never be assigned a bona fide function.

Most bioscientists will also agree that we need to continue advancing from simply cataloging non-coding regions of the human genome toward characterizing ncRNA functions, both elementally and phenotypically, an endeavor of great challenge that requires everyone's input. Solving the enigma of human gene expression, so intricately linked to the regulatory roles of ncRNAs, holds the key to devising personalized medicines to treat most, if not all, human diseases, rendering the stakes high, and unresolved disputes counterproductive.[108] The fact that newly ascendant RNA therapeutics that directly interface with cellular RNAs seem to finally show us a path to success in this challenge[109] only makes the need for deciphering ncRNA function more urgent. Succeeding in this goal would finally fulfill the promise of the human genome project after it revealed so much non-protein coding sequence (Figure 1). As a side effect, it may make updating Wikipedia and encyclopedia entries less controversial.

ACKNOWLEDGMENTS

N.G.W. appreciates support from NIH R35 grant GM131922, as well as thanks two anonymous reviewers for their very valuable feedback and input.

CONFLICT OF INTEREST STATEMENT

The author declares no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Nils G. Walter https://orcid.org/0000-0002-7301-1275

REFERENCES

1. Students, C. (2022). Chem 455–505: Repeated sequence (DNA). : Wikipedia. https://en.wikipedia.org/wiki/Repeated_sequence_(DNA)
2. Moran, L. A. (2022). University of Michigan biochemistry students edit Wikipedia. *Internet Blog*, https://sandwalk.blogspot.com/2022/12/university-of-michigan-biochemistry.html
3. Consortium, E. P., Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E. L., Freese, P., Gorkin, D. U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., … Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, *583*, 699–710.
4. Consortium, E. P., Snyder, M. P., Gingeras, T. R., Moore, J. E., Weng, Z., Gerstein, M. B., Ren, B., Hardison, R. C., Stamatoyannopoulos, J. A., Graveley, B. R., Feingold, E. A., Pazin, M. J., Pagan, M., Gilchrist, D. A., Hitz, B. C., Cherry, J. M., Bernstein, B. E., Mendenhall, E. M., Zerbino, D. R., … Myers, R. M. (2020). Perspectives on ENCODE. *Nature*, *583*, 693–698.
5. Zhang, X. O., Gingeras, T. R., & Weng, Z. (2019). Genome-wide analysis of polymerase III-transcribed Alu elements suggests cell-type-specific enhancer function. *Genome Research*, *29*, 1402–1414.
6. Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J. Y., Cody, N. A. L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L. P. B., Bergalet, J., Duff, M. O., Garcia, K. E., Gelboin-Burkhart, C., … Yeo, G. W. (2020).

A large-scale binding and functional map of human RNA-binding proteins. *Nature*, *583*, 711–719.
7. Consortium, E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, *306*, 636–640.
8. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., & Stamatoyannopoulos, J. A. (2020). Global reference mapping of human transcription factor footprints. *Nature*, *583*, 729–736.
9. Mattick, J. S. (2023). A Kuhnian revolution in molecular biology: Most genes in complex organisms express regulatory RNAs. *BioEssays*, *45*, e2300080.
10. Mattick, J. S., & Dinger, M. E. (2013). The extent of functionality in the human genome. *The HUGO Journal*, *7*, 2.
11. Malpas, J. (2016). The ENCODE Project and the ENCODE Controversy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University: The Metaphysics Research Lab, Philosophy Department.
12. Amaral, P., Carbonell-Sala, S., De La Vega, F. M., Faial, T., Frankish, A., Gingeras, T., Guigo, R., Harrow, J. L., Hatzigeorgiou, A. G., Johnson, R., Murphy, T. D., Pertea, M., Pruitt, K. D., Pujar, S., Takahashi, H., Ulitsky, I., Varabyou, A., Wells, C. A., Yandell, M., … Salzberg, S. L. (2023). The status of the human gene catalogue. *Nature*, *622*, 41–47.
13. Alberts, B., Johnson, A., & Lewis, J. (2002). *Molecular biology of the cell New York*. Garland Science.
14. Mattick, J. S. (2003). Challenging the dogma: The hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays*, *25*, 930–939.
15. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R., Zavolan, M., Davis, M. J., Wilming, L. G., … Hayashizaki, Y. (2005). The transcriptional landscape of the mammalian genome. *Science*, *309*, 1559–1563.
16. Kapranov, P., Willingham, A. T., & Gingeras, T. R. (2007). Genome-wide transcription and the implications for genomic organization. *Nature Reviews Genetics*, *8*, 413–423.
17. Mangan, R. J., Alsina, F. C., Mosti, F., Sotelo-Fonseca, J. E., Snellings, D. A., Au, E. H., Carvalho, J., Sathyan, L., Johnson, G. D., Reddy, T. E., Silver, D. L., & Lowe, C. B. (2022). Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell*, *185*, 4587–4603.e23.
18. Chimpanzee, S., & Analysis, C. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*, 69–87.
19. Suntsova, M. V., & Buzdin, A. A. (2020). Differences between human and chimpanzee genomes and their implications in gene expression, protein functions and biochemical properties of the two species. *BMC Genomics*, *21*, 535.
20. Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74.
21. Astolfi, P. A., Salamini, F., & Sgaramella, V. (2010). Are we genomic mosaics? Variations of the genome of somatic cells can contribute to diversify our phenotypes. *Current Genomics*, *11*, 379–386.
22. Bizzotto, S., & Walsh, C. A. (2022). Genetic mosaicism in the human brain: Fom lineage tracing to neuropsychiatric disorders. *Nature Reviews Neuroscience*, *23*, 275–286.
23. Alles, J., Fehlmann, T., Fischer, U., Backes, C., Galata, V., Minet, M., Hart, M., Abu-Halima, M., Grasser, F. A., Lenhof, H. P., Keller, A., & Meese, E. (2019). An estimate of the total number of true human miRNAs. *Nucleic Acids Research*, *47*, 3353–3364.
24. O'Brien, J., Hayder, H., Zayed, Y., & Peng, C. (2018). Overview of MicroRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology*, *9*, 402.
25. Peng, Y., & Croce, C. M. (2016). The role of MicroRNAs in human cancer. *Signal Transduction and Targeted Therapy*, *1*, 15004.
26. Walter, N. G. (2019). Biological pathway specificity in the cell-does molecular diversity matter? *BioEssays*, *41*, e1800244.
27. Pheasant, M., & Mattick, J. S. (2007). Raising the estimate of functional human sequences. *Genome Research*, *17*, 1245–1253.
28. Eddy, S. R. (2012). The C-value paradox, junk DNA and ENCODE. *Current Biology*, *22*, R898–R899.
29. Doolittle, W. F. (2013). Is junk DNA bunk? A critique of ENCODE. *PNAS*, *110*, 5294–5300.
30. Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A., & Elhaik, E. (2013). On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution*, *5*, 578–590.
31. Niu, D. K., & Jiang, L. (2013). Can ENCODE tell us how much junk DNA we carry in our genome? *Biochemical and Biophysical Research Communications*, *430*, 1340–1343.
32. Guo, J. U., & Bartel, D. P. (2016). RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science*, *353*, aaf5371.
33. Varshney, D., Spiegel, J., Zyner, K., Tannahill, D., & Balasubramanian, S. (2020). The regulation and functions of DNA and RNA G-quadruplexes. *Nature Reviews Molecular Cell Biology*, *21*, 459–474.
34. Song, J., Gooding, A. R., Hemphill, W. O., Love, B. D., Robertson, A., Yao, L., Zon, L. I., North, T. E., Kasinath, V., & Cech, T. R. (2023). Structural basis for inactivation of PRC2 by G-quadruplex RNA. *Science*, *381*, 1331–1337.
35. Marx, V. (2022). How noncoding RNAs began to leave the junkyard. *Nature Methods*, *19*, 1167–1170.
36. Michaelis, L., & Menten, M. L. (1913). Die Kinetik der Invertinwirkung. *Biochemische Zeitschrift*, *49*, 333–369.
37. Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven Symposia in Biology*, *23*, 366–370.
38. Ohno, S. (1972). Simplicity of mammalian regulatory systems. *Developmental Biology*, *27*, 131–136.
39. Carroll, S. B. (2008). Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell*, *134*, 25–36.
40. Mattick, J., & Amaral, P. (2022). *RNA, the Epicenter of Genetic Information: A new understanding of molecular biology*. CRC Press, Abingdon (UK).
41. Gao, F., Cai, Y., Kapranov, P., & Xu, D. (2020). Reverse-genetics studies of lncRNAs-what we have learnt and paths forward. *Genome Biology*, *21*, 93.
42. Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., Dunham, I., Elnitski, L. L., Farnham, P. J., Feingold, E. A., Gerstein, M., Giddings, M. C., Gilbert, D. M., Gingeras, T. R., Green, E. D., … Hardison, R. C. (2014). Defining functional DNA elements in the human genome. *PNAS*, *111*, 6131–6138.
43. Salmena, L., Poliseno, L., Tay, Y., Kats, L., & Pandolfi, P. P. (2011). A ceRNA hypothesis: The Rosetta Stone of a hidden RNA language? *Cell*, *146*, 353–358.
44. Qu, J., Li, M., Zhong, W., & Hu, C. (2015). Competing endogenous RNA in cancer: A new pattern of gene expression regulation. *International Journal of Clinical and Experimental Medicine*, *8*, 17110–17116.
45. Sun, Y., & Ma, L. (2019). New insights into long non-coding RNA MALAT1 in cancer and metastasis. *Cancers*, *11*, 216.
46. Arun, G., Aggarwal, D., & Spector, D. L. (2020). MALAT1 long non-coding RNA: Functional implications. *Noncoding RNA*, *6*, 22.
47. Scherer, M., Levin, M., Butter, F., & Scheibe, M. (2020). Quantitative proteomics to identify nuclear RNA-binding proteins of Malat1. *International Journal of Molecular Sciences*, *21*, 1166.
48. Butler, A. A., Johnston, D. R., Kaur, S., & Lubin, F. D. (2019). Long noncoding RNA NEAT1 mediates neuronal histone methylation and age-related memory impairment. *Science Signaling*, *12*, eaaw9277.

49. Chung, A., Dahan, N., Alarcon, J. M., & Fenton, A. A. (2017). Effects of regulatory BC1 RNA deletion on synaptic plasticity, learning, and memory. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *24*, 646–649.
50. Ramilowski, J. A., Yip, C. W., Agrawal, S., Chang, J. C., Ciani, Y., Kulakovskiy, I. V., Mendez, M., Ooi, J. L. C., Ouyang, J. F., Parkinson, N., Petri, A., Roos, L., Severin, J., Yasuzawa, K., Abugessaisa, I., Akalin, A., Antonov, I. V., Arner, E., Bonetti, A., … Carninci, P. (2020). Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Research*, *30*, 1060–1072.
51. Horlbeck, M. A., Liu, S. J., Chang, H. Y., Lim, D. A., & Weissman, J. S. (2020). Fitness effects of CRISPR/Cas9-targeting of long noncoding RNA genes. *Nature Biotechnology*, *38*, 573–576.
52. Cao, H., Xu, D., Cai, Y., Han, X., Tang, L., Gao, F., Qi, Y., Cai, D., Wang, H., Ri, M., Antonets, D., Vyatkin, Y., Chen, Y., You, X., Wang, F., Nicolas, E., & Kapranov, P. (2021). Very long intergenic non-coding (vlinc) RNAs directly regulate multiple genes in cis and trans. *BMC Biology*, *19*, 108.
53. Statello, L., Guo, C. J., Chen, L. L., & Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology*, *22*, 96–118.
54. Yip, C. W., Hon, C. C., Yasuzawa, K., Sivaraman, D. M., Ramilowski, J. A., Shibayama, Y., Agrawal, S., Prabhu, A. V., Parr, C., Severin, J., Lan, Y. J., Dostie, J., Petri, A., Nishiyori-Sueki, H., Tagami, M., Itoh, M., Lopez-Redondo, F., Kouno, T., Chang, J. C., … Shin, J. W. (2022). Antisense-oligonucleotide-mediated perturbation of long non-coding RNA reveals functional features in stem cells and across cell types. *Cell Reports*, *41*, 111893.
55. Liu, S. J., Horlbeck, M. A., Cho, S. W., Birk, H. S., Malatesta, M., He, D., Attenello, F. J., Villalta, J. E., Cho, M. Y., Chen, Y., Mandegar, M. A., Olvera, M. P., Gilbert, L. A., Conklin, B. R., Chang, H. Y., Weissman, J. S., & Lim, D. A. (2017). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*, *355*, aah7111.
56. Zangouei, A. S., Zangoue, M., Taghehchian, N., Zangooie, A., Rahimi, H. R., Saburi, E., Alavi, M. S., & Moghbeli, M. (2023). Cell cycle related long non-coding RNAs as the critical regulators of breast cancer progression and metastasis. *Biological Research*, *56*, 1.
57. Slack, F. J., & Chinnaiyan, A. M. (2019). The role of non-coding RNAs in oncology. *Cell*, *179*, 1033–1055.
58. Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., & Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*, 182–187.
59. Orom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R., & Shiekhattar, R. (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell*, *143*, 46–58.
60. Arner, E., Daub, C. O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drablos, F., Lennartsson, A., Ronnerblad, M., Hrydziuszko, O., Vitezic, M., Freeman, T. C., Alhendi, A. M., Arner, P., Axton, R., Baillie, J. K., Beckhouse, A., Bodega, B., Briggs, J., Brombacher, F., … Hayashizaki, Y. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, *347*, 1010–1014.
61. Gregory, T. R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews of the Cambridge Philosophical Society*, *76*, 65–101.
62. Rothfels, K., Sexsmith, E., Heimburger, M., & Krause, M. O. (1966). Chromosome size and DNA content of species of anemone L. and related genera (Ranunculaceae). *Chromosoma*, *20*, 54–74.
63. Kapraun, D. F. (2005). Nuclear DNA content estimates in multicellular green, red and brown algae: Phylogenetic considerations. *Annals of Botany*, *95*, 7–44.
64. Thomas Jr, C. A. (1971). The genetic organization of chromosomes. *Annual Review of Genetics*, *5*, 237–256.
65. Sharma, A., Schneider, K. L., & Presting, G. G. (2008). Sustained retrotransposition is mediated by nucleotide deletions and interelement recombinations. *PNAS*, *105*, 15470–15474.
66. Sheen, F. M., Sherry, S. T., Risch, G. M., Robichaux, M., Nasidze, I., Stoneking, M., Batzer, M. A., & Swergold, G. D. (2000). Reading between the LINEs: Human genomic variation induced by LINE-1 retrotransposition. *Genome Research*, *10*, 1496–1508.
67. Ishak, C. A., & De Carvalho, D. D. (2020). Reactivation of endogenous retroelements in cancer development and therapy. *Annual Review of Cancer Biology*, *4*, 159–176.
68. Cost, G. J., Feng, Q., Jacquier, A., & Boeke, J. D. (2002). Human L1 element target-primed reverse transcription in vitro. *Embo Journal*, *21*, 5899–5910.
69. Solyom, S., & Kazazian Jr., H. H. (2012). Mobile elements in the human genome: Implications for disease. *Genome Medicine*, *4*, 12.
70. Richardson, S. R., Morell, S., & Faulkner, G. J. (2014). L1 retrotransposons and somatic mosaicism in the brain. *Annual Review of Genetics*, *48*, 1–27.
71. Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, *10*, 691–703.
72. Bartonicek, N., Rouet, R., Warren, J., Loetsch, C., Rodriguez, G. S., Walters, S., Lin, F., Zahra, D., Blackburn, J., Hammond, J. M., Reis, A. L. M., Deveson, I. W., Zammit, N., Zeraati, M., Grey, S., Christ, D., Mattick, J. S., Chtanova, T., Brink, R., … King, C. (2022). The retroelement Lx9 puts a brake on the immune response to virus infection. *Nature*, *608*, 757–765.
73. Liang, L., Cao, C., Ji, L., Cai, Z., Wang, D., Ye, R., Chen, J., Yu, X., Zhou, J., Bai, Z., Wang, R., Yang, X., Zhu, P., & Xue, Y. (2023). Complementary Alu sequences mediate enhancer-promoter selectivity. *Nature*, *619*, 868–875.
74. Doolittle, W. F., Brunet, T. D., Linquist, S., & Gregory, T. R. (2014). Distinguishing between "function" and "effect" in genome biology. *Genome Biology and Evolution*, *6*, 1234–1237.
75. Palazzo, A. F., & Lee, E. S. (2015). Non-coding RNA: What is functional and what is junk? *Frontiers in Genetics*, *6*, 2.
76. Mudge, J. M., Ruiz-Orera, J., Prensner, J. R., Brunet, M. A., Calvet, F., Jungreis, I., Gonzalez, J. M., Magrane, M., Martinez, T. F., Schulz, J. F., Yang, Y. T., Alba, M. M., Aspden, J. L., Baranov, P. V., Bazzini, A. A., Bruford, E., Martin, M. J., Calviello, L., Carvunis, A. R., … van Heesch, S. (2022). Standardized annotation of translated open reading frames. *Nature Biotechnology*, *40*, 994–999.
77. Wright, B. W., Yi, Z., Weissman, J. S., & Chen, J. (2022). The dark proteome: Translation from noncanonical open reading frames. *Trends in Cell Biology*, *32*, 243–258.
78. Dinger, M. E., Gascoigne, D. K., & Mattick, J. S. (2011). The evolution of RNAs with multiple functions. *Biochimie*, *93*, 2013–2018.
79. Williamson, L., Saponaro, M., Boeing, S., East, P., Mitter, R., Kantidakis, T., Kelly, G. P., Lobley, A., Walker, J., Spencer-Dene, B., Howell, M., Stewart, A., & Svejstrup, J. Q. (2017). UV irradiation induces a non-coding RNA that functionally opposes the protein encoded by the same gene. *Cell*, *168*, 843–855.e13.
80. Grelet, S., Link, L. A., Howley, B., Obellianne, C., Palanisamy, V., Gangaraju, V. K., Diehl, J. A., & Howe, P. H. (2017). A regulated PNUTS mRNA to lncRNA splice switch mediates EMT and tumour progression. *Nature Cell Biology*, *19*, 1105–1115.
81. Podlevsky, J. D., & Chen, J. J. (2016). Evolutionary perspectives of telomerase RNA structure and function. *RNA Biology*, *13*, 720–732.
82. Breaker, R. R. (2022). The biochemical landscape of riboswitch ligands. *Biochemistry*, *61*, 137–149.
83. Jones, C. P., & Ferre-D'Amare, A. R. (2017). Long-range interactions in riboswitch control of gene expression. *Annual Review of Biophysics*, *46*, 455–481.
84. Kavita, K., & Breaker, R. R. (2023). Discovering riboswitches: The past and the future. *Trends in Biochemical Sciences*, *48*, 119–141.

85. Widom, J. R., Nedialkov, Y. A., Rai, V., Hayes, R. L., Brooks 3rd, C. L., Artsimovitch, I., & Walter, N. G. (2018). Ligand modulates cross-coupling between riboswitch folding and transcriptional pausing. *Molecular Cell*, *72*, 541–552.e6.
86. Chatterjee, S., Chauvier, A., Dandpat, S. S., Artsimovitch, I., & Walter, N. G. (2021). A translational riboswitch coordinates nascent transcription-translation coupling. *PNAS*, *118*, e2023426118.
87. Scull, C. E., Dandpat, S. S., Romero, R. A., & Walter, N. G. (2021). Transcriptional riboswitches integrate timescales for bacterial gene expression control. *Frontiers in Molecular Biosciences*, *7*, 607158.
88. Yadav, R., Widom, J. R., Chauvier, A., & Walter, N. G. (2022). An anionic ligand snap-locks a long-range interaction in a magnesium-folded riboswitch. *Nature Communications*, *13*, 207.
89. Chauvier, A., Porta, J. C., Deb, I., Ellinger, E., Meze, K., Frank, A. T., Ohi, M. D., & Walter, N. G. (2023). Structural basis for control of bacterial RNA polymerase pausing by a riboswitch and its ligand. *Nature Structural & Molecular Biology*, *30*, 902–913.
90. Ellinger, E., Chauvier, A., Romero, R. A., Liu, Y., Ray, S., & Walter, N. G. (2023). Riboswitches as therapeutic targets: Promise of a new era of antibiotics. *Expert Opinion on Therapeutic Targets*, *27*, 433–445.
91. Rodgers, M. L., & Woodson, S. A. (2019). Transcription Increases the Cooperativity of Ribonucleoprotein Assembly. *Cell*, *179*, 1370–1381.e12.
92. Rodgers, M. L., & Woodson, S. A. (2021). A roadmap for rRNA folding and assembly during transcription. *Trends in Biochemical Sciences*, *46*, 889–901.
93. Wang, N., Hu, Y., & Wang, Z. (2023). Regulation of alternative splicing: Functional interplay with epigenetic modifications and its implication to cancer. *Wiley Interdisciplinary Reviews RNA*, e1815.
94. Drogalis, L. K., & Batey, R. T. (2020). Requirements for efficient ligand-gated co-transcriptional switching in designed variants of the B. subtilis pbuE adenine-responsive riboswitch in E. coli. *PLoS ONE*, *15*, e0243155.
95. Sender, R., & Milo, R. (2021). The distribution of cellular turnover in the human body. *Nature Medicine*, *27*, 45–48.
96. Shadrin, A. A., & Parkhomchuk, D. V. (2014). Drake's rule as a consequence of approaching channel capacity. *Die Naturwissenschaften*, *101*, 939–954.
97. Yewdell, J. W. (2021). Antigenic drift: Understanding COVID-19. *Immunity*, *54*, 2681–2687.
98. Panigrahi, A., & O'Malley, B. W. (2021). Mechanisms of enhancer action: The known and the unknown. *Genome Biology*, *22*, 108.
99. Finn, E. H., Pegoraro, G., Brandao, H. B., Valton, A. L., Oomen, M. E., Dekker, J., Mirny, L., & Misteli, T. (2019). Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell*, *176*, 1502–1515.e10.
100. Jin, W., Jiang, G., Yang, Y., Yang, J., Yang, W., Wang, D., Niu, X., Zhong, R., Zhang, Z., & Gong, J. (2022). Animal-eRNAdb: A comprehensive animal enhancer RNA database. *Nucleic Acids Research*, *50*, D46–D53.
101. Mattick, J. S. (2023). Enhancers are genes that express organizational RNAs. *Frontiers in RNA Research*, *1*, 1–11.
102. Arnold, P. R., Wells, A. D., & Li, X. C. (2019). Diversity and emerging roles of enhancer RNA in regulation of gene expression and cell fate. *Frontiers in Cell and Developmental Biology*, *7*, 377.
103. Sartorelli, V., & Lauberth, S. M. (2020). Enhancer RNAs are an important regulatory layer of the epigenome. *Nature Structural & Molecular Biology*, *27*, 521–528.
104. Hou, T. Y., & Kraus, W. L. (2021). Spirits in the Material World: Enhancer RNAs in Transcriptional Regulation. *Trends in Biochemical Sciences*, *46*, 138–153.
105. Harrison, L. J., & Bose, D. (2022). Enhancer RNAs step forward: New insights into enhancer function. *Development (Cambridge, England)*, *149*, dev200398.
106. Chen, Q., Zeng, Y., Kang, J., Hu, M., Li, N., Sun, K., & Zhao, Y. (2023). Enhancer RNAs in transcriptional regulation: Recent insights. *Frontiers in Cell and Developmental Biology*, *11*, 1205540.
107. Phillips, K. W. (2014). How diversity works. *Scientific American*, *311*, 42–47.
108. Mathur, S., & Sutton, J. (2017). Personalized medicine could transform healthcare. *Biomedical. Report*, *7*, 3–5.
109. Hastings, M. L., & Krainer, A. R. (2023). RNA therapeutics. *Rna*, *29*, 393–395.
110. Fields, C., Adams, M. D., White, O., & Venter, J. C. (1994). How many genes in the human genome? *Nature Genetics*, *7*, 345–346.
111. Mattick, J. S., Amaral, P. P., Carninci, P., Carpenter, S., Chang, H. Y., Chen, L. L., Chen, R., Dean, C., Dinger, M. E., Fitzgerald, K. A., Gingeras, T. R., Guttman, M., Hirose, T., Huarte, M., Johnson, R., Kanduri, C., Kapranov, P., Lawrence, J. B., Lee, J. T., … Wu, M. (2023). Long non-coding RNAs: Definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology*, *24*, 430–447.
112. Trausch, J. J., Xu, Z., Edwards, A. L., Reyes, F. E., Ross, P. E., Knight, R., & Batey, R. T. (2014). Structural basis for diversity in the SAM clan of riboswitches. *PNAS*, *111*, 6624–6629.
113. Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., Rivas, E., Eddy, S. R., Finn, R. D., Bateman, A., & Petrov, A. I. (2021). Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, *49*, D192–D200.

How to cite this article: Walter, N. G. (2024). Are non-protein coding RNAs junk or treasure?. *BioEssays*, e2300201. https://doi.org/10.1002/bies.202300201