CellPress
REVIEWS

Opinion

Long Noncoding RNAs and Repetitive Elements: Junk or Intimate Evolutionary Partners?

Hyunmin Lee,[1,2] Zhaolei Zhang,[1,2,3] and Henry M. Krause[1,3,*]

Our recent ability to sequence entire genomes, along with all of their transcribed RNAs, has led to the surprising finding that only ~1% of the human genome is used to encode proteins. This finding has led to vigorous debate over the functional importance of the transcribed but untranslated portions of the genome. Currently, scientists tend to assume coding genes are functional until proven not to be, while the opposite is true for noncoding genes. This review takes a new look at the evidence for and against widespread noncoding gene functionality. We focus in particular on long noncoding RNA (noncoding RNAs longer than 200 nucleotides) genes and their 'junk' associates, transposable elements, and satellite repeats. Taken together, the suggestion put forward is that more of this junk DNA may be functional than nonfunctional and that noncoding RNAs and transposable elements act symbiotically to drive evolution.

Highlights

Most lncRNA genes are expressed during spermatogenesis and are localized to many different subcellular locations.

Solubility issues and alternative 3′ end processing have resulted in underestimates of lncRNA abundance.

The diversity of known lncRNA molecular and subcellular functions is growing rapidly and may approach those of proteins.

lncRNA and transposable element expression are coordinated during periods when spermatocyte heterochromatin is removed and replaced by protamines.

Most lncRNAs contain transposable element sequences and many are expressed under the control of transposable element promoters.

Our Genomes Are 98% Noncoding. Why?

There has been a sometimes quiet, sometimes loud, decades-long discussion on the roles of noncoding transcripts and transposable elements (TEs) within our genomes. This discussion escalated when the ENCODE program revealed that the majority of the human genome is transcribed (at least 76% in humans), but that only ~1.2% of this RNA is protein-coding [1,2]. Based on our previous knowledge of bacterial and yeast genomes, these findings were a huge surprise. For most bacteria, over 90% of their DNA is protein-coding. The *Escherichia coli* genome uses an average of 1 × 10^3 bp per coding gene, with the eukaryotic yeast genome using about twice that amount. In contrast, our genomes contain ~1.6 × 10^5 bp/coding gene. Unexpectedly, humans have only approximately five times more coding genes than *E. coli*, three times more than yeast, and fewer than both mice and the roundworm *Caenorhabditis elegans* (Table 1, Key Table).

This new sequence information pushed most scientists to fall into one of two camps. The ENCODE team and some others surmised that the additional transcribed but untranslated portions of mammalian genomes are like the dark matter of the universe: previously unrecognized but likely a major contributor to our relatively complex body plans and cognitive abilities [1,2]. However, based primarily on previous arguments [3,4], those in the other camp believed that these additional sequences are mainly composed of nonfunctional 'junk' or 'selfish' DNA (TEs, repeats, and noncoding DNA), with little or nothing to do with our physical or cognitive complexities [5–7]. Before getting into the arguments supporting the two sides, it is important that we first revisit a seemingly simple but nevertheless contentious issue: what defines function?

What Does 'Functional' Mean?

At the heart of the term 'functional' is the definition of what constitutes a gene. Scientists used to, and a surprising number still do, define a gene as a sequence that encodes a protein. However, following the discovery of noncoding RNA genes, the basic definition was changed to a DNA sequence that encodes a functional product, although more complex definitions have been put forward to deal with issues such as overlapping and unrelated products produced from the same DNA sequence [2,8]. This concept has become even more confusing, as many DNA sequences previously thought of as transcription factor-binding regulatory elements are also transcribed, and in most of the cases studied, the transcribed RNAs play important roles in processes such as transcription factor recruitment, DNA looping, chromatin composition, nuclear localization, etc. [9–11]. So, by the definition above, these sequences would also qualify as genes. These issues are discussed further in [2,12].

[1]Donnelly Centre, University of Toronto, Toronto, ON, Canada

[2]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada

[3]Department of Computer Science, University of Toronto, Toronto, ON, Canada

*Correspondence: h.krause@utoronto.ca

https://doi.org/10.1016/j.tig.2019.09.006

Key Table

Table 1. Comparative Genome Complexities

Organism	kbp	Coding genes	Ratio (bp/gene)
Escherichia coli	4.6×10^3	4.3×10^3	1.1×10^3
Saccharomyces cerevisiae	1.2×10^4	5.8×10^3	2.2×10^3
Caenorhabditis elegans	1.0×10^5	2.2×10^4	4.6×10^3
Drosophila melanogaster	1.2×10^5	1.5×10^4	8.2×10^3
Mouse	2.8×10^9	2.3×10^4	1.2×10^5
Human	3.3×10^9	2.0×10^4	1.7×10^5

This leads us to the term ‘functional’. Historically, for a gene to be considered functional there had to be an observable phenotype when it was mutated. In most cases, however, this was, and continues to be, assessed under optimal laboratory conditions. We now know that this is much more complicated, as many genes with important functions can be compensated for by other genes with related functions (genetic redundancy). In this way, cells and organisms have evolved brilliant and extensive backup systems to deal with the numerous types of genetic and environmental disruptions that they may encounter. In the laboratory, with ideal temperature, food, space, and mating conditions, these backup plans are not needed nearly as often as in the wild where issues such as temperature swings, drought, mate availability, and the presence of predators are all major issues. Most of our inbred ‘wild type’ and genetically altered model organism strains would never survive or propagate effectively in the wild. Yet, there has been a tendency to assume that genes without an obvious ‘in-lab’ phenotype are dispensable or even ‘nonfunctional’.

In yeast, only ~1/3 of coding genes are required for viability under standard growth conditions. While this low number seemed surprising at first, we now know that combining the deletions of two ‘nonrequired’ genes can result in ‘synthetic’ lethality. The numbers of required and synthetically lethal genes also vary depending on factors such as growth media composition, temperature, and a multitude of other stresses. Indeed, a more in-depth analysis of the yeast gene deletion set, grown in a diverse set of conditions, revealed phenotypes for 97% of them [13]. Similar to yeast, the initial estimation of essential gene numbers for *Drosophila* coding genes suggested that only 1/4–1/3 are required for viability under standard laboratory conditions [14,15], and when human cells are grown in culture, only ~10% of coding genes were found to be essential [16]. Clearly though, this does not mean that the genes not required in these utopian environments are nonfunctional. In many cases, the importance of a gene may be better gauged by the depth and sophistication of built-in redundancies. Some examples of backup redundancies include the ability of *Drosophila* segmentation genes to adapt to large changes in the expression of other segmentation genes [17,18], or by paralogous homologues in *Hox* gene clusters [19].

Studies Suggesting Noncoding Genes Are Nonfunctional

There are many kinds of noncoding RNAs (ncRNAs), such as the well-studied ribosomal and transfer RNAs. Among the more recently recognized ncRNAs are the relatively short miRNAs, small interfering RNAs (siRNAs), and piwi-interacting RNAs (piRNAs). Due to similarities in their sizes, structures, and molecular functions, scientists have been able to identify and assess many of these new RNA functions in a relatively quick and efficient manner, showing that the majority serve clear and important roles [20]. However, the most diverse and genetically abundant class of ncRNAs, referred to as long noncoding RNAs (lncRNAs), have not been so easy to tackle. Their diversity in terms of length, numbers, structures, distributions, and functions is partly to blame for the lag in attention and understanding. However, the prior widely held perception that they are predominantly junk is also factored in.

Some examples of the shadow cast by the assumption that most lncRNA genes are nonfunctional are found in early studies of pioneer lncRNAs, in which some groups concluded prematurely that the RNAs themselves were nonfunctional and that it was simply the act of transcribing them that was important. One of the first examples was the *H19* lncRNA, which regulates imprinting and contributes to numerous cancers when misexpressed [21]. Based on a number of deletion constructs and cell-based assays, it was concluded that it was the *H19* promoter and its activity that was required and not the RNA itself [22]. Similar studies initially suggested that the lncRNAs *Kcnq1ot1* and *Airn* were also not themselves required (reviewed in [23,24]). Although a PubMed search reveals that there have since been more than a thousand publications on the functions of these lncRNAs, both in *cis* and *trans*, many are still only aware of the earlier dismissive publications.

Similar studies and conclusions were made with *Hox* locus lncRNAs. Although it was suggested some time ago [25] that these lncRNAs could be used to coordinate the proximal to distal recruitment of epigenetic-regulating transcription factors, similar to the mechanism that *Rox1* had just been shown to recruit chromatin components, several groups conducted experiments that, as with *H19*, *Kcnq1ot1*, and *Airn*, suggested that it was the act of transcription, and not the RNAs themselves, that was important [26–30]. Again, we now know that many mammalian *Hox* complex lncRNAs do in fact have important functions in coordinating the proximal to distal epigenetic states across these complexes with anterior to posterior *Hox* gene expression patterns [31]. It seems likely that this will also turn out to be true for *Drosophila Hox* lncRNAs, as a relatively recent study has shown that simply transcribing through Polycomb response elements (PREs) is not sufficient for mediating subsequent epigenetic states, as had previously been thought [32].

Some of the first CRISPR-based lncRNA knockout studies have also led some authors and readers to negative takes on lncRNA gene functionality. These publications are similar, in that a number of embryonically expressed lncRNAs failed to yield obvious or overtly significant phenotypes when deleted [33,34]. However, as pointed out earlier, these modest effects may be due to functional redundancies and/or the nature of the assays. As discussed below, it may also be related to the fact that most lncRNAs appear to have evolved relatively recently, with the majority of their expression and functions occurring in male reproductive tissues and the brain. The basic animal body design, however, is largely controlled by ancient and highly conserved genetic pathways that were derived before the majority of lncRNAs evolved [35–37]. This likely explains the relative paucity of lncRNA expression and major roles during these earlier stages of development, although this is not to say that lncRNAs do not also have important functions at these times, as clearly evidenced by the roles of *Rox*/*Xist*, *Hox*, and *let-7*, for example.

Other Reasons to Discount lncRNA Functionality

While ENCODE and modENCODE studies have demonstrated high numbers of lncRNAs expressed in metazoan genomes, and suggested relatively high levels of functionality, they also provided data supporting the alternative view, noting that most of the detected lncRNAs were very low in abundance. However, this latter observation can largely be attributed to issues such as the use of cultured cell lines and the isolation methods used. For example, as noted in the ENCODE publication [2], transcripts annotated as very low or undetected in some cell lines could be quite abundant in others, or in tissues or stages of development not represented by the chosen cell lines. RNA-Seq approaches also tend to average out spatially and temporally restricted patterns, which locally, may be relatively high. Another key issue is that these studies made use of poly(A)-containing RNA selection, which discards the 50% or more of lncRNAs that are not polyadenylated [12,38]. *In situ* hybridization (ISH), however, does not discriminate between RNAs with alternative 3′ ends. For example, an analysis of over 100 *Drosophila* lncRNAs, all previously annotated as very low in expression or undetected, showed that all were expressed at similar levels to coding genes, although not usually as broadly or as early in development [39]. ISH-based studies also do not have to deal with RNA extraction issues; one study on the *Neat1* lncRNA, which nucleates paraspeckles, showed that it aggregates in the interphase of the phenol/chloroform mixtures used to extract the RNA, and that >20 times more RNA could be extracted after repeated passage of the extract through a syringe needle [40]. Another recent study has shown that the use of 100 mM ammonium acetate appears to be a more simple and effective way to dissociate similar liquid–liquid phase-separated (LLPS) structures and to fully recover their RNA cargoes [41]. Further analyses will be required to determine how LLPS solubility issues affect the extractability of other RNA cargoes.

In terms of conservation, many junk advocates were expecting functional lncRNAs to exhibit higher levels of conservation, perhaps even approaching those of coding genes [6,42], but this is not a fair expectation given that ncRNAs do not have to deal with protein coding issues such as codon usage, reading frame, folding, stability, or function. Because of this, functional elements within ncRNAs are also relatively free to shuffle around within the gene, which makes aligning their sequences much more challenging. This becomes even more challenging if the complementary sequences of conserved double-strand structures are free to drift far away from one another. Indeed, recent genome-wide RNA–RNA crosslinking studies have shown that up to 40% of double-stranded structures have counterpart sequences that are more than 200 nucleotides away, which is a larger span than what has been used until now to identify potential palindromic counterparts [43]. Importantly, this problem is further magnified for the many interactions that were found in the crosslinking studies to occur in trans. This study also confirmed that many base pairs in these secondary structures are relatively free to change, but are quickly compensated for by corresponding changes in the paired nucleotides. Overall, 25% of the aligned double strand helix-forming sequences were found to be conserved between amniotes (mammals, birds, reptiles), with compensating covariations averaging 46% [43]. These numbers are strikingly high given the ~300 million years of divergence among these species. Previous studies using conventional alignment methods had found that homologous lncRNA sequences or structures are generally undetectable beyond 50 million years of species divergence [44].

Although homologous lncRNA gene sequences are difficult to detect beyond 50 million years, other studies have shown that syntenic lncRNA promoters and exon/intron boundaries were actually similarly or even more conserved than those of coding genes [2,44–46]. Furthermore, some of these orthologs were shown to rescue one another, despite the lack of obvious sequence or structure conservation [44,47]. A particularly good example of lncRNAs with conserved synteny and promoters, and relatively degenerate sequences, are those referred to as topological anchor point RNAs (tapRNAs), which appear to control chromatin looping and topology [11]. As with guide RNAs that have been shown to recruit transcription factors to DNA, this may not be too surprising, as sequences that bind to DNA in *cis* will retain complementarity no matter how much the counterpart DNA sequence changes and the sequences and structures that bind proteins appear to be relatively degenerate and promiscuous [10,11,48]. As mentioned earlier, the linker sequences that join these functional motifs are also unlikely to be constrained by selective evolutionary pressure.

While all of these rationales can be given for poor lncRNA gene collinearity and conservation, a comparison between the relatively well annotated *Drosophila melanogaster* coding and lncRNA genes shows that lncRNA genes are actually surprisingly well conserved between *Drosophila* species (Figure 1). Furthermore, as would be expected if functional, lncRNA gene exons are more conserved than introns. In some cases, the overall degree of lncRNA gene and exon conservation significantly exceed those of the average coding genes. It should be noted though that *Drosophila* contains relatively few lncRNA genes compared with humans, suggesting that their rate of evolution may be much slower and, thus, their conservation and functionality levels may be proportionally higher.

Another possible concern about lncRNA functionality that is seldom brought up is the apparent paucity of reports on natural and forward genetic mutations in lncRNA gene loci that result in phenotypes. However, this can also be attributed to many of the considerations that affect conservation levels, such as the increased ability, relative to coding genes, to incur sequence changes without significantly affecting overall structure or function. In many cases, researchers have also likely concluded that mutations affecting functions in ncDNA were in key regulatory or structural elements associated with nearby coding genes, or simply ignored them due to our past ‘coding gene-centric’ ways of thinking. However, it should be pointed out that the majority of genome-wide association study-identified disease loci are within ncDNA sequences, many of which are or may be transcribed [1,49–53], suggesting that there may be many more mutations in lncRNA genes

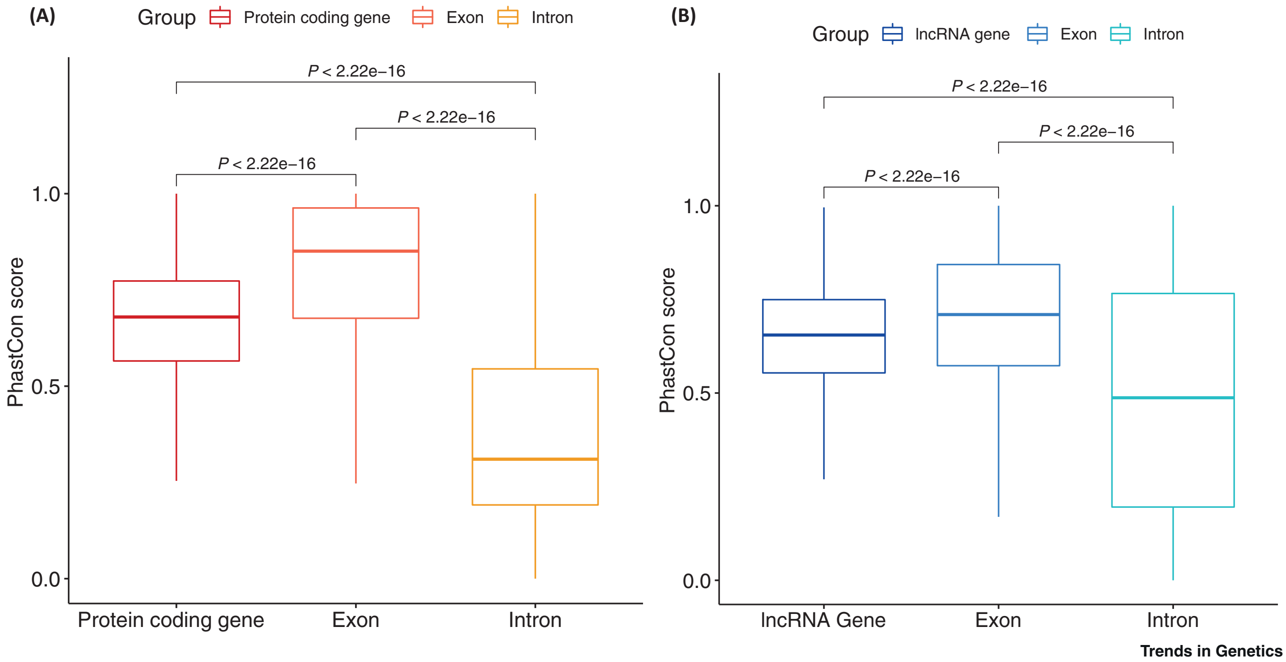

Figure 1. PhastCons Scores for *Drosophila* Coding and Long Noncoding RNA (lncRNA) Genes, Exons, and Introns, Derived from All 16 Available Sequenced *Drosophila* Species.

(A) Coding gene scores. (B) lncRNA gene scores. Note that lncRNA introns are often retained in nuclear-localized lncRNAs and in architectural RNAs (arcRNAs). For the latter, the repetitive sequences therein interact with intrinsically disordered domains of transcription factors. This may account for the somewhat higher range and mean of PhastCons scores for lncRNA introns as compared with coding gene introns. The *Drosophila* species polled diverge by as much as 60 million years of evolution.

than are currently expected. Indeed, the number of lncRNA-related diseases is growing rapidly, with a recent report tallying 3772 validated associations and another 97 998 yet to be tested positive correlations [54].

Another reason many assume that much of the transcribed genome is junk is because much of this RNA is comprised of spliced out introns or products of TEs. In the case of introns, many of the same sequences are also transcribed from independent promoters in the sense and/or antisense directions (approximately 20% of currently annotated *Drosophila* lncRNAs and 35% of human lncRNAs overlap with coding gene introns), suggesting that both are likely to have regulatory roles. Many examples already exist for first intron RNAs that act as guides for the recruitment of transcription factors to counterpart intronic DNA enhancers [55–57]. Similar functions appear to be carried out by stable, intron-derived circular lncRNAs referred to as ciRNAs [58,59]. Consistent with their proposed roles in gene expression, ciRNAs are nuclearly localized, unlike most other circular RNAs (circRNAs). Notably, circRNAs and ciRNAs are not polyadenylated, meaning that they will also have been discounted in profiling assays that use(d) poly(A) selection. Parenthetically, past profiling studies have actually opted for poly(A) selection, partly because it was assumed that intron sequence reads were derived from spliced out intron ‘debris’, which can overwhelm ‘legitimate’ low frequency reads [44]. Again, the use of ISH facilitates the detection of both linear and circular RNAs and so should be useful in determining their levels of expression and cellular/subcellular distributions.

Reasons to Consider That Most lncRNAs Might Be Functional

One of the more general reasons to assume that our genomes have not been accumulating ncRNA genes by random chance is our knowledge of the powers of natural selection and rates of sequence variation. For example, those of us who have worked with heterozygous mutant or transgenic lines know how quickly they can be taken over if a wild type individual gets into the stock. The rapid change in peppered moth coloration in post-industrial age England [60], and a more recent study on bird lice body color [61], are other excellent examples of the powers and speed of natural selection. With this in mind, it seems unlikely that large increases in genome size would be tolerated if they did not provide a positive contribution. The cost in energy, time, and potential errors incurred during DNA replication would also be large and not beneficial in times of famine. Likewise, the cost in terms of transcription would also be wasteful if most of this DNA were transcribed but nonfunctional. That said, there are examples, from protozoa to plants and animals, of species that have genomes inordinately larger than most of their related species counterparts. Examples include the amoeba *Polychaos dubium*, the marbled lungfish, and the plant *Paris japonica*, all of which have genomes orders of magnitude larger than ours. However, these cases are relatively few and are believed to occur when species populations and consequential natural selection impacts are relatively low [62]. In general, increasing genome size in higher eukaryotes usually correlates with lncRNA and TE numbers. As has been pointed out previously, this correlation between organism complexity and lncRNA gene numbers is, so far, better than for coding gene numbers [10,63]. However, this needs to be tested for many more organisms, especially those outliers that contribute to the 'C-value enigma' [64].

Another general reason to think that a large percentage of lncRNAs will have functions is that the vast majority appear to be trafficked within cells to precise subcellular destinations [65–67]. This is also the case for mRNAs and is crucial for determining the location of their protein products [68,69]. This not only ensures that the proteins are produced in the correct locations, but equally importantly, that they are not produced where they could engage in deleterious interactions or functions. The latter is also likely true for lncRNAs. Furthermore, as with their production, localization also requires energy, which makes little evolutionary sense if it does not provide benefits to the host.

In terms of much needed large-scale assessments of lncRNA functionality, a recent CRISPR-based study examined the requirement of 105 lncRNAs expressed exclusively or primarily in *Drosophila* testes for effects on fertility [70]. The authors found that 31% of the knocked out genes are required for full fertility in a laboratory setting. This percentage is similar to the previously mentioned ratios of yeast and *Drosophila* coding genes that are essential for viability under similar conditions. The use of enhancer/suppressor screens and more stringent conditions are certain to reveal significantly higher levels of functionality not revealed in this first analysis.

Spermatogenesis, Evolution, and Repetitive Elements

One of the reasons the researchers in the study above [70] chose to specifically focus on lncRNA functions in testes is because this is where the majority are expressed and are believed to have evolved. This evolutionary process, first referred to as the 'out of the testes' model, postulates that conditions become favorable for new gene expression during spermatogenesis due to the unparalleled accessibility of DNA at this time [44,71]. The relative absence of heterochromatin allows transcription from promoter sequences that are otherwise inaccessible, including any that may have recently evolved. If the RNAs transcribed happen to have a positive effect on sperm function or success, then the genetic changes leading to their expression are likely to be passed on to the next generation. With time and further refinement of promoters, enhancers, and functions, these changes may also lead to expression and benefits in other tissues.

While there is now ample evidence for this mode of lncRNA evolution [44,71], the question of what positive roles these lncRNAs may play, or how new promoters and RNAs become active *de novo*, have not been deeply explored. One of the driving forces behind the latter appears to be the action of transposons and retrotransposons [72], referred to henceforth as TEs. In most cells, TEs tend to be silenced within heterochromatin, but become widely accessible during spermatogenesis. Fortunately, the majority of their expression, products, and activity are suppressed by ncRNAs such as piRNAs and siRNAs, many of which are derived from the TEs themselves [20]. Much of this control is also likely to be exerted by lncRNAs, as 2/3 of human lncRNAs are estimated to contain TE sequences that contribute an average of 30% to overall sequence content [72]. Many of these sequences are derived from 'inactive' TEs or TE remnants. Thus, even inactive TE remnants may play a significant role in DNA rearrangements and new gene evolution. In turn, these lncRNA sequences are likely to facilitate hybridization to homologous TE DNA sequences or RNA products, leading to silencing at multiple levels, as seen with TE-derived pi- and siRNAs [73].

This contribution to lncRNA gene evolution and composition has also been shown to occur with another component of junk DNA, satellite repeats. Like TEs, satellite repeats are a functional component of heterochromatin, also playing important roles in centromere and telomere function. They are also transcribed, with the RNAs localizing to what are normally heterochromatic regions [74] and helping to recruit heterochromatin forming proteins [95]. Removal of *Drosophila* satellite sequences during spermatogenesis disrupts the transition of sperm DNA into properly packaged protamine structures [74]. A number of *Drosophila* genes expressed solely during spermatogenesis from the Y chromosome also contain enormous megabase-sized introns composed almost entirely of satellite repeats. These oversized introns have been shown to play important roles in the timely expression, packaging, and splicing of these transcripts [75].

If there was a delicate balance between the levels of TE and satellite silencing and occasional expression, rearrangements and/or transposition, then, as noted previously in various organisms [72,76,77], some of these events could lead, either directly or indirectly, to the expression of flanking lncRNA sequences (illustrated in Figure 2). Indeed, the upstream regions of lncRNA genes are highly enriched for TEs and many of these have been shown to regulate lncRNAs nearby [72]. The subsequent addition of introns and other useful modifications has also been shown to be regulated by TEs [44,72]. For some of these physically related TEs and lncRNAs, the lncRNAs have been shown to silence the upstream or related TEs. Similar evolutionary processes are also likely to regulate adaptation to stress, as many TEs are expressed during heat shock and are known to contribute to other stress adaptations such as responses to relocation into unfamiliar or hostile environments [78–80].

If *de novo* TE rearrangements were random, they would likely have as many detrimental effects as positive ones, if not more. However, three factors, at least, are likely to compensate for these. The first is the existence of checkpoints that recognize and eliminate damaged sperm [81,82]. The second is the huge numbers of sperm produced, which would allow for a relatively large number of 'failed experiments'. The third is that compromised sperm would be less likely to be competitive in the process of fertilization. Thus, lncRNAs, transposons, and satellite repeats may be involved in a symbiotic relationship that, in proper balance and measure, facilitates evolution. One advantage of this proposed evolutionary process is that it is tightly shut down by chromatin silencing in all subsequent cells and developmental stages of the organism. This is particularly important for long-lived life forms and also provides a potential evolutionary advantage for those with excess sperm but relatively few progeny. Notably, this proposed relationship between the major components of 'junk DNA' during spermatogenesis is probably highly conserved, not only throughout the animal kingdom, but also in the production of pollen, the plant equivalent of sperm [83].

The proposed widespread roles of lncRNAs in policing TE activities is closely related to known roles for lncRNAs in controlling paternal imprinting. For example, *H19* expression also begins during spermatogenesis and controls the subsequent epigenetic state of nearby coding genes such as *Igf2*, which plays a role in progeny size. *H19* silences *Igf2* and other genes in the locus by guiding the histone deacetylase MBD1 and the lysine methyl transferase KMT, to locus target sites [84]. Many other paternally imprinted genes have also been shown to be regulated by lncRNAs. In many cases, the lncRNA gene is located within clusters of imprinted coding genes [85]. In addition to controlling and setting up the epigenetic state of TEs and imprinted genes, we have suggested previously that ncRNAs may also be playing a more general role in epigenetic marking, not only during spermatogenesis, but also under other cellular states where epigenetic markers and chromatin complexes may need to be removed, such as during DNA replication or repair [10].

In addition to supplying a means of modulating the activities of TEs during spermatogenesis, lncRNAs are also likely to regulate many other aspects of sperm differentiation and activity, as evidenced by the extent of lncRNA expression in testes, the diversity of fertility defects observed upon deletion of

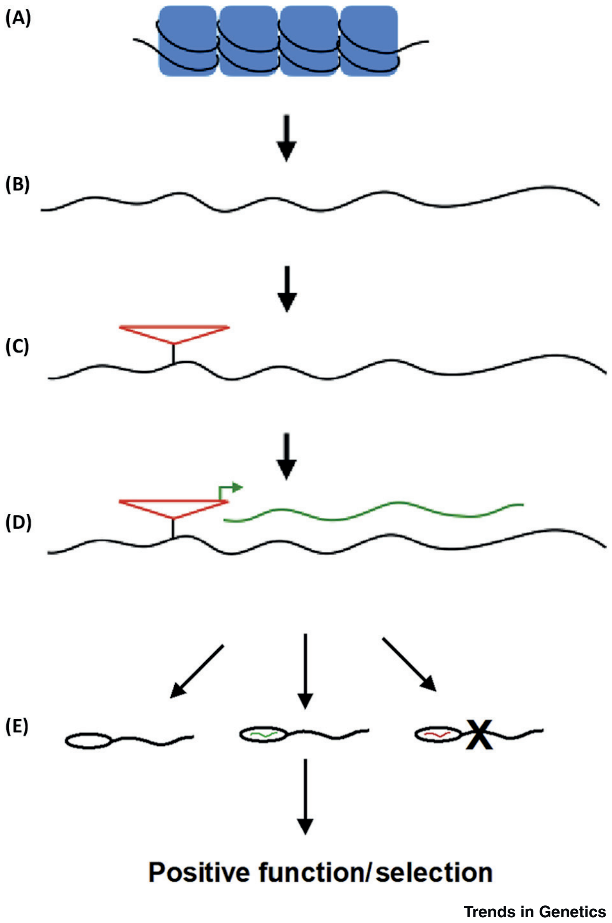

Figure 2. Model for *De Novo* Long Noncoding RNA Evolution in Testes.
(A) General heterochromatic chromatin state preventing DNA access or expression (nucleosomes in blue, DNA black). (B) Chromatin opening during spermatogenesis. (C) Transposition or transposon-mediated insertion (red line) or rearrangement event (D) directly or indirectly results in transcription of flanking noncoding sequences (green line). Additional transposable element-induced recombinations can also introduce additional elements such as repeats, splice sites, or introns. (E) Numbers of sperm allow for new gene experimentation, with negatively affecting new transcripts (red) resulting in sperm elimination or inability to compete and positive effects (green transcript) providing selective advantages.

testes-expressed *Drosophila* lncRNAs [70], and most telling, the high levels of non-nuclear lncRNA subcellular distributions observed in all tissues, including the testes ([66] and unpublished observations).

Diversity of lncRNA Roles and Properties

Among the driving forces of lncRNA evolution are likely their novel abilities to interact highly specifically with both DNA and other RNAs as well as proteins (reviewed in [10,48]). Studies have also shown a general ability to interact with membrane lipids [86–88] and, in at least one case, the ability to interact specifically with phosphatidylinositol-3,4,5-trisphosphate (PIP3), which recruits the lncRNA LINK-A, and bound protein, to activated trans-membrane receptors [89]. These various interactions allow lncRNAs to act as guides, nucleators, and scaffolds for numerous complexes found throughout the cell (reviewed in [10,48]). One of the growing realizations regarding the majority of these complexes is that they are composed of ribonucleoprotein (RNP) compositions that result in LLPS, which physically distinguishes them from their surrounding cellular milieus. In many cases, this property appears to be dependent upon lncRNAs for nucleation, localization, scaffolding, and composition (reviewed in [48]). The number of subcellular structures shown to possess these LLPS properties has grown rapidly, starting first with nuclear and cytoplasmic particles such as paraspeckles and stress granules, and now known to include most other common RNP structures. More recent examples include Cajal bodies, heterochromatin, nuclear pores, centrosomes, and lipid rafts [90]. Fluorescence *in situ* hybridization studies have also indicated many more such subcellular puncta and structures that do not appear to correlate with those that are already known [66]. It will be interesting to see if lncRNAs play similar organizing and structural roles for many or most of these structures.

Dark Matters

In terms of functionality, it is worth returning to the comparisons between lncRNA genes in the genome and dark matter in the universe. In the case of dark matter, individual units play infinitesimally small roles, yet as a whole are absolutely essential for universe integrity. This may also be the case for many lncRNAs that, individually, may not provide essential functions, but in larger numbers and combinations, act to provide critical mass. This may apply, for example, to LLPS structures. The flip side of this is that these abilities, even if nonessential, may lead to serious consequences if these lncRNAs are expressed in the wrong place, time, or numbers, as seen with *H19* and its contributions to cancer [91]. This 'ectopic activity' appears to be the case for most of the lncRNAs that cause or are associated with cancers [92,93]. We have also made a case previously for the involvement of LLPS-organizing lncRNAs in the cause or augmentation of neural degenerative and cognitive diseases that involve the formation of plaque-like structures [48]. Indeed, this may be the downside of lncRNA expression in the brain, the tissue that is second to the testes in terms of lncRNA gene expression levels and diversity [65]. This point also leads back to the issue of function. One of the obvious benefits of all of these new lncRNAs expressed in our brains is that they may contribute to increased neuronal diversity and capabilities, leading to increased cognition, memory, and related abilities. If so, their roles may be difficult to discern in model organisms unless assessed with appropriate assays. Consider, for example, the particular means required to identify flies or mice with autism. A recent example of an important lncRNA function that may have been impossible to discern without the correct assay is the discovery of LDAIR, a lncRNA that oscillates in expression both diurnally and seasonally and plays a major role in coordinating stress-associated self-protective behaviors [94]. Because of their subtlety, or possible compensation by redundancy, many of these functions may be more readily discerned using ectopic expression (as seen naturally with lncRNA-induced cancers) rather than with knockouts. The new CRISPR activation (CRISPRa) approaches are good candidates for testing this. Of course, follow-up studies would be required to determine if or how these ectopic functions, interactions, and outcomes compare with their normal ones.

Concluding Remarks

The discovery of lncRNAs, and their large numbers in metazoa, has been one of the major discoveries of the post genome-sequencing era. Clearly, our study and understanding of their functions in cells, reproduction, evolution, cognition, and disease are in their infancy. However, based on the lessons we have learned thus far, and what we know about the powers and mechanisms of natural selection, we should probably be considering the possibility that the majority of expressed lncRNAs are actively or potentially involved in important cellular processes. The challenge will be to devise approaches that can test this in an efficient and unbiased fashion. Some of these more pressing goals and directions are listed in the Outstanding Questions.

Outstanding Questions

What percent of lncRNA genes are functional?

How many of these functions are hidden by genetic redundancy or obscure phenotypes?

Do lncRNA gene numbers consistently correlate with organism complexity?

What is the range and diversity of lncRNA functions?

How and why do new lncRNA genes originate?

Why are the highest numbers of lncRNAs expressed within the male reproductive system first and the brain second?

Acknowledgments

We thank Andrew Spence and Julie Brill for providing comments and suggestions.

References

1. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74
2. Djebali, S. *et al.* (2012) Landscape of transcription in human cells. *Nature* 489, 101–108
3. Doolittle, W.F. and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284, 601–603
4. Orgel, L.E. and Crick, F.H. (1980) Selfish DNA: the ultimate parasite. *Nature* 284, 604–607

5. Doolittle, W.F. (2013) Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U. S. A.* 110, 5294–5300
6. Graur, D. *et al.* (2013) On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5, 578–590
7. Palazzo, A.F. and Gregory, T.R. (2014) The case for junk DNA. *PLoS Genet.* 10, e1004351
8. Gerstein, M.B. *et al.* (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res.* 17, 669–681
9. Quinodoz, S. and Guttman, M. (2014) Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. *Trends Cell Biol.* 24, 651–663
10. Jandura, A. and Krause, H.M. (2017) The new RNA world: growing evidence for long noncoding RNA functionality. *Trends Genet.* 33, 665–676
11. Amaral, P.P. *et al.* (2018) Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. *Genome Biol.* 19, 32
12. Kellis, M. *et al.* (2014) Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 111, 6131–6138
13. Hillenmeyer, M.E. *et al.* (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* 320, 362–365
14. Spradling, A.C. *et al.* (1999) The Berkeley *Drosophila* Genome Project gene disruption project: single P-element insertions mutating 25% of vital *Drosophila* genes. *Genetics* 153, 135–177
15. Chen, S. *et al.* (2010) New genes in *Drosophila* quickly become essential. *Science* 330, 1682–1685
16. Wang, T. *et al.* (2015) Identification and characterization of essential genes in the human genome. *Science* 350, 1096–1101
17. Namba, R. *et al.* (1997) *Drosophila* embryonic pattern repair: how embryos respond to bicoid dosage alteration. *Development* 124, 1393–1403
18. Hughes, S.C. and Krause, H.M. (2001) Establishment and maintenance of parasegmental compartments. *Development* 128, 1109–1118
19. Wagner, A. (1996) Genetic redundancy caused by gene duplications and its evolution in networks of transcriptional regulators. *Biol. Cybern.* 74, 557–567
20. Bartel, D.P. (2018) Metazoan microRNAs. *Cell* 173, 20–51
21. Lecerf, C. *et al.* (2019) The long non-coding RNA H19: an active player with multiple facets to sustain the hallmarks of cancer. *Cell Mol. Life Sci.* Published online July 23, 2019. https://doi.org/10.1007/s00018-019-03240-z
22. Jones, B.K. *et al.* (1998) Igf2 imprinting does not require its own DNA methylation or H19 RNA. *Genes Dev.* 12, 2200–2207
23. Pauler, F.M. *et al.* (2007) Silencing by imprinted noncoding RNAs: is transcription the answer? *Trends Genet.* 23, 284–292
24. Kornienko, A.E. *et al.* (2013) Gene regulation by the act of long non-coding RNA transcription. *BMC Biol.* 11, 59
25. Nasiadka, A. *et al.* (2002) Anterior-posterior patterning in the *Drosophila* embryo. *Adv. Dev. Biol. Biochem.* 12, 155–204
26. Hogga, I. and Karch, F. (2002) Transcription through the iab-7 *cis*-regulatory domain of the bithorax complex interferes with maintenance of Polycomb-mediated silencing. *Development* 129, 4915–4922
27. Rank, G. *et al.* (2002) Transcription through intergenic chromosomal memory elements of the *Drosophila* bithorax complex correlates with an epigenetic switch. *Mol. Cell. Biol.* 22, 8026–8034
28. Schmitt, S. *et al.* (2005) Intergenic transcription through a polycomb group response element counteracts silencing. *Genes Dev.* 19, 697–708
29. Petruk, S. *et al.* (2006) Transcription of bxd noncoding RNAs promoted by trithorax represses Ubx in *cis* by transcriptional interference. *Cell* 127, 1209–1221
30. Sessa, L. *et al.* (2007) Noncoding RNA synthesis and loss of Polycomb group repression accompanies the colinear activation of the human HOXA cluster. *RNA* 13, 223–239
31. De Kumar, B. and Krumlauf, R. (2016) HOXs and lincRNAs: two sides of the same coin. *Sci. Adv.* 2, e1501402
32. Erokhin, M. *et al.* (2015) Transcriptional read-through is not sufficient to induce an epigenetic switch in the silencing activity of Polycomb response elements. *Proc. Natl. Acad. Sci. U. S. A.* 112, 14930–14935
33. Schor, I.E. *et al.* (2018) Non-coding RNA expression, function, and variation during *Drosophila* embryogenesis. *Curr. Biol.* 28, 3547–3561
34. Goudarzi, M. *et al.* (2019) Individual long non-coding RNAs have no overt functions in zebrafish embryogenesis, viability and fertility. *eLife* 8, e40815
35. Domazet-Loso, T. and Tautz, D. (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468, 815–818
36. Kalinka, A.T. *et al.* (2010) Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468, 811–814
37. Irie, N. and Kuratani, S. (2011) Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat. Commun.* 2, 248
38. Cheng, J. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308, 1149–1154
39. Clark, A.G. *et al.* (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450, 203–218
40. Chujo, T. *et al.* (2017) Unusual semi-extractability as a hallmark of nuclear body-associated architectural noncoding RNAs. *EMBO J.* 36, 1447–1462
41. Jain, A. and Vale, R.D. (2017) RNA phase transitions in repeat expansion disorders. *Nature* 546, 243–247
42. Lindblad-Toh, K. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478, 476–482
43. Lu, Z. *et al.* (2016) RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* 165, 1267–1279
44. Ulitsky, I. (2016) Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* 17, 601–614
45. Carninci, P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563
46. Ponjavic, J. *et al.* (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* 17, 556–565
47. Ulitsky, I. *et al.* (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* 147, 1537–1550
48. Krause, H.M. (2018) New and prospective roles for lncRNAs in organelle formation and function. *Trends Genet.* 34, 736–745
49. Khurana, E. *et al.* (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342, 1235587
50. Brown, J.B. *et al.* (2014) Diversity and dynamics of the *Drosophila* transcriptome. *Nature* 512, 393–399
51. Zhang, F. and Lupski, J.R. (2015) Non-coding genetic variants in human disease. *Hum Mol Genet* 24, R102–R110
52. Schonrock, N. *et al.* (2016) Seq and you will find. *Curr. Gene. Ther.* 16, 220–229

53. Giral, H. *et al.* (2018) Into the wild: GWAS exploration of non-coding RNAs. *Front. Cardiovasc. Med.* 5, 181
54. Ma, L. *et al.* (2019) LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucl. Acids. Res.* 47, 2699
55. Tay, M.L. and Pek, J.W. (2017) Maternally inherited stable intronic sequence RNA triggers a self-reinforcing feedback loop during development. *Curr. Biol.* 27, 1062–1067
56. Jia Ng, S.S. *et al.* (2018) Generation of *Drosophila* sisRNAs by independent transcription from cognate introns. *iScience* 4, 68–75
57. Chan, S.N. and Pek, J.W. (2019) Stable intronic sequence RNAs (sisRNAs): an expanding universe. *Trends Biochem. Sci.* 44, 258–272
58. Zhang, Y. *et al.* (2013) Circular intronic long noncoding RNAs. *Mol. Cell* 51, 792–806
59. Sekar, S. and Liang, W.S. (2019) Circular RNA expression and function in the brain. *Noncoding RNA Res.* 4, 23–29
60. Cook, L.M. and Saccheri, I.J. (2013) The peppered moth and industrial melanism: evolution of a natural selection case study. *Heredity (Edinb)* 110, 207–212
61. Bush, S.E. *et al.* (2010) Evolution of cryptic coloration in ectoparasites. *Am. Nat.* 176, 529–535
62. Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* 302, 1401–1404
63. Mattick, J.S. (2011) The central role of RNA in human development and cognition. *FEBS Lett.* 585, 1600–1616
64. Gregory, T.R. (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev. Camb. Philos. Soc.* 76, 65–101
65. Mercer, T.R. *et al.* (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U. S. A.* 105, 716–721
66. Wilk, R. *et al.* (2016) Diverse and pervasive subcellular distributions for both coding and long noncoding RNAs. *Genes Dev.* 30, 594–609
67. Carlevaro-Fita, J. and Johnson, R. (2019) Global positioning system: understanding long noncoding RNAs through subcellular localization. *Mol. Cell* 73, 869–883
68. Lecuyer, E. *et al.* (2007) Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* 131, 174–187
69. Lecuyer, E. *et al.* (2009) Global implications of mRNA localization pathways in cellular organization. *Curr. Opin. Cell. Biol.* 21, 409–415
70. Wen, K.J. *et al.* (2016) Critical roles of long noncoding RNAs in *Drosophila* spermatogenesis. *Genome Res.* 26, 1233–1244
71. Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20, 1313–1326
72. Kapusta, A. *et al.* (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* 9, e1003470
73. Czech, B. and Hannon, G.J. (2016) One loop to rule them all: the ping-pong cycle and piRNA-guided silencing. *Trends Biochem. Sci.* 41, 324–337
74. Mills, W.K. *et al.* (2019) RNA transcribed from heterochromatic simple-tandem repeats are required for male fertility and histone-protamine exchange in *Drosophila melanogaster*. *bioRxiv*. Published online April 24, 2019. https://doi.org/10.1101/617175.
75. Fingerhut, J.M. *et al.* (2019) Satellite DNA-containing gigantic introns in a unique gene expression program during *Drosophila* spermatogenesis. *PLoS Genet* 15, e1008028
76. Chapman, J.A. *et al.* (2010) The dynamic genome of *Hydra*. *Nature* 464, 592–596
77. Chalopin, D. *et al.* (2014) Evolutionary active transposable elements in the genome of the coelacanth. *J. Exp. Zool. B Mol. Dev. Evol.* 322, 322–333
78. Sun, W. *et al.* (2014) An adaptive transposable element insertion in the regulatory region of the EO gene in the domesticated silkworm, *Bombyx mori*. *Mol. Biol. Evol.* 31, 3302–3313
79. Horvath, V. *et al.* (2017) Revisiting the relationship between transposable elements and the eukaryotic stress response. *Trends Genet.* 33, 832–841
80. Lanciano, S. and Mirouze, M. (2018) Transposable elements: all mobile, all different, some stress responsive, some adaptive? *Curr. Opin. Genet. Dev.* 49, 106–114
81. Mahadevaiah, S.K. *et al.* (2008) Extensive meiotic asynapsis in mice antagonises meiotic silencing of unsynapsed chromatin and consequently disrupts meiotic sex chromosome inactivation. *J. Cell. Biol.* 182, 263–276
82. Subramanian, V.V. and Hochwagen, A. (2014) The meiotic checkpoint network: step-by-step through meiotic prophase. *Cold. Spring Harb. Perspect. Biol.* 6, a016675
83. Cui, X. *et al.* (2015) Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. *Mol. Plant.* 8, 935–945
84. Monnier, P. *et al.* (2013) H19 lncRNA controls gene expression of the imprinted gene network by recruiting MBD1. *Proc. Natl. Acad. Sci. U. S. A.* 110, 20693–20698
85. Autuoro, J.M. *et al.* (2014) Long noncoding RNAs in imprinting and X chromosome inactivation. *Biomolecules* 4, 76–100
86. Khvorova, A. *et al.* (1999) RNAs that bind and change the permeability of phospholipid membranes. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10649–10654
87. Janas, T. and Yarus, M. (2003) Visualization of membrane RNAs. *RNA* 9, 1353–1361
88. Michanek, A. *et al.* (2010) RNA and DNA interactions with zwitterionic and charged lipid membranes - a DSC and QCM-D study. *Biochim. Biophys. Acta* 1798, 829–838
89. Lin, A. *et al.* (2017) The LINK-A lncRNA interacts with PtdIns(3,4,5)P3 to hyperactivate AKT and confer resistance to AKT inhibitors. *Nat. Cell Biol.* 19, 238–251
90. Alberti, S. *et al.* (2019) Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* 176, 419–434
91. Raveh, E. *et al.* (2015) The H19 long non-coding RNA in cancer initiation, progression and metastasis - a proposed unifying theory. *Mol. Cancer* 14, 184
92. Bolha, L. *et al.* (2017) Long noncoding RNAs as biomarkers in cancer. *Dis. Markers* 2017, 7243968
93. Camacho, C.V. *et al.* (2018) Long noncoding RNAs and cancer, an overview. *Steroids* 133, 93–95
94. Nakayama, T. *et al.* (2019) Seasonal regulation of the lncRNA LDAIR modulates self-protective behaviours during the breeding season. *Nat. Ecol. Evol.* 3, 845–852
95. Camacho, O.V. *et al.* (2017) Major satellite repeat RNA stabilize heterochromatin retention of Suv39h enzymes by RNA-nucleosome association and RNA:DNA hybrid formation. *eLife* 6, e25293