REVIEW

doi:10.1038/nature10887

Modular regulatory principles of large non-coding RNAs

Mitchell Guttman[1,2] & John L. Rinn[1,3]

It is clear that RNA has a diverse set of functions and is more than just a messenger between gene and protein. The mammalian genome is extensively transcribed, giving rise to thousands of non-coding transcripts. Whether all of these transcripts are functional is debated, but it is evident that there are many functional large non-coding RNAs (ncRNAs). Recent studies have begun to explore the functional diversity and mechanistic role of these large ncRNAs. Here we synthesize these studies to provide an emerging model whereby large ncRNAs might achieve regulatory specificity through modularity, assembling diverse combinations of proteins and possibly RNA and DNA interactions.

More than half a century after being placed as the central component in the flow of genetic information from gene to protein, it is now accepted that RNA can perform diverse roles. Shortly after the discovery of messenger RNA, a large class of heteronuclear RNAs (hnRNAs)[1] was described, which did not include mRNA or associate with polyribosomes[2]. Following years of sifting through these hnRNAs, the first RNA subfamilies were identified. These included small nuclear RNAs involved in splicing regulation[3] and small nucleolar RNAs involved in ribosome biogenesis[4], as well as the ribosomal RNAs and transfer RNAs involved in translation[5,6].

The world of RNA genes became even more complex with the discovery of RNAs that resembled mRNA in length and splicing structure but did not code for proteins. The first example was H19, which was identified as an RNA that was induced during liver development in the mouse[7]. The mouse *H19* transcript contained no large open reading frames (ORFs), but instead only small sporadic ORFs that were not evolutionarily conserved, did not template translation *in vivo* and did not produce an identifiable protein product[8]. Shortly afterwards, another non-coding RNA (ncRNA), termed XIST, was found to be expressed exclusively from the inactive X chromosome[9] and later demonstrated to be required for X inactivation in mammals[10]. Over the next two decades, more large ncRNA genes were discovered including *Airn*[11], *Tug1* (ref. 12), *NRON*[13] and *HOTAIR*[14]. With the availability of a draft sequence of the human genome, it became clear that much of the mammalian genome is transcribed[15–18]. These transcripts were mapped to discrete loci throughout the genome. Over the next 10 years, both large and small RNA transcripts were discovered at an unprecedented rate[15,17–20]; however, the functional significance of most of these transcripts was unclear. Although some of these could be considered noise[21,22], there are still many large ncRNAs that are known to have diverse functions[23–29].

This Review focuses on the classic examples of large ncRNAs that have helped to form the basis of more recent global studies of coding potential, function and mechanism. We discuss the concepts that have emerged from these examples that provide a framework for understanding the principles of RNA interactions. We propose that by assembling distinct regulatory components, large ncRNAs could produce intricate functional specificity, which is suggestive of a possible modular RNA code.

RNA maps

After the sequencing of the human genome, the next major hurdle was to define the genes it encoded. To do this, several research groups developed tiling microarrays[17,19,20] and complementary DNA sequencing methods[15] to investigate transcriptional activity across the human genome, which led to the observation of widespread transcription of the genome. These studies, although limited to specific tissues and cell types, demonstrated that the mammalian genome encodes many thousands of non-coding transcripts including both short (<200 nucleotides in length) and long (>200 nucleotides in length) transcripts. In this Review, we focus on large ncRNAs produced from long transcripts, including those that originate from intergenic loci or overlapping protein-coding genes.

Dramatic innovations in sequencing technologies have allowed the deep sequencing of cDNAs, known as RNA-Seq[30]; this deep sequencing, coupled with new computational methods for assembling the transcriptome[31], has identified non-coding transcripts across many different cell types and tissues[31,32]. It is now clear that there are thousands of well-expressed large ncRNAs with exquisite cell-type and tissue specificity[31–33].

As the numbers of identified non-coding transcripts increased, so did the uncertainty regarding their function; this led some authors to express concern that many of these transcripts may be just transcriptional noise[21,22] with no function or incidental by-products of transcription from enhancer regions[34,35]. These concerns are supported by the observations that many of these transcripts are expressed at extremely low levels[32,36] and they have lower levels of evolutionary conservation than protein-coding genes[25,31,37]. Although some of these transcripts may indeed be transcriptional noise[21], the remaining transcripts consist of many distinct subclasses, including processed small RNAs[18,29,38], promoter-associated RNAs[18,39], transcripts from enhancer regions[34,35] and functional large ncRNAs[14,23]; each class varies in its expression and conservation properties[31,37]. Distinguishing between these classes of RNA transcripts requires additional biological information including the coding potential of the RNA and the chromatin modifications of the corresponding genomic region (Fig. 1a).

Chromatin signatures

Genomic DNA is wrapped around histone proteins and packaged into higher-order structures termed chromatin[40]. These histones can be modified in different ways that are indicative of the underlying DNA functional state. Advances in sequencing technologies have allowed the comprehensive characterization of the chromatin-modification landscape of mammalian genomes[41–44]. These studies revealed combinations of histone modifications (termed chromatin signatures) that correspond to various gene properties, including a signature for active

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

transcription[41,44]. This signature consists of a short stretch of trimethylation of histone protein H3 at the lysine in position 4 (H3K4me3), which corresponds to promoter regions, followed by a longer stretch of trimethylation of histone H3 at the lysine in position 36 (H3K36me3), which covers the entire transcribed region[41,44] (Fig. 1a).

Chromatin maps revealed that, similar to protein-coding genes, many ncRNA genes also contain a 'K4–K36' signature[44]. By searching for K4–K36 domains that do not overlap with known genes, chromatin signatures revealed approximately 1,600 regions in the mouse genome and approximately 2,500 regions in the human genome that were actively transcribed[25,45]. The vast majority of these intergenic K4–K36 domains produce multi-exonic RNAs that have little capability to encode a conserved protein[25,31]. RNAs expressed from these K4–K36 domains were termed large intergenic ncRNAs (lincRNAs) because

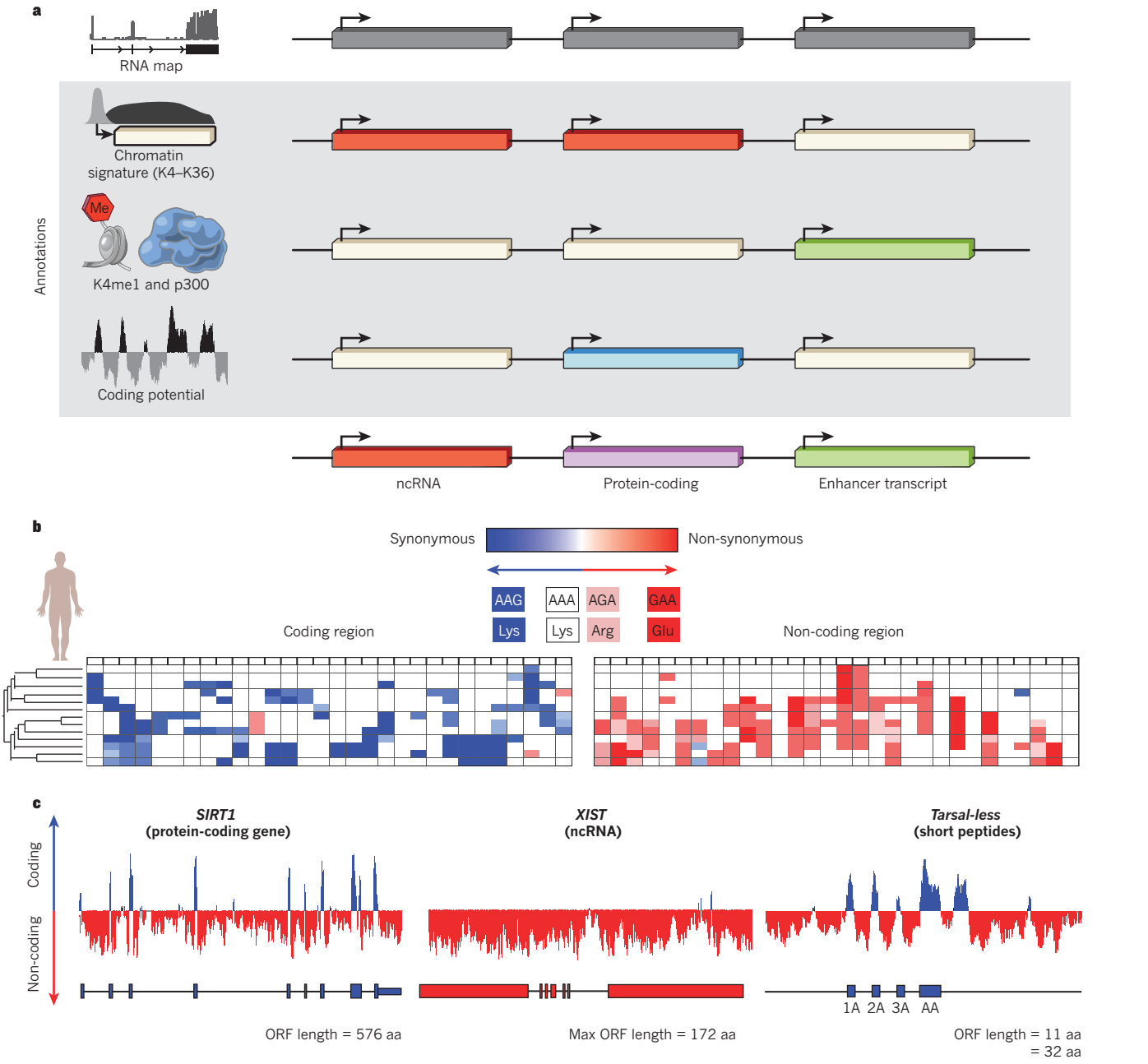


Figure 1 | Layering of genomic regions. **a**, Genomic regions are colour-coded by the presence of different genomic annotations. RNA transcription of a locus (grey), K4–K36 chromatin signature (red), K4me1 modification and transcriptional activator p300 (green) and protein-coding potential (blue). By overlaying this information, distinct transcripts are revealed, including ncRNAs (red), protein-coding genes (purple) and transcripts from enhancer regions (green). **b**, A cross-species alignment of a coding and a non-coding gene. Boxes represent codons, and each row represents a different aligned species. Blue boxes represent mutations that cause a synonymous substitution, and red boxes represent mutations that cause a non-synonymous substitution. A score capturing the coding potential of a sequence across species aligns sequences in all frames and scores mutations that maintain coding potential (blue boxes) relative to mutations that break coding potential (that is, non-synonymous mutations, stop codons and frameshifting insertions or deletions) (red boxes). **c**, The coding potential score is shown for three gene types, *SIRT1* (a protein-coding gene), *XIST* (ncRNA gene) and *tarsal-less* (small-peptide coding gene), in which positive scores represent coding regions (blue) and negative scores represent non-coding regions (red). In each example, the gene structure is shown, where blue boxes represent known protein-coding exons and red boxes represent non-coding exons. *SIRT1* with an ORF length of 576 amino acids (aa) contains a positive score over each coding exon but not the non-coding regions. *XIST* with an ORF length of 172 amino acids contains negative scores over the entire transcribed region. *tarsal-less* with an ORF of 11 and 32 amino acids, contains positive scores over all known small peptides.

identification by this chromatin signature required the RNAs to be contained within the intergenic regions[25]. Similarly, chromatin-state maps revealed that active enhancer regions contained short stretches of H3 lysine 4 monomethylation (H3K4me1) (ref. 43) and the transcriptional coactivator p300 (ref. 42), as well as additional modifications[46] (Fig. 1a). By coupling RNA sequencing and chromatin maps, many of the already identified non-coding transcripts were observed to be transcribed from active enhancers[34,35]. However, lincRNAs and transcripts from enhancer regions are distinct classes, which are marked by different chromatin signatures[25,34]. Although it needs to be determined whether transcripts originating from enhancers have a function[34,35], the functional importance of lincRNAs is becoming clearer[14,23,24,26,28,47]. Several of these lincRNAs have been shown to have enhancer-like functions as they activate the expression of neighbouring genes[24,28].

Coding potential

Determining whether a transcript is non-coding is challenging because a long non-coding transcript is likely to contain an ORF purely by chance[48]. Accordingly, the evidence for the absence of coding potential for the *XIST* and *H19* genes came from the lack of evolutionary conservation of the identified ORFs, the lack of homology to known protein domains and the inability to template significant protein production[8,49]. These principles have been generalized to classify coding potential across thousands of transcripts by scoring conserved ORFs across dozens of species[50,51], by searching for homology in large protein-domain databases[52], and by sequencing RNA associated with polyribosomes[53].

Computational methods such as the 'codon substitution frequency' algorithm[50,51] leverage evolutionary information to determine whether an ORF is conserved across species and provide a general strategy for determining coding potential (Fig. 1b, c). Owing to the large number of available genome sequences, these methods have been used to accurately determine conserved coding potential in regions as small as 5 amino acids[25], which makes them extremely sensitive to the potentially small peptides, such as the 11 amino acid peptide encoded by the *tarsal-less* gene[54,55] (Fig. 1c). Despite their sensitivity, conservation-based methods may fail to detect newly evolved proteins because they do not contain a conserved ORF[50,51]. However, because many ncRNAs show clear evolutionary constraint[25,31,37] but no evolutionarily conserved ORF, this indicates that the observed evolutionary selection is not due to a newly evolved protein.

Experimental methods, such as ribosome profiling, have provided a strategy for identifying ribosome occupancy on RNA, which have been proposed as a method for distinguishing between coding and non-coding transcripts[53]. However, this still needs to be tested because non-coding transcripts that show an association with the ribosome have not been shown to have a protein product[53,56]. Importantly, an association of RNA with a ribosome alone cannot be taken as evidence of protein-coding potential because both the ncRNAs of *H19* and *TUG1* can be detected in the ribosome[53,57] despite having clear roles as ncRNAs[8,45,58,59].

An alternative explanation for these observed associations is 'translational noise', spurious association that may lead to non-functional translation products[22]. Consistent with this, virtually all of the transcripts that have been suggested to encode small peptides by ribosome profiling[53] lack the evolutionary conservation of their proposed coding regions[25,31], which is in striking contrast to almost all known protein-coding genes[60], including the few well-characterized functional small peptides[56,61,62] (Fig. 1c). Accordingly, identification of any new protein-coding gene requires the clear demonstration of the function of the protein product *in vivo*[54,55].

Global identification of ncRNA function

Identifying the functional role of an ncRNA requires direct perturbation experiments, such as loss-of-function and gain-of-function. Individual ncRNAs involved in specific processes have been functionally characterized (see ref. 63 for a review). For example, XIST is crucial for random inactivation of the X chromosome[10]; Air is crucial for imprinting control at the *Igf2r* locus[11]; HOTAIR affects expression of the *HOXD* gene family[14], as well as other genes throughout the genome[45,64,65]; HOTTIP affects expression of the *HOXA* gene family[28]; lincRNA-RoR affects reprogramming efficiency[47]; NRON affects NFAT transcription factor activity[13]; and Tug1 affects retina development through the regulation of the cell cycle[12]. Although there are now many examples of large ncRNAs that are required for the correct regulation of gene expression, this is just one of many functions in which they are involved; ranging from telomere replication[66] to translation[67].

The global characterization of ncRNA function has proved to be challenging because, in most cases, it is unclear which phenotype to investigate[13]. One approach to classifying the putative function of ncRNAs uses 'guilt-by-association'[25]. This approach associates ncRNAs with biological processes based on a common expression pattern across cell types and tissues (Fig. 2a) and can therefore identify groups of ncRNAs that are associated with specific cellular processes (Fig. 2b). This approach has been used to predict roles for hundreds of ncRNAs in diverse biological processes such as stem cell pluripotency, immune responses, neural processes and cell-cycle regulation[25,27,36].

Although these correlations cannot prove that ncRNAs have a function in these processes, they do provide a hypothesis for targeted loss-of-function experiments. For example, lincRNA-p21 was predicted to be associated with the p53-mediated DNA damage response[25], and indeed lincRNA-p21 was found to be a target of p53 and on perturbation was shown to regulate apoptosis in response to DNA damage[26]. In the same way, the ncRNA PANDA (p21 associated ncRNA DNA damage activated) was implicated, and was demonstrated to have a function, in the regulation of apoptosis[27]. Another ncRNA, lincEnc1 (ref. 25), was predicted to have a role in cell-cycle regulation in embryonic stem (ES) cells and has been shown in a separate study to affect the proliferation of ES cells[68].

Alternatively, global approaches can be used to determine function, such as systematic RNA interference (RNAi) knockdown followed by gene-expression profiling. Unlike correlation analysis, these perturbation-based experiments provide evidence for the function of an ncRNA[23]. Methods to classify function using this approach are conceptually similar to guilt-by-association because the function can be inferred on the basis of the genes that are affected by loss of function of ncRNAs[23]. A systematic perturbation study demonstrated that knockdown of the vast majority of lincRNAs expressed in ES cells had a major effect on gene expression[23]. The gene-expression signatures revealed dozens of lincRNAs that block key lineage-commitment programs within ES cells and function in crucial ES cell regulatory and signalling pathways. Importantly, this study also identified 26 lincRNAs that are required to maintain the pluripotent state[23].

Not all non-coding transcripts are functional RNA molecules. Several examples of intergenic transcription have been identified in which the process of transcription alone changes the chromatin- and transcription-factor-binding landscape to allow activation and repression of neighbouring genes[69,70]. Methods that degrade RNA after its transcription, such as RNAi, can distinguish between a functional RNA molecule and the process of transcription, on which there should be no observable effect after RNA degradation. Collectively, the genome-wide guilt-by-association approach and targeted and global perturbation studies have demonstrated that large ncRNAs have a crucial regulatory role in diverse biological processes[23,25–27,32,47].

cis- versus *trans*-regulatory mechanisms

The discovery that the *XIST* product was an ncRNA, led immediately to the suggestion of a model for how it could function in an allele-specific manner[9]. In theory, an ncRNA has an intrinsic *cis*-regulatory capacity because it can function while remaining tethered to its own locus[9,71] (Fig. 2c), whereas an mRNA must be dissociated, exported and translated for it to function. Here we define a *cis*-regulator as one that exerts its function on a neighbouring gene on the same allele from which it is transcribed, and define a *trans*-regulator as one that

does not meet this criterion. Owing to the unique *cis*-regulatory capability of ncRNAs, it has been speculated that *cis*-regulation could be a common mechanism for large ncRNAs[24,71]. However, global functional evidence strongly suggests that this is not the case (Box 1).

To distinguish *cis*- from *trans*-regulatory models, initial studies have used correlation analysis and identified a significant correlation of expression between ncRNAs and their neighbouring protein-coding genes[21,72]. However, several of these cases have been demonstrated to be *trans*-regulatory models, and the apparent correlations are due to shared upstream regulation (such as, lincRNA-p21 (ref. 26) and lincRNA-Sox2 (ref. 25)), positional correlation

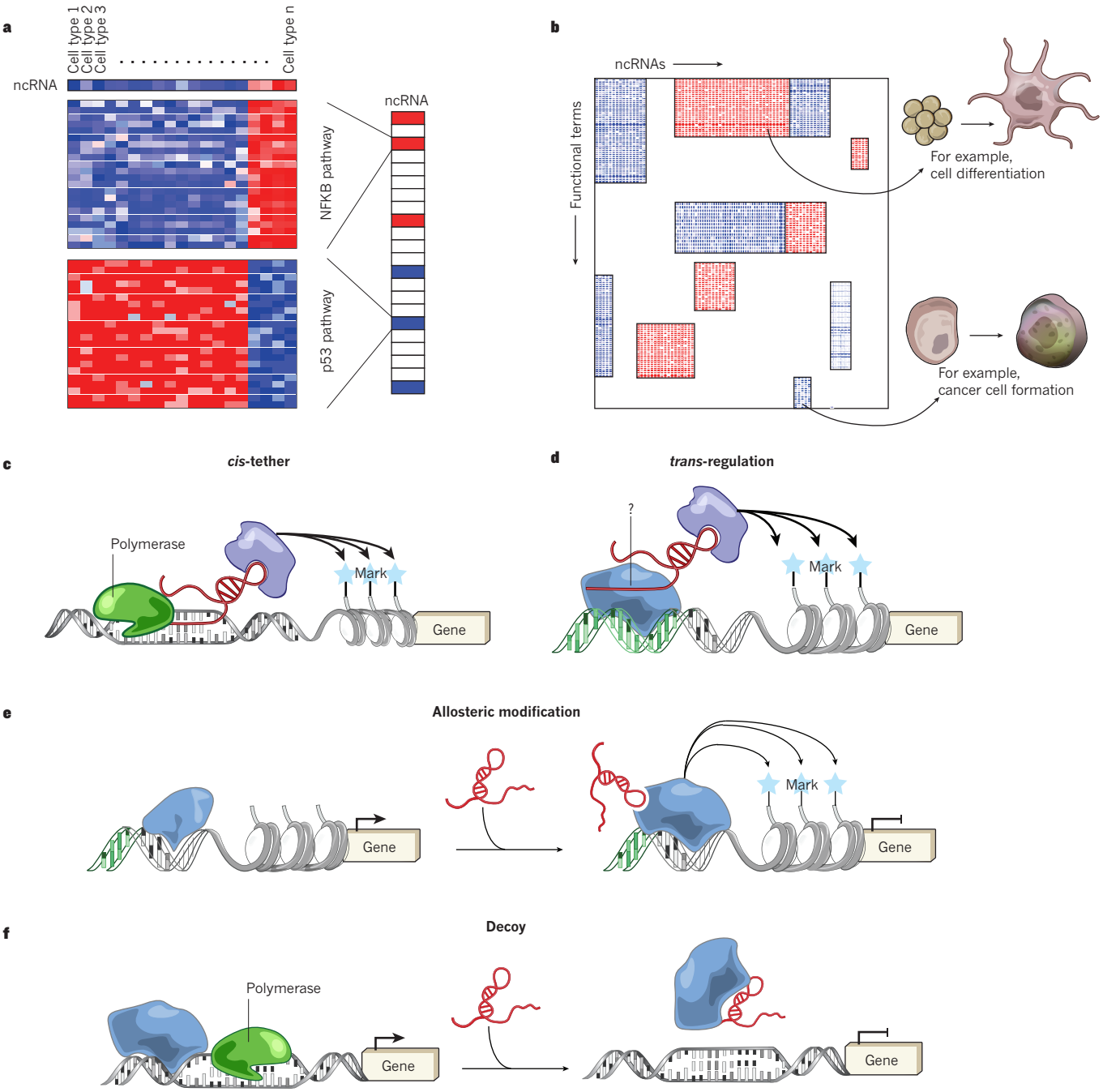
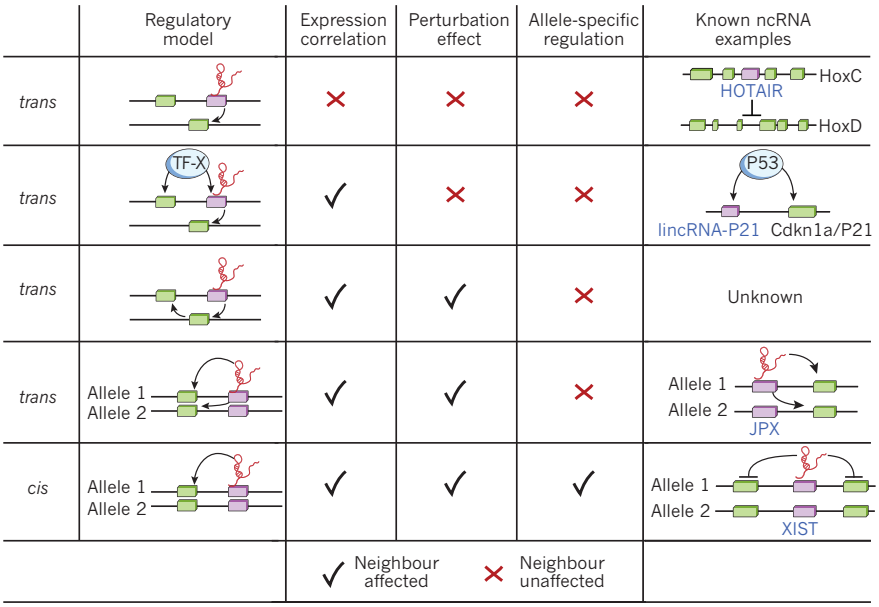


Figure 2 | Classification of ncRNA function. **a**, Illustration of an ncRNA with expression patterns related to the NFκB pathway. Each row represents a gene, and a positive association (red box) is assigned between the ncRNA and the pathway based on the correlation of the genes in the process. Similarly, the ncRNA is assigned negative association (blue box) with the p53 pathway based on anticorrelation with the genes in the process. **b**, The scores for each functional term and ncRNA can be clustered to identify classes of ncRNAs. In this example (adapted, with permission, from ref. 25) each column represents a different ncRNA, and each row represents a different functional term. **c**, A model of ncRNAs that have a *cis*-function by remaining tethered to their site of transcription. In this model, RNA polymerase (green) transcribes an RNA (red), which can associate with regulatory proteins (purple) to affect neighbouring regions, as proposed for XIST[9,71]. **d**, One model for ncRNA *trans*-regulation. In this model an ncRNA can associate with DNA-binding proteins (blue) and regulatory proteins to localize and affect the expression of the targets, as proposed for HOTAIR[64]. **e**, A model for ncRNAs that bind regulatory proteins and change their activity, in this case leading to a change in modification state and expression of the target gene, as proposed for the *CCND1* ncRNAs, which interact with the TLS protein[89]. **f**, A model for ncRNAs that act as 'decoys'. In this model, ncRNAs bind protein complexes and prevent them from binding to their proper regulatory targets, as proposed for GAS5 and PANDA[27].

BOX 1

Distinguishing *cis*- from *trans*-regulation

If an ncRNA is a *cis*-regulator, then several observations will be true: (i) the gene-expression levels of a neighbouring gene will be correlated with the RNA expression across all conditions; (ii) loss-of-function of the RNA would affect expression of a neighbouring gene, and (iii) the ncRNA would affect expression of a neighbouring gene on the same allele that it is expressed from. The absence of any of these criteria supports *trans*-regulation. We illustrate this point using five common regulatory models. The figure shows what would be observed using specific computational and experimental methods for each regulatory model. The boxes with a tick indicate observed effects on neighbouring genes for each method, and boxes with a cross indicate no observed effect on neighbouring genes. Known ncRNA examples of each of these regulatory models are shown to the right of the figure.



(such as, HOTAIR[14]), transcriptional 'ripple effects'[21] and indirect regulation of neighbouring genes (Box 1). Consistent with these explanations, a recent study showed that an increased correlation of expression between ncRNAs and their neighbouring genes is comparable to that observed for protein-coding genes[32].

Recently, loss-of-function experiments have been used to investigate *cis*- versus *trans*-effects of lincRNAs. One study knocked down seven lincRNAs and identified no effects on neighbouring genes but did show an effect on other genes[45]. A second study knocked down 12 lincRNAs, 7 of which had modest effects on some of the genes within a wide genomic neighbourhood[24]. More recently, a systematic study knocked down approximately 150 lincRNAs and identified no effect on the neighbouring genes for about 95% of the lincRNAs, which is similar to that observed for protein-coding genes[23].

Although perturbation experiments can demonstrate that an RNA functions as a *trans*-regulator, evidence for RNA acting as a *cis*-regulator is more difficult to obtain (Box 1). For example, perturbation experiments demonstrated that the ncRNA from *JPX* affects the expression of the neighbouring *XIST* gene, but as a *trans*-regulator[73]. Conclusive proof of *cis*-regulation requires the demonstration that an RNA regulates a neighbouring gene on the same allele (Box 1). So far, few studies have performed this test, and it is unclear what percentage of ncRNAs that are suggested to have a *cis*-function by loss-of-function experiments[24,28] will pass this test. Together, these studies indicate that although some ncRNAs are *cis*-regulators[9,11,74–76], the vast majority, which have been identified and characterized so far, function as *trans*-regulators[14,23,26,45,73,77].

Formation of RNA–protein interactions

The precise mechanism by which ncRNAs function remains poorly understood. However, one emerging theme is the interaction between ncRNAs and protein complexes. The functional importance of many ncRNA–protein interactions for correct transcriptional regulation has been demonstrated[14,23,45,78–81], including several ncRNAs that are required for the correct localization of chromatin proteins to genomic DNA targets[79–83].

The XIST ncRNA is a key example demonstrating that RNA can play a direct role in silencing large genomic regions[81] by physically interacting with the polycomb complex[84], leading to the condensation of chromatin and transcriptional repression of an entire X chromosome[85] (Fig. 2c). Similar to XIST, many ncRNAs have been identified that physically associate with chromatin-regulatory complexes and 'guide' the associated complexes to specific genomic DNA regions, including HOTAIR[14], AIR[86], KCNQ1ot1 (ref. 75) and lincRNA-p21 (ref. 26) (Fig. 2d).

Biochemical evidence has demonstrated that many large ncRNAs interact with chromatin regulators[23,45,87,88]. The precise numbers vary depending on the experimental approach[45,87], but a conservative estimate suggests that at least 30% of lincRNAs associate with at least 1 of 12 distinct chromatin-regulatory complexes, which include readers, writers and erasers of chromatin modifications[23].

Importantly, lincRNAs can provide regulatory specificity to these complexes because the knockdown of these lincRNAs affects a subset of the genes that are normally regulated by these complexes[23,45]. One hypothesis is that ncRNAs provide regulatory specificity by localizing chromatin-regulatory complexes to genomic DNA targets[14,26,28,45,78,86]. Several methods have been developed to generate maps of RNA–DNA proximity[82,83], but it still needs to be determined what percentage of ncRNAs localize to genomic DNA regions and how these interactions occur.

In addition to their role in chromatin regulation, ncRNAs can also modulate the regulatory activity of protein complexes (Fig. 2e). As an example, an ncRNA upstream of cyclin D1 can bind to the TLS (translocation in liposarcoma) RNA-binding protein, which changes it from an inactive to an active state[89]. Similarly, the NRON ncRNA can bind to the NFAT (nuclear factor of activated T cells)-transcription factor rendering it inactive because it prevents nuclear accumulation[13]. ncRNAs can also function as molecular 'decoys' by preventing correct regulation through competitive binding (Fig. 2f). For example, the GAS5 ncRNA binds to the glucocorticoid receptor and prevents the receptor from binding to its correct regulatory elements[90], and the PANDA ncRNA can prevent NF-Y localization, which leads to apoptosis[27]. Similarly, several studies have shown that ncRNAs can function as decoys to other RNA species, such as miRNAs, to control miRNA levels[91,92].

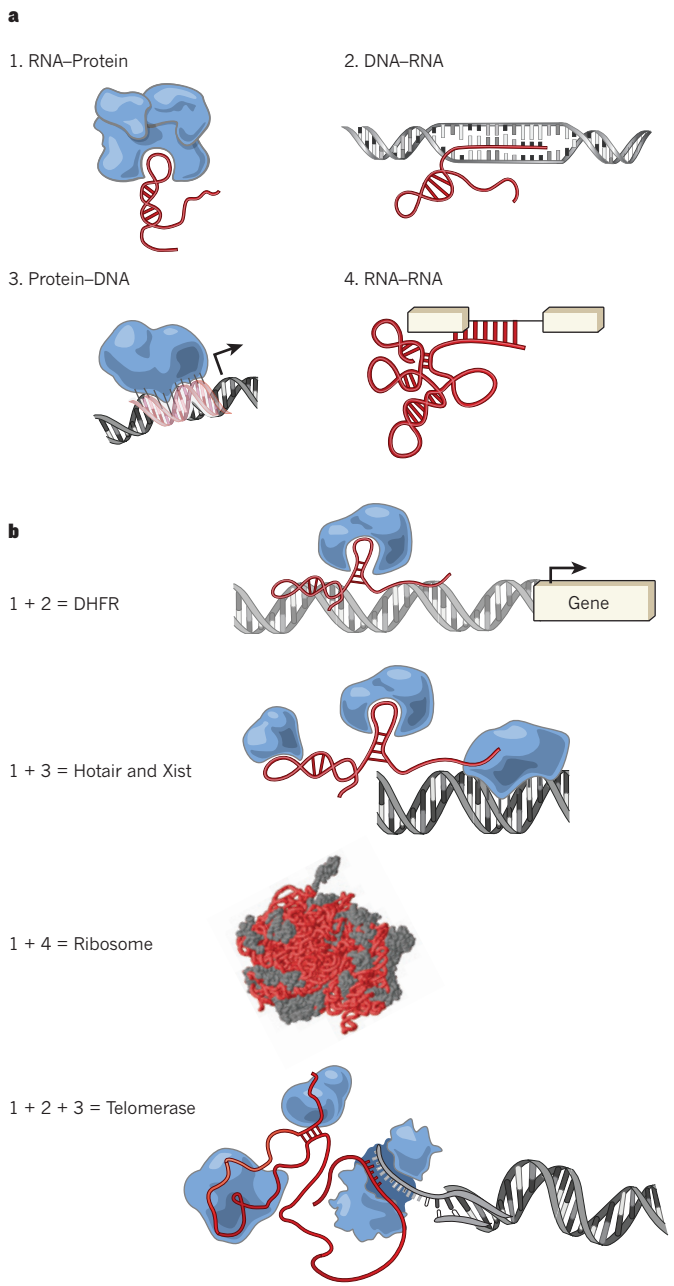


Figure 3 | Modular principles of large ncRNA genes. **a**, The four principles of nucleic acid and protein interactions. (1) RNA–protein interactions, (2) DNA–RNA hybridization-based interactions, (3) DNA–protein interactions and (4) RNA–RNA hybridization based interactions. **b**, Each of these principles can be combined to build distinct complexes. For example, combining RNA–protein and RNA–DNA interactions can localize a protein complex to a specific DNA sequence in an RNA-dependent manner; as has been implicated for the DHFR[99] promoter and localization of DNMT3b[98]. Combining RNA–protein and protein–DNA principles can also localize a diverse set of proteins, which have a molecular scaffold created by RNA, to a specific DNA sequence in a protein-dependent manner. The ribosome is a multifaceted combination of RNA–protein interactions that facilitate correct RNA–RNA interactions for the ribozyme activity of the ribosome. The telomere replication activity of telomerase is an example of combining RNA–protein, RNA–DNA and protein–DNA interactions.

Large ncRNAs as molecular scaffolds of proteins

One emerging theme common to many large ncRNAs is the formation of multiple distinct RNA–protein interactions that are used to carry out their function (Fig. 3). The first indication of this phenomenon came from the discovery of telomerase[93]. Telomerase activity requires a telomerase RNA component (TERC)[94], which serves as a template for telomeric regulation and as a molecular scaffold for the polymerase enzyme around the RNA[95] (Fig. 3b). Importantly, genetic studies demonstrated that TERC plays a modular functional role, as genetically swapping particular domains of TERC retained the overall function[66]. This indicated that TERC was made up of discrete functional modules to bring multiple proteins into the proximity of a protein[66].

More recently, HOTAIR was shown to contain distinct protein-interaction domains that can associate with polycomb repressive complex 2 (PRC2) (ref. 14) and the CoREST–LSD1 complex[64], which together are required for correct function (Fig. 3b). XIST also has discrete functional domains. Through a series of genetic deletions XIST was shown to contain at least two discrete domains that are responsible for silencing (RepA) and localization (RepC)[81] (Fig. 3b). These functional domains could be independently deleted without affecting the role of the other domain, which suggests the modular nature of the *XIST* ncRNA[81]. These functional domains of XIST also interact with discrete proteins; the silencing domain (RepA) binds to PRC2 and the localization domain (RepC) binds to YY1 (ref. 96) and hnRNPU[97]. These examples show that large ncRNAs can function as molecular scaffolds of protein complexes. Importantly, this phenomenon is likely to be a general one because approximately 30% of ES cell lincRNAs associate with multiple regulatory complexes[23].

In addition to interacting with multiple proteins, in several examples, ncRNAs have been shown to interact directly with both DNA and RNA. ncRNAs for example form triplex structures with DNA[98,99] (Fig. 3a) such as a ncRNA that binds to the ribosomal DNA promoter and interacts with the DNMT3b protein to silence expression[98]. Furthermore, RNA can form traditional duplex base-pairing interactions with DNA, a property that has long been speculated for large ncRNAs[71]. Finally, RNA can form base-pair interactions with RNA (Fig. 3a), which are crucial for processes such as tRNA–mRNA anticodon recognition[5], ribonuclease P recognition of pre-tRNAs[5], miRNA targeting[100], ribosome structure as a ribozyme[67] and splicing regulation[6]. Despite these examples, the interactions between large ncRNAs, genomic DNA and other RNAs are not well characterized.

A potential modular RNA code

Collectively, the studies reviewed here suggest an intriguing hypothesis: large ncRNAs are flexible modular scaffolds[23,64,66,81]. In this model, RNA contains discrete domains that interact with specific protein complexes. These RNAs, through a combination of domains, bring specific regulatory components into proximity with each other, which results in the formation of a unique functional complex. These RNA regulatory complexes can include interactions with proteins but can also extend to RNA–DNA and RNA–RNA regulatory interactions.

RNA is well-suited for this role because it is a malleable evolutionary substrate compared with a protein, allowing for the selection of discrete interaction domains[5]. Specifically, RNA can be easily mutated, tested and selected without breaking its core functionality[5]. This model of modular interactions can explain the observation that there are highly conserved 'patches' within large ncRNA genes[25,31,37] that could have evolved for specific protein interactions[26,81,84]. The remaining regions may be more evolutionarily flexible, allowing the formation of new functional domains by random mutation and selection. This is consistent with the observation that non-constrained regions of telomerase are dispensable[66].

The model of RNA as a modular scaffold is not limited to protein interactions. RNA can also base-pair with DNA, which might be used to guide complexes to specific DNA sequences. Alternatively, RNAs might guide complexes by bridging together sets of DNA-binding proteins. Such a model could explain how the same protein complexes are guided to different DNA loci in distinct cell types.

Large ncRNAs can also form RNA–RNA interactions, raising intriguing possibilities for future investigations. For example, two large RNA molecular scaffolds might be linked through RNA–RNA interactions. Another possibility is that RNA–RNA interactions could result in

unique RNA structures that can interact with protein complexes that are not attainable by the individual units. This has been observed in the ribosome, where the combination of RNA–RNA and RNA–protein interactions are required for correct complex formation.

Outlook

We are only beginning to understand the mechanism by which large ncRNAs carry out their regulatory function. A modular RNA regulatory code is an attractive hypothesis but remains to be tested; in particular, the way in which large ncRNAs, and proteins interact, and the underlying molecular principles are still unknown. Understanding these principles will require the identification of the sites of the RNA–protein interactions and the exact RNA-binding proteins *in vivo*. Furthermore, the way in which large ncRNAs localize to their target genes is unknown but could involve direct RNA–DNA interactions (Fig. 3a) or interactions with proteins that contain DNA recognition elements, which has been suggested for XIST[96] and HOTAIR[64]. To gain insight into these processes, it will be important to catalogue the interactions that ncRNAs form with genomic DNA and RNAs. These data will help elucidate the rules that guide these interactions as well as the functional implications of these associations, which can then be tested experimentally.

If large ncRNAs are truly modular, then each individual domain would have a unique function that is independent of other domains. Demonstrating modularity will require the genetic deletion of domains and spacer regions, as well as domain-swapping experiments. Learning these principles would result in a defined 'modular RNA code' for how RNAs can affect cell states. By truly understanding this modular RNA code, it may be possible to create synthetically engineered RNAs that could interact with both nucleic acids and protein modules to carry out engineered regulatory roles. However, at present, it is premature to dismiss the possibility of large ncRNAs having other mechanisms of action that may not fit neatly into this modular RNA code. In the meantime, it is clear that mammalian genomes encode a diverse set of large important ncRNAs. ■

1. Warner, J. R., Soeiro, R., Birnboim, H. C., Girard, M. & Darnell, J. E. Rapidly labeled HeLa cell nuclear RNA. I. Identification by zone sedimentation of a heterogeneous fraction separate from ribosomal precursor RNA. *J. Mol. Biol.* **19,** 349–361 (1966).
2. Salditt-Georgieff, M., Harpold, M. M., Wilson, M. C. & Darnell, J. E., Jr. Large heterogeneous nuclear ribonucleic acid has three times as many 5´ caps as polyadenylic acid segments, and most caps do not enter polyribosomes. *Mol. Cell. Biol.* **1,** 179–187 (1981).
This paper demonstrates an abundant class of RNA species that do not enter polyribosomes.
3. Weinberg, R. A. & Penman, S. Small molecular weight monodisperse nuclear RNA. *J. Mol. Biol.* **38,** 289–304 (1968).
4. Zieve, G. & Penman, S. Small RNA species of the HeLa cell: metabolism and subcellular localization. *Cell* **8,** 19–31 (1976).
5. Gesteland, R. F., Cech, T. & Atkins, J. F. *The RNA World : The Nature of Modern RNA Suggests a Prebiotic RNA World*. 3rd edn (Cold Spring Harbor Laboratory Press, 2006).
6. Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.* **2,** 919–929 (2001).
7. Pachnis, V., Brannan, C. I. & Tilghman, S. M. The structure and expression of a novel gene activated in early mouse embryogenesis. *EMBO J.* **7,** 673–681 (1988).
8. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the *H19* gene may function as an RNA. *Mol. Cell. Biol.* **10,** 28–36 (1990).
This paper was the first report of a large ncRNA showing that the *H19* transcript lacked conserved ORFs and did not make a protein product *in vivo*.
9. Brown, C. J. *et al*. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349,** 38–44 (1991).
10. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for *Xist* in X chromosome inactivation. *Nature* **379,** 131–137 (1996).
11. Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding *Air* RNA is required for silencing autosomal imprinted genes. *Nature* **415,** 810–813 (2002).
12. Young, T. L., Matsuda, T. & Cepko, C. L. The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr. Biol.* **15,** 501–512 (2005).
13. Willingham, A. T. *et al*. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309,** 1570–1573 (2005).
14. Rinn, J. L. *et al*. Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. *Cell* **129,** 1311–1323 (2007).
15. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–1563 (2005).
This paper describes the large-scale cDNA sequencing efforts in the mouse genome and reveals many thousands of non-coding transcripts.
16. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
17. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306,** 2242–2246 (2004).
18. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316,** 1484–1488 (2007).
19. Rinn, J. L. *et al.* The transcriptional activity of human Chromosome 22. *Genes Dev.* **17,** 529–540 (2003).
20. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296,** 916–919 (2002).
21. Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nature Cell Biol.* **10,** 1106–1113 (2008).
22. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Struct. Mol. Biol.* **14,** 103–105 (2007).
23. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477,** 295–300 (2011).
24. Orom, U. A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143,** 46–58 (2010).
25. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (2009).
This paper applied a chromatin signature to identify lincRNAs and used a guilt-by-association approach to classify their likely functions in diverse biological processes.
26. Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142,** 409–419 (2010).
27. Hung, T. *et al*. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nature Genet.* **43,** 621–629 (2011).
28. Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472,** 120–124 (2011).
29. Wilusz, J. E., Freier, S. M. & Spector, D. L. 3´ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **135,** 919–932 (2008).
30. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
31. Guttman, M. *et al. Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnol.* **28,** 503–510 (2010).
32. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25,** 1915–1927 (2011).
33. Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F. & Mattick, J. S. Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl Acad. Sci. USA* **105,** 716–721 (2008).
34. De Santa, F. *et al.* A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biol.* **8,** e1000384 (2010).
35. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
36. Ravasi, T. *et al*. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* **16,** 11–19 (2006).
37. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17,** 556–565 (2007).
38. Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nature Genet.* **41,** 572–578 (2009).
39. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322,** 1849–1851 (2008).
40. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
41. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
42. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858 (2009).
43. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
44. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
45. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* **106,** 11667–11672 (2009).
46. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
47. Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nature Genet.* **42,** 1113–1117 (2010).
48. Dinger, M. E., Pang, K. C., Mercer, T. R. & Mattick, J. S. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* **4,** e1000176 (2008).
49. Brockdorff, N. *et al.* 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 (1992).
50. Lin, M. F., Deoras, A. N., Rasmussen, M. D. & Kellis, M. Performance and scalability of discriminative metrics for comparative gene identification in

12 *Drosophila* genomes. *PLoS Comput. Biol.* **4,** e1000067 (2008).
51. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27,** i275–i282 (2011).
52. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38,** D211–D222 (2010).
53. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802(2011).
54. Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A. & Couso, J. P. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.* **5,** e106 (2007).
This paper demonstrates the existence of functional small peptides within a presumed 'non-coding' transcript through ORF conservation, *in vivo* protein identification and functional analysis.
55. Kondo, T. *et al.* Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* **329,** 336–339 (2010).
56. Jiao, Y. & Meyerowitz, E. M. Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control. *Mol. Syst. Biol.* **6,** 419 (2010).
57. Li, Y. M. *et al.* The *H19* transcript is associated with polysomes and may regulate *IGF2* expression in *trans*. *J. Biol. Chem*. **273,** 28247–28252 (1998).
58. Cai, X. & Cullen, B. R. The imprinted H19 noncoding RNA is a primary microRNA precursor. *RNA* **13,** 313–316 (2007).
59. Yang, L. *et al.* ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. *Cell* **147,** 773–788 (2011).
60. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl Acad. Sci. USA* **104,** 19428–19433 (2007).
61. Kastenmayer, J. P. *et al.* Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res.* **16,** 365–373 (2006).
62. Hanada, K., Zhang, X., Borevitz, J. O., Li, W. H. & Shiu, S. H. A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. *Genome Res* **17,** 632–640 (2007).
63. Mattick, J. S. The genetic signatures of noncoding RNAs. *PLoS Genet.* **5,** e1000459 (2009).
64. Tsai, M. C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329,** 689–693 (2010).
This paper identified multiple protein-interaction domains within HOTAIR that together allowed it to carry out its function, which demonstrated that a large ncRNA can act as a molecular scaffold.
65. Gupta, R. A. *et al.* Long non-coding RNA *HOTAIR* reprograms chromatin state to promote cancer metastasis. *Nature* **464,** 1071–1076 (2010).
66. Zappulla, D. C. & Cech, T. R. Yeast telomerase RNA: a flexible scaffold for protein subunits. *Proc. Natl Acad. Sci. USA* **101,** 10024–10029 (2004).
This paper demonstrated that telomerase RNA can bridge proteins by showing that protein interaction domains can be swapped and spacer regions deleted with minimal impact on the function of the RNA.
67. Korostelev, A. & Noller, H. F. The ribosome in focus: new structures bring new insights. *Trends Biochem. Sci.* **32,** 434–441 (2007).
68. Ivanova, N. *et al*. Dissecting self-renewal in stem cells with RNA interference. *Nature* **442,** 533–538 (2006).
69. Martens, J. A., Laprade, L. & Winston, F. Intergenic transcription is required to repress the *Saccharomyces cerevisiae SER3* gene. *Nature* **429,** 571–574 (2004).
70. Schmitt, S., Prestel, M. & Paro, R. Intergenic transcription through a Polycomb group response element counteracts silencing. *Genes Dev.* **19,** 697–708 (2005).
71. Lee, J. T. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev.* **23,** 1831–1842 (2009).
72. Ponjavic, J., Oliver, P. L., Lunter, G. & Ponting, C. P. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet.* **5,** e1000617 (2009).
73. Tian, D., Sun, S. & Lee, J. T. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* **143,** 390–403 (2010).
74. Koerner, M. V., Pauler, F. M., Huang, R. & Barlow, D. P. The function of non-coding RNAs in genomic imprinting. *Development* **136,** 1771–1783 (2009).
75. Pandey, R. R. *et al. Kcnq1ot1* antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell* **32,** 232–246 (2008).
76. Bertani, S., Sauer, S., Bolotin, E. & Sauer, F. The noncoding RNA Mistral activates *Hoxa6* and *Hoxa7* expression and stem cell differentiation by recruiting MLL1 to chromatin. *Mol. Cell* **43,** 1040–1046 (2011).
77. Feng, J. *et al.* The *Evf-2* noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.* **20,** 1470–1484 (2006).
78. Koziol, M. J. & Rinn, J. L. RNA traffic control of chromatin complexes. *Curr. Opin. Genet. Dev.* **20,** 142–148 (2010).
79. Maison, C. *et al.* Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nature Genet.* **30,** 329–334 (2002).
80. Bernstein, E. *et al.* Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Mol. Cell. Biol.* **26,** 2560–2569 (2006).
81. Wutz, A., Rasmussen, T. P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of *Xist* RNA. *Nature Genet.* **30,** 167–174 (2002).
This paper reported the generation of deletion mutants across the *Xist* locus and identified the discrete domains responsible for the silencing and localization roles of the RNA.
82. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA–chromatin interactions. *Mol. Cell* **44,** 667–678 (2011).
83. Simon, M. D. *et al.* The genomic binding-sites of a non-coding RNA. *Proc. Natl Acad. Sci. USA* **108,** 20497–20502 (2011).
84. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322,** 750–756 (2008).
85. Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A. & Panning, B. *Xist* RNA and the mechanism of X chromosome inactivation. *Annu. Rev. Genet.* **36,** 233–278 (2002).
86. Nagano, T. *et al.* The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322,** 1717–1720 (2008).
87. Zhao, J. *et al.* Genome-wide identification of Polycomb-associated RNAs by RIP-seq. *Mol. Cell* **40,** 939–953 (2010).
88. Kaneko, S. *et al.* Phosphorylation of the *PRC2* component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA. *Genes Dev.* **24,** 2615–2620 (2010).
89. Wang, X. *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in *cis* to inhibit transcription. *Nature* **454,** 126–130 (2008).
90. Kino, T., Hurt, D. E., Ichijo, T., Nader, N. & Chrousos, G. P. Noncoding RNA Gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci. Signal* **3,** ra8 (2010).
91. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta stone of a hidden RNA language? *Cell* **146,** 353–358 (2011).
92. Cesana, M. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147,** 358–369 (2011).
93. Greider, C. W. & Blackburn, E. H. Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. *Cell* **43,** 405–413 (1985).
94. Feng, J. *et al.* The RNA component of human telomerase. *Science* **269,** 1236–1241 (1995).
95. Lingner, J. *et al.* Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* **276,** 561–567 (1997).
96. Jeon, Y. & Lee, J. T. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell* **146,** 119–133 (2011).
97. Hasegawa, Y., Brockdorff, N., Kawano, S., Tsutui, K. & Nakagawa, S. The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev. Cell* **19,** 469–476 (2010).
98. Schmitz, K. M., Mayer, C., Postepska, A. & Grummt, I. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev.* **24,** 2264–2269 (2010).
99. Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445,** 666–670 (2007).
100. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136,** 215–233 (2009).

Acknowledgements We thank M. Cabili, J. Engreitz, M. Garber, P. McDonel and A. Pauli for their reading and suggestions; T. Cech for comments and suggestions; E. Lander for helpful discussions and ideas; and S. Knemeyer and L. Gaffney for assistance with figures in this Review.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to M.G. (mguttman@mit.edu) and J.L.R. (john_rinn@harvard.edu).