Cell

Leading Edge
Review

lincRNAs: Genomics, Evolution, and Mechanisms

Igor Ulitsky[1,2] and David P. Bartel[1,2,*]
[1]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA
[2]Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
*Correspondence: dbartel@wi.mit.edu
http://dx.doi.org/10.1016/j.cell.2013.06.020

Long intervening noncoding RNAs (lincRNAs) are transcribed from thousands of loci in mammalian genomes and might play widespread roles in gene regulation and other cellular processes. This Review outlines the emerging understanding of lincRNAs in vertebrate animals, with emphases on how they are being identified and current conclusions and questions regarding their genomics, evolution and mechanisms of action.

Introduction

The conventional view of the mammalian genome was that ~20,000 protein-coding genes were dispersed within mostly repetitive and largely nontranscribed sequence. Over the past decade, this view has been challenged by increasingly thorough examinations of the RNA species in mammalian cells. These studies have revealed the fascinating complexity of the transcriptome, in which protein-coding genes produce many alternative products, and genomic regions previously thought to be transcriptionally silent give rise to a range of processed and regulated transcripts that do not appear to code for functional proteins. A few of these transcripts are precursors for small regulatory RNAs, such as microRNAs, but the vast majority have no recognizable purpose.

A sensible hypothesis is that most of the currently annotated long (typically >200 nt) noncoding RNAs are not functional, i.e., most impart no fitness advantage, however slight. Like all biochemical processes, the transcription machinery is not perfect and can produce spurious RNAs that have no purpose (Struhl, 2007). Due to the intrinsic properties of RNA, these transcripts would have a collapsed fold (Schultes et al., 2005). Because chromatin states vary across cell fates, cryptic promoters would be differentially accessible in different cellular contexts, and thus many spurious transcripts would also have tissue-specific expression. Because of the underlying transcriptional processes and chance occurrence of splice sites, many would also be capped, spliced, and polyadenylated. Thus, none of these features offer an informative indicator of function. Moreover, many of these spurious RNA species that confer no fitness advantage would also impose minimal fitness cost, in which case, simply tolerating them would be more feasible than evolving and maintaining more rigorous control mechanisms that could prevent their production. A second source of nonfunctional RNAs would be those generated during regulatory events in which the act of transcription matters, whereas the product of transcription does not. These would include RNAs generated during transcriptional interference, which involves transcription of noncoding loci that overlap regulatory regions and is known to regulate gene expression in both prokaryotes and eukaryotes (Shearwin et al., 2005). Against this backdrop of many nonfunctional transcripts, some long noncoding RNAs, including the *Xist* RNA, which is required for mammalian dosage compensation (Penny et al., 1996), clearly are functional, and the roster of biological processes in which long noncoding RNAs are reported to play key roles is rapidly growing and now includes cell-cycle regulation, apoptosis, and establishment of cell identity (reviewed in Ponting et al., 2009; Pauli et al., 2011; Rinn and Chang, 2012).

Despite general agreement that some long noncoding RNAs are functional and others are not, opinions vary widely as to the fraction that is functional (Kowalczyk et al., 2012). Because of their marginal sequence conservation and a sense that spurious transcripts would impose minimal fitness cost, we suspect that most are not functional. However, even a scenario in which only 10% are functional implies the existence of more than a thousand human loci generating noncoding transcripts with biological roles. These enigmatic RNAs will consume decades of effort for many labs undertaking molecular, mechanistic, and phenotypic analyses. And regardless of function, long noncoding RNAs might have diagnostic applications, with changes in their expression already associated with cancer and several neurological disorders (Prensner et al., 2011; Brunner et al., 2012; Ziats and Rennert, 2013).

To identify noncoding RNAs and their corresponding genes cleanly, and to simplify their analysis by avoiding the complications arising from overlap with other types of genes, recent focus has been on long intervening noncoding RNAs (lincRNAs, also called long ''intergenic'' noncoding RNAs even though the lincRNAs derive from genes and are thus genic), which do not overlap exons of either protein-coding or other non-lincRNA types of genes. Here, we also focus on this subgroup, as lincRNA gene expression patterns, sequence conservation and perturbation outcomes are easier to interpret than those of transcripts from loci overlapping other gene classes. We presume that the features of lincRNAs will also apply to many other long noncoding RNA transcripts that were excluded from lincRNA lists

Table 1. Large-Scale Efforts to Catalog lincRNA Loci and Transcripts

Reference	Data for Transcript Reconstruction	Genomic Features and Filters	Coding-Potential Filters	Number of lincRNAs
Mouse				
Ravasi et al., 2006	cDNAs		Manual curation, ORF length, CRITICA	13,502 transcripts
Ponjavic et al., 2007	cDNAs, CAGE		Manual curation, ORF length, BLAST, CRITICA	3,122 transcripts
Guttman et al., 2009	Chromatin marks, tiling arrays	Collection of approximate exonic regions, chromatin domain ≥5 kb	CSF	1,675 loci (1,250 conservatively defined)
Guttman et al., 2010	RNA-seq	Multi-exon only	CSF	1,140 lincRNA transcripts
Sigova et al., 2013	RNA-seq, cDNAs, chromatin marks,	Antisense overlap with mRNA introns allowed, ≥100 nt mature length	CPC	1,664 loci
Human				
Khalil et al., 2009	Chromatin marks, tiling arrays	Collection of approximate exonic regions, chromatin domain ≥ 5 kb	CSF	3,289 loci
Jia et al., 2010	cDNAs	Overlap with mRNAs allowed		5,446 transcripts
Ørom et al., 2010	cDNAs	Restricted to loci >1 kb away from known protein-coding genes, ≥200 nt mature length	Manual curation based on length, conservation and other characteristics of the ORFs	3,019 transcripts from 2,286 loci
Cabili et al., 2011	RNA-seq	Multi-exon only, ≥200 nt mature length	PhyloCSF, Pfam	8,195 transcripts (4,662 in the stringent set)
Derrien et al., 2012	cDNAs	Overlap with mRNAs allowed (intergenic transcripts reported separately), ≥200 nt mature length	Manual curation based on length, conservation and other characteristics of the ORFs	14,880 transcripts from 9,277 loci, including 9,518 intergenic transcripts
Sigova et al., 2013	RNA-seq, cDNAs, chromatin marks,	Antisense overlap with mRNA introns allowed, ≥100 nt mature length	CPC	3,548 loci from embryonic stem cells, and 3,986 loci from endodermal cells
Frog				
Tan et al., 2013	RNA-Seq	>25 kb away from known protein-coding genes or on a different strand from the neighboring genes, ≥200 nt mature length	ORF length, BLAST, Pfam	6,686 transcripts from 3,859 loci
Zebrafish				
Ulitsky et al., 2011	RNA-seq, cDNAs, 3P-seq, chromatin marks	Antisense overlap with mRNA introns allowed, ≥200 nt mature length	CPC	691 transcripts from 567 loci
Pauli et al., 2012	RNA-seq	Stringent criteria for single exon, intron overlap with mRNA allowed, ≥160 nt mature length	ORF length, PhyloCSF, BLAST, Pfam	397 intergenic and 184 intronic overlapping transcripts
Fly				
Tupy et al., 2005	cDNA		Manual curation based on ORF length, conservation and other characteristics, K_a/K_s test, QRNA	17 transcripts
Young et al., 2012	RNA-seq	≥200 nt locus length		1,119 trancripts
Nematode				
Nam and Bartel, 2012	RNA-seq, 3P-seq	≥100 nt mature length	CPC, RNAcode, ribosome profiling, polysome association	262 lincRNA transcripts from 170 loci

(*Continued on next page*)

Table 1. *Continued*

Reference	Data for Transcript Reconstruction	Genomic Features and Filters	Coding-Potential Filters	Number of lincRNAs
Arabidopsis				
Liu et al., 2012a	cDNA, tiling arrays, RNA-seq	In part a collection of approximate exonic regions, >500 bp away from protein-coding genes, no overlap with transposable elements allowed, ≥200 nt mature length	ORF length	6,480 transcription units from tiling arrays, 278 transcripts from RNA-seq
Maize				
Boerner and McGinnis, 2012	cDNA	Both sense overlap with introns and antisense overlap with mRNA or introns allowed, ≥200 nt mature length	ORF length	2,492 transcripts
Plasmodium falciparum				
Broadbent et al., 2011	Tiling arrays	Collection of approximate exonic regions, ≥200 nt mature length	BLAST	60 transcripts

Transcripts overlapping protein-coding sequences on either strand were excluded unless noted otherwise. Coding-potential filters included: ORF length; similarity to known protein-coding regions (BLAST); substitution patterns in whole-genome alignments, quantified by CRITICA (Badger and Olsen, 1999), CSF (Lin et al., 2007), PhyloCSF (Lin et al., 2011), QRNA (Rivas and Eddy, 2001; Rivas et al., 2001), or RNAcode (Washietl et al., 2011), as indicated; the CPC algorithm, which evaluates ORF properties and similarity to known proteins (Kong et al., 2007); the HMMER algorithm, which tests for potential to encode a known protein domain (Pfam); ribosome profiling, and polyribosome association. Criteria used to define the lincRNA collection (and not those used only for characterization) are listed.

because of complicating (albeit, often functionally inconsequential) overlap with other annotations.

At the outset, we emphasize that lincRNA classification differs from that of other RNAs, in that lincRNAs are defined more by what they are not than by what they are. As is typical of stable RNA polymerase II products, lincRNAs are nearly always capped and polyadenylated, and are frequently spliced. But aside from this positive descriptor of being Pol II products, lincRNAs are defined using negative descriptors, i.e., *not* coding for proteins and *not* overlapping transcripts of certain other types of genes. Reliance on these negative descriptors risks grouping together a hodgepodge of transcripts with very diverse properties and mechanisms of action. In many ways the lincRNA field faces challenges similar to those faced by early biologists trying to categorize and contemplate the diverse array of life forms that were not plants and not animals. We suspect that there might be dozens of distinct functional noncoding RNA classes that have transcripts currently grouped into the catch-all class of lincRNAs. Until these classes are understood and differentiated, insights from the study of one lincRNA will be difficult to apply to others, and attempts to understand the general features of lincRNAs will at best reflect only the more populated classes. With these caveats in mind, we review the current understanding of vertebrate lincRNAs, focusing on their identification, genomics, evolution and mechanisms of action.

lincRNA Identification

lincRNAs and lincRNA candidates have been cataloged in human, mouse, zebrafish, frog, fly, nematode, *Arabidopsis*, maize, and *Plasmodium* (Table 1). Interrogation of lincRNA function or mechanisms depends on high-quality transcript models of lincRNA genes, including accurate genomic positions of the start site, splice sites, and polyadenylation site of each transcript. Useful collections of lincRNAs are those that capture full-length transcripts and avoid those encoding functional peptides. Methodological advances and increased throughput are continuously improving the ability to meet these goals and help explain the diversity of annotation criteria and cutoffs (Table 1), which in turn might be one of the reasons lincRNA lists from different studies do not have more overlap.

Because of their poly(A) tails and other mRNA-like features, lincRNAs are represented in typical cDNA cloning, tiling array, and RNA-seq data sets. The first large-scale catalog of putatively noncoding transcripts came from the FANTOM project (Okazaki et al., 2002; Carninci et al., 2005), which used cDNA cloning followed by Sanger sequencing and reported >34,000 long noncoding RNAs expressed in different mouse tissues, of which 3,652 had confident support (Ravasi et al., 2006). Subsequent studies refined EST- and cDNA-based lincRNA catalogs in mouse and human, which comprise the current RefSeq and Ensembl lincRNA annotations (Derrien et al., 2012; Pruitt et al., 2012). In parallel, tiling microarrays were used to detect transcribed regions (Bertone et al., 2004; Guttman et al., 2009; Khalil et al., 2009), which was potentially more sensitive than cloning but suffered from reduced dynamic range and difficulties in defining splice junctions and connecting transcribed regions into transcript models (Agarwal et al., 2010). More recently, high-throughput sequencing of millions of short RNA fragments (RNA-seq) is enabling transcript models to be reconstructed, either with the aid of a reference genome (Trapnell et al., 2010; Cabili et al., 2011) or without it (Grabherr et al., 2011). RNA-seq has yielded billions of strand-specific paired-end reads of

~100 nt each, and those can be sufficient for reconstruction of even very lowly expressed transcripts (Cabili et al., 2011; Pauli et al., 2012). Furthermore, even rarer transcripts can be specifically enriched using array-based capture methods prior to sequencing (Mercer et al., 2012).

Despite the advantages of RNA-seq in terms of sensitivity and accessibility, assembly of transcript models from short reads still has limitations, stemming primarily from the relatively small portion of the full transcript accounted for by each read and from sequence redundancies in the genome. It remains difficult to determine which exon combinations co-occur in long multiply spliced transcripts and to discriminate between independent lincRNAs and fragments of alternative mRNA isoforms or pseudogenes. Focusing only on spliced transcripts helps improve specificity (Cabili et al., 2011) but misses some bona fide single-exon lincRNAs, such as *Malat1* and *Neat1* (Hutchinson et al., 2007). Therefore, curated lincRNA databases (e.g., RefSeq and Ensembl) still rely primarily on cDNA sequences obtained using Sanger sequencing (Derrien et al., 2012), but we expect that this will change soon, as read lengths for high-throughput sequencing methods continue to improve and as multiple data sets are more effectively integrated to build models.

Additional data sets that can improve transcript models include chromatin maps and data from methods used to identify transcript start and polyadenylation sites (Figure 1A). Trimethylation of lysine 4 and lysine 36 in histone H3 (H3K4me3 and H3K36me3 marks), which characterize regions of Pol II transcription initiation and elongation, respectively, were used in conjunction with tiling arrays for building some lincRNA collections (Guttman et al., 2009; Khalil et al., 2009). These maps have limitations, however, as peaks of H3K4me3 can be broad and also occur at the first exon-intron junction (Bieberstein et al., 2012) (Figure 1A), and H3K36me3 enrichment is dependent on splicing and typically extends beyond the polyadenylation site (de Almeida et al., 2011) (Figure 1A). Other sources of supporting data have come from high-throughput sequencing experiments tailored to identify specific regions within RNA molecules. These include methods for high-resolution mapping of transcription start sites, e.g., using cap analysis of gene expression (CAGE) (Kodzius et al., 2006), and genome-wide annotation of polyadenylation sites, e.g., using 3P-seq (Jan et al., 2011; Ulitsky et al., 2012) (Figure 1A). A combination of independent evidence for transcription initiation, termination and exon-intron structure can enable confident identification of both multiple- and single-exon lincRNAs (Ulitsky et al., 2011).

Criteria for Distinguishing between Coding and Noncoding Transcripts

Perhaps the most challenging aspect of lincRNA discovery is that the concept of a noncoding RNA is loosely defined. Most long transcripts with known noncoding functions typically contain multiple potential open reading frames (ORFs). These ORFs might not be translated, might be translated inefficiently, or might be translated to produce a protein that has no functional consequences, e.g., because it is rapidly degraded. Due to their considerable lengths, many lincRNAs should by chance contain an ORF of at least 100 aa (Dinger et al., 2008). A clear binary separation between coding and noncoding transcripts is thus impossible, and the best that can be done is to use graded and imperfect criteria that preferentially identify transcripts that are *unlikely* to code for *functional* proteins.

Several features of bona fide protein-coding genes can be used as criteria to distinguish them from lincRNAs (Figure 1B, Table 1): (1) coding regions tend to be much longer than expected by chance (Dinger et al., 2008); (2) nucleotide frequencies of functional ORFs are dictated by nonrandom codon usage; (3) during evolution, selective pressures bias nucleotide substitutions in coding sequences (e.g., giving rise to a higher substitution rates in the silent positions of codons); (4) protein-coding genes typically contain known protein domains (e.g., present in the Pfam database); (5) coding regions are likely to bear sequence similarities to entries in protein databases. Different studies use different combinations of these five criteria in attempts to exclude protein-coding genes. The underlying assumption across these criteria is that short, recently evolved yet functional proteins are relatively rare. In support of this assumption, the current protein databases list very few functional peptides that originate from short ORFs—disregarding pseudogenes, Ensembl 68 lists only 11 human protein-coding genes that have a known function (described in Gene Ontology annotations) and an ORF < 50 aa, and none of these are shorter than 30 aa. (Note that most short peptides with known functions arise from longer ORFs because they are processed from longer precursors.)

Each of the criteria for predicting coding potential is of limited utility when used in isolation. For instance, presence of an ORF of at least 300 nt (100 aa) is commonly used for defining a transcript as coding. However, a transcript of 2 kb is expected to have an ORF of about 200 nt, and an ORF of 300 nt is only one standard deviation longer than expected (Dinger et al., 2008). Indeed, well characterized human lincRNAs, such as *H19*, *Xist*, *Meg3*, *Hotair*, and *Kcnq1ot1* all have ORFs of at least 100 aa (Dinger et al., 2008). Even significant similarity to "known" protein-coding genes might be misleading, as protein databases contain large numbers of protein sequences predicted by translation of the longest ORF in sequenced cDNAs but without any further functional evidence. Using a combination of filters can address some of these problems (Badger and Olsen, 1999; Liu et al., 2006; Kong et al., 2007), though the scarcity of standards (in particular, long RNAs known to have exclusively noncoding functions) makes calibration of these difficult. An interim solution is to assemble two collections of transcript models, one with confidently predicted lincRNAs and another for which the evidence is less conclusive (referred to as transcripts of unknown coding potential or TUCPs) (Cabili et al., 2011).

Methods for focused experimental interrogation of the coding potential of a lincRNA include testing whether the transcript can yield peptides when translated in vitro (Lanz et al., 1999; Galindo et al., 2007), testing whether it associates with polysomes (Brockdorff et al., 1992), and checking if its ORFs can yield a protein when fused to a sequence coding for a peptide for which antibodies are available (Anguera et al., 2011).

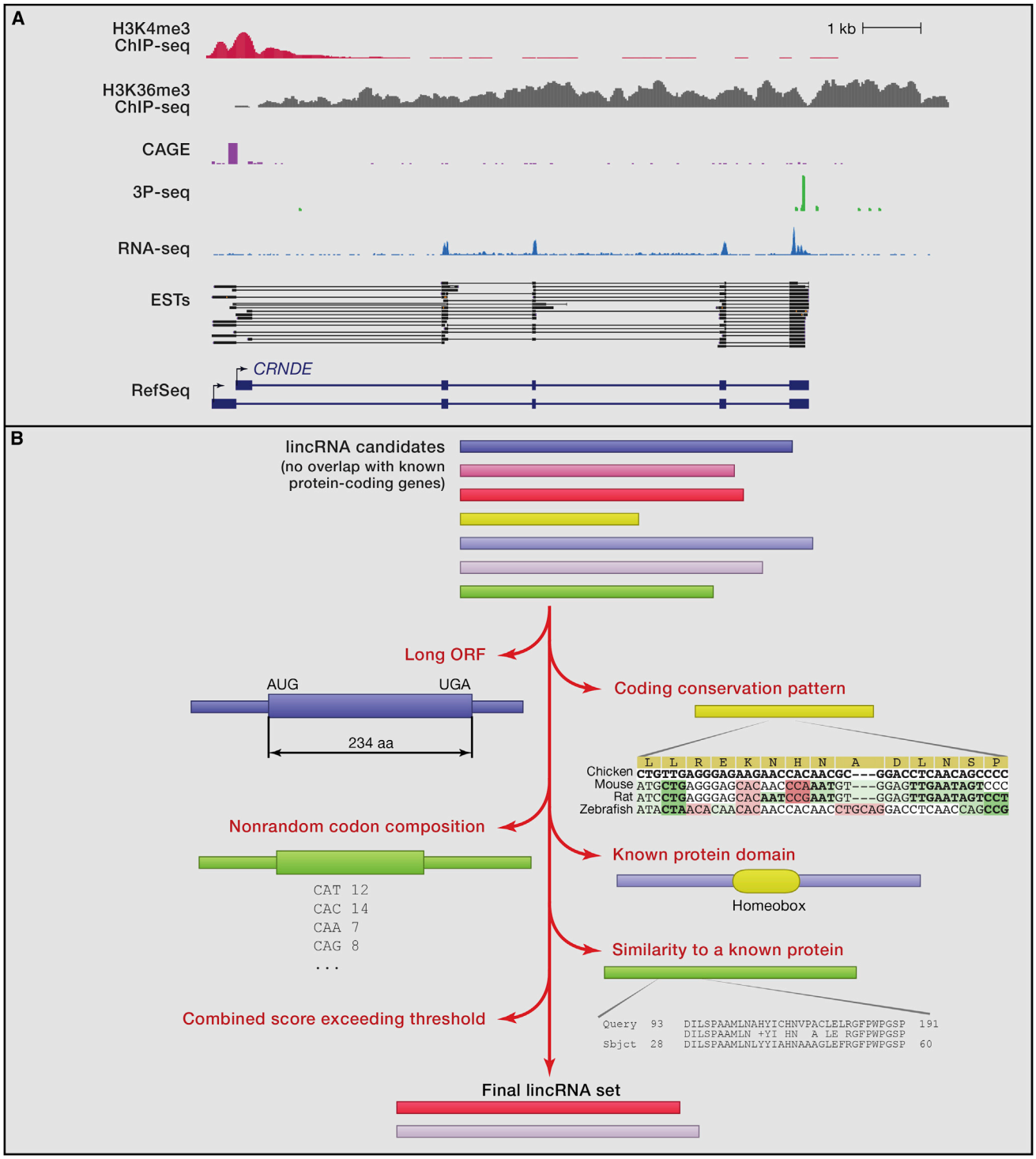

Figure 1. Assembling lincRNA Collections

(A) Data sets useful for constructing lincRNA transcript models. Information from the indicated genome-wide data sets are plotted for the *CRDNE* lincRNA locus (chr16:54,950,197-54,963,922 in the human hg19 assembly). A subset of ESTs from GenBank and the corresponding RefSeq annotations are also shown. ChIP-seq and CAGE (ENCODE project, HeLaS3 cells), 3P-Seq (HeLa cells, C. Jan and D.P.B., unpublished data), RNA-seq (HeLa cells; Guo et al., 2010) were plotted using the UCSC genome browser.

(B) A generic lincRNA annotation pipeline, illustrating criteria used to filter potential mRNAs from the list of candidates.

However, an ability to recruit the ribosome and be translated would not preclude a noncoding function. If the gene function can be assayed, the best approach is to introduce changes that perturb the ORF, such as those inducing frameshifts, and test for retention of the function (Hu et al., 2011; Ulitsky et al., 2011).

Global approaches can also show which transcripts are translated. Particularly useful is ribosome profiling, which utilizes high-throughput sequencing to map RNA regions associated with translating ribosomes (Ingolia et al., 2011). Analysis of ribosome profiling of mouse embryonic stem cells suggests that as many as half of the lincRNAs expressed in these cells are significantly associated with ribosomes (Ingolia et al., 2011). One interpretation of this observation is that the assumption of very few genes with short ORFs coding for functional peptides is wrong and that many of the currently annotated lincRNAs are in fact coding for short functional peptides. An example frequently cited in support of this interpretation is the *Drosophila tarsal-less/polished rice* transcript, which was originally thought to function as a long noncoding RNA but subsequently shown to code for very short functional peptides (Tupy et al., 2005; Kondo et al., 2010).

Although other examples of unrecognized functional peptides will undoubtedly be found, several lines of evidence suggest that this interpretation does not explain most of the ribosome association. First, as mentioned above, the algorithms used for generating lincRNA collections typically use sequence alignment to detect signatures of coding sequence conservation, and would detect at least those short coding regions that are highly conserved. Second, ribosomes are associated with some lincRNAs known to be enriched and function in the nucleus, such as *Malat1* and *Neat1*, suggesting that those transcripts have some background engagement with ribosomes (presumably when they occasionally reach the cytoplasm) even though their known nuclear functions are noncoding. Third, a recent proteomics study that specifically focused on identifying short endogenous peptides detected peptides from only eight (0.4%) of the lincRNAs expressed in the human K562 cell line, and the extent to which even these peptides are functional is unknown (Slavoff et al., 2013). Fourth, and perhaps most important, is the concept of lincRNA upstream ORFs (uORFs; see below).

lincRNA uORFs

Engagement with the translating ribosome can serve purposes that have nothing to do with the translation product. Indeed, the ribosome profiling study that reported ribosome engagement in many lincRNAs reported similar engagement in annotated 5′UTRs of thousands of mRNAs, yet in contrast to translation in lincRNAs, translation of these short uORFs was not proposed to produce functional peptides (Ingolia et al., 2011). uORF translation typically plays regulatory roles, affecting translation of downstream ORFs or mRNA stability (Calvo et al., 2009; Wethmar et al., 2010). Consistent with the idea that the act of uORF translation, which can be the basis of the regulatory mechanism, is more important than the product of this translation, short peptides translated from uORFs are rarely conserved in sequence (Crowe et al., 2006), can be very unstable (Hackett et al., 1986) and are rarely detectable in mass-spectrometry-based proteomic data (Menschaert et al., 2013). We suggest that the same might be true for lincRNAs. The translated ORFs in lincRNAs might act as uORFs to prevent ribosome scanning or translation in downstream regions of the transcripts, thereby enabling the lincRNAs to perform noncoding functions in the cytoplasm without interference from the ribosome (Figures 2A and 2B). lincRNA uORFs might also tether factors to ribosomes (Figure 2C) or modulate the stability of the lincRNA by influencing RNA decay pathways, some of which depend on translation (Figure 2D).

At the molecular level, most lincRNAs appear indistinguishable from mRNAs, with 5′-m^{7}GpppN cap structures, poly(A) tails, and exon-exon splice junctions, all of which stimulate mRNA translation (Shoemaker and Green, 2012). When considering these mRNA-like features, combined with the realization that most lincRNAs have a significant presence in the cytoplasm (see Subcellular Localization, below), the question is not: why are so many lincRNAs associated with ribosomes? The relevant question is: why are only half of the annotated lincRNAs associated with ribosomes? An important focus of future research will be determining how lincRNA export from the nucleus is regulated and how the cytoplasmic lincRNAs that do not depend on uORFs manage to avoid the translation machinery.

Bifunctional RNAs

The hypothesis that many lincRNAs have uORFs, which produce peptides, albeit nonfunctional ones, takes some liberties with the concept of noncoding RNA (although perhaps not as great as the liberties taken when speaking of uORFs falling in 5′UTRs, i.e., "untranslated regions"). Classification of noncoding transcripts is further complicated by the fact that some transcripts can have both coding and noncoding functions (Dinger et al., 2008). *Xenopus* and *E. coli* each provide an example in which the identical mature RNA embodies both coding and noncoding functions (Kloc et al., 2005; Wadler and Vanderpool, 2007). However, known examples of mRNAs moonlighting as long noncoding RNAs are still scarce, perhaps because of the challenges in identifying which mRNAs also have noncoding functions. When the coding and noncoding functions emerge at different times during evolution or when the noncoding function outlives the loss of ancestral coding potential of bifunctional mRNA, noncoding and coding transcripts with similar sequence might be found in different contemporary species, and the identification of such instances could potentially expedite the discovery of some bifunctional transcripts (Ulitsky et al., 2011; Marques et al., 2012).

lincRNA Genomics

As expected for a mixture of multiple classes of noncoding RNAs, lincRNAs lack defining sequence or structure characteristics. Nonetheless, several general features of lincRNAs in vertebrates are apparent in recent catalogs of human and zebrafish lincRNAs (Cabili et al., 2011; Ulitsky et al., 2011; Derrien et al., 2012; Pauli et al., 2012).

lincRNA genes are typically shorter than protein-coding genes (Ulitsky et al., 2011; Derrien et al., 2012; Pauli et al., 2012) and

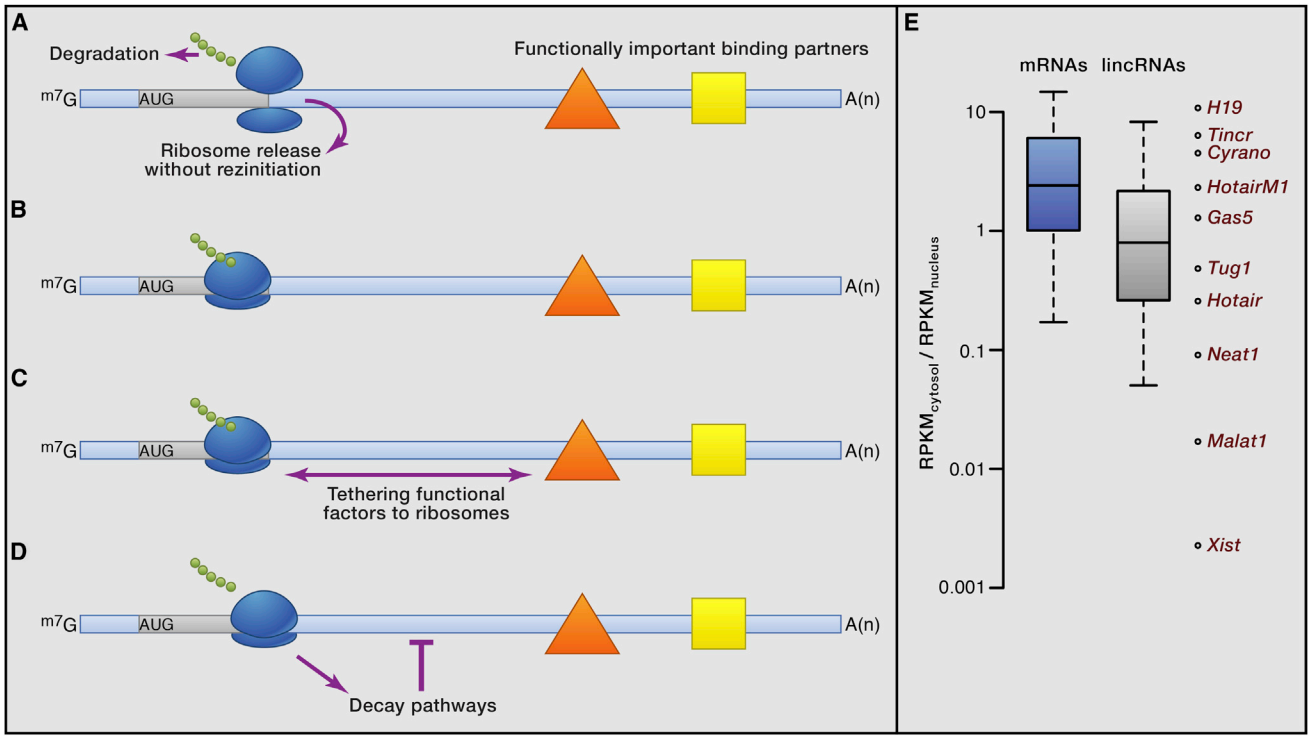

Figure 2. Ribosomal Association and Subcellular Localization of lincRNAs

(A) A potential role for a lincRNA uORF. Translation of a uORF into a peptide that is rapidly degraded would prevent ribosomal scanning of downstream regions, thereby protecting downstream binding factors from displacement by scanning ribosomes.

(B) Translating a nascent peptide sequence that induces ribosomal stalling would achieve an effect similar to that described in (A).

(C) The uORF can recruit a ribosome, which might be important for downstream lincRNA function.

(D) The translation of a uORF might influence the susceptibility of the lincRNA to different RNA decay pathways, such as nonsense-mediated decay (NMD).

(E) Relative subcellular localization of mRNAs and lincRNAs in MCF-7 cells. mRNA annotations were from Ensembl, and lincRNA annotations were from Ensembl, Refseq and (Cabili et al., 2011). RPKM (reads per kilobase per million mapped reads) values were computed with Cufflinks (Trapnell et al., 2010) using RNA-seq data for nuclear and cytoplasmic fractions of MCF-7 cells (Djebali et al., 2012). Ratios for selected lincRNAs are indicated.

have fewer exons, typically only 2–3 (Cabili et al., 2011; Derrien et al., 2012; Pauli et al., 2012). Exons in lincRNA genes are on average slightly longer than exons in protein-coding genes (Ravasi et al., 2006; Derrien et al., 2012), presumably because the average estimate is skewed by typically longer first and last exons (Zhu et al., 2009). Transcriptional regulation, chromatin-modification patterns, and splicing signals of lincRNAs are similar to those of protein-coding genes (Ponjavic et al., 2007; Cabili et al., 2011; Derrien et al., 2012; Pauli et al., 2012), although lincRNA transcripts seem somewhat less efficiently spliced (Tilgner et al., 2012).

Most annotated lincRNAs are polyadenylated, although alternative 3′-end topologies are also occasionally observed. In humans, there are ~80 lincRNAs with circular isoforms—far fewer than the nearly 2,000 human mRNAs with circular isoforms identified in the same study (Memczak et al., 2013). A few other lincRNAs are stabilized by a triple-helical structure at their 3′ end (Brown et al., 2012; Wilusz et al., 2012) or by snoRNAs at both ends (Yin et al., 2012).

lincRNAs from human, mouse, and zebrafish are significantly more likely than mRNAs to overlap repetitive elements (Ulitsky et al., 2011; Kelley and Rinn, 2012), perhaps because lincRNA functions are more tolerant of retrotransposon insertions. Repetitive elements are also reported to play important mechanistic roles in lincRNAs, by facilitating base pairing with other RNAs containing repeats from the same family (Gong and Maquat, 2011) or through other, less understood mechanisms (Carrieri et al., 2012). Tandem repeats are also prevalent and occasionally functionally important in lincRNA genes: at least eight different tandem-repeat groups are found in *Xist*, seven in the first and functionally important exon (Nesterova et al., 2001; Elisaphenko et al., 2008; Zhao et al., 2008), and repetitive regions were also found within the functional domains of *Miat* (Tsuiji et al., 2011), *DBE-T* (Cabianca et al., 2012), *CDR1as/ciRS-7* (Hansen et al., 2013; Memczak et al., 2013), and other lincRNAs.

lincRNA genes are preferentially found within 10 kb of protein-coding genes (Bertone et al., 2004; Ponjavic et al., 2009; Jia et al., 2010; van Bakel et al., 2010; Cabili et al., 2011; Sigova et al., 2013), which has led to the suggestion that many lincRNAs are byproducts of mRNA biogenesis (van Bakel et al., 2010). Countering this idea are analyses showing that (1) genomic colocalization persists in collections of lincRNAs supported by independent evidence for transcription initiation and termination, and (2) the distribution of distances between lincRNAs and their closest protein-coding genes resembles that of adjacent protein-coding genes (Ulitsky et al., 2011).

Studies in human, mouse, and zebrafish suggested that large gene deserts flanking transcription-factor (TF) genes, particularly those with roles in embryonic development, preferentially harbor lincRNAs (Mercer et al., 2008; Guttman et al., 2009; Ulitsky et al., 2011; Pauli et al., 2012; Wamstad et al., 2012). In vertebrates, developmental TF genes are preferentially surrounded by long intergenic regions (Ovcharenko et al., 2005), and these regions are enriched in regulatory elements, such as highly conserved noncoding elements (HCNEs), which frequently correspond to transcriptional enhancers (Ovcharenko et al., 2005). The extent to which lincRNAs found in gene deserts near developmental TFs are functional or fundamentally different from other lincRNAs is unclear. lincRNAs might preferentially fall in these regions because (1) these lincRNAs regulate gene expression in *cis*, as observed for *HOTTIP* (Wang et al., 2011) and *Mistral* (Bertani et al., 2011); (2) the colocalized lincRNA and TF genes might act in concert and thus benefit from coregulation, as observed for *Six3* and *Six3os* (Rapicavoli et al., 2010); or (3) the multiplicity of enhancer elements around TFs might provide an accommodating environment for the emergence of new lincRNA genes. In offering the third possibility, we are not suggesting that a significant number of lincRNAs can be attributed to the transcription observed within many enhancer elements (De Santa et al., 2010; Kim et al., 2010); these enhancer transcripts are not typically polyadenylated, and lincRNA genes overlap enhancers no more frequently than do protein-coding genes (Cabili et al., 2011).

Secondary Structure

Secondary structure is important for most noncoding RNA classes, including some long noncoding RNA (Kino et al., 2010; Maenner et al., 2010; Novikova et al., 2012; Wilusz et al., 2012), but the prevalence of secondary structure-mediated roles in lincRNA biology remains unknown. Indeed, when the whole transcript is considered, lincRNAs are not predicted to be more structured than mRNAs. The fraction of paired nucleotides in the predicted optimal folds of the human and mouse lincRNA transcripts resembles that of mRNAs (Managadze et al., 2011). The amount of predicted secondary structure correlates positively with lincRNA expression levels, perhaps because more structured lincRNAs are more stable, or because both structure and expression correlate with G/C content (Kudla et al., 2006). In any case, no correlation is observed between the amount of predicted secondary structure and evolutionary conservation (Managadze et al., 2011).

If many lincRNAs contained short, highly structured regions critical for function, then these lincRNAs would have regions with evolutionary conserved secondary structures. Given alignable sequences, several computational tools (reviewed in Gorodkin et al., 2010) can detect such regions. Surprisingly, depending on the lincRNA set studied, such predicted structures are either depleted or only mildly enriched in lincRNA exons (Marques and Ponting, 2009; I.U. and D.P.B., unpublished data). As discussed below, it is unlikely that many additional conserved structures have been missed due to an inability to align their corresponding primary sequences. Conserved secondary structures thus seem to occupy only a small fraction of the vertebrate lincRNA transcriptome. Similar observations were made in *C. elegans*, where the overlap between a set of noncoding RNA candidates generated using predicted-structure-based criteria and a set of transcript models generated using RNA-seq data was even smaller than that expected by chance (Nam and Bartel, 2012).

These results should not be interpreted to indicate that lincRNAs are devoid of secondary structure. Even randomly generated RNA sequences have compact folds with secondary structure (Schultes et al., 2005), and there is no reason to suspect that lincRNAs would differ. Thus, the presence of a computationally predicted or an experimentally supported structured region in a lincRNA is not informative for judging whether the structure is functionally important. The emerging picture is that for most regions of most lincRNAs, the collapse characteristic of arbitrary RNA sequences is sufficient for lincRNA function, with specific, evolutionarily conserved structural elements occupying only a very small fraction of the lincRNA real estate. Known examples of such elements include the proposed PRC2-binding elements in *Xist* and the triple-helical elements that can impart lincRNA stability (Maenner et al., 2010; Brown et al., 2012; Wilusz et al., 2012). With additional study and improved tools, additional examples presumably will be found.

Expression Levels

Compared to mRNA expression, lincRNA expression is typically more variable between tissues (Cabili et al., 2011; Derrien et al., 2012; Pauli et al., 2012), with many lincRNAs preferentially expressed in brain and testis (Ravasi et al., 2006; Cabili et al., 2011; Derrien et al., 2012). Expression similarity between a lincRNA gene and its closest protein-coding neighbor is generally not greater than that between two adjacent protein-coding genes (Cabili et al., 2011; Ulitsky et al., 2011; Pauli et al., 2012).

The median lincRNA level is only about a tenth that of the median mRNA level (Ravasi et al., 2006; Guttman et al., 2009; Guttman et al., 2010; Cabili et al., 2011; Ulitsky et al., 2011; Derrien et al., 2012; Pauli et al., 2012; Sigova et al., 2013). The extent to which the lower level is caused by less efficient transcription or more efficient degradation of lincRNAs remains unknown. Two studies, one using a transcription inhibitor and the other using pulse-chase analysis, both concluded that mRNAs and long noncoding RNAs (including lincRNAs) have similar half-life distributions (Clark et al., 2012; Tani et al., 2012). Thus, at least the lincRNAs that accumulate to sufficient levels for quantification in such studies are not preferentially destabilized by pathways that degrade aberrant mRNA molecules. When comparing different lincRNAs, the characteristics associated with increased stability include those associated with increased mRNA stability, such as splicing, cytoplasmic localization and G/C-rich nucleotide composition (Clark et al., 2012).

Subcellular Localization

Perhaps the most common misperception of lincRNAs is that they are predominantly localized in the nucleus. Some of the best-studied lincRNAs, such as *Xist*, *Malat1*, *Neat1*, and *Miat*, are almost exclusively in the nucleus (Brown et al., 1992; Hutchinson et al., 2007; Sone et al., 2007) and even define specific nuclear domains (Hutchinson et al., 2007; Sone et al., 2007; Clemson et al., 2009). However, other studied lincRNAs are found mostly in the cytoplasm (Coccia et al., 1992; Kino et al., 2010;

Yoon et al., 2012). When RNA is sequenced from nuclear and cytoplasmic fractions, lincRNAs have a ~2-fold enrichment in the nuclear fraction relative to mRNAs in five of the six human cell types examined (Derrien et al., 2012). In the remaining cell type, NHEK cells, the lincRNA distribution is no different than that of mRNAs. Similarly, we observe a 3-fold relative enrichment in the nucleus using data from MCF-7 cells (Figure 2E). However, because polyadenylated RNA species in the cell (dominated by cytoplasmic mRNAs) are not equally distributed between nucleus and cytoplasm, these relative enrichments do not accurately represent absolute enrichments. Therefore, although many lincRNAs are exclusively or predominantly nuclear (Figure 2E), the observed $\leq$3-fold nuclear enrichments of lincRNAs relative to mRNAs refute the notion that as a group, currently annotated lincRNAs are predominantly localized in the nucleus. Consider, for example, cells in which the typical mRNA is six times more abundant in the cytoplasm than in the nucleus. With 3-fold relative nuclear enrichment, the typical lincRNA would still be two times more abundant in the cytoplasm than in the nucleus. Bearing in mind that some lincRNAs might act in the nucleus before making their way to the cytoplasm, the current picture is that most lincRNAs spend most of their time in the cytoplasm. The more specific localization of lincRNAs within either the cytoplasm or nucleus, as well as the factors and sequence elements that dictate this localization, remain largely unexplored.

lincRNA Evolution

Our understanding of other noncoding RNAs has been greatly advanced by studying conservation patterns within their genes and between the noncoding RNAs and their interaction partners (Woese et al., 1980; Michel and Westhof, 1990; Bartel, 2009). Likewise, analyzing the natural selection pressures acting on noncoding RNAs can identify elements and structures important for function. This analysis can also suggest which lincRNAs are functional, provide important clues to their modes of action and identify relevant model organisms for studying the biology of human lincRNAs.

Rapid Evolutionary Turnover of lincRNA Sequences

In stark contrast to mRNAs and many classes of noncoding RNAs, mammalian lincRNAs lack known orthologs in species outside of vertebrates. One possible exception is the Telomeric repeat-containing RNA (*Terra*), which is conserved between human and yeast but is a nonconventional lincRNA in that only a small fraction of its transcripts is polyadenylated (reviewed in Feuerhahn et al., 2010).

Compared to protein-coding sequences, most of which are highly conserved throughout vertebrates, lincRNA sequences evolve very rapidly. Less than 6% of zebrafish lincRNAs have detectable sequence conservation with human or mouse lincRNAs (Ulitsky et al., 2011), and only ~12% of human and mouse lincRNAs appear to be conserved in the other species (Church et al., 2009; Cabili et al., 2011). Within rodents, only ~60% of the lincRNAs (compared to >90% of mRNAs) expressed in *Mus musculus* liver have alignable counterparts expressed in the livers of *Mus castaneus* and rat (Kutter et al., 2012), which shared common ancestors with *M. musculus* only ~1 and ~15 million years ago, respectively. Interestingly, the presence of a lineage-specific lincRNA gene correlates with higher expression of adjacent protein-coding genes in that lineage (Kutter et al., 2012).

Despite their rapid evolution, lincRNA sequences display detectable, albeit weak, signatures of natural selection. Members of an initial lincRNA catalog in mouse (Okazaki et al., 2002) were poorly conserved when evaluated using mouse-rat and mouse-human genome alignments (Wang et al., 2004). More recently, improved identification and filtering of lincRNA candidates and improved methods for estimating conservation have led to evidence that lincRNA exons are more conserved than intergenic regions but significantly less than either coding or noncoding portions of mRNA exons (Ponjavic et al., 2007; Guttman et al., 2009; Khalil et al., 2009; Marques and Ponting, 2009; Ulitsky et al., 2011; Derrien et al., 2012). Interestingly, fly lincRNAs (which are much shorter than mammalian lincRNAs) appear better conserved at the sequence level, evolving faster than ORFs but slower than 3′UTRs and intergenic regions (Young et al., 2012) (I.U., unpublished data).

Is lincRNA Sequence Conservation Currently Overestimated or Underestimated?

Even the modest magnitude of the sequence conservation reported within lincRNA exons might be overestimated. Conservation scores and substitution rates used to evaluate lincRNA sequence conservation are derived from whole-genome alignments, which compare genome rather than lincRNA sequences. For example, the presence of a segment homologous to a human lincRNA exon in the chicken genome does not necessary imply that the homologous segment is part of a chicken lincRNA. In chicken, this segment might be transcribed as part of an mRNA or might not be transcribed at all. Indeed, when exons of human or mouse lincRNAs are traced to the zebrafish genome through whole-genome alignments, the corresponding regions rarely overlap zebrafish lincRNAs, and in about a third of the cases they overlap zebrafish mRNAs (Ulitsky et al., 2011). In another example, although both potentially functional regions in the human *Hotair* lincRNA appear to be conserved in the mouse genome (He et al., 2011) only the 3′ region appears to be part of the murine *Hotair* homolog (Schorderet and Duboule, 2011). Possible explanations for mapping to non-lincRNA annotations include annotation errors, interconversion between coding and noncoding transcripts during evolution (discussed below), or selective pressures on DNA elements, such as transcriptional enhancers, that overlap lincRNA genes. To the extent that any of these explanations are relevant, even the modest sequence conservation reported in lincRNA exons might overestimate the selective pressures acting to preserve lincRNA function. Obtaining more informative conservation estimates will require more comprehensive lincRNA catalogs in multiple vertebrate species so that lincRNAs can be compared to lincRNAs rather than to genomic alignments.

Why are lincRNA sequences so poorly conserved? Perhaps the fraction of lincRNAs that are nonfunctional is large, and thus changes in most lincRNA sequences exact no fitness cost. Alternatively, existing approaches for comparing genomic sequences, which rely heavily on stretches of high sequence

conservation, might be poorly suited for detecting homology between lincRNAs. One idea is that lincRNAs might be under pressure to conserve structure but not sequence, and thus homologs would be missed with methods that focus on primary-sequence homology. However, pressures to conserve secondary structure also substantially slow down changes in the corresponding primary sequence, such that the evolutionary time needed to erase primary-sequence similarity within a conserved secondary structure is probably far too long to have occurred within the mammalian clade. Nonetheless, as illustrated below, detailed comparative analyses of specific lincRNAs supports the notion that lincRNA conservation has been systematically underestimated for other reasons.

Because finding optimal alignments between long sequences is time and resource consuming, the BLAST heuristic is typically used to identify sequence homologs or generate whole-genome alignments. BLAST accelerates search of similar sequences by identifying short regions of high sequence conservation and then refining the sequence alignments around these regions (Altschul et al., 1997). This approach is very powerful in many cases, and for the past 15 years BLAST has served as a major bioinformatics workhorse. However, BLAST as well as more sensitive tools often fail to identify sequence conservation in cases for which synteny and other genomic evidence strongly indicate that the corresponding lincRNAs are orthologous. Some improvements to BLAST designed to detect homology among RNA genes have been proposed (Bussotti et al., 2011), but more substantial increases in sensitivity await better understanding of the nature of selective pressures acting on lincRNA loci. Described below are case studies for six lincRNAs (*Xist*, *Cyrano*, *Megamind*, *Miat*, *Malat1*, and *PAN*), which illustrate the challenges of using existing methods for examining lincRNA evolution.

X-inactive specific transcript (Xist) is a master regulator of X chromosome inactivation in eutherian mammals (Brockdorff et al., 1992; Brown et al., 1992; Penny et al., 1996). Although poorly conserved throughout most of its sequence, *Xist* is conserved in its exon-intron structure, with a consensus of ten exons (Nesterova et al., 2001; Elisaphenko et al., 2008). *Xist* and at least three additional lincRNAs in the X-inactivation center descended from protein-coding genes still present in other amniotes (Duret et al., 2006). Although regions of sequence similarity are observed between at least four mammalian *Xist* exons and six chicken *Lnx3* mRNA exons (Elisaphenko et al., 2008), none of these are evident in current whole-genome alignments. *Xist* sequences in contemporary species contain multiple ancient and conserved repeats alongside young and species-specific repeats originating from mobile elements, as the repetitive fraction of *Xist* increased from about 4.4% in the eutherian ancestor to as much as 12.4% in the human (Elisaphenko et al., 2008). Interestingly, the first exon of *Xist*, which contains most of the known functional repetitive elements (Beletskii et al., 2001; Wutz et al., 2002; Sarma et al., 2010), is characterized by low PhastCons scores, perhaps because some of these repeats contain short functional sequences interspersed among poorly conserved spacers (Wutz et al., 2002). In contrast, although the most obvious sequence conservation resides in exon 4, deleting this exon does not affect X inactivation (Caparros et al., 2002). *Xist* thus illustrates significant challenges for comparative analysis; due to its size and sequence divergence among mammals, and despite its functional importance, *Xist* appears quite poorly conserved when inspected through the lens of whole-genome alignments.

The *Cyrano* lincRNA is conserved throughout vertebrates (with the potential exception of lizards) and is required for proper morphogenesis and neurogenesis in zebrafish (Ulitsky et al., 2011). Within the most conserved region of *Cyrano* is a 26 nt site that pairs to the miR-7 miRNA and is perfectly conserved in at least 55 vertebrates from human to lamprey (Ulitsky et al., 2011). In addition to this conserved site, *Cyrano* orthologs share similar exon-intron architectures (Figure 3A) and multiple shorter (<10 nt) highly conserved sites (I.U. and D.P.B., unpublished data). Although the human ortholog can rescue the *Cyrano* knockdown in zebrafish, the human and fish genes do not align with each other in whole-genome alignments (Figure 3A). This alignment failure occurs because the signal for sequence similarity does not exceed detection thresholds when considered in the context of full-genome pairwise comparisons, even though BLASTN detects a conserved 67 nt segment when the human and zebrafish *Cyrano* genes are directly compared.

Megamind is also conserved throughout vertebrates and required for proper brain development in zebrafish (Ulitsky et al., 2011). Unlike *Cyrano*, *Megamind* lacks stretches of consecutive highly conserved bases but instead contains 40 positions with at least 90% identity in over 50 vertebrates, which appear at phased positions within a 95 nt region. Even with the most permissive parameter settings, BLASTN fails to identify *Megamind* homologs in EST collections from some fish. These homologs are nonetheless identified with high statistical significance using a hidden Markov model trained using the *Megamind* conserved regions (Ulitsky et al., 2011). The reliance of BLAST on contiguous stretches of high conservation is thus a substantial limitation when comparing sequences in which highly conserved positions are intermingled with rapidly evolving ones.

Miat (also called *Gomafu* or *Rncr2*) was originally discovered as a lincRNA highly enriched in specific neurons in mouse retina (Blackshaw et al., 2004; Sone et al., 2007) and later found to be more widely expressed in the nervous system and cultured neurons, where it specifies cell identify (Sone et al., 2007; Rapicavoli et al., 2010). *Miat* sequence variants are also associated with increased risk of myocardial infarction (Ishii et al., 2006). *Miat* is retained in the nucleus in mammalian and avian cells, and defines a subnuclear domain that does not overlap with other nuclear bodies (Sone et al., 2007; Tsuiji et al., 2011). Although *Miat* appears to be restricted to mammals in whole-genome alignments based on the human and mouse genomes, orthologs are present in syntenic positions of chicken and frog (Figure 3B) (Rapicavoli et al., 2010; Tsuiji et al., 2011). These homologs all contain a relatively short region with multiple copies of the (U)ACUAAC(C) motif, which resembles the intron branch point and can bind to Splicing factor 1 (Sf1) (Rapicavoli et al., 2010; Tsuiji et al., 2011). This region maps to the last exon within *Miat* orthologs but is nested in rapidly evolving sequence, and apart from the motif repeats, sequence similarity within the region is sparse (Figure 3B). Indeed, BLASTN finds no significant similarity between human and frog *Miat* and only a short

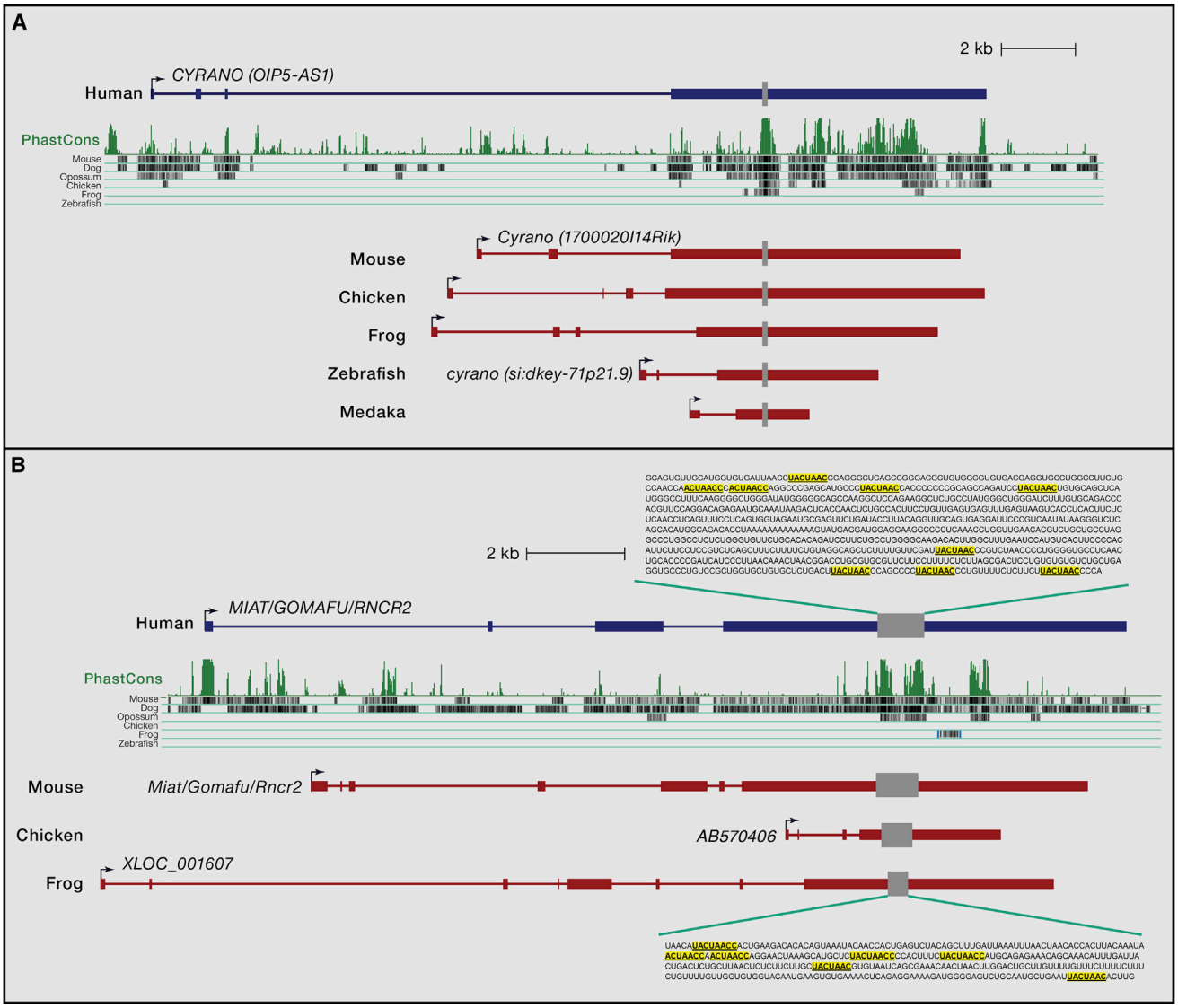

Figure 3. Evolution of *Cyrano* and *Miat* lincRNAs

(A) *Cyrano*. Gene models from the indicated species are shown, together with the PhastCons track. The gray bar indicates a ~70 nt region of homology detected in a focused search, starting with the zebrafish ortholog.

(B) *Miat*. Gene models from the indicated species are shown, together with the PhastCons track. The gray box indicates a region in the last exon that contains multiple copies of the (U)ACUAAC(C) motif, as shown for human and frog.

(<30 bp) region of similarity between human and chicken sequences.

Malat1 is an exceptionally highly expressed, nuclear-retained, single-exon lincRNA that was originally identified in metastatic tumors (Ji et al., 2003). Although *Malat1* helps organize nuclear speckle domains, which contain splicing factors (Tripathi et al., 2010), it is not essential for life and development in mouse (Eißmann et al., 2012; Nakagawa et al., 2012; Zhang et al., 2012). The most abundant *Malat1* isoform is not polyadenylated, and its 3′ end instead forms a triple-helical RNA structure (Brown et al., 2012; Wilusz et al., 2012). This 3′ end is generated by RNase P cleavage of the nascent transcript, which releases the 61 nt *Malat1-associated small cytoplasmic RNA* (*mascRNA*). *Malat1* was originally considered a mammalian-specific lincRNA (Hutchinson et al., 2007; Tripathi et al., 2010) and only more recently found in other vertebrates (Stadler, 2010; Ulitsky et al., 2011). Although the entire genomic region appears to have been lost in the avian clade, *Malat1* orthologs appear in syntenic genomic positions near *Scyl1* in mammals, frogs, and fish. The zebrafish *malat1* shares striking features with the mammalian *Malat1*, including similar length of ~7 kb, very high expression levels, no apparent introns, a noncanonical 3′ end, and a canonical yet inefficient polyadenylation site ~4 kb after the transcription start site (Figure 4A). However, apart from its

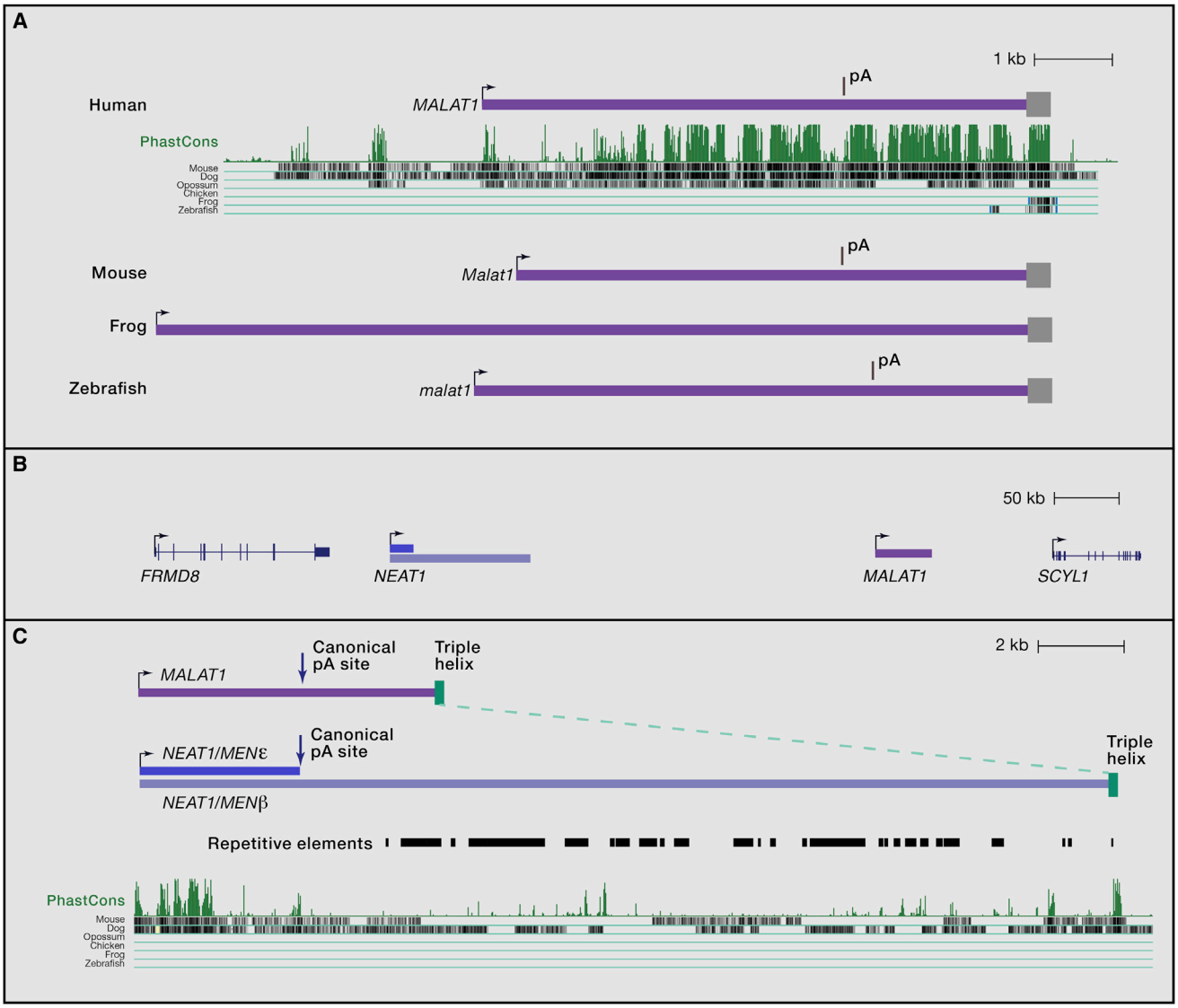

Figure 4. Evolution of the *Malat1* and *Neat1* lincRNAs

(A) *Malat1* gene models from the indicated species are shown, together with the PhastCons track indicating homology to the human genome detected in the whole-genome alignments. The gray box corresponds to the region of sequence similarity at the 3′ end of *Malat1*.

(B) The human *NEAT1/MALAT1* locus.

(C) *Neat1* and its similarities with *Malat1*. The human gene models are shown, together with annotated repeats and the PhastCons track for *Neat1*.

3′ terminal region, which includes the *mascRNA* and another short (<70 bases) segment of homology (Figure 4A), the mammalian *Malat1* gene has no recognizable sequence similarity with its fish counterpart.

Several features of the *PAN* lincRNA from Kaposi's sarcoma-associated herpesvirus (KSHV) resemble those of *Malat1* (Sun et al., 1996; Tycowski et al., 2012). Like *Malat1*, *PAN* is a long, unspliced, very abundantly expressed lincRNA that ends with triple-helical RNA element essential for its accumulation (Conrad et al., 2006; Mitton-Fry et al., 2010). A computational approach that relied on sequence and structure similarity identified homologous elements in six other viral genomes, including two additional gammaherpesviruses (Tycowski et al., 2012). Moreover, the elements in the other gammaherpesviruses occur at ends of lincRNAs that have similar lengths and syntenic positions with *PAN* but share little to no other detectable sequence similarity with *PAN*. These presumed homologs could be identified using a tailored bioinformatics approach but not a conventional sequence-homology search.

As the previous examples each illustrate, sequence-homology search tools often fail to detect known lincRNA orthologs. To the extent that orthologs are missed, metrics that depend on whole-genome alignments or other output from these tools will underestimate lincRNA conservation. Countering this

underestimate are the false-positive orthologs arising from alignments to nonlincRNA sequences, described at the beginning of this section. Thus, the question as to whether lincRNA sequence conservation is currently overestimated or underestimated remains open, with the answer awaiting improved tools and more comprehensive lincRNA catalogs from more species.

lincRNA Synteny despite Undetectable Sequence Conservation

Some lincRNAs are at conserved genomic locations, with conserved exon-intron structures yet no detectable sequence conservation. For example, protein-coding genes adjacent to a lincRNA gene in zebrafish are more likely to have orthologs adjacent to lincRNA genes in human or mouse, even when all lincRNAs with sequence homology are excluded from the analysis (Ulitsky et al., 2011). Importantly, this enrichment remains significant after controlling for the fact that some genes (particularly those of developmental transcriptional regulators) tend to be far from other protein-coding genes and are therefore more likely to be adjacent to lincRNA genes. Perhaps these lincRNAs have conserved sequence-dependent functions, yet their sequences are too divergent to be detected with existing tools. The examples of conserved lincRNAs with limited sequence conservation listed above suggest that this scenario is relevant for at least some lincRNAs. Alternatively, the act of transcription rather than the identity of the transcribed RNA might be important, in which case, the inability to detect lincRNA sequence conservation would accurately reflect an absence of sequence-based posttranscriptional function.

Evolutionary Trajectories of lincRNA Genes

The low levels of sequence conservation observed in vertebrates point to either rapid sequence evolution or frequent gain and loss of lincRNA genes (Ulitsky et al., 2011). With respect to the gain of new genes, three evolutionary scenarios might be considered. New lincRNA genes might originate from either ancestral protein-coding genes; duplication and divergence of other lincRNA genes; or de novo, from intergenic DNA (Ponting et al., 2009). Although the origins of most mammalian lincRNAs are unknown, examples below illustrate the first two of these three evolutionary possibilities.

As mentioned previously, *Xist* evolved from a protein-coding gene *Lnx3* that is still present in noneutherian vertebrates (Duret et al., 2006). Because pseudogenization is a rather common event, and many pseudogenes are transcribed (Pink et al., 2011; Pei et al., 2012), other lincRNAs might have similar origins. Because analyses of expression and conservation patterns of pseudogenes are complicated by their sequence-similar protein-coding relatives, pseudogenes are typically excluded from lincRNA collections. Nevertheless, the sequences of at least 68 human pseudogenes appear to be under selection in mammals (Khachane and Harrison, 2009), and an increasing number of pseudogenes are reported to have noncoding functions. Some contain inverted repeats or are transcribed in the antisense orientation, triggering RNAi-mediated repression of their protein-coding cousins in the oocyte (Tam et al., 2008; Watanabe et al., 2008). Others are proposed to influence mRNA regulation by binding and depleting *trans*-acting factors (reviewed in Pink et al., 2011), although this mechanism is often implausible when considering the unfavorable stoichiometry between the pseudogene transcripts and the factors (Ebert and Sharp, 2010). The emergence of new lincRNA genes from protein-coding genes might often occur through neofunctionalization of the pseudogene. In addition, the observation of transcripts possessing both coding and noncoding functions opens the alternative possibility for duplication and subfunctionalization of bifunctional ancestral genes.

New genes can also emerge from the opposite direction, with ancestral noncoding transcripts serving as raw material for the birth of novel protein-coding genes. Candidates for such an event include 24 predicted human protein-coding genes of at least 50 aa that in other primates have homologous genes that do not appear to code for sufficiently homologous proteins (Xie et al., 2012), with similar phenomena observed in other species (Cai et al., 2008; Carvunis et al., 2012). Although detecting most of the older protein-coding gene birthing events will be more difficult, examples might be detected if the coding transcript retained a noncoding function that constrained its sequence. Indeed, a zebrafish lincRNA gene conserved in teleosts and chondrichthyes appears to have acquired a functional protein-coding region in the tetrapod lineage (Ulitsky et al., 2011). The conserved noncoding region of these genes has a conserved predicted secondary structure (I.U. and D.P.B., unpublished data), which further supports the model of a conserved noncoding element residing within an ancient lincRNA that later evolved a short, functional protein-coding region to become a bifunctional mRNA.

Within a species, lincRNA sequences are rarely similar to each other (Ulitsky et al., 2011; Derrien et al., 2012), and with few exceptions (e.g., *megamind*; Ulitsky et al., 2011) most studied lincRNAs appear in single copies in vertebrate genomes. Thus, lincRNAs rarely originate from duplication of other lincRNAs, or their similarity becomes undetectable rapidly after duplication. Support for the latter explanation is found in one of the few clear examples of lincRNA duplication. In mammalian genomes, *Neat1* appears immediately upstream of *Malat1*, in tandem orientation suggestive of an ancestral gene duplication (Figure 4B) (Stadler, 2010). *Neat1* has two isoforms that resemble the two *Malat1* isoforms (Figure 4C). These are the 3.7 kb *Men*ε, which ends with a canonical polyadenylation site, and the 22.7 kb *Men*β, which shares its 5′ end with *Men*ε and the mechanism of its 3′-end formation and a triple-helical terminal structure with the longer *Malat1* isoform (Brown et al., 2012; Wilusz et al., 2012). *Malat1* and *Neat1* lincRNAs each localize to specific nuclear domains, *Malat1* to the nuclear speckles and *Neat1* to the paraspeckles (Hutchinson et al., 2007). Despite these many lines of evidence for shared ancestry, comparison of the human *Neat1* and *Malat1* sequences reveals no homology beyond a short stretch at the very 3′ end, which includes the triple-helical element and downstream structure required for RNase P cleavage. Presumably other duplicated lincRNA genes also underwent similarly rapid divergence following their duplication, thereby obscuring their common origins.

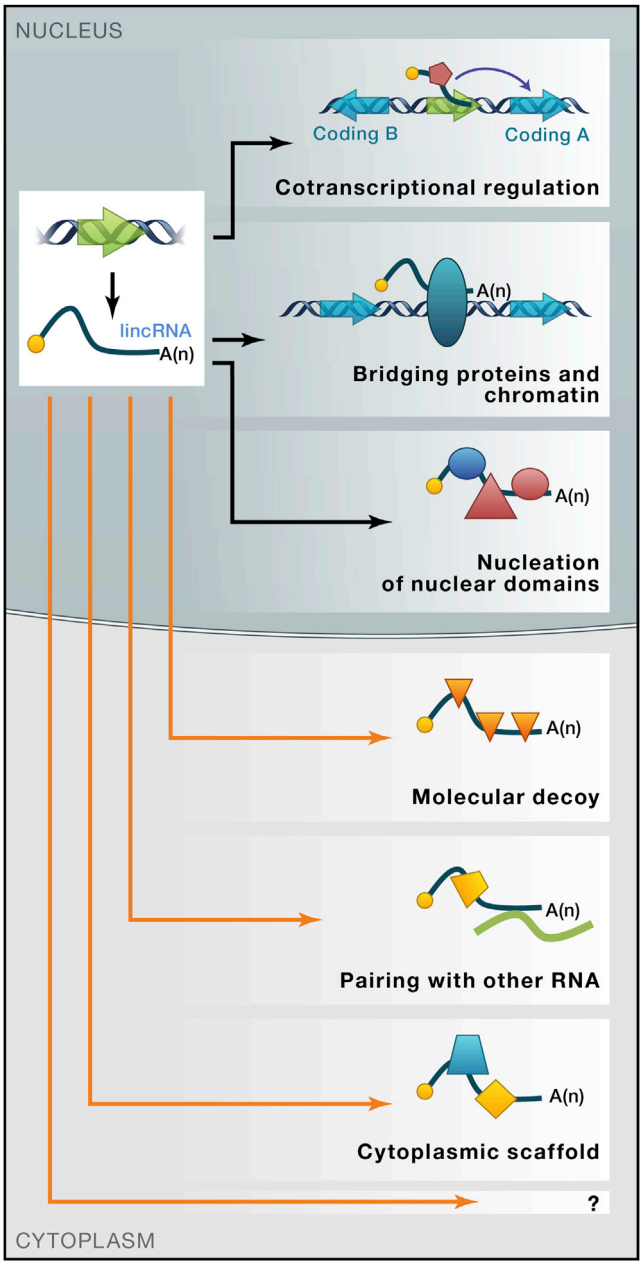

Figure 5. Diverse Mechanisms Proposed for lincRNA Function
Modes of action include cotranscriptional regulation (e.g., through either the interaction of factors with the nascent lincRNA transcript or the act of transcribing through a regulatory region), regulation of gene expression in *cis* or in *trans* through recruitment of proteins or molecular complexes to specific loci, scaffolding of nuclear or cytoplasmic complexes, titration of RNA-binding factors, and pairing with other RNAs to trigger posttranscriptional regulation. The two latter mechanisms are illustrated in the cytoplasm (where they are more frequently reported) but could also occur in the nucleus. Additional mechanisms will presumably be proposed as additional functions of lincRNAs are discovered.

Mechanisms of Action

Little is known about the biological roles of lincRNAs, and even less about how they carry out those roles, but several potential mechanisms for nuclear and cytoplasmic lincRNAs have been suggested based on the few relatively well-studied examples (Figure 5). lincRNAs might act through a broad array of mechanisms, which would be consistent with the wide variety of subcellular localizations, expression levels, and stabilities observed for lincRNAs in mammalian cells.

The potential mechanisms of lincRNA function can be divided into three groups: (1) those that rely solely on the act of transcription or on the nascent RNA; (2) those that require the processed RNA yet depend on the site of transcription; and (3) those that are independent of the site of transcription. A major difference between the first two groups and the last one is in whether the direct targets of the lincRNA activity are found only in proximity to the lincRNA gene (*cis* targets, groups 1 and 2), or anywhere in the cell (*trans* targets, group 3).

The well-studied examples of *cis*-acting chromatin-associated lincRNAs include some of the lincRNAs transcribed from and acting at the X-inactivation center (reviewed in Lee, 2009; Augui et al., 2011). Which features of these lincRNAs are unique to X-inactivation biology and which are relevant to other lincRNAs is unclear. Examples of other cis-regulatory lincRNAs include *ncRNA-a1-7*, *Hottip*, and *Mistral*, the perturbation of which leads to decreased expression of some nearby genes (Ørom et al., 2010; Bertani et al., 2011; Wang et al., 2011; Lai et al., 2013).

A single *cis*-acting molecule might be able to target a neighboring locus, which would explain the relatively low expression levels of many lincRNAs. A prevalence of *cis*-regulatory lincRNAs would also explain the significant synteny of lincRNA loci from distant vertebrates and their generally limited sequence conservation. A potential mechanism by which *cis*-acting lincRNAs might function without performing any sequence-specific activities would be for the nascent lincRNA transcripts to flag regions of open, transcriptionally competent chromatin through the recruitment of promiscuous RNA-binding proteins.

Despite known *cis*-acting examples and the above-mentioned arguments favoring the prevalence of *cis*-acting function, other observations challenge the notion that most lincRNAs act in *cis*-regulatory circuits. lincRNA knockdown in mouse embryonic stem cells rarely changes the expression of neighboring genes, with mRNA levels of one of the 20 closest neighbors of the lincRNA affected in <10% of the cases examined (Guttman et al., 2011). Moreover, only about 3% of the human lincRNAs have expression profiles strongly correlated with those of their neighbors (compared with 1.5% for mRNAs), and strong negative correlations are exceedingly rare (Derrien et al., 2012), arguing against widespread effects of lincRNA expression on neighboring regulatory programs. Further evidence favoring *trans* functions is the observation that most lincRNA are predominantly cytoplasmic (Figure 2E), which suggests that many might function in the cytosol and thus would not be *cis*-acting. More information on the relative prevalence of *cis* and *trans* mechanisms will come from genome-wide approaches to study lincRNA chromatin occupancy as well as focused studies of additional lincRNAs.

Interactions between lincRNAs and Other Cellular Factors

As expected, increasing evidence suggests that many lincRNAs function through specific interactions with other cellular factors,

namely proteins, DNA, and other RNA molecules. Much effort is being devoted to finding these interacting partners as a strategy for gaining insight into molecular mechanism.

A popular view is that many lincRNAs regulate gene expression by directing chromatin-modification complexes to specific target regions (Rinn and Chang, 2012). This view is based on observations from some well-studied lincRNAs, such as *Xist* (Penny et al., 1996), *Hotair* (Tsai et al., 2010), *Hottip* (Wang et al., 2011), and *Mistral* (Bertani et al., 2011), and the mechanistic understanding of long RNAs that overlap the protein-coding regions of their targets (and hence are not classified as lincRNAs), such as *Air* (Sleutels et al., 2002), *Kncq1ot1* (Pandey et al., 2008), and *Anril* (Yap et al., 2010). Accordingly, most studies of lincRNA-associated proteins have focused on chromatin factors. For example, lincRNAs are reported to associate with CTCF (Yao et al., 2010), YY1 (Jeon and Lee, 2011), Mediator (Lai et al., 2013), WDR5 (Wang et al., 2011; Gomez et al., 2013; Grote et al., 2013), LDS1 (Tsai et al., 2010), and the polycomb complexes PRC1 (Schoeftner et al., 2006) and PRC2 (Rinn et al., 2007; Zhao et al., 2008; Tsai et al., 2010; Grote et al., 2013; Klattenhoff et al., 2013), although the extent to which some of these interactions are direct and specific remains controversial (Brockdorff, 2013). Conversely, searches for transcripts associated with PRC2 detect significant fractions (~20% in human and ~10% in mouse) of annotated lincRNAs (Khalil et al., 2009; Zhao et al., 2010; Guttman et al., 2011). The functional outcomes of these binding events are unclear, as lincRNAs account for a relatively small fraction of the PRC2-RNA interactome, and lincRNAs reported to be associated with PRC2 in human and mouse have no overlap (Zhao et al., 2010). Another large-scale study found that as many as 30% of lincRNAs expressed in mouse embryonic stem cells are associated with at least one of 11 chromatin regulators (Guttman et al., 2011), although some of these interactions may be indirect and mediated by protein-protein interactions (Brockdorff, 2013). The nature of the lincRNA-protein recognition, and whether it relies primarily on RNA primary sequence or on structural features, remains largely unknown, as regions mediating lincRNA-protein interactions have been identified in only a few cases, and these regions are currently too large to suggest how binding specificity is achieved (Huarte et al., 2010; Murthy and Rangarajan, 2010).

Part of the appeal of lincRNAs acting to direct chromatin-modifying complexes to DNA is that it would help solve the mystery of how protein complexes without intrinsic sequence-specific DNA-binding ability, such as the polycomb complex, find their DNA targets. However, this model pushes to the fore the questions of how these proteins recognize RNA, how the low abundance of most lincRNAs can be reconciled with roles in recruiting protein complexes to hundreds or thousands of genomic loci, and how lincRNAs might recognize DNA targets.

lincRNAs might recognize specific regions in genome through direct interactions with the DNA. One way to do this would be to act as a nascent transcript, while still tethered to the DNA by the RNA polymerase, as occurs for transcripts targeted by the endogenous small interfering RNAs (siRNAs) that direct chromatin silencing in fission yeast (Moazed, 2009). In theory, lincRNAs might also directly recognize DNA by other mechanisms, either through triplex interactions with the Hoogstein face of purine stacks within the DNA duplex (Frank-Kamenetskii and Mirkin, 1995) or through base-pairing interactions with single strands within an unwound region of the DNA. Such interactions might be facilitated by proteins that could either help stabilize the base triples or help melt the DNA to enable RNA pairing. Alternatively, lincRNAs might recognize specific genomic regions through indirect interactions, either base pairing with nascent transcripts or interacting with DNA-binding proteins or complexes. Identification of principles that guide lincRNAs to specific chromatin regions will benefit from methods for high-throughput identification of target regions akin to the recent genome-wide isolation and sequencing of DNA associated with an RNA of interest (Chu et al., 2011; Simon et al., 2011).

Many lincRNAs presumably have functions unrelated to chromatin modification. An appealing way for these lincRNAs to form interactions is through base pairing with other RNA molecules, as this is the way that members of other classes of noncoding RNAs (e.g., tRNAs, snRNAs, snoRNAs, and microRNAs) interact with their targets and partners. For example, *antisense Uchl1* regulates *Uchl1* translation by pairing to a segment of its 5′UTR (transcribed from an overlapping genomic region) (Carrieri et al., 2012), and the *TINCR* lincRNA is reported to pair with and stabilize mRNAs containing a 25 nt motif (Kretz et al., 2013). Formation of double-stranded RNA by a lincRNA and its target might also activate downstream pathways. For example, a group of Alu repeat-containing RNAs are reported to repress targets with sequence-similar complementary Alu elements in their 3′UTRs via the Staufen 1 (STAU1)-mediated mRNA decay pathway (Gong and Maquat, 2011). Another proposed function of mammalian lincRNAs is to pair to microRNAs and titrate them away from their mRNA targets, as can be done using artificial "sponge" RNAs (Ebert et al., 2007) and as observed for select plant and viral RNAs (Franco-Zorrilla et al., 2007; Cazalla et al., 2010). In mammals, however, nearly all of the proposed "competing endogenous RNAs" fail to reach levels sufficiently high to achieve consequential miRNA titration. The most notable exception is *CDR1as*/*ciRS-7*, a highly expressed circular RNA with more than 70 conserved miR-7 target sites (Hansen et al., 2013; Memczak et al., 2013). The paucity of other highly expressed noncoding RNAs with many target sites argues against the widespread function of lincRNAs as microRNA sponges. Nonetheless, *Cyrano* illustrates that lincRNA function can require microRNA pairing, presumably for purposes other than titration (Ulitsky et al., 2011).

A compelling idea is that many lincRNAs might make use of interactions with protein, DNA, and other RNAs to act as scaffolds to bring together different proteins or bridging protein complexes and specific chromatin regions (Guttman and Rinn, 2012). For example, *Neat1*/*Menβ* and *Malat1* bind multiple proteins localizing to the paraspeckles and nuclear speckles, respectively, and *Menβ* is essential for paraspeckle formation (Clemson et al., 2009; Sunwoo et al., 2009; Murthy and Rangarajan, 2010; Souquere et al., 2010; Tripathi et al., 2010). With the recognition that most lincRNAs are mostly cytoplasmic, we suggest that this scaffolding mechanism might also play important roles in the cytosol. The binding of a lincRNA to a protein might also regulate the protein activity. For example, lincRNA binding was shown to affect the action of some transcription regulators,

including Tsl (Wang et al., 2008) and Nfat (Willingham et al., 2005). One possible mechanism is for the lincRNA to act as a decoy that titrates the protein away from its potential targets, as has been reported for lincRNA *Gas5* and glucocorticoid receptor (Kino et al., 2010), *PANDA* and NF-Y (Hung et al., 2011), *sno-lncRNAs* and Fox2 (Yin et al., 2012), and *Gadd7* and TDP-43 (Liu et al., 2012b). However, when considering that most proteins accumulate to many more molecules per cell than do their corresponding mRNAs and that the typical mRNA is still expressed at higher levels than the typical lincRNA, the titration mechanism seems possible for only a small subset of lincRNAs.

Concluding Remarks

lincRNA research is at a very interesting juncture—thousands of lincRNA genes have been identified, and the diverse functional and mechanistic underpinnings of a few well-studied examples suggest that many of these (hundreds, if not more) might participate in important and diverse aspects of biology. Recent observations regarding lincRNA genomics and evolution, such as their frequently cytoplasmic accumulation or their frequently syntenic loci despite undetectable sequence conservation, only add to the mysteries of lincRNA function and mechanism. With all this intrigue, biologists with diverse interests and backgrounds are exploring how lincRNAs might participate in the biological processes that they study. To do so, some are also expanding the experimental toolbox for interrogating lincRNA function and mechanism by developing improved tools for comparative genomics and for high-throughput identification of binding partners. The insights on the horizon will help separate this rag-tag set of transcripts into coherent, well-defined subclasses, thereby enabling the information gained from the study of one lincRNA to be more reliably leveraged for the understanding of many others, and ultimately providing a firm grasp on how many of the thousands of lincRNA genes found in the cell are functional.

ACKNOWLEDGMENTS

We thank A. Shkumatava, M. Cabili, B. Kleaveland, L. Boyer, and other colleagues for stimulating discussions and for helpful comments on this manuscript. Our work on lincRNAs is supported by NIH grant GM067031. D.B. is an Investigator of the Howard Hughes Medical Institute.

REFERENCES

Agarwal, A., Koppstein, D., Rozowsky, J., Sboner, A., Habegger, L., Hillier, L.W., Sasidharan, R., Reinke, V., Waterston, R.H., and Gerstein, M. (2010). Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics *11*, 383.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Anguera, M.C., Ma, W., Clift, D., Namekawa, S., Kelleher, R.J., 3rd, and Lee, J.T. (2011). *Tsx* produces a long noncoding RNA and has general functions in the germline, stem cells, and brain. PLoS Genet. *7*, e1002248.

Augui, S., Nora, E.P., and Heard, E. (2011). Regulation of X-chromosome inactivation by the X-inactivation centre. Nat. Rev. Genet. *12*, 429–442.

Badger, J.H., and Olsen, G.J. (1999). CRITICA: coding region identification tool invoking comparative analysis. Mol. Biol. Evol. *16*, 512–524.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

Beletskii, A., Hong, Y.K., Pehrson, J., Egholm, M., and Strauss, W.M. (2001). PNA interference mapping demonstrates functional domains in the noncoding RNA *Xist*. Proc. Natl. Acad. Sci. USA *98*, 9215–9220.

Bertani, S., Sauer, S., Bolotin, E., and Sauer, F. (2011). The noncoding RNA *Mistral* activates *Hoxa6* and *Hoxa7* expression and stem cell differentiation by recruiting MLL1 to chromatin. Mol. Cell *43*, 1040–1046.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242–2246.

Bieberstein, N.I., Carrillo Oesterreich, F., Straube, K., and Neugebauer, K.M. (2012). First exon length controls active chromatin signatures and transcription. Cell Rep *2*, 62–68.

Blackshaw, S., Harpavat, S., Trimarchi, J., Cai, L., Huang, H., Kuo, W.P., Weber, G., Lee, K., Fraioli, R.E., Cho, S.H., et al. (2004). Genomic analysis of mouse retinal development. PLoS Biol. *2*, E247.

Boerner, S., and McGinnis, K.M. (2012). Computational identification and functional predictions of long noncoding RNA in *Zea mays*. PLoS ONE *7*, e43047.

Broadbent, K.M., Park, D., Wolf, A.R., Van Tyne, D., Sims, J.S., Ribacke, U., Volkman, S., Duraisingh, M., Wirth, D., Sabeti, P.C., and Rinn, J.L. (2011). A global transcriptional analysis of *Plasmodium falciparum* malaria reveals a novel family of telomere-associated lncRNAs. Genome Biol. *12*, R56.

Brockdorff, N. (2013). Noncoding RNA and Polycomb recruitment. RNA *19*, 429–442.

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., and Rastan, S. (1992). The product of the mouse *Xist* gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515–526.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992). The human *XIST* gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527–542.

Brown, J.A., Valenstein, M.L., Yario, T.A., Tycowski, K.T., and Steitz, J.A. (2012). Formation of triple-helical structures by the 3′-end sequences of MALAT1 and MENβ noncoding RNAs. Proc. Natl. Acad. Sci. USA *109*, 19202–19207.

Brunner, A.L., Beck, A.H., Edris, B., Sweeney, R.T., Zhu, S.X., Li, R., Montgomery, K., Varma, S., Gilks, T., Guo, X., et al. (2012). Transcriptional profiling of long non-coding RNAs and novel transcribed regions across a diverse panel of archived human cancers. Genome Biol. *13*, R75.

Bussotti, G., Raineri, E., Erb, I., Zytnicki, M., Wilm, A., Beaudoing, E., Bucher, P., and Notredame, C. (2011). BlastR—fast and accurate database searches for non-coding RNAs. Nucleic Acids Res. *39*, 6886–6895.

Cabianca, D.S., Casa, V., Bodega, B., Xynos, A., Ginelli, E., Tanaka, Y., and Gabellini, D. (2012). A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. Cell *149*, 819–831.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Cai, J., Zhao, R., Jiang, H., and Wang, W. (2008). De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. Genetics *179*, 487–496.

Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. USA *106*, 7507–7512.

Caparros, M.L., Alexiou, M., Webster, Z., and Brockdorff, N. (2002). Functional analysis of the highly conserved exon IV of *XIST* RNA. Cytogenet. Genome Res. *99*, 99–105.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al.; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., et al. (2012). Long non-coding antisense RNA controls *Uchl1* translation through an embedded SINEB2 repeat. Nature *491*, 454–457.

Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., et al. (2012). Proto-genes and de novo gene birth. Nature *487*, 370–374.

Cazalla, D., Yario, T., and Steitz, J.A. (2010). Down-regulation of a host microRNA by a *Herpesvirus saimiri* noncoding RNA. Science *328*, 1563–1566.

Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., and Chang, H.Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol. Cell *44*, 667–678.

Church, D.M., Goodstadt, L., Hillier, L.W., Zody, M.C., Goldstein, S., She, X., Bult, C.J., Agarwala, R., Cherry, J.L., DiCuccio, M., et al.; Mouse Genome Sequencing Consortium. (2009). Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. *7*, e1000112.

Clark, M.B., Johnston, R.L., Inostroza-Ponta, M., Fox, A.H., Fortini, E., Moscato, P., Dinger, M.E., and Mattick, J.S. (2012). Genome-wide analysis of long noncoding RNA stability. Genome Res. *22*, 885–898.

Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., and Lawrence, J.B. (2009). An architectural role for a nuclear noncoding RNA: *NEAT1* RNA is essential for the structure of paraspeckles. Mol. Cell *33*, 717–726.

Coccia, E.M., Cicala, C., Charlesworth, A., Ciccarelli, C., Rossi, G.B., Philipson, L., and Sorrentino, V. (1992). Regulation and expression of a growth arrest-specific gene (*gas5*) during growth, differentiation, and development. Mol. Cell. Biol. *12*, 3514–3521.

Conrad, N.K., Mili, S., Marshall, E.L., Shu, M.D., and Steitz, J.A. (2006). Identification of a rapid mammalian deadenylation-dependent decay pathway and its inhibition by a viral RNA element. Mol. Cell *24*, 943–953.

Crowe, M.L., Wang, X.Q., and Rothnagel, J.A. (2006). Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. BMC Genomics *7*, 16.

de Almeida, S.F., Grosso, A.R., Koch, F., Fenouil, R., Carvalho, S., Andrade, J., Levezinho, H., Gut, M., Eick, D., Gut, I., et al. (2011). Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. Nat. Struct. Mol. Biol. *18*, 977–983.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. *8*, e1000384.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Dinger, M.E., Pang, K.C., Mercer, T.R., and Mattick, J.S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput. Biol. *4*, e1000176.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Duret, L., Chureau, C., Samain, S., Weissenbach, J., and Avner, P. (2006). The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science *312*, 1653–1655.

Ebert, M.S., and Sharp, P.A. (2010). Emerging roles for natural microRNA sponges. Curr. Biol. *20*, R858–R861.

Ebert, M.S., Neilson, J.R., and Sharp, P.A. (2007). MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat. Methods *4*, 721–726.

Eißmann, M., Gutschner, T., Hämmerle, M., Günther, S., Caudron-Herger, M., Groß, M., Schirmacher, P., Rippe, K., Braun, T., Zörnig, M., and Diederichs, S. (2012). Loss of the abundant nuclear non-coding RNA *MALAT1* is compatible with life and development. RNA Biol. *9*, 1076–1087.

Elisaphenko, E.A., Kolesnikov, N.N., Shevchenko, A.I., Rogozin, I.B., Nesterova, T.B., Brockdorff, N., and Zakian, S.M. (2008). A dual origin of the *Xist* gene from a protein-coding gene and a set of transposable elements. PLoS ONE *3*, e2521.

Feuerhahn, S., Iglesias, N., Panza, A., Porro, A., and Lingner, J. (2010). TERRA biogenesis, turnover and implications for function. FEBS Lett. *584*, 3812–3818.

Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I., Puga, M.I., Rubio-Somoza, I., Leyva, A., Weigel, D., García, J.A., and Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. Nat. Genet. *39*, 1033–1037.

Frank-Kamenetskii, M.D., and Mirkin, S.M. (1995). Triplex DNA structures. Annu. Rev. Biochem. *64*, 65–95.

Galindo, M.I., Pueyo, J.I., Fouix, S., Bishop, S.A., and Couso, J.P. (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol. *5*, e106.

Gomez, J.A., Wapinski, O.L., Yang, Y.W., Bureau, J.F., Gopinath, S., Monack, D.M., Chang, H.Y., Brahic, M., and Kirkegaard, K. (2013). The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-γ locus. Cell *152*, 743–754.

Gong, C., and Maquat, L.E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. Nature *470*, 284–288.

Gorodkin, J., Hofacker, I.L., Torarinsson, E., Yao, Z., Havgaard, J.H., and Ruzzo, W.L. (2010). De novo prediction of structured RNAs from genomic sequences. Trends Biotechnol. *28*, 9–19.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., Macura, K., Bläss, G., Kellis, M., Werber, M., and Herrmann, B.G. (2013). The tissue-specific lncRNA *Fendrr* is an essential regulator of heart and body wall development in the mouse. Dev. Cell *24*, 206–214.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835–840.

Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non-coding RNAs. Nature *482*, 339–346.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., et al. (2010). *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. *28*, 503–510.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

Hackett, P.B., Petersen, R.B., Hensel, C.H., Albericio, F., Gunderson, S.I., Palmenberg, A.C., and Barany, G. (1986). Synthesis *in vitro* of a seven amino acid peptide encoded in the leader RNA of Rous sarcoma virus. J. Mol. Biol. *190*, 45–57.

Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. Nature *495*, 384–388.

He, S., Liu, S., and Zhu, H. (2011). The sequence, structure and evolutionary features of HOTAIR in mammals. BMC Evol. Biol. *11*, 102.

Hu, W., Yuan, B., Flygare, J., and Lodish, H.F. (2011). Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. Genes Dev. *25*, 2573–2578.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., et al. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell *142*, 409–419.

Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P., et al. (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat. Genet. *43*, 621–629.

Hutchinson, J.N., Ensminger, A.W., Clemson, C.M., Lynch, C.R., Lawrence, J.B., and Chess, A. (2007). A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC Genomics *8*, 39.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

Ishii, N., Ozaki, K., Sato, H., Mizuno, H., Saito, S., Takahashi, A., Miyamoto, Y., Ikegawa, S., Kamatani, N., Hori, M., et al. (2006). Identification of a novel non-coding RNA, *MIAT*, that confers risk of myocardial infarction. J. Hum. Genet. *51*, 1087–1099.

Jan, C.H., Friedman, R.C., Ruby, J.G., and Bartel, D.P. (2011). Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature *469*, 97–101.

Jeon, Y., and Lee, J.T. (2011). YY1 tethers *Xist* RNA to the inactive X nucleation center. Cell *146*, 119–133.

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., et al. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene *22*, 8031–8041.

Jia, H., Osak, M., Bogu, G.K., Stanton, L.W., Johnson, R., and Lipovich, L. (2010). Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA *16*, 1478–1487.

Kelley, D.R., and Rinn, J.L. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. *13*, R107.

Khachane, A.N., and Harrison, P.M. (2009). Assessing the genomic evidence for conserved transcribed pseudogenes under selection. BMC Genomics *10*, 435.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. USA *106*, 11667–11672.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Kino, T., Hurt, D.E., Ichijo, T., Nader, N., and Chrousos, G.P. (2010). Noncoding RNA *gas5* is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. Sci. Signal. *3*, ra8.

Klattenhoff, C.A., Scheuermann, J.C., Surface, L.E., Bradley, R.K., Fields, P.A., Steinhauser, M.L., Ding, H., Butty, V.L., Torrey, L., Haas, S., et al. (2013). *Braveheart*, a long noncoding RNA required for cardiovascular lineage commitment. Cell *152*, 570–583.

Kloc, M., Wilk, K., Vargas, D., Shirato, Y., Bilinski, S., and Etkin, L.D. (2005). Potential structural role of non-coding and coding RNAs in the organization of the cytoskeleton at the vegetal cortex of *Xenopus* oocytes. Development *132*, 3445–3457.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., et al. (2006). CAGE: cap analysis of gene expression. Nat. Methods *3*, 211–222.

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. Science *329*, 336–339.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. *35*(Web Server issue), W345-9.

Kowalczyk, M.S., Higgs, D.R., and Gingeras, T.R. (2012). Molecular biology: RNA discrimination. Nature *482*, 310–311.

Kretz, M., Siprashvili, Z., Chu, C., Webster, D.E., Zehnder, A., Qu, K., Lee, C.S., Flockhart, R.J., Groff, A.F., Chow, J., et al. (2013). Control of somatic tissue differentiation by the long non-coding RNA *TINCR*. Nature *493*, 231–235.

Kudla, G., Lipinski, L., Caffin, F., Helwak, A., and Zylicz, M. (2006). High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol. *4*, e180.

Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., and Marques, A.C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. *8*, e1002841.

Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497–501.

Lanz, R.B., McKenna, N.J., Onate, S.A., Albrecht, U., Wong, J., Tsai, S.Y., Tsai, M.J., and O'Malley, B.W. (1999). A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. Cell *97*, 17–27.

Lee, J.T. (2009). Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Genes Dev. *23*, 1831–1842.

Lin, M.F., Carlson, J.W., Crosby, M.A., Matthews, B.B., Yu, C., Park, S., Wan, K.H., Schroeder, A.J., Gramates, L.S., St Pierre, S.E., et al. (2007). Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. Genome Res. *17*, 1823–1836.

Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*, i275–i282.

Liu, J., Gough, J., and Rost, B. (2006). Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet. *2*, e29.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C., and Chua, N.H. (2012a). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. Plant Cell *24*, 4333–4345.

Liu, X., Li, D., Zhang, W., Guo, M., and Zhan, Q. (2012b). Long non-coding RNA *gadd7* interacts with TDP-43 and regulates *Cdk6* mRNA decay. EMBO J. *31*, 4415–4427.

Maenner, S., Blaud, M., Fouillen, L., Savoye, A., Marchand, V., Dubois, A., Sanglier-Cianférani, S., Van Dorsselaer, A., Clerc, P., Avner, P., et al. (2010). 2-D structure of the A region of *Xist* RNA and its implication for PRC2 association. PLoS Biol. *8*, e1000276.

Managadze, D., Rogozin, I.B., Chernikova, D., Shabalina, S.A., and Koonin, E.V. (2011). Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. Genome Biol. Evol. *3*, 1390–1404.

Marques, A.C., and Ponting, C.P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol. *10*, R124.

Marques, A.C., Tan, J., Lee, S., Kong, L., Heger, A., and Ponting, C.P. (2012). Evidence for conserved post-transcriptional roles of unitary pseudogenes and for frequent bifunctionality of mRNAs. Genome Biol. *13*, R102.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature *495*, 333–338.

Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., Crappe, J., Gevaert, K., and Van Damme, P. (2013). Deep proteome coverage based on ribosome profiling aids MS-based protein and peptide discovery and provides

evidence of alternative translation products and near-cognate translation initiation events. Mol. Cell. Proteomics. Published online February 21, 2013.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. (2008). Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. USA *105*, 716–721.

Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S., and Rinn, J.L. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat. Biotechnol. *30*, 99–104.

Michel, F., and Westhof, E. (1990). Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J. Mol. Biol. *216*, 585–610.

Mitton-Fry, R.M., DeGregorio, S.J., Wang, J., Steitz, T.A., and Steitz, J.A. (2010). Poly(A) tail recognition by a viral RNA element through assembly of a triple helix. Science *330*, 1244–1247.

Moazed, D. (2009). Small RNAs in transcriptional gene silencing and genome defence. Nature *457*, 413–420.

Murthy, U.M., and Rangarajan, P.N. (2010). Identification of protein interaction regions of VINC/NEAT1/Men ε RNA. FEBS Lett. *584*, 1531–1535.

Nakagawa, S., Ip, J.Y., Shioi, G., Tripathi, V., Zong, X., Hirose, T., and Prasanth, K.V. (2012). *Malat1* is not an essential component of nuclear speckles in mice. RNA *18*, 1487–1499.

Nam, J.W., and Bartel, D.P. (2012). Long noncoding RNAs in *C. elegans*. Genome Res. *22*, 2529–2540.

Nesterova, T.B., Slobodyanyuk, S.Y., Elisaphenko, E.A., Shevchenko, A.I., Johnston, C., Pavlova, M.E., Rogozin, I.B., Kolesnikov, N.N., Brockdorff, N., and Zakian, S.M. (2001). Characterization of the genomic *Xist* locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. Genome Res. *11*, 833–849.

Novikova, I.V., Hennelly, S.P., and Sanbonmatsu, K.Y. (2012). Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. Nucleic Acids Res. *40*, 5034–5051.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al.; FANTOM Consortium; RIKEN Genome Exploration Research Group Phase I & II Team. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563–573.

Ørom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46–58.

Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W., and Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. Genome Res. *15*, 137–145.

Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D., and Kanduri, C. (2008). *Kcnq1ot1* antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol. Cell *32*, 232–246.

Pauli, A., Rinn, J.L., and Schier, A.F. (2011). Non-coding RNAs as regulators of embryogenesis. Nat. Rev. Genet. *12*, 136–149.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. *22*, 577–591.

Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X.J., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., et al. (2012). The GENCODE pseudogene resource. Genome Biol. *13*, R51.

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorff, N. (1996). Requirement for *Xist* in X chromosome inactivation. Nature *379*, 131–137.

Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L., and Carter, D.R. (2011). Pseudogenes: pseudo-functional or key regulators in health and disease? RNA *17*, 792–798.

Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. *17*, 556–565.

Ponjavic, J., Oliver, P.L., Lunter, G., and Ponting, C.P. (2009). Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS Genet. *5*, e1000617.

Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. Cell *136*, 629–641.

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., et al. (2011). Transcriptome sequencing across a prostate cancer cohort identifies *PCAT-1*, an unannotated lincRNA implicated in disease progression. Nat. Biotechnol. *29*, 742–749.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. *40*(Database issue), D130–D135.

Rapicavoli, N.A., Poth, E.M., and Blackshaw, S. (2010). The long noncoding RNA *RNCR2* directs mouse retinal cell specification. BMC Dev. Biol. *10*, 49.

Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M., et al. (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res. *16*, 11–19.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., and Chang, H.Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.

Rivas, E., and Eddy, S.R. (2001). Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics *2*, 8.

Rivas, E., Klein, R.J., Jones, T.A., and Eddy, S.R. (2001). Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Curr. Biol. *11*, 1369–1373.

Sarma, K., Levasseur, P., Aristarkhov, A., and Lee, J.T. (2010). Locked nucleic acids (LNAs) reveal sequence requirements and kinetics of *Xist* RNA localization to the X chromosome. Proc. Natl. Acad. Sci. USA *107*, 22196–22201.

Schoeftner, S., Sengupta, A.K., Kubicek, S., Mechtler, K., Spahn, L., Koseki, H., Jenuwein, T., and Wutz, A. (2006). Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. EMBO J. *25*, 3110–3122.

Schorderet, P., and Duboule, D. (2011). Structural and functional differences in the long non-coding RNA hotair in mouse and human. PLoS Genet. *7*, e1002071.

Schultes, E.A., Spasic, A., Mohanty, U., and Bartel, D.P. (2005). Compact and ordered collapse of randomly generated RNA sequences. Nat. Struct. Mol. Biol. *12*, 1130–1136.

Shearwin, K.E., Callen, B.P., and Egan, J.B. (2005). Transcriptional interference—a crash course. Trends Genet. *21*, 339–345.

Shoemaker, C.J., and Green, R. (2012). Translation drives mRNA quality control. Nat. Struct. Mol. Biol. *19*, 594–601.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., and Young, R.A. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl. Acad. Sci. USA *110*, 2876–2881.

Simon, M.D., Wang, C.I., Kharchenko, P.V., West, J.A., Chapman, B.A., Alekseyenko, A.A., Borowsky, M.L., Kuroda, M.I., and Kingston, R.E. (2011). The genomic binding sites of a noncoding RNA. Proc. Natl. Acad. Sci. USA *108*, 20497–20502.

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat. Chem. Biol. *9*, 59–64.

Sleutels, F., Zwart, R., and Barlow, D.P. (2002). The non-coding *Air* RNA is required for silencing autosomal imprinted genes. Nature *415*, 810–813.

Sone, M., Hayashi, T., Tarui, H., Agata, K., Takeichi, M., and Nakagawa, S. (2007). The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. J. Cell Sci. *120*, 2498–2506.

Souquere, S., Beauclair, G., Harper, F., Fox, A., and Pierron, G. (2010). Highly ordered spatial organization of the structural long noncoding NEAT1 RNAs within paraspeckle nuclear bodies. Mol. Biol. Cell *21*, 4020–4027.

Stadler, P.F. (2010). Evolution of the Long Non-coding RNAs *MALAT1* and *MEN*. In Advances in Bioinformatics and Computational Biology, C.E. Ferreira, S. Miyano, and P.F. Stadler, eds. (Rio de Janeiro, Brazil: Springer).

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol. *14*, 103–105.

Sun, R., Lin, S.F., Gradoville, L., and Miller, G. (1996). Polyadenylylated nuclear RNA encoded by Kaposi sarcoma-associated herpesvirus. Proc. Natl. Acad. Sci. USA *93*, 11883–11888.

Sunwoo, H., Dinger, M.E., Wilusz, J.E., Amaral, P.P., Mattick, J.S., and Spector, D.L. (2009). MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. Genome Res. *19*, 347–359.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M., and Hannon, G.J. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature *453*, 534–538.

Tan, M.H., Au, K.F., Yablonovitch, A.L., Wills, A.E., Chuang, J., Baker, J.C., Wong, W.H., and Li, J.B. (2013). RNA sequencing reveals a diverse and dynamic repertoire of the *Xenopus tropicalis* transcriptome over development. Genome Res. *23*, 201–216.

Tani, H., Mizutani, R., Salam, K.A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., and Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res. *22*, 947–956.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. *22*, 1616–1625.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., et al. (2010). The nuclear-retained noncoding RNA *MALAT1* regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol. Cell *39*, 925–938.

Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. Science *329*, 689–693.

Tsuiji, H., Yoshimoto, R., Hasegawa, Y., Furuno, M., Yoshida, M., and Nakagawa, S. (2011). Competition between a noncoding exon and introns: *Gomafu* contains tandem UACUAAC repeats and associates with splicing factor-1. Genes Cells *16*, 479–490.

Tupy, J.L., Bailey, A.M., Dailey, G., Evans-Holm, M., Siebel, C.W., Misra, S., Celniker, S.E., and Rubin, G.M. (2005). Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA *102*, 5495–5500.

Tycowski, K.T., Shu, M.D., Borah, S., Shi, M., and Steitz, J.A. (2012). Conservation of a triple-helix-forming RNA stability element in noncoding and genomic RNAs of diverse viruses. Cell Rep *2*, 26–32.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

Ulitsky, I., Shkumatava, A., Jan, C.H., Subtelny, A.O., Koppstein, D., Bell, G.W., Sive, H., and Bartel, D.P. (2012). Extensive alternative polyadenylation during zebrafish development. Genome Res. *22*, 2054–2066.

van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most "dark matter" transcripts are associated with known genes. PLoS Biol. *8*, e1000371.

Wadler, C.S., and Vanderpool, C.K. (2007). A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. Proc. Natl. Acad. Sci. USA *104*, 20454–20459.

Wamstad, J.A., Alexander, J.M., Truty, R.M., Shrikumar, A., Li, F., Eilertson, K.E., Ding, H., Wylie, J.N., Pico, A.R., Capra, J.A., et al. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. Cell *151*, 206–220.

Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., Samudrala, R., Yu, J., and Wong, G.K. (2004). Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. Nature *431*, 1 p following 757; discussion following 757.

Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M.G., Glass, C.K., and Kurokawa, R. (2008). Induced ncRNAs allosterically modify RNA-binding proteins in *cis* to inhibit transcription. Nature *454*, 126–130.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Washietl, S., Findeiss, S., Müller, S.A., Kalkhof, S., von Bergen, M., Hofacker, I.L., Stadler, P.F., and Goldman, N. (2011). RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA *17*, 578–594.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., et al. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature *453*, 539–543.

Wethmar, K., Smink, J.J., and Leutz, A. (2010). Upstream open reading frames: molecular switches in (patho)physiology. Bioessays *32*, 885–893.

Willingham, A.T., Orth, A.P., Batalov, S., Peters, E.C., Wen, B.G., Aza-Blanc, P., Hogenesch, J.B., and Schultz, P.G. (2005). A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science *309*, 1570–1573.

Wilusz, J.E., JnBaptiste, C.K., Lu, L.Y., Kuhn, C.D., Joshua-Tor, L., and Sharp, P.A. (2012). A triple helix stabilizes the 3′ ends of long noncoding RNAs that lack poly(A) tails. Genes Dev. *26*, 2392–2407.

Woese, C.R., Magrum, L.J., Gupta, R., Siegel, R.B., Stahl, D.A., Kop, J., Crawford, N., Brosius, J., Gutell, R., Hogan, J.J., and Noller, H.F. (1980). Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. Nucleic Acids Res. *8*, 2275–2293.

Wutz, A., Rasmussen, T.P., and Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of *Xist* RNA. Nat. Genet. *30*, 167–174.

Xie, C., Zhang, Y.E., Chen, J.Y., Liu, C.J., Zhou, W.Z., Li, Y., Zhang, M., Zhang, R., Wei, L., and Li, C.Y. (2012). Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. PLoS Genet. *8*, e1002942.

Yao, H., Brick, K., Evrard, Y., Xiao, T., Camerini-Otero, R.D., and Felsenfeld, G. (2010). Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. Genes Dev. *24*, 2543–2555.

Yap, K.L., Li, S., Muñoz-Cabello, A.M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M.J., and Zhou, M.M. (2010). Molecular interplay of the noncoding RNA *ANRIL* and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of *INK4a*. Mol. Cell *38*, 662–674.

Yin, Q.F., Yang, L., Zhang, Y., Xiang, J.F., Wu, Y.W., Carmichael, G.G., and Chen, L.L. (2012). Long noncoding RNAs with snoRNA ends. Mol. Cell *48*, 219–230.

Yoon, J.H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J.L., De, S., Huarte, M., Zhan, M., Becker, K.G., and Gorospe, M. (2012). LincRNA-p21 suppresses target mRNA translation. Mol. Cell *47*, 648–655.

Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.L., and Ponting, C.P. (2012). Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. Genome Biol. Evol. *4*, 427–442.

Zhang, B., Arun, G., Mao, Y.S., Lazar, Z., Hung, G., Bhattacharjee, G., Xiao, X., Booth, C.J., Wu, J., Zhang, C., and Spector, D.L. (2012). The lncRNA *Malat1* is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. Cell Rep *2*, 111–123.

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J., and Lee, J.T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science *322*, 750–756.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell *40*, 939–953.

Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J.Q., and Tian, D. (2009). Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics *10*, 47.

Ziats, M.N., and Rennert, O.M. (2013). Aberrant expression of long noncoding RNAs in autistic brain. J. Mol. Neurosci. *49*, 589–593.