NON-CODING RNA

Evolution to the rescue: using comparative genomics to understand long non-coding RNAs

Igor Ulitsky

Abstract | Long non-coding RNAs (lncRNAs) have emerged in recent years as major players in a multitude of pathways across species, but it remains challenging to understand which of them are important and how their functions are performed. Comparative sequence analysis has been instrumental for studying proteins and small RNAs, but the rapid evolution of lncRNAs poses new challenges that demand new approaches. Here, I review the lessons learned so far from genome-wide mapping and comparisons of lncRNAs across different species. I also discuss how comparative analyses can help us to understand lncRNA function and provide practical considerations for examining functional conservation of lncRNA genes.

Expressed sequence tag (EST). Typically 3′-biased Sanger-sequencing read of approximately 700 nucleotides.

Full-length cDNA
A cDNA that ideally captures a full-length mRNA transcript from the 5′ cap to the 3′ polyadenylated tail; sequenced by multiple Sanger sequencing runs.

Department of Biological Regulation, Weizmann Institute of Science, 234 Herzl Street, Rehovot 76100, Israel.
igor.ulitsky@weizmann.ac.il

doi:10.1038/nrg.2016.85
Published online 30 Aug 2016; corrected online 6 Sep 2016

Long non-coding RNAs (lncRNAs) are defined as RNAs of at least 200 nucleotides (nt) in length that are independently transcribed, and that molecularly resemble mRNAs, yet do not have recognizable potential to encode functional proteins. The 200 nt cutoff excludes most canonical ncRNAs, such as small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs) and tRNAs, and roughly corresponds to the retention threshold of protocols for the purification of long RNAs. Genomic studies based on expressed sequence tag (EST) and full-length cDNA sequencing, tiling microarrays and RNA sequencing (RNA-seq) identified thousands of lncRNAs in diverse animal and plant genomes. One recent study that combined RNA-seq data from multiple sources reported over 58,000 lncRNA loci in the human genome[1]. Future studies will plausibly increase this number, as lncRNAs are more tissue-specific and expressed at lower levels than mRNAs[2–4], and many cell types (in particular those that are rare or found in early embryonic stages) have not yet been thoroughly interrogated by RNA-seq. The fraction of annotated lncRNAs that are functional — that is, have any recordable impact on a molecular, cellular or organismal level — is still unknown. The questions of whether lncRNAs are functional and how they perform their functions are of particular interest considering the rapidly increasing number of lncRNAs that are implicated in changing expression or losing sequence integrity in different instances of human disease[5,6].

Comparative analysis of genes across species can be a powerful tool for studying their functions and modes of action, as it has been for other non-coding RNAs and proteins. For instance, the discovery that the let-7 microRNA (miRNA) is conserved from human to nematodes ignited major interest in miRNAs in 2000 (REF. 7), and subsequent comparative analysis has been instrumental in identifying miRNA genes, predicting miRNA targets in mRNAs and for revealing features that are important for miRNA biogenesis[8–10]. Comparative approaches require two main ingredients: sets of genes or genomes that can be compared, and algorithms for matching and evaluating the similarity. Applying comparative sequence analysis to lncRNAs is challenging on both fronts. Until recently, only a few lncRNAs had been annotated in species other than human and mouse, and lncRNAs typically lack long regions with high constraint on sequence (which are needed by tools that have been developed for comparing protein-coding genes) or regions with strong constraint on secondary structure (which is a key ingredient used by tools that have been developed for studying shorter RNAs). In addition, as our understanding of the modes of action of lncRNAs is still very rudimentary and the 'rules' underlying their functions remain unknown, it is a major challenge to develop models that will accurately capture evolutionary constraints on lncRNA loci (similar to models that use the ratio of non-synonymous to synonymous changes (dN/dS) to study the constraints on preserving a particular protein-coding sequence[11]). Nevertheless, recent studies have begun to take the first steps towards mapping and comparing lncRNAs across mammals and other vertebrates[2,12–14] (TABLE 1), and have uncovered constant turnover of lncRNA genes in evolution alongside extensive sequence changes in those lncRNAs that are conserved. In parallel, as detailed below, researchers have tested conservation of function among

Table 1 | **Databases and data sets of lncRNAs annotated in multiple species**

Study or database	Species	Raw data	Comparable data and methodology across species	Allows retrieval of lncRNA homologues across species	Web sites
PLAR[2]	17 vertebrates	RNA-seq from multiple tissues and 3P-seq in two species	Yes	Yes	http://webhome.weizmann.ac.il/home/igoru/PLAR
Necsulea *et al.*[12]	11 vertebrates	RNA-seq from multiple tissues	Yes	Yes	http://www.nature.com/nature/journal/v505/n7485/full/nature12943.html#supplementary-information
Washietl *et al.*[13]	6 mammals	RNA-seq from multiple tissues	Yes	Yes	http://genome.cshlp.org/content/early/2014/01/15/gr.165035.113/suppl/DC1
PhyloNONCODE[14]	10 vertebrates	RNA-seq	Yes	Yes	http://www.bioinfo.org/phyloNoncode
lncRNAdb	69 species; 12 species with ≥5 lncRNAs	Manual curation	No	Yes	http://lncrnadb.org
RNAcentral	13 species	Combination of 22 databases, including GENCODE	No	No	http://rnacentral.org
NONCODE	16 species	Literature mining and GenBank	No	Yes	http://www.noncode.org
PLNlncRbase	43 plant species	Manual curation	No	No	http://bioinformatics.ahau.edu.cn/PLNlncRbase
GreeNC	37 plant species and 6 algal species	Transcriptomes analysed for coding potential	Similar methodologies; different data	No	http://greenc.sciencedesigners.com

3P-seq, poly(A)-position profiling by sequencing; lncRNA, long non-coding RNA; RNA-seq, RNA sequencing.

homologues of specific lncRNAs. Although most lncRNA studies were conducted in vertebrate species, studies in other clades have so far reported a surprisingly similar picture, suggesting that although no common lncRNA genes have been found so far between species separated by more than 500 million years of evolution, the principles guiding lncRNA evolution across eukaryotes are similar.

In this Review, I survey the main methods that have been used to identify and compare lncRNAs across species, and summarize the shared conclusions of studies on lncRNA evolution in vertebrates, insects, sponges and plants. I then discuss the current understanding of the evolutionary origins of lncRNAs and the mechanisms through which the complexity of their loci has increased during evolution. Last, I discuss recent studies of the evolution of function in specific lncRNAs. Throughout this Review (and particularly in BOX 1), I provide practical guidelines for identifying and studying homologues of lncRNAs of interest.

Identification of lncRNA genes

A typical lncRNA is biochemically identical to an mRNA: it harbours a 5′ cap and a 3′ polyadenylated (poly(A)) tail, and is thus easily sequenced by standard RNA-seq protocols[15]. In recent years, researchers have been using increasingly deep RNA-seq to map the transcriptomes of various tissues and conditions across eukaryotes, and have identified numerous new lncRNAs in each system[2,12,13,16–18]. These efforts built on earlier studies that were based on ESTs and full-length cDNA sequencing, which yielded fewer transcript models that were more accurately annotated owing to longer read length[19]. A rough scheme of the RNA-seq-based lncRNA identification is outlined in FIG. 1. Until recently, the most common sequencing methods used oligo(dT)-based enrichment for poly(A) RNAs, which include the vast majority of functionally characterized lncRNAs. More recently, protocols that deplete only the rRNAs and sequence the rest of the 'total RNA', including non-poly(A) transcripts, are being adopted increasingly[20]. In my experience, the use of total RNA does not add a substantial number of lncRNAs, and it is important to keep in mind the drawbacks of using total RNA; namely, a lower per cent of usable reads and a higher per cent of reads mapping to introns[21], which are features that make transcript model assembly and expression-level quantification more challenging than from poly(A)-enriched data. For either protocol, the most popular tools for read mapping and transcript assembly are TopHat[22] and Cufflinks[23] from the Tuxedo suite (HISAT and StringTie are recent successors of those tools and have improved performance[24,25]). Recent benchmarking efforts showed that the Tuxedo tools are comparable in performance to others[26,27], and that full-length transcript assembly using short-read data is a challenging task. Therefore, although the transcript models reconstructed from short-read RNA-seq are certainly useful, they are not necessarily accurate across all exons.

Once a transcriptome is assembled, a computational pipeline is needed for the filtering, annotation and discovery of those transcripts that meet the lncRNA

Homologues
A pair of genes that descended from a common ancestral gene.

criteria. Some of the major differences between the computational approaches are whether they consider single-exon transcripts (that are notoriously enriched with artefacts), whether they allow some degree of overlap between lncRNAs and other known genes (for example, overlap with introns of protein-coding genes on the same strand), and how they distinguish between coding and non-coding genes[28]. These factors heavily influence the numbers of identified lncRNAs.

Databases of lncRNA annotations

Systematic curation efforts have enabled the development of several lncRNA databases (TABLE 1). Reference Sequence (RefSeq) and GENCODE (accessible through Ensembl) are widely used databases of transcript structures that are based mostly on curated EST and cDNA data; these databases contain few, but relatively accurate, isoforms. Primarily based on deep RNA-seq, other databases hold almost an order of magnitude more transcript isoforms than RefSeq; for example, approximately 60,000 lncRNA genes have been identified in the MiTranscriptome data set[1]. The complexity of alternative splicing, along with alternative promoters and polyadenylation sites (and to a lesser extent, algorithmic difficulties), contributes to the large number of isoforms that are reconstructed for individual lncRNA genes[29]. Importantly, even lncRNAs that are expressed at low levels are reproducibly detected across individuals[13], indicating that their annotation is unlikely to be erroneous.

Box 1 | Identifying homologues of a lncRNA of interest in other species

The extent of conservation is increasingly regarded as a key question in evaluating the impact of studied long non-coding (lncRNAs). If a lncRNA is implicated in a human condition, it is important to know whether it can be studied in model organisms; conversely, if a lncRNA is discovered in a model organism, evidence of conservation is important for establishing relevance to human biology. Several approaches that are available for identifying homologues of a lncRNA are discussed below.

Sequence conservation in whole-genome alignments

The easiest way to look for homology is to use whole-genome alignments (WGAs), such as those available in the University of California, Santa Cruz (UCSC) Genome Browser or in Ensembl, to compare either the whole lncRNA locus or individual exons across species. Alignability in WGAs requires an extent of conservation that reaches significance when comparing whole genomes, leading to potentially reduced power. An open question is whether there are many functionally conserved lncRNAs that are not alignable in the WGAs. These are probably rare among mammals, as the number of positionally conserved lncRNAs is similar to the number of those having sequence conservation[2,30], but when considering more distal species, there are more position-conserved lncRNAs (after subtracting the number expected by chance) than sequence-conserved ones[2]. Therefore, in such comparisons, cases in which sequence homology has eroded to a point at which it does not reach significance on a genome-wide level are likely to be more common.

lncRNA sequence conservation by direct comparison with sequences in other species

An alternative to WGAs that also addresses the difference between DNA conservation and lncRNA conservation is to directly align the query lncRNA with lncRNAs from other species (that is, from the data sets in TABLE 1) using BLAST or other algorithms[118]. This approach is less computationally intensive than WGA and the level of similarity required to reach significance is lower. If a lncRNA has several isoforms, each can be compared separately, or the exonic coordinates of all the isoforms can be merged into a single 'meta-transcript' that contains all the exonic bases, and then the meta-transcripts can be compared across species.

Structure or profile conservation

When comparing lncRNAs across more-distant species, sequence conservation might be too subtle for homologue detection. If sequences from more closely related species are available, the pattern of changes in a specific short (<200 nucleotide) region can be used to build a sequence profile (for example, using HMMER) or a structure-based profile (for example, using Infernal[119]). Such structure-based profiles were used for detecting distant homologues of the polyadenylated nuclear non-coding RNA (PAN) lncRNA in viruses[69] and RNA on the X (roX) lncRNAs in *Drosophila* species[68].

Positional conservation

Positional conservation occurs when lncRNAs in different species are found flanking orthologous genes (within a certain distance), and have the same relative orientation[2,73]. In practice, if a WGA with the species in question is available, one can inspect increasingly larger regions around the lncRNA of interest, project them to the other species and inspect the corresponding locus (see TABLE 1 for resources of lncRNA annotations in various species). If a WGA is not available, it is possible to use a database of protein-coding gene orthologues (such as Ensembl Compara and HomoloGene) to identify the orthologues of the neighbours of the query lncRNA in the target genome, and then to inspect their neighbourhood for potential positional homologues. The strength of evidence for position conservation depends on the distance between the lncRNA and the gene, and on the size and characteristics of the intergenic region. Because some protein-coding genes are flanked by multiple lncRNAs, it is sometimes difficult to assign specific pairs of lncRNAs as positionally conserved. To the best of my knowledge, there are currently no methods that can assign statistical significance to the level of evidence of positional conservation of a specific pair of lncRNAs. Given a set of criteria (for example, that the lncRNA-encoding locus is immediately flanking the conserved protein-coding gene and appears within 50 kb of it) and a pair of species, an estimate can be made of how many positionally conserved lncRNA pairs are expected by chance and how many are observed, to estimate an empirical false-discovery rate (FDR) for the criteria used[2]. It is important to emphasize that significant sequence similarity (for example, BLASTN *E*-value $<10^{-5}$) is a far stronger indicator of potential functional homology between lncRNAs than is positional conservation.

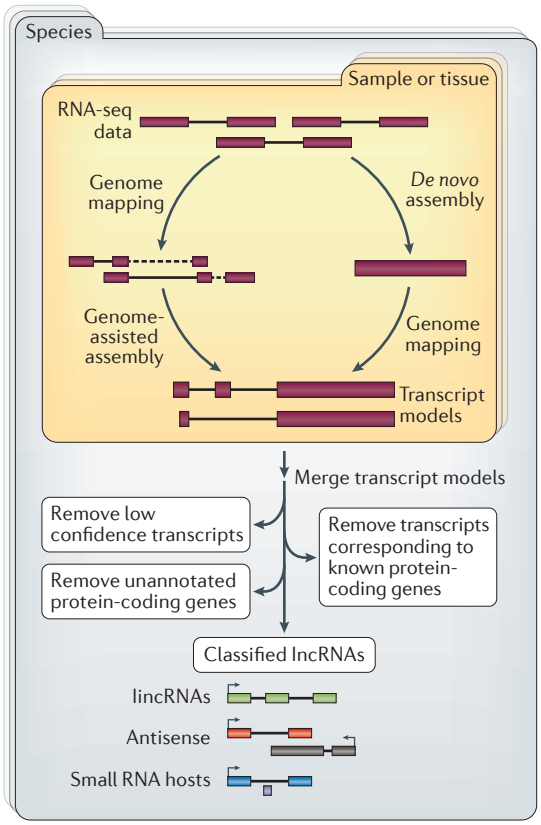


Figure 1 | **A generic pipeline for the identification of lncRNAs from RNA-seq data.** Long non-coding RNAs (lncRNAs) are identified separately in each species and in each tissue or sample. RNA sequencing (RNA-seq) reads are either first mapped to the genome and then assembled into transcripts (genome-guided assembly, such as that performed by Cufflinks[120]), or first assembled into transcripts (*de novo* assembly, such as that performed by Trinity[121]) and then mapped to the genome. Transcripts from all samples are then merged, multiple filtering steps remove various artefacts and protein-coding genes, and the remaining transcripts are classified into one of the lncRNA classes. lincRNAs, long intergenic non-coding RNAs.

This diversity of resources raises the question of which data set is best to use for lncRNA analysis. For practical purposes, it is usually desirable to focus on the major isoforms of each gene — which in my experience are easier to study by using slimmer transcript databases (such as RefSeq or GENCODE) — and to quantify expression on the gene level rather than on the isoform level. However, to study lncRNAs that are highly tissue-specific or expressed at low levels, or that have rare alternative splicing isoforms, more comprehensive databases will be more suitable.

Systematic comparisons of lncRNAs across species

By using various methodologies for identifying lncRNA homologues (described in BOX 1), recent studies have explored evolutionary trajectories of lncRNAs in vertebrate[2,12–14,30], insect[16,18,31], plant[32–34] and basal animal species[35,36] (TABLE 1). lncRNA loci from various species can be compared on multiple levels, as discussed below.

Primary sequence conservation. If genomes of closely related species are available, the parameter that is easiest to measure is the turnover of the DNA sequence, which can be deduced from whole-genome alignments and compared to that of other genomic features to assess the degrees of contribution of primary sequence to fitness. Such comparisons showed that lncRNA exons evolve faster than exons of protein-coding genes across bilateria[12,37–40] and plants[41]. Within the lncRNA loci, there is slightly higher conservation in exons compared with introns, indicating that the mature RNA products of some lncRNAs may be functional. With the exception of *Drosophila melanogaster*, in which lncRNA exons are highly conserved[40], there have been some inconsistencies between reports about the difference in conservation between the exons of lncRNAs and the introns of protein-coding genes or random intergenic sequences[38]. The differences between these studies mostly stem from the disparities in the set of lncRNAs that were analysed. In conservatively selected sets of lncRNAs, which are enriched with more robustly expressed and accurately annotated isoforms, exons appear to evolve more slowly than introns of protein-coding genes and more slowly than other intergenic regions[4,37,38,42], but this difference is much smaller than the difference in conservation between protein-coding exons and lncRNA exons, indicating that the vast majority of the lncRNA sequence evolves under little to no selective constraint. With broader and less-filtered lncRNA collections, the mean conservation erodes and eventually approaches that of non-transcribed intergenic regions[1,40,43].

The lengths of alignable sequences among lncRNA homologues are approximately five times shorter than in protein-coding genes[2]. A typical lncRNA conserved between human and mouse will exhibit only 20% inter-species homology, and homology drops to 5% in lncRNAs conserved between human and fish[2]. Therefore, a subset of lncRNAs (enriched with those that are relatively highly expressed)[44] evolves under constraints on their mature sequences, but these constraints are much weaker and span a shorter fraction of the gene when compared with those acting on coding sequences or miRNAs.

Constraint on lncRNA sequence can also be evaluated among members of the same species[40]. Surprisingly, there is no evidence for purifying selection acting on lncRNA exons in the human population, but there is strong evidence for such selection in fruitflies. This difference can be explained by the vast differences in the effective population sizes of these species: if lncRNAs contain sites that evolve with small selection coefficients, constraint will be virtually invisible in human genomes owing to the small effective population size[40].

Conservation of transcription status and splicing patterns. A key assumption made when using DNA sequence alignments to study lncRNA evolution is that lncRNA exons in one species align to lncRNA exons in the other species. However, transcription typically evolves faster than the underlying DNA sequence and thus, in many cases, lncRNA loci are homologous to

Purifying selection
(Also called negative selection). Selective removal of deleterious alleles.

Effective population size
The size of an idealized population that would experience genetic drift in a similar way to the actual population.

non-transcribed sequences in the other species[2,12,13]. Therefore, it is important to study lncRNAs by directly comparing lncRNA-producing loci, and such studies in multiple species have uncovered rapid turnover of lncRNA loci[2,12,13]. For example, my laboratory found that in 17 vertebrates, more than 70% of lncRNAs have appeared in the past 50 million years[2]. Splicing patterns also evolve rapidly, with only approximately 20% of splicing events in human lncRNAs conserved outside of primates[13]. lncRNA loci are thus commonly gained and lost in evolution, and those lncRNAs that are retained drastically change their exon–intron architecture and their sequences across species in which the lncRNA is present.

One potential caveat of studies comparing lncRNAs in bulk and using heterogeneous data is that the conservation of the lncRNAs expressed only in specific cell types might be underestimated if the compared tissues are not carefully matched. This does not seem to be a major concern, as studies that focused on a specific tissue in a few species reached similar conclusions. For example, one study identified and compared lncRNAs expressed in the liver in three rodents and found that only 60% (160 out of 268) of the lncRNAs expressed in mouse liver had homologues that are expressed in rat liver and only 27% (76 out of 273) had homologues that are expressed in human liver[45]. Similar results were seen when comparing human and mouse islets of Langerhans[46], eye[47] and pluripotent stem cells[30]; even in carefully matched systems, most human lncRNAs do not have recognizable homologues in mice and vice versa.

The rapid evolution of most lncRNAs is inconsistent with many having functions that depend on specific sequence throughout their loci. It is possible that many lncRNAs carry no function, or that lncRNA functions may rely on short elements for which the surrounding sequence context has limited importance. One possible type of such sites comprises binding sites for miRNAs or RNA-binding proteins, which may allow some lncRNAs to act as competing endogenous RNAs (ceRNAs)[48], although the mechanisms that allow low-abundance lncRNAs to compete for binding with hundreds to thousands of more abundant mRNAs remain unclear in many cases (see REF. 49 for a review of the current understanding of the ceRNA hypothesis).

Secondary structure and its conservation. An open and debated question is whether secondary structure plays an important part in lncRNA biology, as it does in other non-coding RNAs, which rely heavily on structured elements for their biogenesis and functions. Two main practical aspects of the importance of secondary structure are whether selection acting on structure rather than primary sequence explains the rapid rate of lncRNA sequence evolution, and whether focusing on regions with stable or conserved structures assists in homing in on functionally important regions.

As with any long RNA, lncRNAs fold into secondary structures, many of which are stable, but that fact alone does not imply that the secondary structure is important for function. On average, lncRNA transcripts are slightly less structured than mRNAs *in vitro*[50], but significantly more structured than mRNAs *in vivo*[51]. Surprisingly, there is no correlation between the amount of secondary structure and overall sequence conservation[44,50]. The experimental evidence for lncRNAs broadly acting through specific structures is scarce. Notable exceptions are triplex elements that stabilize the 3′ ends of MALAT1 (metastasis-associated lung adenocarcinoma transcript 1) and NEAT1 (nuclear enriched abundant transcript 1) lncRNAs[52], the roX-box stem-loop structures in the *D. melanogaster* roX (RNA on the X) lncRNAs[53,54] and possibly the RepA repeat in the XIST (X-inactive specific transcript) RNA[55–57].

For cases in which using only primary sequence conservation to define homology has not identified human homologues of mouse lncRNAs, can structure-only conservation lead us to these 'missing' homologues, as proposed by one study[58]? To the best of my knowledge, there are no examples of such cases that have been shown experimentally. Furthermore, sequence alignability between mammalian species does not require strong purifying selection over long stretches[59], and pressure to preserve structured elements should in most cases be sufficient for maintaining alignability. Indeed, elements in which the structure but not the sequence is thought to be important, such as basal stems of miRNA hairpins, are easily alignable between mammals. Additional evidence suggesting that structure conservation without sequence alignability is rare among mammals comes from comparing the number of lncRNAs that are syntenic between humans and other mammals (after subtracting background expectation) to the number of lncRNAs that have sequence similarity. The gap between these two numbers is small (not more than a couple of dozen lncRNAs for humans and each other tested mammal)[2,30], so it is unlikely that many lncRNAs have conserved structures between species as distant as human and mouse yet remain invisible in whole-genome alignments.

Short regions of sequence evolving under selection to preserve secondary structure can be predicted across the genome using methods based on scanning whole-genome alignments, such as EvoFold and RNAz (reviewed in REF. 60), and loci of some functional human lncRNAs, such as *MALAT1*, *NEAT1* (REF. 61) and *NORAD* (non-coding RNA activated by DNA damage)[62,63], overlap such regions. Surprisingly, the overlap between lncRNA exons and segments predicted to evolve under constraints on secondary structures is small in the human genome[38] as well as in the genomes of other species[64]. A study using a different background model recently reported more than 4 million regions that are evolving under selection to preserve secondary structure[65], but a vast number of those regions overlap regions that do not appear to be transcribed at appreciable levels. Another recent study has mapped the secondary structure of the HOTAIR (HOX transcript antisense intergenic RNA) lncRNA[66] and highlighted some structures as evolutionarily conserved, but a recent preliminary statistical analysis of the levels of conservation suggested that in this and potentially other cases there is no evidence for selection on preservation of

Triplex
An RNA structure formed by three strands of RNA, two that form a Watson–Crick duplex and a third that binds in the major groove of the duplex forming Hoogsteen and reverse Hoogsteen hydrogen bonds.

Syntenic
Preserving order and orientation of genes or other genomic elements between species.

specific structures[67]. Genome-wide analysis thus provides limited support for widespread pressure to preserve secondary structures in lncRNAs. This does not imply that structure-based homology searches cannot sometimes be very useful for lncRNA homology detection; for example, elegant and carefully tailored structure-based approaches were used to detect homologues of roX lncRNAs in *Drosophila* species[68] and the viral PAN (polyadenylated nuclear non-coding) RNA[69] in distant species in which primary-sequence homology approaches have failed (it is noteworthy that the species in these studies are separated by many more generations than humans and mice).

Overall, although there is still much remaining to be learned about the structure–function axis in lncRNA genes, most current evidence suggests that regions where specific secondary structures are important for conserved functions occupy a much smaller fraction of the lncRNA sequences compared with those of canonical ncRNAs, such as rRNAs, snoRNAs, tRNAs and snRNAs.

Positional conservation. It has been proposed that in many cases, transcription through a lncRNA locus (or part of it) is important, whereas the RNA product plays a secondary part, if any[70]. For example, transcription through the region of the *AIRN* (antisense of *IGF2R* non-protein-coding RNA) lncRNA locus that overlaps the promoter of insulin-like growth factor 2 receptor (*IGF2R*) is important for *IGF2R* silencing, whereas the rest of the 118 kb *AIRN* RNA is dispensable for this purpose[71]. In such lncRNAs, one can expect that the position of the region that is transcribed would be conserved, whereas the exon positions and the bulk of the mature lncRNA sequence would evolve neutrally, with the exception of elements that are required for continued transcriptional elongation, such as short splicing motifs. Indeed, splicing motifs are preferentially conserved in many lncRNAs[72]. Furthermore, it has been observed that when comparing distant species, a significant number of lncRNAs are 'positionally conserved' — that is, found in the same relative orientation to orthologous protein-coding genes and/or other conserved regions[2,32,73,74] — and many of those do not share detectable sequence conservation. Such pairs may correspond to lncRNAs that have conserved functional sequences that are too short or degenerate to be detected, or to lncRNAs in which only the act of transcription is under selective pressure. In many of the lncRNAs with deep positional conservation, such as *PVT1* and *DEANR1* (definitive endoderm-associated lncRNA 1; also known as *linc-FOXA2*)[2,75], the length of the transcribed locus and the exon–intron architecture also evolve rapidly, indicating that the second scenario (a role for transcription itself) may be more common.

Classes of lncRNA evolutionary trajectories

The analysis of lncRNA conservation at the different levels presented above[30] gives rise to the classification system proposed here in which each class corresponds to a different level of conservation and distinct lncRNA features, and probably different mechanisms of action as well (FIG. 2).

'Class I' lncRNAs are conserved lncRNAs in which exon–intron structure and multiple sequences along the length of the lncRNA are conserved among species. A representative of this class is MIAT (myocardial infarction associated transcript; also known as GOMAFU)[76], which contains 5–7 exons in both human and mouse, 4 of which are conserved (FIG. 2b). At present, we know that this class constitutes a minority of conserved lncRNAs but includes some of the better-studied ones, such as XIST, cyrano (also known as OIP5-AS1), NEAT1, MALAT1 and NORAD. It is expected that many of the *trans*-acting lncRNAs will belong to this group and indeed some of the better-studied Class I lncRNAs are enriched in the cytoplasm, and therefore probably act independently of their sites of transcription.

'Class II' lncRNAs are those in which the act of transcription and some RNA elements (biased towards the 5′ end of the RNA) are conserved, whereas the majority of the locus experienced drastic changes in exon–intron structure and length. For example, such a conserved lncRNA is found downstream of the *ONECUT1* gene in human, mouse and other vertebrates (FIG. 2c). In Class II lncRNAs, only a few splice sites, if any, are conserved, and transposable elements (TEs) contributed heavily to locus diversification across species (see below). These lncRNAs are more likely to be *cis*-acting and to regulate gene expression in regions surrounding their loci.

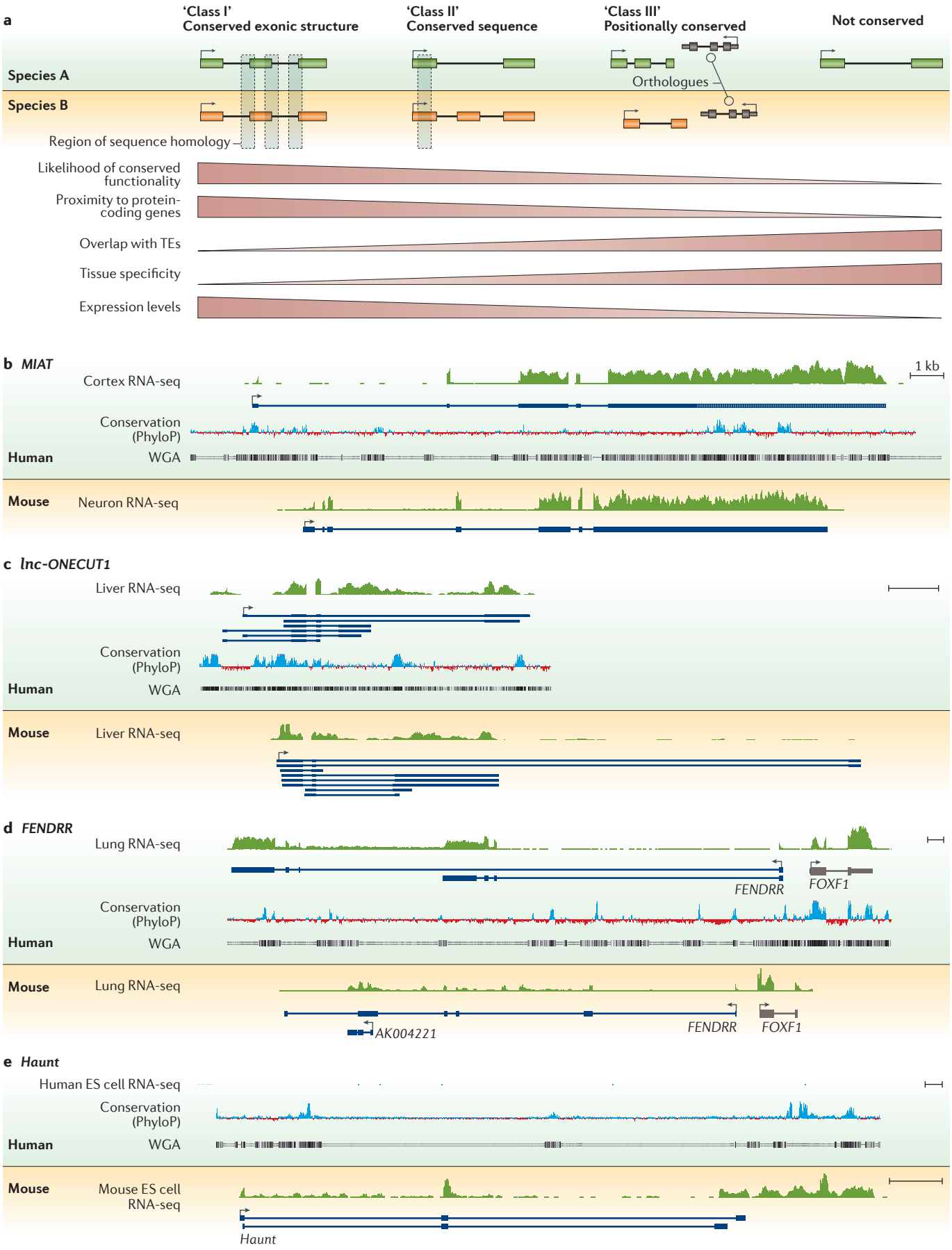
'Class III' lncRNAs are conserved lncRNAs in which, beyond conservation of promoter sequences and the act of transcription of the specific region, there are no

Figure 2 | **Classes of lncRNA conservation. a** | Proposed classes of sequence conservation among long non-coding RNAs (lncRNAs) and their correlation with genomic features. See the main text for a description of the individual features and references to the publications supporting the positive and negative correlations with the level of conservation. **b** | High conservation of exon–intron structure; for example, the *MIAT* (myocardial infarction associated transcript; also known as *GOMAFU*) lncRNA locus in human and mouse. The RNA sequencing (RNA-seq) track shows the coverage of reads from the human cortex from the Human Proteome Atlas (HPA) transcriptome database[122] and the mouse cerebellar granular neurons[123]. Phylogenetic *P* value (PhyloP) scores[124], which describe base-wise conservation during vertebrate evolution, were taken from the University of California, Santa Cruz (UCSC) Genome Browser. Whole-genome alignment (WGA) track shows alignable regions between human and mouse genomes. **c** | A lncRNA with conserved sequence, but divergent exon-intron structure; for example, a lncRNA found downstream of the *ONECUT1* gene in human and mouse. Human adult liver RNA-seq is from the HPA and mouse adult liver RNA-seq is from the Encyclopedia of DNA Elements (ENCODE) project. **d** | A lncRNA with a conserved position and very limited sequence conservation: the forkhead box F1 (*FOXF1*) gene and the *FOXF1* adjacent non-coding developmental regulatory RNA (FENDRR) lncRNA. RNA-seq from adult lung from the HPA and ENCODE projects. **e** | A mouse lncRNA with no evidence of expression in human, the *Haunt* (also known as *Halr1* or *linc-Hoxa1*) locus. RNA-seq from human[125] and mouse[126] embryonic stem (ES) cells. TEs, transposable elements.

Orthologous
Pertains to homologous genes in different species that have evolved from a common ancestral gene by speciation.

Trans-acting
Regulation that is not *cis* acting; for example, regulation by diffusible factors that can comparably regulate both homologous loci in a diploid organism.

Cis-acting
Acting from the same molecule, typically interpreted as regulation occurring on the same physical chromosome.



regions with recognizable sequence similarity and there is typically no conservation of gene structure. Some of these lncRNAs might be transcribed from conserved enhancer elements, with limited or no function of the RNA product or the act of transcription, such as in the case of the *Lockd* lncRNA[77]. In several lncRNAs, such as *FENDRR* (*FOXF1* adjacent non-coding developmental regulatory RNA) (FIG. 2d), there is conservation of the promoter and of the first splice site, but not of the rest of the exons, suggesting that it is the act of transcriptional elongation (supported by productive splicing; see below) that is important.

Notably lncRNAs that host conserved small RNAs, such as miRNAs and snoRNAs, evolve under a separate set of pressures and therefore can be defined as a separate class[30]. Importantly, most human or mouse lncRNAs are not found in the other species, such as *Haunt* (also known as *Halr1* or *linc-Hoxa1*)[78] (FIG. 2e). It is possible that some human lncRNAs perform primate-specific functions, and that others independently evolved functions similar to those of lncRNAs in other species, but it is plausible that many of them are simply not functional.

As illustrated in FIG. 2a, various genomic and functional features are correlated with the degree of conservation. For example, lncRNAs with enhancer-like chromatin marks at their promoters (a high ratio of histone H3 lysine 4 monomethylation (H3K4me1) compared to trimethylation (H3K4me3)) are less conserved than those with more canonical promoters (enrichment for H3K4me3 relative to H3K4me1)[79]. Most conserved lncRNAs are also closer to protein-coding genes, less likely to overlap transposable elements, and are more broadly and highly expressed[2,13].

Rapid turnover of lncRNAs in other phyla

The high prevalence of lncRNAs is not unique to vertebrate genomes. Thousands of lncRNAs have now been described in the much smaller *D. melanogaster* genome[18], as well as in mosquito and bee genomes[16,31] (see REF. 80 for a review on lncRNAs in insects), and over 10,000 lncRNAs may be present in some species of plants with larger genomes, such as maize[81] and cotton[82]. More than 2,000 lncRNA loci were annotated in sponges, which are non-bilaterian animal species with simple morphology[35,36].

The number of lncRNAs identified in individual species is strongly influenced by the breadth and depth of the available RNA-seq data, as well as by the annotation criteria and filters (for example, whether single-exon or intron-overlapping transcripts were considered), and so it remains difficult to correlate genomic characteristics with the propensity of the genome to give rise to lncRNAs. However, it is interesting to note that the main features associated with vertebrate lncRNAs — short length with few exons, and low and tissue-specific expression — also appear in these other species.

In many species it has been difficult to measure lncRNA conservation owing to a lack of sufficiently close species with sequenced genomes and/or transcriptomes; for example, the closest sequenced relatives of some sponges diverged more than 450 million years ago. In species in which the comparison of lncRNAs across a set of species with a reasonable gradient of evolutionary distances is possible, the emerging picture is of rapid turnover similar to the one observed in vertebrates. For example, only 20% of the lncRNAs in the mosquito *Anopheles gambiae* have alignable sequences in the genome of *Anopheles minimus*, whereas 90% of *A. gambiae* proteins are alignable between the two species[16] (these species diverged less than 80 million years ago). In plants, a large excess of positionally conserved lncRNAs, compared with sequence-conserved lncRNAs, was found among the genomes of nine Brassicaceae and Cleomaceae plants[32], and between rice and maize[33].

The overall features of lncRNAs observed in vertebrates are thus probably applicable to lncRNAs from other clades and vice versa. However, to date, no clear homologues and no lncRNAs with clearly analogous mechanisms have been identified between vertebrate lncRNAs and those of other species. Therefore, it is not clear to what extent the mechanisms used by lncRNAs in other clades are also used in vertebrates and vice versa.

Evolutionary origins of new lncRNAs

The observation that most lncRNAs in vertebrate genomes do not have homologues in species separated by more than 50 million years of evolution[2] suggests a high frequency of new lncRNA origination. Several mechanisms for such events are described below and in FIG. 3Aa–Ae.

Duplication. Protein-coding genes evolve by duplication and subfunctionalization[83], with few exceptions[84]. If this route were common in lncRNAs, we would expect to see some sequence similarity among lncRNAs within the same species (although lncRNA paralogues are expected to be less similar to each other than protein paralogues owing to the faster sequence evolution). In practice, such intra-species similarity among lncRNAs is rare[2,73,85], and when it does occur, it can often be attributed to unannotated fragments of TEs[2]; therefore, whole-locus duplication only rarely contributes to the evolution of new lncRNAs. Still, specific lncRNA pairs (such as the two or three paralogues of *megamind* (also known as *TUNA*) found in most vertebrates[73]) have probably evolved by duplication, as have *MALAT1* and *NEAT1* (REF. 61), which have apparently unrelated functions but maintain certain common features, such as nuclear retention and stabilization by a triple-helical element at their 3′ end[52].

Loss of coding potential of protein-coding genes. Mutations, TE insertions and genomic rearrangements in protein-coding loci can lead to nonsense mutations and loss of protein-coding function. If these events do not lead to a loss of transcription (or if transcription is later regained) and if nonsense-mediated decay is either not triggered or not very efficient, then a new lncRNA gene can be formed at the same locus. Three of the lncRNA loci in the eutherian X-inactivation centre — *XIST*, *JPX* (also known as *ENOX*) and *FTX* — originated through this mechanism and were retained across mammals[86,87].

Paralogues
Homologous genes related by duplication within a genome.

Nonsense mutations
Mutations in which a codon encoding an amino acid is mutated into a stop codon.

Formation of new transcriptional units following integration of TEs. TEs are potent rewirers of genomes and have made major contributions to innovation in gene regulation in mammals[88]. Mammalian lncRNAs heavily overlap TEs; for example, ~40% of lncRNA sequences are recognizable as TE-derived, ~80% of lncRNAs overlap at least one TE, and ~25% of promoters and polyadenylation sites of human lncRNAs are TE-derived[89,90]. The insertion of a TE containing a functional promoter (such as endogenous retroviruses (ERVs)[90]) can be sufficient to drive transcription initiation at a previously non-transcribed locus. If the locus contains or gains splicing and polyadenylation elements downstream of the new promoter a new lncRNA will be formed. Notably, both splicing and polyadenylation depend on relatively short sequence elements that occur frequently by chance. ERV promoters are also typically regulated and act within relatively narrow developmental time windows, such as in pluripotent cells[89] or testis[91], so lncRNAs formed in this fashion share specific temporal and spatial expression patterns.

Stabilization of cryptic transcripts by mutations that enhance splicing. Recent studies have shown that divergent transcription occurs at most active promoters and enhancers in mammals[92]. The products of these events are predominantly cryptic unspliced and non-poly(A) RNAs of varying length (~1 kb on average) that are rapidly degraded by the exosome and potentially other complexes[93]. A functional 5′ splice site recognized by the U1 small nuclear ribonucleoprotein (snRNP) can suppress early polyadenylation. One or more of these suppression events in combination with a functional 3′ splice site can favour splicing over polyadenylation and lead to the production of a stable transcript. Therefore, point mutations or TE insertions that introduce U1 binding sites can easily transform cryptic transcripts into stable RNAs, which can then acquire functions as lncRNAs or as new protein-coding genes[94,95].

Exaptation of previously non-coding sequence. lncRNA origination events that do not result from the mechanisms listed above probably arise from a series of mutations that create a favourable combination of promoters, splice sites and polyadenylation elements, leading to exaptation of a previously non-transcribed locus into a lncRNA. These new lncRNAs will be expressed under the control of enhancer elements acting in spatial proximity to them and, as elegant experiments using random insertions of weak promoters in mice have shown[96], the output of such promoters will often be highly tissue-specific. Prevalence of this scenario can help to explain why lncRNAs that are found away from protein-coding genes are typically more tissue-specific than those expressed from divergent promoters with protein-coding genes.

Estimation of the rate of lncRNA gain and loss in evolution is challenging, as it is difficult to prove that a certain sequence is entirely missing or not transcribed in a given species, or that the lncRNA was not present in ancestral species. Regardless of the origin, new lncRNAs appear to be predominantly expressed in the germ line, particularly in the vertebrate testis[12]. The permissive chromatin environment in the testis allows transcription of a wide range of genomic elements in meiotic spermatocytes and postmeiotic spermatids[97], and it is likely that most of these elements carry no function. An intriguing alternative hypothesis (without current experimental support) is that the permissive expression landscape in germ cells is important, as it allows for efficient selection against new genes that are deleterious on the cellular level, thus preventing the genetic changes that favour the production of toxic RNAs from being passed on to the next generation.

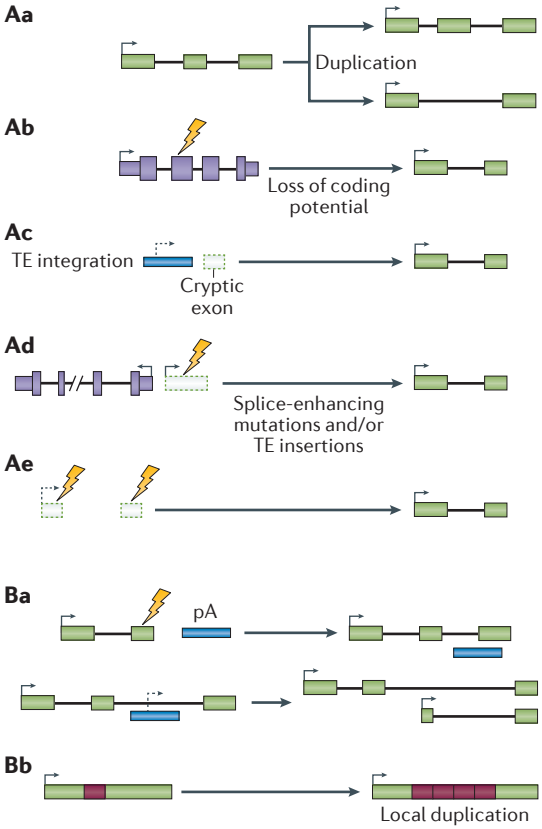


Figure 3 | **Pathways for origination and diversification of lncRNA loci.** Possible scenarios for the formation of new long non-coding RNA (lncRNA) loci. An ancestral lncRNA locus can be duplicated (part **Aa**). An ancestral protein-coding gene can lose its coding potential owing to a sequence change, but the transcriptional programme in the locus can be retained (part **Ab**). A transposable element (TE) carrying a functional promoter, or sequences resembling one, can be integrated next to sequences encoding cryptic exons (part **Ac**). An unstable transcript product of bidirectional transcription can be stabilized by changes favouring splicing and the formation of a stable product (part **Ad**). Last, a combination of genetic changes occurring in the vicinity of each other can lead to the formation of promoter and RNA processing elements in an orientation that is required for lncRNA production (part **Ae**). Two main known mechanisms for lncRNA locus complexity increase, exonization of TEs (part **Ba**) and local sequence duplications (part **Bb**). Lightning signs indicate a series of mutations and the blue rectangles indicate newly integrated TEs; pA indicates a polyadenylation signal.

Exaptation
Co-option of a functionally unrelated DNA sequence for a novel function.

Routes for increased complexity in lncRNA loci

Interestingly, most 'young' (that is, <50 million years old) lncRNAs across vertebrates and other species share common genomic features that sometimes set them apart from 'old' lncRNAs[2,12,13], suggesting that rather than being associated with specific functionality, the features of young lncRNAs characterize new genes as they emerge in evolution from unstable transcripts or from non-transcribed regions. In vertebrates, such genes typically have two or three exons (compared with ~8 exons for protein-coding genes in vertebrates) and are ~1,000 nt long (compared with ~2,000 nt for mRNAs)[2]. Interestingly, some functionally characterized lncRNAs are much longer or much more highly expressed. For example, XIST is a 17 kb transcript that is spread across 8 exons, and one of the isoforms of ANRIL (antisense non-coding RNA in the *INK4* locus; also known as CDKN2B-AS1) is ~4 kb long and spread across 19 exons. There is a mild but significant correlation between the evolutionary age of a lncRNA and its length[2]. Longer lncRNAs are in many cases extensively alternatively spliced, producing multiple isoforms. The locus diversification does not appear to be completely random, with conserved elements exhibiting 5′ bias[2] and TE insertions preferentially occurring closer to the 3′ end[90].

Two main routes are known to contribute to an increase in the complexity of lncRNA loci during evolution, adoption of TEs and local duplications (FIG. 3Ba,Bb).

Adoption of TEs. TE insertions are typically heavily selected against within protein-coding exons, as they usually lead to disruption of the protein sequence. By contrast, depletion of TEs within lncRNA sequences is very weak[2], and numerous TE insertions are found in relatively highly conserved lncRNAs such as PVT1 (REF. 2) and cyrano[73,90]. A reasonable expectation is that most of the TE-derived sequences in lncRNAs are not functional, but are also not deleterious to lncRNA function, leading to weak selection against TE insertions. An intriguing question that will require extensive experimental studies is whether some TE-derived exons also correspond to functional domains acquired following TE exonization[98]; to date, functional elements in several lncRNAs[98], including XIST[99], UCHL1-AS1 (REF. 100) and ANRIL[101], have been mapped to regions derived from TEs.

Local duplications. Local repeats have been reported to be enriched in lncRNA loci[102], and several well-studied functional lncRNAs, such as XIST, FIRRE (functional intergenic repeating RNA element)[102,103], NORAD[62,63],

Table 2 | **Examples of lncRNAs with studied functions in multiple species**

lncRNAs	Level of conservation	Assay and function	Evidence of cross-species functionality	Refs
XIST	Conserved across eutherian mammals	Required for X chromosome inactivation in XX female cells	Introduction of the human X-inactivation region (~480 kb transgene) is sufficient for XIST induction, coating of the mouse autosomes and silencing of the mouse X chromosome	127,128
roX1 and roX2	Conserved in *Drosophila* species with subtle sequence and structure conservation	Required for dosage compensation by increasing expression from the X chromosome in males. *roX1* and *roX2* compound null is embryonic lethal in males	*roX* genes from other *Drosophila* species can rescue male lethality in *roX*-null *Drosophila melanogaster*	54,68
cyrano (also known as OIP5-AS1) and megamind (also known as TUNA)	Short sequence stretches conserved across vertebrates	Morpholino-mediated knockdown causes embryonic development defects	Zebrafish embryos can be rescued by co-injection of the human and mouse homologues of those lncRNAs*	73
Terminator and punisher	Sequence conservation between human and zebrafish	Morpholino-mediated knockdown causes embryonic development defects	Zebrafish embryos can be rescued by co-injection of the human homologues	129
LncMyoD	Positional conservation and subtle sequence homology between human and mouse (detectable in genome alignment but not by BLAST)	Knockdown in myoblasts causes differentiation defects and gene expression changes	Syntenic human lncRNA could rescue the expression of MHC after knockdown of the endogenous mouse LncMyoD	130
HID1	Conserved across land plants	Transcriptional regulation of the *PIF3* gene in *trans*	Expression of rice *HID1* gene rescued the hid1 elongated hypocotyl phenotype in *Arabidopsis thaliana*	131
Rsx	Conserved across metatherians (marsupials)	Associated with paternal X chromosome inactivation	Integration of an opossum *Rsx* transgene into an autosome in mouse embryonic stem cells leads to gene silencing in *cis*	132

lncRNA, long non-coding RNA; MHC, myosin heavy chain; roX, RNA on the X ; Rsx, RNA on the silent X; XIST, X-inactive specific transcript. *A recent study has shown that morpholino-mediated knockdown of *megamind* causes the same phenotype in wild-type fish and *megamind*-knockout fish, suggesting that this phenotype may result from an off-target effect[133]. Such an off-target effect would be surprising, as multiple independent morpholinos lead to the same phenotype. This observation may possibly be explained by the presence of two sequence-related paralogues of *megamind* in the zebrafish genome that may compensate for *megamind* loss, and targeting of these paralogues by morpholino may lead to the *megamind* phenotype in *megamind*-knockout fish.

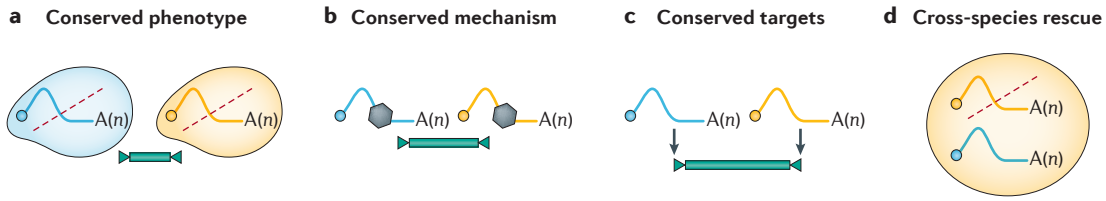


Figure 4 | **Manifestations of conserved functionality in lncRNA genes.** **a** | Loss of a homologous long non-coding RNA (lncRNA) in different species can result in the same phenotype. **b** | Homologous lncRNAs can act through a conserved mechanism. **c** | Target genes regulated by the lncRNAs can be the same. **d** | The loss of function of a lncRNA in one species can be rescued by the exogenous expression of the homologue from a different species. lncRNAs are shown as curved lines, with a 5′ cap (circle) and 3′ polyadenylated tail (A(*n*)). lncRNAs from different species are shown in blue versus yellow. Conserved function is indicated by the green bar and triangles; red dashed lines indicate experimental loss-of-function of a lncRNA; and the black hexagon represents an RNA-binding protein.

CDR1as[104], and roX1 and roX2 (REFS 53,68), harbour short repeated sequences. These repeats span a range of sequence similarities; for example, the repeats in FIRRE and XIST are highly similar to each other, whereas those in NORAD are very diverged[63]. These differences might reflect functional constraints on preserving inter-repeat similarity (which may facilitate higher-order structures[57]). Functionally, sequence duplications can endow a lncRNA with multiple platforms for the binding of factors; for example, miRNAs in the case of CDR1as[104], HNRNPU (heterogeneous nuclear ribonucleoprotein U) proteins by FIRRE[102], and PUM (pumilio) proteins by NORAD[62].

Conservation of lncRNA function

Do lncRNAs that have conserved sequences also act in similar ways across species? One easily quantifiable yet very crude proxy for the physiological function is the expression domain. When the spatial expression patterns of lncRNA homologues are compared across species, they are typically as conserved as those of mRNAs. Several studies have found that lncRNA tissue specificity, as well as specific expression patterns, are generally highly conserved[2,13]. Such conservation was also found when individual lncRNAs were compared with higher resolution of spatial expression using fluorescence *in situ* hybridization (FISH)[105]. Conserved lncRNAs are thus likely to act in similar contexts in different species.

Several lncRNAs have been tested for conservation of their functionality across species (TABLE 2). Although the number of tested cases remains too small to reach universal conclusions — or to understand when to expect conservation or divergence of function — the emerging picture is that relatively minor sequence conservation can be sufficient for maintaining conserved functions. Functional conservation can exhibit itself in different ways (FIG. 4): the loss-of-function phenotype of the lncRNA homologues can be similar; the molecular functionality can be conserved; the target genes affected by the lncRNA can be similar; or the lncRNA in one species can functionally replace its orthologue in another species. The last scenario is particularly useful, as cases in which the homologues from different species and artificial constructs are capable of rescuing a genetic null for a lncRNA[68] can be used to distil essential functional features of lncRNAs and validate predictions from comparative genomics.

The functional interrogation of lncRNA function *in vivo* is still in its infancy, and there are multiple methodological issues that need to be considered (see REF. 106 for a detailed discussion of the pros and cons of the available methods). Still, several lncRNAs were shown to have related loss-of-function phenotypes across species. For example, XIST is required for X inactivation in both human and mouse cells[107], loss of NEAT1 causes loss of paraspeckles across species[108,109], and CARMEN (cardiac mesoderm enhancer-associated non-coding RNA) is required for cardiomyogenesis in both humans and mice[110]. In the case of HOTAIR, a functional discrepancy between human and mouse lncRNAs has been reported: the human orthologue of HOTAIR was shown to regulate the expression of the *HOXD* (homeobox D) cluster in primary human fibroblasts[111], whereas *HoxD* expression was unaffected in mice in which the entire *HoxC* cluster (part of which encodes HOTAIR) has been deleted[112]. Interpretation of cross-species differences in this case is hindered by the use of different cells in human and mouse, and the fact that subsequently published targeted deletion of mouse HOTAIR did lead to a specific phenotype and upregulation of several *HoxD* genes[113].

When a lncRNA retains functionality across species, does it act through the same mechanism or targets? In most cases this remains unknown. In perhaps the most extensive study of conservation of lncRNA function, the genomic binding sites of the roX1 and roX2 lncRNAs were mapped in four *Drosophila* species, and it was found that although the functionality of the lncRNAs is conserved, their binding sites differ drastically across species, while maintaining some features such as proximity to genes[68]. When roX lncRNAs from other species were tested in *roX*-null *D. melanogaster*, they bound to the *D. melanogaster* binding sites, explaining the ability of those homologues to rescue the *roX*-null *D. melanogaster* mutants.

Notably, alongside these examples of lncRNAs with conserved functionality over large evolutionary distances, there are also numerous highly expressed and functional lncRNAs in mouse for which no clear human orthologue has been identified to date, including Braveheart[114] and Haunt[78,115], as well as functional primate-specific lncRNAs such as BDNFAS (BDNF antisense RNA)[116] and HPAT5 (human pluripotency-associated transcript 5)[117] that have no known mouse orthologues.

Concluding remarks

A rich experimental and computational toolbox is essential for tackling the multitude of questions about the extent and nature of lncRNA functions. Comparative genomics is an essential and increasingly used part of this toolbox, and comparative analyses have already yielded numerous insights into lncRNA biology. Better understanding of the molecular determinants of lncRNA action, improvements in the coverage and depth of lncRNA catalogues across species, new algorithms for identifying short islands of conservation in rapidly evolving loci, and systematic experimental evaluation of the functions of lncRNA homologues across species are all likely to increase substantially the utility of comparative analysis and its accessibility to researchers interested in individual lncRNAs.

Many of the emerging dogmas of lncRNA evolution are fragile and should be treated with the appropriate scepticism. Specifically, many of the following crucial questions will be resolved only through experiments. Are positionally conserved lncRNAs often functionally equivalent? Do functionally equivalent lncRNAs maintain short sequences or structural elements that are conserved but missed by current tools? Are there lncRNAs that are functionally conserved between vertebrates and other species, and did those independently evolve similar mechanisms of action? Answers to these questions will help to answer the bigger question of whether we are currently underestimating the extent of lncRNA conservation, and if we are not, and only few lncRNAs are conserved between distant species, to what extent do lncRNAs underlie phenotypic differences between species? Time will tell.

1. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
2. Hezroni, H. *et al.* Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* **11**, 1110–1122 (2015).
This study compares features and loci of lncRNAs across various vertebrates and shows rapid lncRNA turnover combined with conservation of expression patterns, and positional conservation without sequence conservation across large evolutionary distances.
3. Cabili, M. N. *et al.* Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* **16**, 20 (2015).
4. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
This study provides the first comprehensive RNA-seq-based catalogue of human lncRNAs and characterizes their features.
5. Gong, J., Liu, W., Zhang, J., Miao, X. & Guo, A. Y. lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse. *Nucleic Acids Res.* **43**, D181–D186 (2015).
6. Wapinski, O. & Chang, H. Y. Long noncoding RNAs and human disease. *Trends Cell Biol.* **21**, 354–361 (2011).
7. Pasquinelli, A. E. *et al.* Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. *Nature* **408**, 86–89 (2000).
8. Auyeung, V. C., Ulitsky, I., McGeary, S. E. & Bartel, D. P. Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell* **152**, 844–858 (2013).
9. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
10. Berezikov, E. Evolution of microRNA diversity and regulation in animals. *Nat. Rev. Genet.* **12**, 846–860 (2011).
11. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573 (1998).
12. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).
13. Washietl, S., Kellis, M. & Garber, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* **24**, 616–628 (2014).
References 12 and 13 are studies that comprehensively compare lncRNA sequence and expression evolution in various tetrapods.
14. Bu, D. *et al.* Evolutionary annotation of conserved long non-coding RNAs in major mammalian species. *Sci. China Life Sci.* **58**, 787–798 (2015).
15. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26–46 (2013).
16. Jenkins, A. M., Waterhouse, R. M. & Muskavitch, M. A. Long non-coding RNA discovery across the genus *Anopheles* reveals conserved secondary structures within and beyond the Gambiae complex. *BMC Genomics* **16**, 337 (2015).
17. Liu, J. *et al.* Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. *Plant Cell* **24**, 4333–4345 (2012).
18. Brown, J. B. *et al.* Diversity and dynamics of the *Drosophila* transcriptome. *Nature* **512**, 393–399 (2014).
19. Ravasi, T. *et al.* Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* **16**, 11–19 (2006).
20. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
21. Zhao, W. *et al.* Comparison of RNA-seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15**, 419 (2014).
22. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
23. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
24. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
25. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
26. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
27. Engstrom, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10**, 1185–1191 (2013).
28. Housman, G. & Ulitsky, I. Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochim. Biophys. Acta* **1859**, 31–40 (2015).
29. Kanitz, A. *et al.* Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16**, 150 (2015).
30. Chen, J. *et al.* Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol.* **17**, 19 (2016).
This study demonstrates a new methodology for detailed comparison of lncRNAs expressed in pluripotent stem cells in several species and suggests a classification of lncRNAs into groups based on their evolutionary histories.
31. Jayakodi, M. *et al.* Genome-wide characterization of long intergenic non-coding RNAs (lincRNAs) provides new insight into viral diseases in honey bees *Apis cerana* and *Apis mellifera*. *BMC Genomics* **16**, 680 (2015).
32. Mohammadin, S., Edger, P. P., Pires, J. C. & Schranz, M. E. Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. *BMC Plant Biol.* **15**, 217 (2015).
33. Wang, H. *et al.* Analysis of non-coding transcriptome in rice and maize uncovers roles of conserved lncRNAs associated with agriculture traits. *Plant J.* **84**, 404–416 (2015).
34. Paytuvi Gallart, A., Hermoso Pulido, A., Anzar Martinez de Lagran, I., Sanseverino, W. & Aiese Cigliano, R. GREENC: a Wiki-based database of plant lncRNAs. *Nucleic Acids Res.* **44**, D1161–D1166 (2016).
35. Bråte, J., Adamski, M., Neumann, R. S., Shalchian-Tabrizi, K. & Adamska, M. Regulatory RNA at the root of animals: dynamic expression of developmental lincRNAs in the calcisponge *Sycon ciliatum*. *Proc. Biol. Sci.* **282**, 20151746 (2015).
36. Gaiti, F. *et al.* Dynamic and widespread lncRNA expression in a sponge and the origin of animal complexity. *Mol. Biol. Evol.* **32**, 2367–2382 (2015).
37. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
This is the first study to use chromatin marks to improve the identification of lncRNAs in mouse and provides a detailed description of a set of lncRNAs that were better conserved than background.
38. Marques, A. C. & Ponting, C. P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.* **10**, R124 (2009).
39. Gardner, P. P. *et al.* Conservation and losses of non-coding RNAs in avian genomes. *PLoS ONE* **10**, e0121797 (2015).
40. Haerty, W. & Ponting, C. P. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol.* **14**, R49 (2013).
41. Zhang, Y. C. *et al.* Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. *Genome Biol.* **15**, 512 (2014).
42. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17**, 556–565 (2007).
43. Wang, J. *et al.* Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature* http://dx.doi.org/10.1038/nature03016 (2004).
44. Managadze, D., Rogozin, I. B., Chernikova, D., Shabalina, S. A. & Koonin, E. V. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol. Evol.* **3**, 1390–1404 (2011).
45. Kutter, C. *et al.* Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* **8**, e1002841 (2012).
This study compares in detail lncRNAs that are expressed in the liver in three rodents and reports rapid evolutionary turnover of lncRNAs, even when the same tissue is compared across closely related species.
46. Morán, I. *et al.* Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell. Metab.* **16**, 435–448 (2012).
47. Mustafi, D. *et al.* Evolutionarily conserved long intergenic non-coding RNAs in the eye. *Hum. Mol. Genet.* **22**, 2992–3002 (2013).

48. Tan, J. Y. *et al.* Extensive microRNA-mediated crosstalk between lncRNAs and mRNAs in mouse embryonic stem cells. *Genome Res.* **25**, 655–666 (2015).
49. Thomson, D. W. & Dinger, M. E. Endogenous microRNA sponges: evidence and controversy. *Nat. Rev. Genet.* **17**, 272–283 (2016).
50. Yang, J. R. & Zhang, J. Human long noncoding RNAs are substantially less folded than messenger RNAs. *Mol. Biol. Evol.* **32**, 970–977 (2015).
51. Spitale, R. C. *et al.* Structural imprints *in vivo* decode RNA regulatory mechanisms. *Nature* **519**, 486–490 (2015).
52. Wilusz, J. E. *et al.* A triple helix stabilizes the 3′ ends of long noncoding RNAs that lack poly(A) tails. *Genes Dev.* **26**, 2392–2407 (2012).
53. Ilik, I. A. *et al.* Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in *Drosophila*. *Mol. Cell* **51**, 156–173 (2013).
54. Park, S. W., Kuroda, M. I. & Park, Y. Regulation of histone H4 Lys16 acetylation by predicted alternative secondary structures in *roX* noncoding RNAs. *Mol. Cell. Biol.* **28**, 4952–4962 (2008).
55. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750–756 (2008).
56. Maenner, S. *et al.* 2D structure of the A region of *Xist* RNA and its implication for PRC2 association. *PLoS Biol.* **8**, e1000276 (2010).
57. Lu, Z. *et al.* RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* **165**, 1267–1279 (2016).
58. Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M. & Gorodkin, J. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.* **16**, 885–889 (2006).
59. Miller, W. *et al.* 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.* **17**, 1797–1808 (2007).
60. Gorodkin, J. *et al. De novo* prediction of structured RNAs from genomic sequences. *Trends Biotechnol.* **28**, 9–19 (2010).
61. Stadler, P. F. in *Advances in Bioinformatics and Computational Biology* (eds Ferreira, C. E. *et al.*) 1–12 (Springer, 2010).
62. Lee, S. *et al.* Noncoding RNA *NORAD* regulates genomic stability by sequestering PUMILIO proteins. *Cell* **164**, 69–80 (2016).
63. Tichon, A. *et al.* A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nat. Commun.* **7**, 12209 (2016).
64. Nam, J. W. & Bartel, D. P. Long noncoding RNAs in *C. elegans*. *Genome Res.* **22**, 2529–2540 (2012).
65. Smith, M. A., Gesell, T., Stadler, P. F. & Mattick, J. S. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* **41**, 8220–8236 (2013).
66. Somarowthu, S. *et al.* HOTAIR forms an intricate and modular secondary structure. *Mol. Cell* **58**, 353–361 (2015).
67. Rivas, E., Clements, J. & Eddy, S. R. Lack of evidence for conserved secondary structure in long noncoding RNAs. Preprint at http://eddylab.org/publications/RivasEddy16/RivasEddy16-preprint.pdf (2016).
68. Quinn, J. J. *et al.* Rapid evolutionary turnover underlies conserved lncRNA-genome interactions. *Genes Dev.* **30**, 191–207 (2016).
This study uses a novel computational approach for the sensitive detection of lncRNA homologues in insects and vertebrates based on a combination of synteny, sequence and structural information, and includes the first comparison of genomic binding sites of lncRNAs across species.
69. Tycowski, K. T., Shu, M. D., Borah, S., Shi, M. & Steitz, J. A. Conservation of a triple-helix-forming RNA stability element in noncoding and genomic RNAs of diverse viruses. *Cell Rep.* **2**, 26–32 (2012).
This study describes a sensitive approach for using a specific sequence-structure pattern to identify lncRNA homologues among extensively divergent viral genomes.
70. Kornienko, A. E., Guenzl, P. M., Barlow, D. P. & Pauler, F. M. Gene regulation by the act of long non-coding RNA transcription. *BMC Biol.* **11**, 59 (2013).
71. Latos, P. A. *et al. Airn* transcriptional overlap, but not its lncRNA products, induces imprinted *Igf2r* silencing. *Science* **338**, 1469–1472 (2012).
This is the most comprehensive study to date of a lncRNA for which only the act of transcription, and not any particular part of the sequence, is important for function.
72. Haerty, W. & Ponting, C. P. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* **21**, 333–346 (2015).
73. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537–1550 (2011).
74. He, Y. *et al.* The conservation and signatures of lincRNAs in Marek's disease of chicken. *Sci. Rep.* **5**, 15184 (2015).
75. Jiang, W., Liu, Y., Liu, R., Zhang, K. & Zhang, Y. The lncRNA *DEANR1* facilitates human endoderm differentiation by activating *FOXA2* expression. *Cell Rep.* **11**, 137–148 (2015).
76. Sone, M. *et al.* The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. *J. Cell Sci.* **120**, 2498–2506 (2007).
77. Paralkar, V. R. *et al.* Unlinking an lncRNA from its associated *cis* element. *Mol. Cell* **62**, 104–110 (2008).
78. Yin, Y. *et al.* Opposing roles for the lncRNA *Haunt* and its genomic locus in regulating *HOXA* gene activation during embryonic stem cell differentiation. *Cell Stem Cell* **16**, 504–516 (2015).
79. Marques, A. C. *et al.* Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biol.* **14**, R131 (2013).
This paper describes a classification of currently annotated lncRNAs into two groups (promoter-associated and enhancer-associated) with different features based on the chromatin signatures at their transcription start sites.
80. Legeai, F. & Derrien, T. Identification of long non-coding RNAs in insects genomes. *Curr. Opin. Insect Sci.* **7**, 37–44 (2015).
81. Li, L. *et al.* Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biol.* **15**, R40 (2014).
82. Wang, M. *et al.* Long noncoding RNAs and their proposed functions in fibre development of cotton (*Gossypium* spp.). *New Phytol.* **207**, 1181–1197 (2015).
83. Long, M., VanKuren, N. W., Chen, S. & Vibranovski, M. D. New gene evolution: little did we know. *Annu. Rev. Genet.* **47**, 307–333 (2013).
84. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).
85. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
This article provides a comprehensive description of lncRNA features and subcellular localization based on the Encyclopedia of DNA Elements (ENCODE) project data.
86. Duret, L., Chureau, C., Samain, S., Weissenbach, J. & Avner, P. The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312**, 1653–1655 (2006).
The paper is the first example of a lncRNA that evolved from a loss of coding potential of an ancestral protein-coding gene.
87. Romito, A. & Rougeulle, C. Origin and evolution of the long non-coding genes in the X-inactivation center. *Biochimie* **93**, 1935–1942 (2011).
88. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
89. Kelley, D. R. & Rinn, J. L. Transposable elements reveal a stem cell specific class of long noncoding RNAs. *Genome Biol.* **13**, R107 (2012).
90. Kapusta, A. *et al.* Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9**, e1003470 (2013).
91. Young, J. M. *et al.* DUX4 binding to retroelements creates promoters that are active in FSHD muscle and testis. *PLoS Genet.* **9**, e1003947 (2013).
92. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
93. Jensen, T. H., Jacquier, A. & Libri, D. Dealing with pervasive transcription. *Mol. Cell* **52**, 473–484 (2013).
94. Wu, X. & Sharp, P. A. Divergent transcription: a driving force for new gene origination? *Cell* **155**, 990–996 (2013).
95. Gotea, V., Petrykowska, H. M. & Elnitski, L. Bidirectional promoters as important drivers for the emergence of species-specific transcripts. *PLoS ONE* **8**, e57323 (2013).
96. Ruf, S. *et al.* Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat. Genet.* **43**, 379–386 (2011).
97. Soumillon, M. *et al.* Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3**, 2179–2190 (2013).
98. Johnson, R. & Guigo, R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* **20**, 959–976 (2014).
99. Elisaphenko, E. A. *et al.* A dual origin of the *Xist* gene from a protein-coding gene and a set of transposable elements. *PLoS ONE* **3**, e2521 (2008).
100. Carrieri, C. *et al.* Long non-coding antisense RNA controls *Uchl1* translation through an embedded SINEB2 repeat. *Nature* **491**, 454–457 (2012).
101. Holdt, L. M. *et al.* Alu elements in *ANRIL* non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through *trans*-regulation of gene networks. *PLoS Genet.* **9**, e1003588 (2013).
102. Hacisuleyman, E., Shukla, C. J., Weiner, C. L. & Rinn, J. L. Function and evolution of local repeats in the *Firre* locus. *Nat. Commun.* **7**, 11021 (2016).
103. Hacisuleyman, E. *et al.* Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat. Struct. Mol. Biol.* **21**, 198–206 (2014).
104. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).
105. Chodroff, R. A. *et al.* Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol.* **11**, R72 (2010).
106. Bassett, A. R. *et al.* Considerations when investigating lncRNA function *in vivo*. *eLife* **3**, e03058 (2014).
This paper provides important practical guidelines for choosing methods for perturbing lncRNA functions and interpreting the results.
107. Goto, T. & Monk, M. Regulation of X-chromosome inactivation in development in mice and humans. *Microbiol. Mol. Biol. Rev.* **62**, 362–378 (1998).
108. Sasaki, Y. T., Ideue, T., Sano, M., Mituyama, T. & Hirose, T. MENε/β noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proc. Natl Acad. Sci. USA* **106**, 2525–2530 (2009).
109. Cornelis, G., Souquere, S., Vernochet, C., Heidmann, T. & Pierron, G. Functional conservation of the lncRNA NEAT1 in the ancestrally diverged marsupial lineage: evidence for NEAT1 expression and associated paraspeckle assembly during late gestation in the opossum *Monodelphis domestica*. *RNA Biol.* http://dx.doi.org/10.1080/15476286.2016.1197482 (2016).
110. Ounzain, S. *et al. CARMEN*, a human super enhancer-associated long noncoding RNA controlling cardiac specification, differentiation and homeostasis. *J. Mol. Cell Cardiol.* **89**, 98–112 (2015).
111. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
112. Schorderet, P. & Duboule, D. Structural and functional differences in the long non-coding RNA *Hotair* in mouse and human. *PLoS Genet.* **7**, e1002071 (2011).
113. Li, L. *et al.* Targeted disruption of *Hotair* leads to homeotic transformation and gene derepression. *Cell Rep.* **5**, 3–12 (2013).
114. Klattenhoff, C. A. *et al. Braveheart*, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* **152**, 570–583 (2013).
115. Maamar, H. & Cabili, M. N., Rinn, J. & Raj, A. *linc-HOXA1* is a noncoding RNA that represses *Hoxa1* transcription in *cis*. *Genes Dev.* **27**, 1260–1271 (2013).
116. Lipovich, L. *et al.* Activity-dependent human brain coding/noncoding gene regulatory networks. *Genetics* **192**, 1133–1148 (2012).
117. Durruthy-Durruthy, J. *et al.* The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nat. Genet.* **48**, 44–52 (2016).
118. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
119. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).

120. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
121. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
122. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell Proteom.* **13**, 397–406 (2014).
123. Lerch, J. K. *et al.* Isoform diversity and regulation in peripheral and central neurons revealed through RNA-seq. *PLoS One* **7**, e30417 (2012).
124. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
125. Schwartz, M. P. *et al.* Human pluripotent stem cell-derived neural constructs for predicting neural toxicity. *Proc. Natl Acad. Sci. USA* **112**, 12516–12521 (2015).
126. Bergmann, J. H. *et al.* Regulation of the ESC transcriptome by nuclear long noncoding RNAs. *Genome Res.* **25**, 1336–1346 (2015).
127. Migeon, B. R. *et al.* Human X inactivation center induces random X chromosome inactivation in male transgenic mice. *Genomics* **59**, 113–121 (1999).
128. Heard, E. *et al.* Human *XIST* yeast artificial chromosome transgenes show partial X inactivation center function in mouse embryonic stem cells. *Proc. Natl Acad. Sci. USA* **96**, 6841–6846 (1999).
129. Kurian, L. *et al.* Identification of novel long noncoding RNAs underlying vertebrate cardiovascular development. *Circulation* **131**, 1278–1290 (2015).
130. Gong, C. *et al.* A long non-coding RNA, *LncMyoD*, regulates skeletal muscle differentiation by blocking IMP2-mediated mRNA translation. *Dev. Cell* **34**, 181–191 (2015).
131. Wang, Y. *et al. Arabidopsis* noncoding RNA mediates control of photomorphogenesis by red light. *Proc. Natl Acad. Sci. USA* **111**, 10359–10364 (2014).
132. Grant, J. *et al. Rsx* is a metatherian RNA with *Xist*-like properties in X-chromosome inactivation. *Nature* **487**, 254–258 (2012).
133. Kok, F. O. *et al.* Reverse genetic screening reveals poor correlation between morpholino-induced and mutant phenotypes in zebrafish. *Dev. Cell* **32**, 97–108 (2015).

Acknowledgements
The author thanks A. Shkumatava, A. Mallory, M. Garber, E. Hornstein, H. Hezroni and N. Gil for discussions and comments on the manuscript. I.U. is the Sygnet Career Development Chair for Bioinformatics and recipient of an Alon Fellowship from The Council for Higher Education of Israel. Work in the Ulitsky laboratory is supported by grants to I.U. from the European Research Council (Project lincSAFARI), the Israeli Science Foundation (1242/14 and 1984/14), the Israeli Centers of Research Excellence (I-CORE) Program of the Planning and Budgeting Committee and The Israel Science Foundation (1796/12), the Minerva Foundation, the Fritz-Thyssen Foundation and by research grants from Lapon Raymond and the Abramson Family Center for Young Scientists.

Competing interests statement
The author declares no competing interests.

DATABASES
Ensembl Compara: http://ensembl.org/info/genome/compara/index.html
GreeNC: http://greenc.sciencedesigners.com
HMMER: http://hmmer.org
lncRNAdb: http://lncrnadb.org
NONCODE: http://www.noncode.org
phyloNONCODE: http://www.bioinfo.org/phyloNoncode
PLAR: http://webhome.weizmann.ac.il/home/igoru/PLAR
PLNlncRbase: http://bioinformatics.ahau.edu.cn/PLNlncRbase
RNAcentral: http://rnacentral.org
UCSC Genome Browser: https://genome.ucsc.edu

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

ERRATUM

Evolution to the rescue: using comparative genomics to understand long non-coding RNAs

Igor Ulitsky

Nature Reviews Genetics http://dx.doi.org/10.1038/nrg.2016.85

In the original version of this article, the sentence "A study using a different background model recently reported more than 4 million regions that are evolving under selection to preserve secondary structure" (section 'Secondary structure and its conservation') was missing a citation of reference 65 (Smith, M. A., Gesell, T., Stadler, P. F. & Mattick, J. S. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* **41**, 8220–8236 (2013)). This citation dropped out during journal typesetting of the article and has now been reinstated. The editors apologize for this error.