ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review

Chris P. Ponting[1] and Wilfried Haerty[2]

[1]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom; email: chris.ponting@ed.ac.uk

[2]Earlham Institute, Norwich, United Kingdom; email: wilfried.haerty@earlham.ac.uk

Annu. Rev. Genom. Hum. Genet. 2022. 23:153–72

First published as a Review in Advance on April 8, 2022

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-112921-123710

Keywords

molecular mechanism, evolutionary constraint, transcriptional noise, RNA structure, knockout phenotype, subcellular localization

Abstract

Do long noncoding RNAs (lncRNAs) contribute little or substantively to human biology? To address how lncRNA loci and their transcripts, structures, interactions, and functions contribute to human traits and disease, we adopt a genome-wide perspective. We intend to provoke alternative interpretation of questionable evidence and thorough inquiry into unsubstantiated claims. We discuss pitfalls of lncRNA experimental and computational methods as well as opposing interpretations of their results. The majority of evidence, we argue, indicates that most lncRNA transcript models reflect transcriptional noise or provide minor regulatory roles, leaving relatively few human lncRNAs that contribute centrally to human development, physiology, or behavior. These important few tend to be spliced and better conserved but lack a simple syntax relating sequence to structure and mechanism, and so resist simple categorization. This genome-wide view should help investigators prioritize individual lncRNAs based on their likely contribution to human biology.

INTRODUCTION

Long noncoding RNAs (lncRNAs): mature transcripts that are at least 200 nucleotides in length and have reduced protein-coding capacity

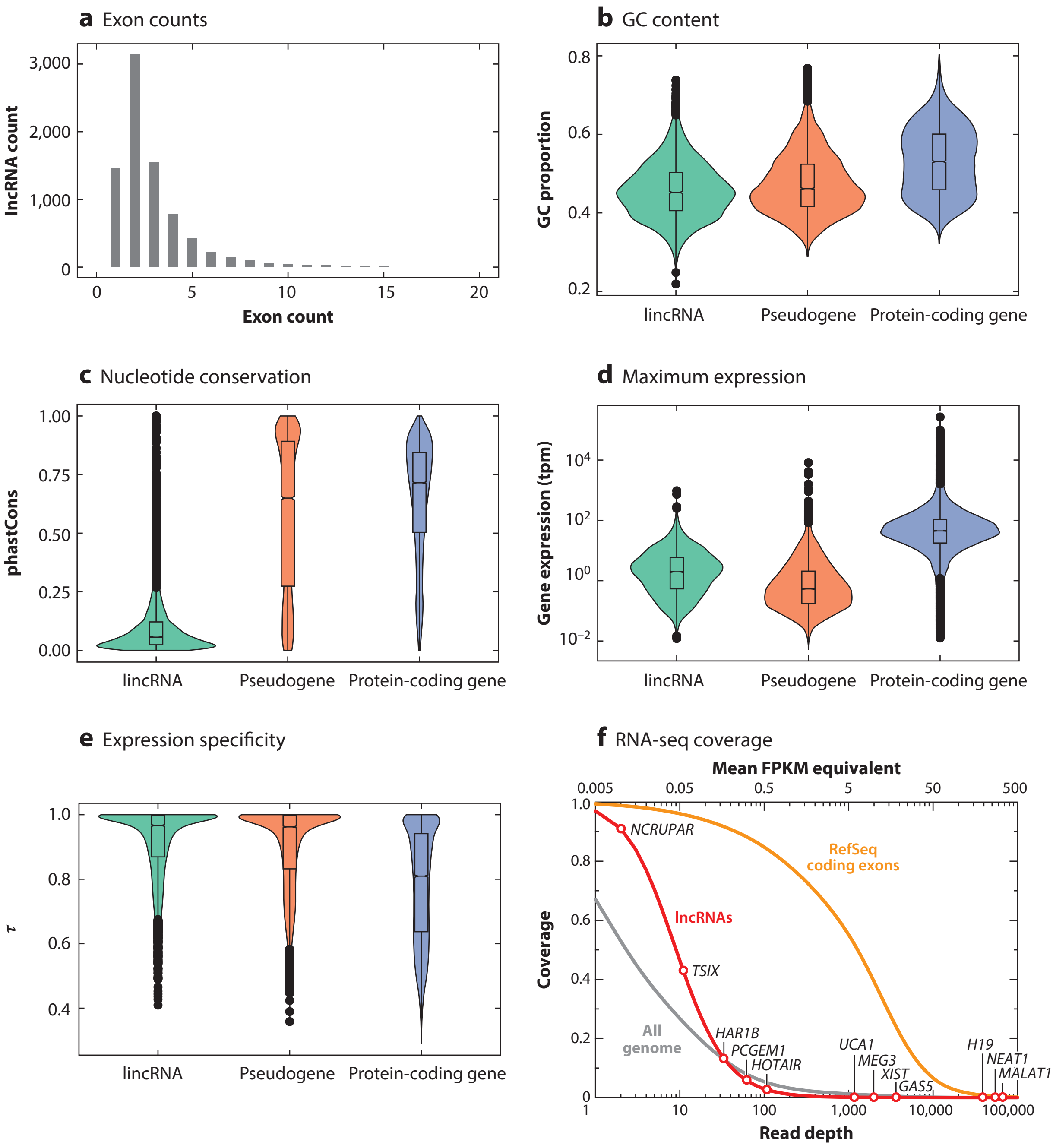
Imagine that your task today is to investigate a single human long noncoding RNA (lncRNA). This locus is unexceptional, even average, in all respects: Its sequence contains on average nearly three exons (GENCODE version 38), but there is otherwise no specific feature that illuminates whether it is functional or what its molecular mechanisms might be. In the few tissues in which this lncRNA has been identified, it is expressed at low levels, and it is most often absent from other mammals, including closely related species (8, 10, 22) (**Figure 1**). If, instead, your task had been to investigate the average protein-coding gene, then this would have been considerably easier, because you would be drawing upon a wealth of highly curated annotations and extensive experimental observations, using sequence features that accurately predict its molecular mechanism and expression data across many tissues, individuals, and species. The absence of such information for lncRNAs leaves you adrift, without a specific hypothesis to investigate. Unfortunately, you cannot make use of model organisms such as mouse because this lncRNA is absent from it, so you decide to adopt the trusted methods of reverse genetics and disrupt this lncRNA in human cells (31). Yet lack of annotation and a mechanistic understanding for this locus causes you to be uncertain what strategy to apply—whether to delete the entire locus or a small portion of it, or to ablate its transcription entirely. Even then, questions remain regarding what cellular phenotypes you should measure, how much these might change, and how you should interpret such observations (see the sidebar titled DNA-Versus RNA-Mediated Function).

This is the average experience of a lncRNA biologist. Our aim in this review is to draw conclusions from genome-scale investigations of lncRNAs in the hope that they help biologists either to improve their experimental designs or to choose a different locus to study, one that is more likely to yield robust experimental observations. Other perspectives, which dwell more on molecular mechanisms and functions of individual lncRNAs, have been reviewed extensively elsewhere (88, 99).

Taking a gene-centric perspective on lncRNAs raises the problem that a lesson learned from one locus is rarely relevant to others. Our deep functional understanding of *Xist*—the master regulator of X chromosome inactivation (reviewed in 88)—for example, has not aided investigation of tens of thousands of annotated lncRNAs (**Table 1**). Whenever researchers propose a lncRNA's mechanism, or its involvement in a pathology such as cancer, almost inevitably they herald this as revealing a new paradigm, one that possibly explains the mode of action of many other lncRNAs. Hundreds of publications state that lncRNAs are emerging as important regulators, elements, or components, and 30% of published reviews on lncRNAs since 2012 employed the term emerging. The implication is that lncRNAs are now being revealed almost as bright-colored butterflies, rather than plain-colored chrysalises. Nevertheless, very few lncRNAs have high-quality evidence for such colorful claims. Instead, low-quality evidence abounds, in part because the lncRNA literature has been contaminated by hundreds of paper-mill publications (106) but also because molecular and cellular observations—such as RNA–molecule interactions and gene expression changes—are often deemed important without sufficient evidence.

As well as acclaiming hard-won advances in human lncRNA biology, it is critical that we recognize the field's substantial knowledge gaps. The ubiquity of lncRNAs within and across eukaryotic species has led some to describe lncRNAs as major actors that contribute substantially to most cellular processes and whose RNA sequence variation will ultimately be recognized as greatly altering human traits and disease susceptibility (70). Faced with the same evidence, others view the vast majority of lncRNAs as nonfunctional, spurious by-products of transcription (82). The truth lies across these two extremes: Some transcripts will lack RNA sequence–dependent function, whereas others will harbor variants that predispose individuals to disease.

The existing review literature has focused mostly on experimental and computational methods that elucidate function for individual lncRNAs. This review focuses on approaches applied across the human genome or lncRNA transcriptome, yet it will also have broader relevance for other animals' lncRNAs. Our purpose was to traverse this spectrum of opinion and ultimately rest



(Caption appears on following page)

Figure 1 (*Figure appears on preceding page*)

Genomic and transcriptomic characteristics of lncRNAs in human. (*a*) Distribution of exon counts for lincRNAs. (*b*–*e*) Comparisons of exonic GC content (panel *b*), nucleotide conservation across vertebrates (phastCons, University of California, Santa Cruz) (panel *c*), maximum expression across GTEx tissues (36) (panel *d*), and expression specificity (τ, where 0 indicates broad expression and 1 indicates tissue-specific expression) (panel *e*) for lincRNAs, pseudogenes, and protein-coding genes. (*f*) RNA-seq coverage across lncRNAs and RefSeq models. GTEx RNA-seq experiments were carried out in a nonstranded manner, leading to imprecise expression estimates for overlapping genes. Transcribed pseudogenes are a class of noncoding RNAs whose evolutionary origin is more recent than those of other noncoding RNAs. Abbreviations: FPKM, fragments per kilobase of transcript per million mapped reads; GTEx, Genotype-Tissue Expression; lincRNA, long intergenic noncoding RNA; lncRNA, long noncoding RNA; RefSeq, Reference Sequence; RNA-seq, RNA sequencing; TPM, transcripts per million. Panel *f* adapted with permission from Reference 25; annotations for lncRNAs, pseudogenes, and protein-coding genes were extracted from GENCODE version 38 (29).

where accumulated evidence provides greatest support. We hope that this assists researchers in their design of experiments that definitively test the molecular mechanism of selected lncRNA loci. We take a precautionary approach, covering how some computational and experimental observations that are interpretable as indicating lncRNA functionality have alternative and often more mundane explanations.

NUMBER AND SEQUENCE

The lncRNA biologist has any number of lncRNAs to investigate. GENCODE version 38 lists 17,944 lncRNA loci, whereas other catalogs contain vastly more, up to 270,044 lncRNA transcripts (65). These numbers have been compared favorably with the stable number (approximately 20,000) of human protein-coding genes. To elevate the importance of lncRNA loci even further, some researchers claim that "the large majority of the human genome is transcribed into non-protein-coding RNAs," whereas "only ~1.2% of the human genome encodes for protein-coding genes" (e.g., 49, p. 1063). The truth, however, is less impressive: Human lncRNA exons span at most 2.3% of the human genome (82), and most intergenic RNA arises from transcription that is initiated within protein-coding genes (1). Moreover, most lncRNAs are expressed at low levels (61, 87) (**Figure 1**). These low levels mean that even if, very optimistically, the number of lncRNA loci is 10-fold greater than the number of protein-coding genes, their molecular output is considerably smaller. A claim that "around 98% of all transcriptional output in humans is non-coding RNA" (69, p. 986) is plausible only when this includes intronic nucleotides of protein-coding gene transcripts.

The size and complexity of the human noncoding transcriptome have been proposed to explain human evolution, development, and cognition (15, 71). This anthropocentric argument is undermined by observations that there are many species whose genomes, and thus transcriptomes,

DNA- VERSUS RNA-MEDIATED FUNCTION

Observation of a phenotype resulting from the ablation or disruption of a lncRNA locus does not necessarily provide information on its mechanism of action. Such an observation can prosaically be the consequence of deleting functional DNA elements overlapping the annotated locus (including promoter and enhancer regions) and other conserved noncoding sequences. It is often helpful to adopt the null hypothesis that phenotypic effects arising from disrupting a lncRNA locus result from DNA-dependent rather than RNA-dependent function. Experiments whose results might lead to rejection of this hypothesis include attempted rescue of phenotypes following reintroduction of the lncRNA transcript or measurement of phenotypes resulting from this transcript's knockdown in a sequence-specific manner.

Table 1 Logical fallacies present in the literature on long noncoding RNAs (lncRNAs)

Fallacy	Claim and counterclaim
The lonely fact	Claim: *Xist* is functional, so all lncRNAs are functional. Alternative: The fraction of all lncRNAs that are functional remains unknown.
Missing the point	Claim: Deletion of a large lncRNA locus results in mouse phenotypes, so its lncRNA must have RNA-dependent function. Alternative: DNA functional elements irrelevant to the lncRNA may have been removed.
False cause	Claim: Adjacent lncRNA and protein-coding genes are transcribed together, so the protein-coding gene is regulated by the lncRNA. Alternative: Correlation does not imply causation.
Slippery slope	Claim: lncRNA *X* binds protein *Y* that regulates gene *Z*. So lncRNA *X* must modulate *Z*'s expression and its effects on cells and organisms. Alternative: The effect of *X* on *Y* need not influence its effect on *Z*.
Weak analogy	Claim: lncRNAs and mRNAs have similar sequence features, so they must share similar mechanisms. Alternative: These features are irrelevant to mechanism.
Appeal to authority	Claim: Hundreds of publications state that lncRNAs are emerging as important regulators, elements, or components, so this must be true. Alternative: Generalities could be untrue, and irrelevant to a specific lncRNA.
Appeal to ignorance	Claim: There is no conclusive evidence that lncRNAs are nonfunctional, so you should concede that all of them could be functional. Alternative: Absence of evidence is not evidence of presence.
False dichotomy	Claim: Either all lncRNAs are functional or none of them are. Alternative: Some are functional; others are not.
Circular reasoning	Claim: lncRNA *Z* is functional. *Z*'s function depends on its transcription. Alternative: Without functional evidence, *Z*'s mechanism remains unknown.
Fallacy of sunk costs	Claim: "I've invested so much in this lncRNA. It's got to have a function!" Alternative: Without evidence to the contrary, any lncRNA should be considered to lack function.
Ambiguity	Claim: Virtually all transcriptional output is noncoding. Alternative: Most transcribed sequence is intronic; very little is of lncRNA exons.
Bandwagon fallacy	Claim: It is generally accepted that lncRNAs interact directly with gene promoters. Alternative: Without robust evidence, such statements have no predictive value.

are more extensive than humans' (81). Furthermore, the human transcriptome currently appears to be more complex than other species' only because it has been more extensively sequenced.

An additional misapprehension is that lncRNAs are biased toward containing two exons (22), whereas the median exon count is only one when transcriptomes are assembled from short reads. Human lncRNAs are not completely devoid of informative sequence features, however. Short sequence motifs (*k*-mers) within a lncRNA show modest power to predict its subcellular localization and protein-binding capability (37, 55). Longer sequence patterns are generally attributable to transposable elements (TEs), which account for approximately 30% of lncRNA sequence, the majority of *Xist* exons (26), and approximately half of the human genome overall (53). TE sequence has been proposed to contribute RNA domains that are essential for lncRNA function (48). Support for this proposal comes from transcripts of one TE, human endogenous retrovirus subfamily H, being required for human embryonic stem cell identity (48) and from a fragment of another TE enhancing lncRNA localization to the nucleus (63). Further support was proposed from enrichments of particular TE subtypes in exons versus introns (12), but these results were not adjusted to account for multiple tests. It has also been pointed out that purifying selection of TE insertions would be more consistent with a depletion of TEs than with their enrichment (52). Finally, there is no distinction between the evolution of TE sequence within lncRNA loci

Transposable element (TE): a high-copy-number DNA sequence, such as short or long interspersed nuclear elements, that arose by transposition and the evolution in sequence adjacent to them (84), which again signifies that most TE insertions are nearly neutral with respect to selective pressure and are not strong predictors of lncRNA functionality. It is difficult to disagree with others' view that "functionality should not be lightly attributed to biochemical activities on the genome, including transposable elements, without proper experimental evidence" (21, p. 1248).

lncRNA annotations are not always perfect. A relatively small number (∼100) of lncRNA annotations are misclassifications, being instead protein-coding genes that contain functional small open reading frames (2, 13, 66, 68, 79, 80). Such annotations are continually being corrected, and hence many fewer lncRNA misclassifications are expected to remain in current databases.

EXPRESSION

The properties of lncRNAs are generally less pronounced than those of protein-coding mRNAs. Their transcripts tend to be shorter (22), and their promoters are weaker (73) and contain fewer complex transcription factor motifs (73). Cotranscriptional splicing is less efficient (97), and transcription often terminates prematurely (91). Their transcripts tend to be less stable (16), and their abundance is more often tissue and cell type specific (8, 61). Overall, these features result in a level of expression that is typically 10-fold lower than mRNAs' (8, 22, 61, 87). These general trends, however, vary widely and cannot reliably predict an individual lncRNA's molecular mechanism (77), with one exception: Transcripts that are rarely transcribed and quickly degraded are less likely to possess RNA sequence–dependent function (91).

These insights have been gleaned by measuring lncRNA expression in bulk samples containing many cells. To observe lncRNA expression from individual promoters, it was necessary to measure allele-specific transcription in single cells (50). Doing so revealed that not only is a lncRNA's expression level lower than mRNAs', but its variability among cells is higher (50). Measuring transcription dynamics showed that lncRNAs have a burst frequency that is fourfold lower than that of mRNAs', and a burst size that is twofold lower (50). For approximately one-third of lncRNA loci, an allele failed to produce a transcriptional burst over a 24-hour period.

Expression of a lncRNA does not guarantee that its RNA sequence is functional. Many regions of the human genome are transcribed yet are rapidly degraded by the RNA exosome (92). RNA polymerase II needs to have low DNA sequence specificity to transcribe many genes from diverse promoters. Transcription thus often initiates from nucleosome-depleted regions before terminating prematurely at cryptic polyadenylation sites, yielding unstable RNA by-products. Such transient RNAs can contribute RNA-sequencing reads at a sufficient abundance to surpass arbitrarily set thresholds. For example, one study predicted 53,864 human intergenic lncRNAs, each expressed at approximately one copy per cell or higher in at least one of 127 data sets (44). Although these lncRNAs are quite modestly enriched in functional features (such as histone modifications and evolutionary conservation; see below), a large fraction could represent rarely expressed and unstable RNAs. The inclusion of rare unstable RNAs in lncRNA sets inevitably overestimates the number of stable lncRNA loci. As studies increasingly investigate diverse tissues and cells, in particular by using deeper RNA-sequencing coverage, the overall tally of proposed human lncRNA loci will inevitably rise yet further.

One strategy to separate high- from low-confidence lncRNAs exploits the principle of replication, a cornerstone of the scientific method. lncRNAs with the highest confidence are those observed as expressed in multiple different samples. The authors of one study, for example, proposed that only 25% of lncRNA loci expressed in granulocytes are robust, on the basis that only these showed expression across all 21 granulocyte samples acquired at three time points from seven donors (56). They further concluded that lncRNAs display significantly greater interindividual expression variability compared with mRNAs.

Even though low-confidence lncRNA transcripts arise from rare transcriptional events, they may still be products of interesting cellular processes. DNA damage repair, for example, is facilitated by recruitment of repair factors by RNA to the damaged site (3). The origin of this RNA is debated, but it could result from transcription of the site of damage just prior to, or soon after, the DNA double-strand break. Similarly, opportunities for rapid RNA polymerase II–mediated transcription at sites of open chromatin will arise as chromatin compartments, loops, and domains are more slowly lost in mitosis and re-formed in the G1 phase.

RNA STRUCTURE

It is sometimes claimed that lncRNAs commonly fold into thermodynamically stable tertiary structures. If so, then these might represent functional domains, akin to the structural and functional units of proteins. Indeed, in light of lncRNAs' very modest primary sequence conservation, some suggest that secondary and tertiary structure conservation instead is critical for their function (86). It has not been possible to prove or disprove this conjecture, because although high-resolution structures can be determined for short (<30 bases) sequence segments or some longer sequences whose structures are stabilized by their association with RNA-binding proteins, doing so for full-length lncRNAs is not currently technically feasible. This lack of structural data results from the fact that lncRNAs—in common with other RNAs—do not adopt a single conformation in isolation (30). Rather, each lncRNA samples from a very large number of conformations, ranging from fully unfolded states to more compact structures. Structures that form more rapidly are more common, whereas slowly forming folds are rarer. Rather than adopting a single structure, therefore, lncRNAs form an ensemble of structures, defined as the population-weighted distribution of all their conformations (30). As a lncRNA encounters other molecules, its ensemble shifts its distribution, altering the time-averaged accessibility of binding sites.

Lacking experimental tertiary structure data, studies began to predict RNA secondary structure content using sequence information only. One such study predicted 35,985 structured RNA elements across the human genome, with an expected false-positive rate of 19.2% (102); another predicted more than 4 million conserved structures at a false-discovery rate of 5–22% (95). Two issues, however, cast doubt on such studies' conclusions. The first is the high proportion (55% in 102) of predicted RNA structures falling outside of transcribed sequence. The second is uncertainty in the reliability of these studies' false-discovery rates (25).

When lncRNA secondary structures form, they are likely stabilized by incorporation within larger molecular complexes. Many sequencing-based methods have been developed to probe RNA interactions and structure (reviewed in 101). These either use small molecules to modify solvent-accessible bases or particular base pairs or use cross-linking and proximity ligation to infer intra- and intermolecular RNA–RNA interactions. As with all methods exploiting high-throughput sequencing as a last step, rare RNA species are underrepresented among observations. RNA in situ conformation sequencing (RIC-seq), for example, predicts 10-fold fewer ncRNA–ncRNA interactions than mRNA–mRNA interactions, and only 5% of 642 hub RNAs (those with relatively high fractions of RNA–RNA interactions) originate from lncRNAs or pseudogenes (9).

Even so, these methods are increasingly being used to predict intramolecular interactions and secondary structures for human lncRNAs such as *Xist*, *HOTAIR*, and *SRA* (101). Such predictions may suffer from unknown technical biases. Also, their predicted structures, even if they occur in vivo, may not confer functionality. Rivas et al. (89) recently investigated whether pairwise covariation in multiple sequence alignments, a reliable indicator of RNA secondary structure, supports the evolutionary validity of these predicted structures. They expected interacting bases within a functional RNA structure to have accumulated compensatory base-pair substitutions over long evolutionary time. They found no evidence for such paired substitutions within proposed structures in *Xist*, *HOTAIR*, and *SRA* lncRNAs, despite their method and data having sufficient power to do so. They cautioned that the "lack of covariation signal in high-power RNA sequence alignments for these lncRNAs suggests that whatever structure they adopt is not detectably constraining their evolution, and thus may not be relevant for their function" (89, p. 3074). Experimentally defined RNA structures, therefore, should not be considered conclusive until experimental evidence of their functional validity is available.

Subcellular Localization

To prioritize mechanistic hypotheses for a specific lncRNA, we soon wish to know its subcellular localization. lncRNAs located only in the chromatin fraction may regulate gene transcription or be by-products of transcriptional noise, whereas cytoplasmic lncRNAs are more likely to act posttranscriptionally (11, 99). Unfortunately, large-scale studies of lncRNA subcellular localization (e.g., 7, 22, 77) have not always agreed on relative nuclear versus cytoplasmic localization, perhaps because of contamination across subcellular fractions or the absence of the nuclear envelope during mitosis. Furthermore, recently developed methods that localize RNAs at subcellular resolution and at a transcriptome scale have been informative of only the most highly abundant lncRNAs, such as *MALAT1* (14, 27, 104).

For larger numbers of lncRNAs, help is at hand from APEX-RIP. This method, which combines engineered ascorbate peroxidase (APEX)–catalyzed proximity biotinylation of endogenous proteins with RNA immunoprecipitation (RIP), identified 81 and 618 intergenic lncRNAs as being enriched in the cytosol and nucleus, respectively, of HEK293T cells (51). In addition, 11 and 28 intergenic lncRNAs were associated with the nuclear lamina and endoplasmic reticulum, respectively.

Large numbers of lncRNAs reside in the nucleus, among which will be by-products of transcriptional noise (see the sidebar titled Transcriptional Noise), newly transcribed RNAs awaiting export from the nucleus, and RNAs regulating transcriptional bursts from proximal protein-coding genes (50). CoT-1 RNAs are surprisingly abundant RNAs that are highly enriched in TE sequences, mostly LINE-1 elements (17). In interphase, these single-stranded RNAs remain tightly associated with their parental chromosome of origin specifically within euchromatin, but not heterochromatin, and appear to promote more open chromatin packaging (42). CoT-1 RNAs represent only one type of a larger class of chromatin-associated RNAs (caRNAs) proposed to form an RNA mesh that helps to assemble large-scale chromatin structure and to regulate

TRANSCRIPTIONAL NOISE

The cellular transcriptional machinery does not perfectly discriminate cryptic promoters from functional gene promoters. This machinery is abundant and so can engage sites momentarily depleted of nucleosomes and rapidly initiate transcription. The chance occurrence of splice sites can then facilitate the capping, splicing, and polyadenylation of long transcripts. A very large number of such rare RNA species are detectable in RNA-sequencing experiments whose properties are virtually indistinguishable from those of bona fide lncRNAs. Consequently, "a sensible [null] 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" (99, p. 26).

chromosome function (76). These caRNAs are derived mostly from pre-mRNAs rather than lncRNAs. To explain caRNA function, it is tempting to invoke the known ability of *Xist* to spread in *cis* from its site of synthesis. However, an individual caRNA's chromosomal location and its amount are unlikely to predict its function there (43). An alternative proposal of caRNA function is that "thousands of transcriptional events that simultaneously occur in each cell" (74, p. 662) may organize a cell's nuclear architecture. Observations, however, argue against lncRNAs having such a role, including the rarity of lncRNA transcriptional bursts (50), their very low abundance, and the lack of evidence that lncRNAs are transcribed coordinately.

Enhancer: a short stretch of DNA that acts to increase transcription of one or more nearby genes; it is often transcribed into an enhancer RNA

Active enhancers are often transcribed, yielding mostly short-lived RNA species that are short or long and poly- and/or unpolyadenylated (20, 54) and thus, in part, can be defined as lncRNAs (58). Some of these RNAs will be inconsequential, resulting from RNA polymerase II–mediated transcription from regions of transiently open chromatin. As reviewed elsewhere (58), however, enhancer activity can be mediated by the resultant enhancer RNAs. Individual enhancer RNAs have been proposed to promote looping between enhancer and promoter, to bind and regulate transcription factors and coregulators, to promote histone acetylation, and to facilitate transcription elongation (58). Independent confirmation of these observations, however, is often lacking, which limits their generalizability and confidence that they are correct. Furthermore, these different mechanisms are not predictable a priori from, for example, sequence- or chromatin-based signatures. General principles of enhancer RNA mechanisms might be revealed by investigating the molecular consequences of guiding large numbers of these RNAs to their cognate enhancers and/or promoters.

Histone Modification

Enhancer RNAs tend to be short (<150 nucleotides), rapidly turned over by the RNA exosome complex, and capped but not polyadenylated or spliced (90). Their transcribed loci, however, can also yield longer transcripts [>200 nucleotides, i.e., enhancer lncRNAs (elncRNAs)] that are polyadenylated, spliced, and more stable. Those elncRNAs with longer half-lives have greater opportunity to have RNA-dependent function, and most will enact this function locally, in *cis*, rather than in *trans*. *Trans*-acting lncRNAs will need to have even greater stability, and thus few will be transcribed from enhancer regions.

RNA stability, function, and subcellular localization are poorly predicted by sequence features. Instead, Marques et al. (67) defined two lncRNA classes by their relative levels of histone H3K4 mono- and trimethylation at transcriptional initiation regions. Those with higher levels of monomethylation (H3K4me1), a canonical marker of enhancer regions, were classified as elncRNAs; those with higher levels of trimethylation (H3K4me3), a canonical marker of promoters, were classified as promoter lncRNAs (plncRNAs) (33, 67). The two lncRNA subtypes are indistinguishable with respect to their length, number of exons, and transcriptional orientation relative to their closest neighboring gene (67). Distinguishing elncRNAs from plncRNAs based on chromatin marks is necessarily specific to each tissue or cell type, yet is relatively robust because elncRNAs are infrequently categorized as plncRNAs (or vice versa) in a second tissue or cell type (6, 67).

What separates elncRNAs from plncRNAs is their lower and more tissue-specific expression and a strong depletion of CpG islands at their transcriptional initiation regions (67). Furthermore, altered expression of elncRNAs, but not plncRNAs, correlates with expression levels of neighboring protein-coding genes (6, 67), indicating that the elncRNA locus and/or its RNA enhances this gene's activity. Because elncRNAs tend to lack sequence conservation, however, it is more likely their act of transcription, rather than their RNA transcripts, that mediates enhancer activity in *cis*

(67). By contrast, plncRNAs show modest sequence conservation, implying that some act in *trans*. In summary, elncRNAs and plncRNAs are distinguished by their H3K4me1 and H3K4me3 marks, respectively, at their transcriptional initiation regions and tend to be involved in transcriptional and posttranscriptional regulation, respectively.

EVOLUTION: CONSERVATION AND CONSTRAINT

After the discovery of lncRNAs, some investigators claimed that they lack conservation (83) whereas others saw them as being highly conserved (38). Both could be true, of course, should each lncRNA contain mostly poorly conserved, yet also some richly conserved, sequence. Resolution of this evolutionary question was important: Mutation of conserved lncRNA sequence would be expected to bring functional and phenotypic consequences, including disease; conversely, mutation within nonconserved lncRNA sequence could have no functional or phenotypic effect.

On one side of this argument are lncRNA enthusiasts who propose that all lncRNAs are functional and that evolutionary arguments opposing this view are unreliable. In 2013, Mattick & Dinger (72) wrote, "[N]oncoding RNAs usually show evidence of biological function in different developmental and disease contexts, with, by our estimate, hundreds of validated cases already published and many more en route, which is a big enough subset to draw broader conclusions about the likely functionality of the rest" (p. 2). Arguing against this conclusion, however, are observations that lncRNAs are mostly dispensable for viable vertebrate development (31) (discussed further below).

On the other side of this debate are evolutionary biologists who hold that a century-old theoretical evolutionary framework can be trusted to provide deep insight into molecular structure, function, and disease. With a neutral model of evolution, lncRNAs were estimated to contain only a small fraction (4.1–5.5%) of functional sequence, implying that mutations in the remaining sequence would not alter reproductive fitness (84). Mattick & Dinger (72) responded that this model's notion of selective neutrality was highly questionable. This was despite the model being founded on only one assumption—specifically, that mutations (in this case insertions or deletions) occur randomly within neutrally evolving sequence (64). Rather than assuming selective neutrality within ancient TE sequence, as Mattick & Dinger claimed, the model predicted that more than 99% of such sequence evolved neutrally (64).

These arguments focus on species-level sequence conservation that becomes evident after the removal of many deleterious variants over millions of years since these species' last common ancestor. Functionality is not always conserved among species, however, because lineage-specific biology can emerge and ancestral biology can be lost (see the sidebar titled Rapid Turnover). To infer

RAPID TURNOVER

When compared with lncRNAs from other species, human lncRNAs are unexceptional in their length, exon count, tissue specificity, and expression level (46). Like lncRNAs from other species, human lncRNAs have not evolutionarily persisted over many tens of millions of years (46, 78). They thus arose ("were born") and were lost ("died") rapidly (57)—indeed, faster than any other functional element type (85). The rapid evolutionary turnover of sequence and transcription leaves few transcribed lncRNA loci in positionally equivalent (i.e., orthologous) genomic locations (46). Rapid evolution of lncRNA loci is consistent with a small contribution to reproductive fitness and thus with absent or relatively minor organismal functions. Thus, the wider the phyletic range of a human lncRNA is (i.e., its evolutionary spread across divergent animal species), the greater the likelihood is that it plays a greater role in human biology.

the functionality of sequence lacking between-species conservation requires a complementary approach using within-species (i.e., population) data. This approach's signature of functionality is the shift to low population frequency of newly emergent alleles. This shift indicates evolutionary constraint and a tendency for deleterious variants to be purged from this population.

One of two polar opposite outcomes was expected from applying this constraint approach to the human population. In one, lncRNA sequence would be highly constrained even if it was poorly conserved in other species, indicative of important human-specific functions; in the other, human lncRNA sequence would be poorly constrained, consistent with its weak conservation over longer evolutionary intervals. Population data provided compelling evidence for this second outcome—specifically, that newly arising mutations in human lncRNAs are seldom deleterious (24, 40). Recent evidence shows that strong selection is almost entirely absent in human lncRNAs whose sequence is not conserved in other species (24).

Constraint: the evolutionary property of reduced reproductive fitness when a sequence is mutated

Molecular function: a property of a molecule that when changed alters fitness

Although not definitive, evolutionary methods can suggest whether conservation and constraint are due to either DNA- or RNA-dependent function. Conservation tends to be strongest close to the lncRNAs' 5′ ends (46) and their promoters (84). Because elncRNA exons tend to lack conservation (67), such loci with conserved promoters could act in *cis* via the process of transcription rather than having RNA sequence–dependent function. By contrast, plncRNAs—whose exons also exhibit modest conservation (67) and whose CpG-associated promoters are well conserved (84)—are more likely to act in *trans* in an RNA sequence–dependent manner.

SPLICING

Splicing patterns also evolve rapidly. Fewer than one-third of human lncRNA splicing events are conserved in rodents, for example—much less than the fraction for human mRNAs (~90%) (103). The human–rodent conservation of GT-AG dinucleotides, which are necessary for efficient splicing, is modest, which implies that lncRNAs' intron location and/or splicing contribute functionally (84). Other sequence features of a spliced transcript, such as exonic splice enhancers (ESEs), also facilitate efficient expression and splicing. ESEs are purine-rich hexamers that bind splicing regulator proteins to aid recognition of splice sites. Unexpectedly, ESEs are unusually frequent near lncRNA splice junctions, occurring at a density comparable to that of ESEs at human mRNAs' splice junctions (41, 93).

ESEs have evolved unusually slowly under purifying selection, with splicing motifs accounting for virtually all selection on human lncRNA sequence (41, 93). This implies that splicing of multiexon lncRNAs is critical to their molecular function. Furthermore, multiexon elncRNAs are more likely than single-exon elncRNAs to be conserved over mammalian evolution (96). Exonic sequence in multiexon lncRNAs also tends to have a higher GC nucleotide content, relative to their introns, which could reflect selection on G or C alleles to improve the efficiency or robustness of splicing and/or transcription (41).

These evolutionary observations begin to explain why efficient elncRNA splicing is associated with increased enhancer activity for nearby protein-coding genes (28, 33, 96). However, how lncRNA processing strengthens enhancer activity for this neighboring gene remains unclear. Models include recruitment, during elncRNA splicing, of transcription, spliceosome, and/or chromatin-regulatory factors to the protein-coding gene via chromosomal looping or chromatin remodeling or via the short-lived elncRNA transcript itself (33, 96). Much investigation is ongoing to determine the various mechanisms underlying enhancer activity, including those involving RNAs. These mechanisms are likely diverse, involving multiple protein or RNA factors, enhancers, and chromatin states, and lie beyond the scope of this review.

MOLECULAR INTERACTIONS

To address a mechanistic hypothesis for a particular lncRNA, an investigator usually employs experiments targeting its transcript, rather than others surveying whole transcriptomes. A single lncRNA can be tested for its interaction with protein or DNA or another RNA class, such as microRNA (miRNA). Such a small-scale study has advantages of greater feasibility, lower cost, and reduced statistical testing burden over higher-throughput approaches. By contrast, a transcriptome-wide approach, interrogating protein, DNA, or RNA interactions for all lncRNAs, contextualizes each interaction among them all. Despite their higher cost and statistical burden, such methods can open up previously unanticipated lines of inquiry that better explain existing observations.

Experiments need well-chosen controls to account for nonphysiological interactions. Mammalian lncRNA studies have used bacterial RNA controls to account for nonspecific binding. These revealed, for example, that Polycomb repressive complex 2 (PRC2) binds bacterial RNA promiscuously (18), as do many mouse genomic loci (100). Investigators might be wise to explicitly distinguish terms: Colocalization may not involve direct contact, interaction may be fleeting and/or inconsequential, and binding implies mechanistic function.

Once there is sufficient evidence of a lncRNA's molecular interaction, the subsequent challenge is to determine whether and how it contributes to cellular and organismal biology. We present four counterexamples, illustrative of nonfunctional interactions. First, engagement of a lncRNA by the ribosome does not result in its translation (39). Second, despite an interaction between the lncRNA *Hotair* and PRC2 protein, there is little evidence that this interaction modifies PRC2's set of *Polycomb* target genes (94). Third, some lncRNA-bound chromatin regions fall outside of regulatory elements (including promoters or enhancers), and the activity of genes near to these regions is not always altered (e.g., 100). Finally, of the $\sim 10^6$ cataloged miRNA–lncRNA interactions (59), which were mostly identified in cell lines, many will not be relevant to human biology.

Experiments are more likely to uncover molecular mechanisms if they carefully employ genetic deletions rather than using only knockdown approaches because the latter suffer from off-target effects (23). Results will also be more compelling if the molecular, cellular, or organismal effects of perturbing a lncRNA correlate with its dosage (100) or are similar from experiments applied to different species.

PHENOTYPES

Review articles commonly state that many lncRNAs are "associated with," "linked to," or "involved in" human diseases. Experimental support for associations, links, or involvements has been collated into databases (75, 105). These observations, however, can be interpreted differently, yielding alternative explanations—for example:

- lncRNA abundance: A lncRNA is differentially expressed between normal and cancer cells, and its expression predicts poor overall patient survival. Nevertheless, if its expression changes only as a consequence of cellular transformation, then the effect on the lncRNA is a by-product, not a causal driver, of oncogenesis.
- Genetic association: A lncRNA locus contains single-nucleotide polymorphisms (SNPs) significantly associated with disease risk. Nevertheless, if these SNPs fail to correlate with this lncRNA's abundance or molecular mechanism in disease-relevant cells, then it does not causally alter disease risk. Even when these SNPs correlate with lncRNA expression, they may not be causal if they also correlate with altered expression or function of other genes nearby, including in other cell types or developmental stages.

- Molecular interaction: A lncRNA interacts with a protein that is mutated in human disease. Nevertheless, this interaction may have no cellular or organismal consequences or have no effect on disease processes, and so it is not conclusive that the lncRNA's interaction modulates disease risk.
- Disruption of predicted binding site: A somatic or inherited disease-linked mutation alters a lncRNA's predicted binding affinity of DNA, miRNA, or protein. Such predictions, however, typically suffer from high false-positive rates (type 1 errors). Moreover, even if binding occurs, this need not contribute to disease.

Presently, the syntax relating lncRNA variants to function or structure is unknown. We do not know how to glean the few variants that alter lncRNA function from among the vast majority that either are nonfunctional or alter function for other molecules. Current approaches rely on chancing on a functional variant and investigating a narrow set of all possible mechanistic hypotheses. For lncRNA biology to advance, more principled and higher-throughput experiments will be required.

Priority locations for deciphering this syntax are lncRNA splice sites and/or exons, owing to their concentrated evolutionary conservation (see above). A large-scale CRISPR-Cas9 screen used paired single guide RNAs (sgRNAs) targeting splice sites to excise exons from 10,996 human lncRNA loci. Four percent of these lncRNAs were initially proposed as being essential for cellular growth in three cell lines (62), although at least one-third of these are likely false-positive observations (47). A CRISPR interference (CRISPRi) screen, which represses transcription rather than removing DNA, predicted that 3% of 16,401 lncRNA loci are required for cellular growth (60). Nevertheless, these findings need to be treated with caution because most phenotypic observations did not reproduce across multiple cell lines. Predictions that 3–4% of lncRNAs show a growth phenotype in at least one cell line may actually be overestimates because CRISPR and other technologies are not immune to off-target effects (31, 34, 98) and because CRISPRi targeting of lncRNA loci can inadvertently repress DNA elements that regulate expression of nearby protein-coding genes.

Studying lncRNA biology using human cell lines is pragmatic. Nevertheless, the range of these lines' phenotypes is limited mostly to growth, which is not directly relevant to most human traits and disease phenotypes. Instead, a model organism is required to study lncRNAs' contributions over a broad spectrum of physiological and behavioral phenotypes, across a wide range of conditions, developmental stages, and experimental stimuli. A frequent choice of model organism, because of its phyletic proximity to human, ease of maintenance, and genetic homozygosity, is the laboratory mouse. Nevertheless, this choice immediately limits investigation to only approximately 10% of human lncRNAs, because only these possess a single ortholog in mouse (78).

Phenotypes are known for mouse protein-coding genes at the genome scale. Thousands have been individually disrupted and phenotyped using standardized protocols, with a large majority yielding discernible phenotypes (5). By contrast, a genome-wide investigation of mouse lncRNA loci has not been attempted, and so their phenotypes have not been elucidated using standardized approaches. The central question of whether disruption of a mammalian lncRNA locus commonly results in an overt phenotype thus remains unanswered.

Smaller-scale studies have reported mouse phenotypes for several dozen in vivo lncRNA knockouts. Altered phenotypes range across various physiologies and behaviors, yielding conclusions that mammalian lncRNAs contribute to diverse cellular and physiological processes. Nevertheless, these conclusions are often controversial even for well-known lncRNAs (31). Alternative mechanistic explanations of loss-of-function phenotypes are possible. These include removal of a functional DNA element irrelevant to the lncRNA; unintended effects arising from read-through transcripts; and introduced reporter genes, sites, or transcriptional terminators (4, 31). Optimal evidence for RNA-dependent lncRNA function derives from loss of function, followed by complementation approaches (e.g., 35).

Other studies report a lack of phenotypic change following disruption of a lncRNA locus (reviewed in 31). These loci may contribute functions whose disruption causes subtle phenotypes that are unobserved in experimental conditions, or are evident only under particular environmental conditions or after stimulus. From discussions with other lncRNA biologists, we believe that when disruption of a lncRNA locus fails to yield a phenotype, this important observation is often not reported in the published literature. If so, then this file-drawer effect introduces a publication bias that lays down false expectation of the likely success of future experiments.

In summary, among mouse lncRNA loci that have been targeted for disruption and phenotypic scrutiny, many have yielded either no in vivo phenotypes or effects that are not always replicated when different strategies to disrupt the locus are adopted. In the absence of strong evidence to the contrary, therefore, the expectation should be that natural mutations within human lncRNAs only rarely cause overt phenotypes.

TRAITS AND DISEASES

A transcriptome-wide association study (TWAS) can yield evidence that a lncRNA contributes to a human disease or trait. In this approach, a genetic association signal for a lncRNA's abundance in a particular tissue is first estimated. Subsequently, this signal is compared with the genetic association signal for a disease or trait. If the two association signals—one for lncRNA expression, the other for the disease or trait—are concordant across a chromosomal region, then they are colocalized, and both are explicable by a single causal DNA variant. Colocalization provides evidence of the lncRNA's role in causing the disease or trait.

De Goede et al. (19) recently used TWAS and colocalization analysis to determine whether genetically determined expression of 14,100 lncRNA loci across 49 tissues might contribute to 101 distinct complex genetic traits. They identified 83 lncRNA and disease or trait pairs for which colocalization evidence indicated that the lncRNA was more likely to causally alter disease risk or trait than any nearby protein-coding gene. First in this list of 83 pairs, for example, was *CYLD-AS1*, whose genetic signal of expression in testis or esophageal mucosa was colocalized with the genetic association signal for Crohn's disease; nearby protein-coding genes failed to be colocalized in this manner. Their overall conclusion was that "these colocalization events represent robust connections between genetic variation, lncRNA gene expression, and complex traits" (19, p. 2644).

Nevertheless, these authors acknowledged that their large expression data set is not comprehensive over all cell types, developmental stages, and environmental stimuli. This means that their study cannot be considered complete over all association signals for protein-coding gene or lncRNA expression and thus that a missing association signal for a protein-coding gene may explain a disease association signal better than the available lncRNA expression signal. For the *CYLD-AS1* prediction discussed above, for example, other expression data analyzed by Open Targets Genetics (32) deprioritized *CYLD-AS1* as causally altering risk of Crohn's disease and prioritized one or more of six neighboring protein-coding genes (in this example, *BRD7*, *ADCY7*, *CYLD*, *NKD1*, *TENT4B*, and *SNX20*). Use of many expression data sets is recommended, therefore, when prioritizing lncRNAs. Results from such analyses are provided by Open Targets Genetics (32) and are facilitated by platforms such as MR-Base (45).

SUMMARY POINTS

Prioritize human lncRNAs that:

1. Show sequence conservation and are transcribed in other mammals, such as mouse, thus removing approximately 90% of long noncoding RNAs (lncRNAs) from further consideration (78). *Rationale*: Since the last common ancestor of humans and mice, any mutation that ablated lncRNA function and thereby lowered reproductive output will have been disadvantageous. *Caveat*: Some lncRNA loci, despite being nonconserved, nevertheless will be functional.
2. Are abundant in multiple samples of primary cells. *Rationale*: lncRNA models without expression replication are less likely to be relevant to human biology and disease. *Caveat*: Some bona fide lncRNA are expressed at low levels, except in rare cell types or in narrowly defined developmental windows.
3. Show specific subcellular localization. *Rationale*: Localization helps to prioritize functional hypotheses. *Caveat*: Localization information does not directly indicate a lncRNA's mechanism.
4. Interact with other molecules. *Rationale*: An experimentally observed interaction helps to prioritize functional hypotheses. *Caveat*: Interactions do not necessarily result in cellular or organismal effects.

FUTURE ISSUES

1. Large-scale mouse lncRNA mutagenesis: Existing data are compatible with most human lncRNAs lacking RNA sequence–dependent function. Yet they are equally compatible with most lncRNAs making modest contributions to human biology. To distinguish these two scenarios requires a large-scale mutagenesis program. Such a program would minimally alter several sites, such as splice sites and promoters, of mouse orthologs of human lncRNAs, followed by detailed and standardized phenotyping.
2. Targeted search for human disease mutations: Promoters and splice sites of evolutionarily conserved lncRNAs are perhaps likely to contain de novo mutations contributing to developmental disorders.
3. Comprehensive transcriptome-wide association studies using large numbers of diverse tissues and cell types: Such studies would identify lncRNA–trait pairs for which all available evidence indicates that the lncRNA is more likely to causally alter the trait than any nearby protein-coding gene.
4. The sharing and publishing of negative results of analyses assessing the molecular mechanism and biological importance of lncRNAs: Encouraging this would greatly contribute to our general knowledge and also help refine strategies to prioritize loci for experimental validation.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was funded by the Medical Research Council (MC_UU_00007/15), Wellcome (106956/Z/15/Z), and the Biotechnology and Biological Sciences Research Council Core Strategic Program (BB/CSP1720/1 and BBS/E/T/000PR9818) and Institute Strategic Program (BB/P016855/1 and BBS/E/T/000PR9783). We apologize for any works not cited here due to space limitations. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any author accepted manuscript version arising from this submission.

LITERATURE CITED

1. Agostini F, Zagalak J, Attig J, Ule J, Luscombe NM. 2021. Intergenic RNA mainly derives from nascent transcripts of known genes. *Genome Biol*. 22:136
2. Anderson DM, Anderson KM, Chang C-L, Makarewich CA, Nelson BR, et al. 2015. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* 160:595–606
3. Bader AS, Hawley BR, Wilczynska A, Bushell M. 2020. The roles of RNA in DNA double-strand break repair. *Br. J. Cancer* 122:613–23
4. Bassett AR, Akhtar A, Barlow DP, Bird AP, Brockdorff N, et al. 2014. Considerations when investigating lncRNA function in vivo. *eLife* 3:e03058
5. Birling M-C, Yoshiki A, Adams DJ, Ayabe S, Beaudet AL, et al. 2021. A resource of targeted mutant mouse lines for 5,061 genes. *Nat. Genet*. 53:416–19
6. Bogu GK, Vizán P, Stanton LW, Beato M, Di Croce L, Marti-Renom MA. 2015. Chromatin and RNA maps reveal regulatory long noncoding RNAs in mouse. *Mol. Cell. Biol.* 36:809–19
7. Cabili MN, Dunagin MC, McClanahan PD, Biaesch A, Padovan-Merhar O, et al. 2015. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol*. 16:20
8. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*. 25:1915–27
9. Cai Z, Cao C, Ji L, Ye R, Wang D, et al. 2020. RIC-seq for global in situ profiling of RNA-RNA spatial interactions. *Nature* 582:432–37
10. Cao H, Wahlestedt C, Kapranov P. 2018. Strategies to annotate and characterize long noncoding RNAs: advantages and pitfalls. *Trends Genet*. 34:704–21
11. Carlevaro-Fita J, Johnson R. 2019. Global positioning system: understanding long noncoding RNAs through subcellular localization. *Mol. Cell* 73:869–83
12. Carlevaro-Fita J, Polidori T, Das M, Navarro C, Zoller TI, Johnson R. 2019. Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome Res*. 29:208–22
13. Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, et al. 2020. Pervasive functional translation of noncanonical human open reading frames. *Science* 367:1140–46
14. Cho C-S, Xi J, Si Y, Park S-R, Hsu J-E, et al. 2021. Microscopic examination of spatial transcriptome using Seq-Scope. *Cell* 184:3559–72.e22
15. Clark BS, Blackshaw S. 2017. Understanding the role of lncRNAs in nervous system development. *Adv. Exp. Med. Biol.* 1008:253–82
16. Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, et al. 2012. Genome-wide analysis of long noncoding RNA stability. *Genome Res*. 22:885–98
17. Creamer KM, Kolpa HJ, Lawrence JB. 2021. Nascent RNA scaffolds contribute to chromosome territory architecture and counter chromatin compaction. *Mol. Cell* 81:3509–25.e5
18. Davidovich C, Wang X, Cifuentes-Rojas C, Goodrich KJ, Gooding AR, et al. 2015. Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA. *Mol. Cell* 57:552–58
19. de Goede OM, Nachun DC, Ferraro NM, Gloudemans MJ, Rao AS, et al. 2021. Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell* 184:2633–48.e19
20. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, et al. 2010. A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLOS Biol*. 8:e1000384

21. de Souza FSJ, Franchini LF, Rubinstein M. 2013. Exaptation of transposable elements into novel *cis*-regulatory elements: Is the evidence always strong? *Mol. Biol. Evol.* 30:1239–51

22. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22:1775–89

23. Dimitrova N, Zamudio JR, Jong RM, Soukup D, Resnick R, et al. 2014. *LincRNA-p21* activates *p21* in *cis* to promote Polycomb target gene expression and to enforce the G1/S checkpoint. *Mol. Cell* 54:777–90

24. Dukler N, Mughal MR, Ramani R, Huang Y-F, Siepel A. 2021. Extreme purifying selection against point mutations in the human genome. bioRxiv 2021.08.23.457339. **https://doi.org/10.1101/2021.08.23.457339**

25. Eddy SR. 2014. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu. Rev. Biophys.* 43:433–56

26. Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB, et al. 2008. A dual origin of the *Xist* gene from a protein-coding gene and a set of transposable elements. *PLOS ONE* 3:e2521

27. Eng C-HL, Lawson M, Zhu Q, Dries R, Koulena N, et al. 2019. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* 568:235–39

28. Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, et al. 2016. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* 539:452–55

29. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, et al. 2021. GENCODE 2021. *Nucleic Acids Res.* 49:D916–23

30. Ganser LR, Kelly ML, Herschlag D, Al-Hashimi HM. 2019. The roles of structural dynamics in the cellular functions of RNAs. *Nat. Rev. Mol. Cell Biol.* 20:474–89

31. Gao F, Cai Y, Kapranov P, Xu D. 2020. Reverse-genetics studies of lncRNAs—what we have learnt and paths forward. *Genome Biol.* 21:93

32. Ghoussaini M, Mountjoy E, Carmona M, Peat G, Schmidt EM, et al. 2021. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* 49:D1311–20

33. Gil N, Ulitsky I. 2018. Production of spliced long noncoding RNAs specifies regions with increased enhancer activity. *Cell Syst.* 7:537–47.e3

34. Goyal A, Myacheva K, Groß M, Klingenberg M, Duran Arqué B, Diederichs S. 2017. Challenges of CRISPR/Cas9 applications for long non-coding RNA genes. *Nucleic Acids Res.* 45:e12

35. Grote P, Wittler L, Hendrix D, Koch F, Währisch S, et al. 2013. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev. Cell* 24:206–14

36. GTEx Consort. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369:1318–30

37. Gudenas BL, Wang L. 2018. Prediction of LncRNA subcellular localization with deep learning from sequence features. *Sci. Rep.* 8:16385

38. Guttman M, Amit I, Garber M, French C, Lin MF, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458:223–27

39. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. 2013. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154:240–51

40. Haerty W, Ponting CP. 2013. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol.* 14:R49

41. Haerty W, Ponting CP. 2015. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* 21:333–46

42. Hall LL, Carone DM, Gomez AV, Kolpa HJ, Byron M, et al. 2014. Stable C0T-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes. *Cell* 156:907–19

43. Hall LL, Lawrence JB. 2016. RNA as a fundamental component of interphase chromosomes: Could repeats prove key? *Curr. Opin. Genet. Dev.* 37:137–47

44. Hangauer MJ, Vaughn IW, McManus MT. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLOS Genet.* 9:e1003569

45. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, et al. 2018. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* 7:e34408

46. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. 2015. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep*. 11:1110–22

47. Horlbeck MA, Liu SJ, Chang HY, Lim DA, Weissman JS. 2020. Fitness effects of CRISPR/Cas9-targeting of long noncoding RNA gene. *Nat. Biotechnol.* 38:573–76

48. Johnson R, Guigó R. 2014. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* 20:959–76

49. Johnsson P, Lipovich L, Grandér D, Morris KV. 2014. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim. Biophys. Acta Gen. Subj.* 1840:1063–71

50. Johnsson P, Ziegenhain C, Hartmanis L, Hendriks G-J, Hagemann-Jensen M, et al. 2020. Transcriptional kinetics and molecular functions of long non-coding RNAs. bioRxiv 2020.05.05.079251. **https://doi.org/10.1101/2020.05.05.079251**

51. Kaewsapsak P, Shechner DM, Mallard W, Rinn JL, Ting AY. 2017. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *eLife* 6:e29224

52. Kannan S, Chernikova D, Rogozin IB, Poliakov E, Managadze D, et al. 2015. Transposable element insertions in long intergenic non-coding RNA genes. *Front. Bioeng. Biotechnol.* 3:71

53. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, et al. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLOS Genet*. 9:e1003470

54. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465:182–87

55. Kirk JM, Kim SO, Inoue K, Smola MJ, Lee DM, et al. 2018. Functional classification of long non-coding RNAs by *k*-mer content. *Nat. Genet.* 50:1474–82

56. Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, et al. 2016. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol*. 17:14

57. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, et al. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLOS Genet*. 8:e1002841

58. Lam MTY, Li W, Rosenfeld MG, Glass CK. 2014. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.* 39:170–82

59. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. 2014. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*. 42:D92–97

60. Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, et al. 2017. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* 355:aah7111

61. Liu SJ, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, et al. 2016. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol*. 17:67

62. Liu Y, Cao Z, Wang Y, Guo Y, Xu P, et al. 2018. Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites. *Nat. Biotechnol.* 36:1203–10

63. Lubelsky Y, Ulitsky I. 2018. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* 555:107–11

64. Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLOS Comput. Biol.* 2:e5

65. Ma L, Cao J, Liu L, Du Q, Li Z, et al. 2019. LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res*. 47:D128–34

66. Magny EG, Pueyo JI, Pearl FMG, Cespedes MA, Niven JE, et al. 2013. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* 341:1116–20

67. Marques AC, Hughes J, Graham B, Kowalczyk MS, Higgs DR, Ponting CP. 2013. Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biol*. 14:R131

68. Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, et al. 2017. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* 541:228–32

69. Mattick JS. 2001. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep*. 2:986–91

70. Mattick JS. 2009. The genetic signatures of noncoding RNAs. *PLOS Genet*. 5:e1000459

71. Mattick JS. 2011. The central role of RNA in human development and cognition. *FEBS Lett*. 585:1600–16

72. Mattick JS, Dinger ME. 2013. The extent of functionality in the human genome. *HUGO J.* 7:2

73. Mattioli K, Volders P-J, Gerhardinger C, Lee JC, Maass PG, et al. 2019. High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. *Genome Res*. 29:344–55

74. Melé M, Rinn JL. 2016. "Cat's cradling" the 3D genome by the act of lncRNA transcription. *Mol. Cell* 62:657–64

75. Miao Y-R, Liu W, Zhang Q, Guo A-Y. 2018. lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Res*. 46:D276–80

76. Michieletto D, Gilbert N. 2019. Role of nuclear RNA in regulating chromatin structure and transcription. *Curr. Opin. Cell Biol.* 58:120–25

77. Mukherjee N, Calviello L, Hirsekorn A, de Pretis S, Pelizzola M, Ohler U. 2017. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat. Struct. Mol. Biol.* 24:86–96

78. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, et al. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505:635–40

79. Niu L, Lou F, Sun Y, Sun L, Cai X, et al. 2020. A micropeptide encoded by lncRNA MIR155HG suppresses autoimmune inflammation via modulating antigen presentation. *Sci. Adv.* 6:eaaz2059

80. Ouspenskaia T, Law T, Clauser KR, Klaeger S, Sarkizova S, et al. 2022. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat. Biotechnol.* 40:209–17

81. Palazzo AF, Gregory TR. 2014. The case for junk DNA. *PLOS Genet*. 10:e1004351

82. Palazzo AF, Lee ES. 2015. Non-coding RNA: What is functional and what is junk? *Front. Genet.* 6:2

83. Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. *Trends Genet*. 22:1–5

84. Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res*. 17:556–65

85. Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLOS Genet*. 10:e1004525

86. Ransohoff JD, Wei Y, Khavari PA. 2018. The functions and unique features of long intergenic non-coding RNA. *Nat. Rev. Mol. Cell Biol.* 19:143–57

87. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno 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

88. Rinn JL, Chang HY. 2020. Long noncoding RNAs: molecular modalities to organismal functions. *Annu. Rev. Biochem.* 89:283–308

89. Rivas E, Clements J, Eddy SR. 2020. Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics* 36:3072–76

90. Sartorelli V, Lauberth SM. 2020. Enhancer RNAs are an important regulatory layer of the epigenome. *Nat. Struct. Mol. Biol.* 27:521–28

91. Schlackow M, Nojima T, Gomes T, Dhir A, Carmo-Fonseca M, Proudfoot NJ. 2017. Distinctive patterns of transcription and RNA processing for human lincRNAs. *Mol. Cell* 65:25–38

92. Schmid M, Jensen TH. 2018. Controlling nuclear RNA levels. *Nat. Rev. Genet.* 19:518–29

93. Schüler A, Ghanbarian AT, Hurst LD. 2014. Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.* 31:3164–83

94. Selleri L, Bartolomei MS, Bickmore WA, He L, Stubbs L, et al. 2016. A Hox-embedded long noncoding RNA: Is it all hot air? *PLOS Genet*. 12:e1006485

95. Smith MA, Gesell T, Stadler PF, Mattick JS. 2013. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res*. 41:8220–36

96. Tan JY, Biasini A, Young RS, Marques AC. 2020. Splicing of enhancer-associated lincRNAs contributes to enhancer activity. *Life Sci. Alliance* 3:e202000663

97. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, et al. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* 22:1616–25

98. Tycko J, Wainberg M, Marinov GK, Ursu O, Hess GT, et al. 2019. Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-coding elements. *Nat. Commun.* 10:4063

99. Ulitsky I, Bartel DP. 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell* 154:26–46

100. Vance KW, Sansom SN, Lee S, Chalei V, Kong L, et al. 2014. The long non-coding RNA *Paupar* regulates the expression of both local and distal genes. *EMBO J.* 33:296–311

101. Wang X-W, Liu C-X, Chen L-L, Zhang QC. 2021. RNA structure probing uncovers RNA structure-dependent biological functions. *Nat. Chem. Biol.* 17:755–66

102. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* 23:1383–90

103. Washietl S, Kellis M, Garber M. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* 24:616–28

104. Xia C, Fan J, Emanuel G, Hao J, Zhuang X. 2019. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *PNAS* 116:19490–99

105. Zhao H, Shi J, Zhang Y, Xie A, Yu L, et al. 2020. LncTarD: a manually-curated database of experimentally-supported functional lncRNA-target regulations in human diseases. *Nucleic Acids Res.* 48:D118–26

106. Zhou B, Ji B, Liu K, Hu G, Wang F, et al. 2021. EVLncRNAs 2.0: an updated database of manually curated functional long non-coding RNAs validated by low-throughput experiments. *Nucleic Acids Res.* 49:D86–91

Annual Review of Genomics and Human Genetics

Volume 23, 2022

Contents

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* articles may be found at http://www.annualreviews.org/errata/genom