The dimensions, dynamics and relevance of the noncoding transcriptome

Ira W. Deveson[1,2], Simon A. Hardwick[1,3], Tim R. Mercer[1,3] & John S. Mattick[1,3°]

[1] *Genomics and Epigenetics Division, Garvan Institute of Medical Research, NSW, Australia*

[2] *School of Biotechnology and Biomolecular Sciences, Faculty of Science, UNSW Sydney, NSW, Australia*

[3] *St Vincent's Clinical School, UNSW Sydney, NSW, Australia*

°*Correspondence: j.mattick@garvan.org.au*

Keywords

Pervasive transcription, lncRNAs, long noncoding RNAs, alternative splicing, RNA regulation

Abstract

The combination of pervasive transcription and prolific alternative splicing produces a mammalian transcriptome of great breadth and diversity. The majority of transcribed genomic bases are intronic, antisense or intergenic to protein-coding genes, yielding a plethora of short and long non-protein-coding regulatory RNAs. Long noncoding RNAs (lncRNAs) share most aspects of their biogenesis, processing and regulation with mRNAs. However, lncRNAs are typically expressed in more restricted patterns, frequently from enhancers, and exhibit almost universal alternative splicing. These features are consistent with their role as modular epigenetic regulators. Here we describe the key studies and technological advances that have shaped our understanding of the dimensions, dynamics and biological relevance of the mammalian noncoding transcriptome.

Appreciating transcriptome diversity

Mammals possess roughly the same number and a similar repertoire of protein-coding genes as nematode worms. By contrast, the intergenic and intronic regions of the mammalian genome are far greater. Indeed, while the number of protein-coding genes is largely static across the animal kingdom, noncoding genome content increases in size with developmental complexity [1].

Initial studies of the mammalian transcriptome were prefaced on the assumption that most genes encode proteins and that mRNAs constitute the bulk of non-ribosomal RNA in cells. It was therefore a surprise to discover that there are many transcripts, albeit usually of lower abundance, that are not protein-coding. In mammals, almost the entire genome is pervasively transcribed to generate not only mRNAs but many small and large non-protein-coding RNAs that are antisense, intronic or intergenic to protein-coding genes [2]. The mammalian transcriptome is further diversified by prolific alternative splicing of both protein-coding and noncoding RNAs.

The breadth and complexity of mammalian transcription was not obvious before scalable cDNA hybridization [3] and sequencing [4], and the subsequent incorporation of next-generation sequencing to create modern RNA sequencing (RNA-Seq) [5]. The proliferation and evolution of RNA-Seq, including the advent of methods for targeted [6], single-molecule [7,8] and single-cell sequencing [9], continues to enlarge our understanding of transcriptional diversity. Nevertheless, the true dimensions of the mammalian transcriptome remain unknown and the spatiotemporal dynamics of gene expression and splicing demand further attention.

This is especially true for the noncoding transcriptome. Long noncoding RNAs (lncRNAs) constitute a large portion of the mammalian transcriptome but are mostly poorly cataloged and characterized. Moreover, the relatively weak evolutionary constraint on their primary sequences, compared to protein-coding genes (noting enigmatic exceptions, such as transcribed ultraconserved elements [10,11]), their low abundance in tissue samples and their incompatibility with a purely protein-centric model of gene regulation has caused many to question the biological relevance of lncRNAs.

The rapid evolution of lncRNA sequences has been reviewed extensively elsewhere [12,13] and is not necessarily indicative of non-functionality, since it is also consistent with plastic structure-function relationships in regulatory molecules and positive selection for phenotypic variation during adaptive radiation [14]. Likewise, accumulating evidence shows that lncRNAs are not simply lowly expressed but transcribed in highly specific patterns [15,16]. This, in addition to their complex alternative splicing [17], suggests that many lncRNAs may fulfill regulatory roles in the mammalian developmental program.

While the number of well characterized lncRNAs is relatively small (but growing), the rise of CRISPR-Cas9 targeted genome manipulation, the development of high-throughput forward genetic screens based on this technology [18], and the proliferation of precise, scalable methods for resolving RNA structure and RNA-protein interactions [19] are enabling the community to address longstanding challenges in lncRNA biology.

The reality of pervasive transcription

The first clear evidence that the mammalian transcriptome included large numbers of non-protein-coding intergenic and antisense RNAs, as well as many stable intronically-derived RNAs, came from genome-wide tiling arrays [20–22] and sequencing of cloned cDNAs [4,23–25]. These unexpected findings garnered controversy and when very few reads obtained in early RNA-Seq experiments aligned outside of known protein-coding genes the evidence for 'pervasive transcription' was questioned [26].

However, such claims stemmed from misinterpretation: the reason RNA-Seq reads from intergenic regions were scarce in the relatively low-depth assays of that time is because sequencing fragments are competitively sampled from a common pool, wherein transcripts of varied abundance are proportionally represented [27,28]. Highly abundant mRNAs dominate the pool and obscure noncoding transcripts, which are generally less abundant [29] (or rather, more precisely expressed; see below). To detect lowly expressed genes and rare isoforms and, even more so, to accurately resolve their spliced architectures, a sample must be sequenced deeply (**Box 1**).

As increasing numbers of cell types and tissues have been profiled at increasing depth it has become clear that the majority of the mammalian genome is dynamically transcribed [4,24,30–33]. Activity was recorded at 75% of genomic bases in a survey of 15 human cell lines by the ENCODE consortium [30]. Moreover, only around half of this activity was observed in any individual sample, implying further activity would be observed in additional samples [30]. Comparable results were obtained for mouse [34].

The advent of targeted RNA-Seq has allowed the noncoding transcriptome to be surveyed at higher resolution [6]. This technique has unearthed widespread, regulated transcription in intergenic regions (previously considered 'gene deserts') below the limit of detection for genome-wide RNA-Seq [6,17]. Hence, even the detailed transcriptional profiles generated by ENCODE are incomplete.

Nonetheless, in the 27 years since the first documentation of a discrete and biologically relevant lncRNA, H19 [35], the catalog of known lncRNAs has rapidly grown [36–39]. Mammalian lncRNA loci now comfortably exceed protein-coding genes in number, with the MiTranscriptome annotation alone listing 58,648 lncRNA loci, compared to 21,313 protein-coding genes [11]. Moreover, the descriptors 'gene' and 'locus' are not entirely appropriate for lncRNAs, as the mammalian transcriptional landscape is largely continuous, containing densely interleaved clusters of noncoding and protein-coding transcripts [4,22].

The similar life histories of mRNAs and lncRNAs

Aside from several minor idiosyncrasies (reviewed elsewhere [40]), many if not most lncRNAs are regulated, transcribed and processed in a similar fashion to mRNAs [41].

LncRNAs and mRNAs are roughly comparable in size and structure [37,42], though some lncRNAs are very large, in excess of 100kb [43]. Like mRNAs, many lncRNAs are transcribed by RNA Pol II, regulated by morphogens and conventional transcription factors, dysregulated in disease, capped at their 5'ends and polyadenylated at their 3'ends [12,41], although some lncRNAs are transcribed by RNA Pol III [44] or processed from intronic sequences [22]. Like protein-coding genes, lncRNA transcription arises from recognizable promoters, which show strong sequence conservation in many cases [4]. LncRNA promoters are enriched for transcription factor binding sites [45,46] and the canonical marks of active gene expression, H3K4me3, H3K9ac and H3K27ac [36,37,47].

Expressed lncRNA promoters are also enriched for the repressive H3K9me3 mark and exhibit lower transcription factor binding densities than protein-coding gene promoters [46,48]. This may be consistent with a recent report suggesting that most intergenic lncRNAs originate from enhancer-type transcription start sites rather than conventional promoters [39].

HuR and U1snRNP (regulators of transcript stability) associate with similar frequency to lncRNAs and mRNAs at matched expression levels, which also exhibit comparable stabilities in human cell lines following transcriptional inhibition [46]. Other studies have shown that lncRNAs have a lower average but similar range of half-lives as mRNAs [49,50].

LncRNA expression is highly precise and dynamic

One of the key concerns about the biological relevance of lncRNAs has been their low abundance in tissue samples, sometimes argued to be simply the manifestation of ‘transcriptional noise’ [51]. However, accumulating evidence suggests that this reflects heightened spatiotemporal precision, rather than low background expression.

It is clear that, while some such as *MALAT1* and *NEAT1* are widely expressed [52], most lncRNAs are highly tissue-specific, more so than protein-coding genes [4,30,36,37,53]. For example, one survey classified 78% of detectable lncRNAs as tissue-specific, compared to just 19% of mRNAs [36]. Importantly, this difference was observed for lncRNAs and mRNAs at matched expression levels, and highly expressed lncRNAs in fact displayed the strongest tissue specificity [36].

In this survey, lncRNAs were detected at around an order of magnitude lower, on average, than mRNAs [36]. Expression measurements from homogenized tissue (analogous to analyzing a smoothie) report a population average among pooled cells, regardless of differences between, or even within, specific cell populations (**Box 2**). Such heterogeneity might be biologically relevant, especially in tissues where well-defined substructures exist.

The brain is the most complex organ and harbors the largest transcriptional diversity of any somatic tissue. Using *in situ* hybridization to visualize the spatial distribution of transcription in mouse brains, an early study showed lncRNAs to be highly abundant in specific cells but spatially precise, often restricted to particular brain regions, structures or cell types [15]. The authors proposed, therefore, that the low abundance of lncRNAs observed in bulk tissue experiments reflects their highly cell-specific expression.

With the emergence of single-cell RNA-seq [9], this matter has been scrutinized in more detail (**Box 2**). In one recent investigation, individual transcriptional profiles were obtain from 276 cells in developing human neocortex [16]. On average, detectable lncRNAs were lower in abundance than mRNAs by an order of magnitude, consistent with measurements from whole tissue [36]. However, the median lncRNA/mRNA ratio in single cells among detectable transcripts exceeded one (i.e. lncRNAs were present at greater median abundance than mRNAs) in one-third of the cells, even approaching the levels of housekeeping genes in some instances [16]. Hence, many lncRNAs were expressed at low levels in pooled cells but high levels in a subset of individual cells.

This phenomenon was not recapitulated in cultured K562 cells, which are more uniform [16].

Similarly, 61 lncRNAs profiled by RNA fluorescence *in situ* hybridization showed no more heterogeneity than mRNAs within any of three cell-lines [54]. That lncRNAs and mRNAs exhibit equivalent cell-to-cell heterogeneity in cultured cell lines suggests that the heightened heterogeneity of lncRNA expression in the neocortex reflects biological differences between cells therein, rather than sporadic expression among homogenous cells.

Even more so than mRNAs, lncRNAs also show precise patterns of subcellular localization (reviewed elsewhere [55]). *NEAT1*, for instance, can be found in nuclear paraspeckles (and is essential for their formation) [56,57], while *XIST* localizes to the inactive X-chromosome (and is essential for its silencing) [58]. As a population, lncRNAs show stronger nuclear localization than mRNAs; 17% of lncRNAs and 15% of mRNAs show relative enrichment in the nucleus, compared to 4% and 26% respectively in the cytoplasm [37]. LncRNAs that are retained in the nucleus may accumulate at their own sites of transcription, or localise elsewhere. *HOTAIR*, for instance, is transcribed from the mammalian *HOXC* locus but accumulates in *trans* at the *HOXD* locus (where it facilitates gene silencing) [59].

LncRNA expression is also highly dynamic during development. This has been demonstrated in differentiating embryonic stem cells [60], muscle [57], T cells [60], mammary gland [61] and neurons [62–64], among other systems. One recent study leveraged single-cell RNA-Seq to resolve the transcriptional repertoire of early human embryo development [65]. Compared to mRNAs, lncRNA abundances were found to be higher within individual cells than in pooled data from multiple cells and developmental stages. However, at the 4-cell or 8-cell stage, when the cells of an embryo are highly similar, a large proportion of detected lncRNAs were expressed in every cell, further indicating that lncRNA expression was not simply ‘leaky’ [65].

Prolific alternative splicing diversifies the transcriptome

Extensive alternative splicing of human mRNAs was recognized many years ago [66,67], but the scope of its influence on the mammalian transcriptome was not fully appreciated before the advent of RNA-Seq. Early systematic analyses of alternative splicing with RNA-Seq showed that 92-94% [68] or 92-97% [69] (i.e. probably all) multi-exon human protein-coding genes undergo alternative splicing. Unique isoforms may be deployed in specific contexts, remolding the transcriptome during development and evolution [70–72].

Most genes express a dominant spliced isoform, accounting for around 80% of transcription in an individual tissue, and multiple minor alternative isoforms [30,73]. In the ENCODE survey of 15 cell lines, an average of 10-12 isoforms were detected per gene, per cell line [30]. However, because high sequencing coverage is required to resolve low-level alternative isoforms (**Box 1**), this is a conservative estimate of isoform diversity. Targeted RNA-Seq highlights this limitation, in one instance unearthing novel isoforms encoding up to three new open reading frames (ORFs) of *TP53,* which is among the most extensively studied of all human genes [6].

An additional limitation is imposed by RNA-Seq read length, since the computational assembly of full-length alternative isoforms from short reads is difficult. With the emergence of long-read sequencing technologies it is now possible to read full-length isoforms as single molecules, negating the challenges of transcript assembly (**Box 1**). Leading studies in this space have resolved complex and precisely organized alternative splicing events, including the coordinated inclusion/exclusion of distant exons and allele-specific isoform expression [7,8,74].

Early evidence from single-cell RNA-Seq experiments suggests that such features reflect the organization, as opposed to random distribution, of alternative isoforms within a heterogeneous tissue, potentially reflecting biological differences between cells [75–78]. In one example, a single alternative splicing event in *NINEIN* is sufficient to trigger differentiation of individual human neural progenitors from a purified population into neurons [78]. We anticipate that the confluence of single-cell and single-molecule sequencing, now feasibly executed in tandem [79], will profoundly advance our understanding of splicing organization.

Many alternative splicing events at protein-coding loci generate isoforms that lack an extended ORF. For example, one survey found that the major isoform for up to 20% of protein-coding genes was noncoding (though we note that this result is highly dependent on the quality of transcript assembly) [73]. Noncoding isoforms are often the product of intron retention, a regulated process that may dampen gene expression by inducing nonsense mediated decay [80,81] or nuclear transcript detention [82]. However, noncoding (or even coding) RNAs derived from protein-coding loci can also transact regulatory functions. For instance, the human β-globin mRNA can convey epigenetic information independently of its translation [83] and a UV-induced noncoding isoform of *ASCC3* facilitates transcriptional recovery after DNA damage in a manner independent of, and antagonistic to, this gene's protein-coding function [84].

Near-universal alternative splicing of noncoding exons

LncRNAs also undergo alternative splicing, though their relatively low abundance in homogenized tissues hinders accurate resolution of these events. The GENCODE v7 noncoding RNA catalog lists alternative isoforms for only around a quarter of lncRNA loci, and indicates that lncRNAs generally have fewer exons and shorter mature transcripts than mRNAs [37]. However, a subsequent detailed characterization of 398 lncRNAs from the same catalog by rapid amplification of cDNA ends and long-read sequencing showed these to be at least equivalent to protein-coding genes in splicing complexity, indicating that insufficient coverage was the reason for the previous underestimate [42].

Targeted RNA-Seq has been similarly applied to obtain more complete models of lncRNA architecture. By targeting exons of annotated lncRNAs, many previously unassembled exons were incorporated into existing lncRNA loci and many were shown to be fragments, with splicing unifying multiple annotated loci [85]. Even the extensively studied lncRNA *HOTAIR* exhibited alternative splicing events that were undetected by conventional RNA-Seq [6].

The splicing of lncRNAs and mRNAs is regulated by local sequence elements that are highly similar, with canonical splice donor (GT) and acceptor (AG) dinucleotides demarcating intron-exon boundaries in both [37,85]. The binding motif for U1snRNP, which initiates spliceosome recruitment [86], also has the same density and positional distribution in both [46]. Canonical poly-pyrimidine enrichments upstream of splice acceptor sites tend to be slightly weaker in pre-lncRNAs than pre-mRNAs, and branch point nucleotides are slightly more distant [46], features that correlate with heightened alternative over constitutive splice site selection. CLIP-seq data also shows a relative depletion of U2AF65, which promotes branch point selection [86], near lncRNA splice acceptor sites, compared to those in nascent mRNA [46].

The latter features may explain, at least in part, a global reduction in splicing efficiency [46,87] and/or splice site selection [79] that more clearly distinguishes lncRNAs from protein-coding genes. Retarded splicing kinetics observed in lncRNAs similarly distinguish spliced exons in untranslated regions (UTRs) of protein-coding genes from those within ORFs [87]. In this context it is interesting that many protein-coding loci also express their 3'UTRs separately from their normally associated protein-coding sequences, in highly cell-specific patterns and with clear genetic activity in *trans* [88].

Recently, a high-resolution transcriptional cross-section of human and mouse chromosome 21 was generated by targeted transcript enrichment, followed by single-molecule and saturating short-read RNA-Seq [17]. This approach revealed that lncRNAs are, contrary to the impression from more shallow surveys, enriched for alternative splicing, with internal exons being near-universally classified as alternative. Human splicing profiles were recapitulated in the Tc1 mouse strain (which carries a copy of human chromosome 21), indicating that they are robustly encoded in the local chromosome sequence, not the manifestation of non-specific splicing activity. Extensive alternative splicing was also observed for untranslated exons at protein-coding loci, suggesting that this is a general feature of noncoding regulatory RNA [17].

LncRNAs therefore exhibit a highly modular or exon-centric architecture; unlike protein-coding genes, whose central exons are constrained by the requirement to maintain continuous ORFs, lncRNA exons behave as discrete units, recombined with maximum flexibility. It is currently unclear whether this reflects precisely organized cell-to-cell heterogeneity (as for mRNAs; above) or uniformly promiscuous usage of alternative isoforms. Single-cell (probably in conjunction with single-molecule) approaches will be required to resolve this dichotomy.

Functional characterization of lncRNAs: unique challenges and emerging solutions

Many well-characterized lncRNAs function as regulatory molecules in the epigenetic control of gene expression and fulfill roles in differentiation and development [2,89]. These roles are easily reconciled with the distinctive features of lncRNA biology just described, namely their precise expression and complex alternative splicing, providing a conceptual framework to guide further discovery and characterization.

We owe much of what we know about lncRNA function to the characterization of flagship examples, such as *XIST* and *NEAT1*, whose biological roles and modes of action are now relatively well understood. While the knowledge garnered has been vital to the development of the field, it should be borne in mind that many of these well-known representatives are atypical. *XIST*, for instance, is more highly conserved than most [13], partly reflecting its derivation from an ancestral protein-coding gene [90], and both *XIST* and *NEAR1* are highly and constitutively expressed, unlike most lncRNAs.

This consideration is important because the characteristic aspects of lncRNA biology described above, in addition to their mechanistic diversity and the subtle and/or context-specific phenotypes that many lncRNAs exhibit, pose challenges to their functional characterization. Strategies that have worked well for protein-coding genes are often inapplicable for lncRNAs (reviewed elsewhere [91]) and the relatively small number of examples for which clear biological roles have been determined probably represent the lowest hanging fruit.

Knock down of lncRNAs in culture, using si/shRNAs, has frequently resulted in altered cell growth or behavior (see [41]), suggesting that perturbation of lncRNAs disturbs epigenetic state.

Several studies have generated lncRNA deletions *in vivo*, with varied success (see [92]). Using classical gene replacement techniques, and targeting mouse lncRNAs with identifiable human orthologs, one investigation reported developmental defects for 5 out of 18 knockout mice [93]. The relatively low frequency of gross phenotypes observed (even for conserved lncRNAs), may reflect a combination of dispensable exons, redundancy in regulatory systems and/or more subtle phenotypes, especially cognitive phenotypes, which are not usually polled [94]. Indeed, since most lncRNAs are expressed in the brain and many are primate-specific [37], it may be that much of the lncRNA-mediated genetic information in mammals is devoted to brain function, and not easily detectable in developmental screens. For example, knockout of the lncRNA *BC1* causes no visible anatomical consequences but leads to a behavioral phenotype that would be lethal in the wild [95].

We anticipate that the key to rapidly expanding our understanding of lncRNA biology lies in high-throughput forward genetics. However, until recently there has been no robust, scalable strategy for agnostic phenotype-to-genotype interrogation of lncRNAs (**Box 3**).

Recently, two breakthrough investigations have answered this call [96,97]. Both did so by pairing the CRISPR-Cas9 system with large guide-RNA libraries (**Box 3**). In one approach, paired guide-RNAs were used to induce large genomic deletions in 700 individual lncRNA loci, of which 51 had a positive or negative influence on cancer cell growth [96]. Reasoning that genomic deletions may affect local regulatory elements, the second study instead used CRISPR interference (CRISPRi; wherein transcription is locally inhibited at targeted sites) to knock down 16,401 lncRNA loci across seven human cell lines. Perturbation of 499 of these targets affected cell growth and the overwhelming majority (89%) of these expressed a phenotype in just a single cell type, emphasizing the context-specificity of lncRNA activity [97]. Both results imply widespread lncRNA functionality, and that the design of elegant, well-directed paradigms for phenotypic screening will lead to further success (**Box 3**).

The CRISPR-Cas9 system also enables detailed examination of individual lncRNAs by targeted genomic manipulations and/or transcriptional perturbation. The removal or modification of specific promoter elements, splice sites, sequence motifs or RNA domains is now relatively facile. Additionally, CRISPRi (and related techniques) can be used to disentangle regulatory effects enacted in *cis* and *trans* to be clearly distinguished. In a recent investigation, 12 lncRNA loci were dissected, individually deleting promoters, exons and introns or inserting premature polyadenylation signals to prevent transcript elongation. The effects of each modification on local gene expression were assessed [98]. Illustrating the power of the CRISPR-Cas9 system for lncRNA characterization, this approach identified instances in which: (1) the act of transcription but not the mature lncRNA molecule itself and; (2) DNA sequences in the lncRNA promoter but not the act of transcription; were sufficient to elicit a *cis*-regulatory effect on a neighboring gene. While these examples highlight the ability for lncRNAs to enact *cis*-regulatory effects, we note that these do not preclude independent *trans*-regulatory functions.

The community also now possesses increasingly precise and scalable methods for resolving RNA structure, RNA-protein and RNA-chromatin interactions (reviewed elsewhere [19]). Advances in the biology of *XIST*, for which the secondary structure [99], protein binding partners [100–102] and sites of chromatin localization during X-inactivation [103,104] are now known, demonstrate how these techniques could be used to help resolve the structure-function relationships and the mode of action for any lncRNA.

LncRNAs as modular epigenetic regulators

The phenomena interrogated by the techniques just mentioned are central to understanding lncRNA functionality; namely, the ability of lncRNAs to form specific and multilateral RNA-protein, RNA-DNA and RNA-RNA binding interactions. Their diverse binding properties and flexibility in size and structure, means lncRNAs are ideally suited to facilitate interactions between other biomolecules, and thereby organize and regulate cellular processes [105].

It is unsurprising then that many lncRNAs participate in the epigenetic regulation of gene expression. LncRNAs commonly interact with chromatin remodeling enzymes/complexes, including PRC2 [106] and MLL/TrxG [107] complexes, histone demethylase LSD1 [108], DNA methyltransferase DNMT1 [109] and demethylation regulator GADD45a [110]. These possess enzymatic activity to alter chromatin state but lack the capacity for site-specific DNA/chromatin-binding. In numerous examples, lncRNAs confer this site-specificity, guiding effector complexes to appropriate genomic targets. LncRNAs may select genomic targets by interacting with additional protein binding partners (e.g. [111]), via direct interaction between lncRNA and DNA/chromatin (e.g. [52]), or in *cis* at the site of lncRNA transcription (e.g. [112]).

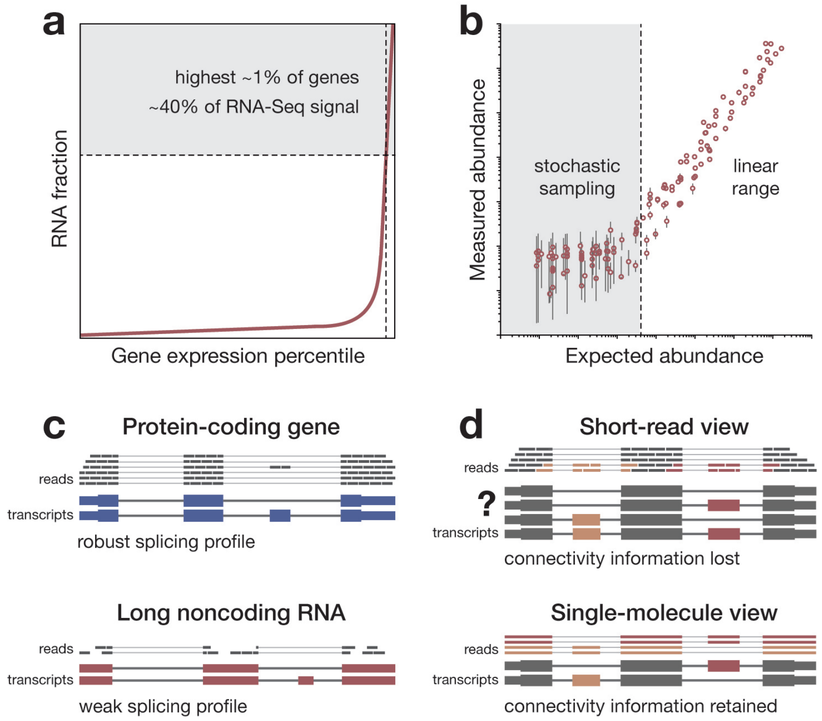
Consistent with guidance of epigenetic processes, many lncRNAs play roles in cell-fate determination (reviewed elsewhere [89]). However, the more profound challenge for the genetic programming of complex organisms is to organize the growth and differentiation of trillions of cells into precisely structured organs and tissues, including bones, muscles and the brain. While protein factors clearly play a role, epigenetic remodeling, guided by regulatory RNAs, appears increasingly important to the developmental program [2,89]. The fact that the native program can be trumped by ectopic expression of transcription factors is not inconsistent with this model; indeed, developmental programs can also be changed by ectopic expression of regulatory RNAs [113].

We are far from understanding how the mammalian developmental program operates but one can imagine a scenario in which every cell (or local group of cells) expresses a unique molecular profile that defines its identity and position during ontogeny and at maturity. This fits with the observations of ordered, not random or sporadic, cell-specific lncRNA expression [15,16]. Indeed, recent data from single-cell RNA-Seq suggest that few (if any) cells in a mammalian tissue would express a lncRNA population redundant with a neighbor (see above).

The exon-centric architecture of lncRNAs, in which exons are recombined into a dizzying diversity of isoforms, can also be reconciled with an RNA-driven developmental program. This implies that exons may act as discrete functional domains, each with a unique and specific affinity for external biomolecules (specific protein domains, DNA motifs etc.). Modular recombination of lncRNA exons may enable diverse and dynamic interactions, for instance, delivering a particular chromatin remodeler to particular sites in the genome at specific moments in development.

Moreover, the evolution of RNA modification and editing, which expands markedly in mammalian and especially primate evolution, may have provided the means to superimpose epigenetic plasticity on an otherwise hardwired genome and transcriptome, enabling physiological and cognitive adaptation [114].

Box 1: A matter of length and depth - limitations and advances in RNA-Seq



RNA-Seq provides an unbiased global snapshot of transcription. Sequencing fragments are sampled from a pool, in which transcripts of different abundances are proportionally represented. Crucially, this enables quantitative measurements of expression and/or splicing.

However, due to the immense size of the transcriptome and its wide range of expression levels, competitive sampling means that highly expressed transcripts obscure lowly expressed transcripts. The structure of the transcriptome is such that, in a typical human sample, the top 1% most highly expressed protein-coding genes commonly soak up around 40% of sequencing reads (**A**) [27]. For this reason, RNA-Seq carries an inherent expression-dependent bias that affects detection, quantification and assembly of RNA transcripts. This is best illustrated by analysis of spike-in controls for RNA-Seq, which are formulated into a staggered mixture spanning the quantitative range of the human transcriptome [27,28]. Typically, spike-in transcripts at high and moderate abundances are robustly quantified. However, among spike-ins of lower abundances, stochastic sampling leads to quantitative variability and, ultimately, the loss of linearity between expected and observed abundances (**B**). Due to their low abundance, compared to mRNAs, lncRNAs are detected with lower sensitivity and quantified with lower accuracy [28].

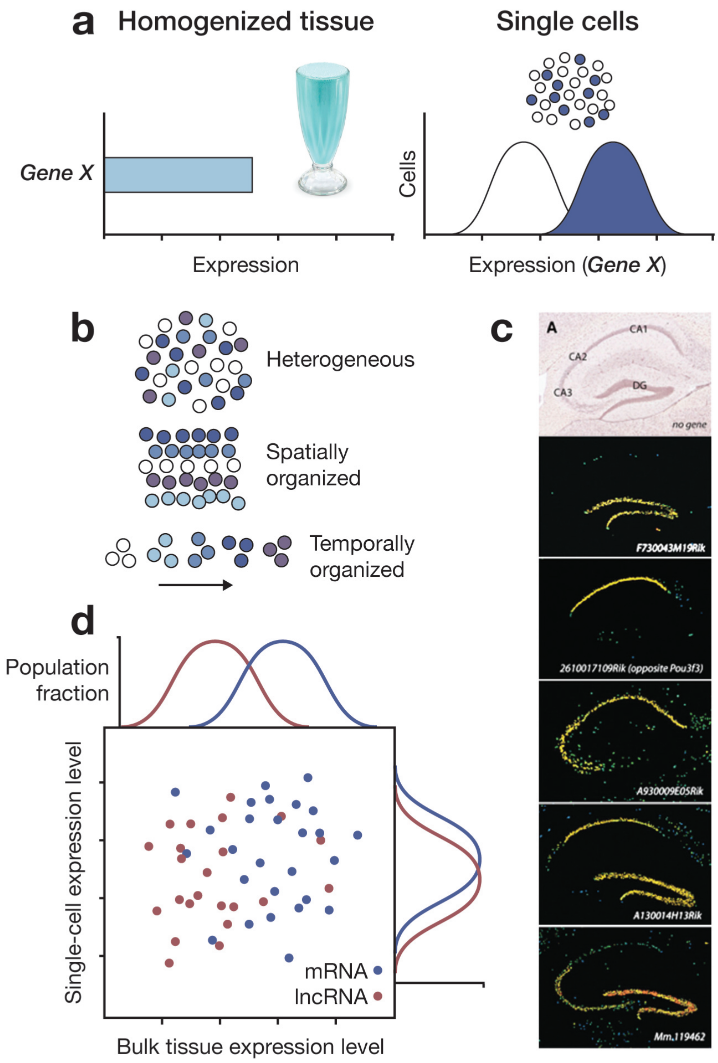
Targeted RNA-Seq may be used to alleviate this issue. By magnifying coverage in specific genome regions, targeted sequencing enables more sensitive gene/isoform discovery and more precise measurements of expression than is feasible with conventional RNA-Seq. This enables improved detection and quantification of lowly expressed genes and lncRNAs [6,85].

Another limitation of traditional RNA-Seq is the reliance on computational assembly of full-length isoforms from short (~100-150 bp) sequencing reads. This is a difficult task, particularly when alternative splicing generates multiple partially redundant isoforms at an individual locus. Because saturating coverage is required for high-quality assembly, the expression dependent bias of RNA-Seq strongly affects this process [28], ensuring rare transcripts, such as lncRNAs, are often poorly resolved (**C**). Targeted RNA-Seq [85] and the coupling of sequencing to rapid amplification of cDNA ends (RACE-Seq) [42] have both been used to better resolve the spliced architecture of specific lncRNAs.

However, even with saturating coverage, long-range exon connectivity within an alternatively spliced locus cannot be established unambiguously using short-read RNA-Seq. Short reads may be used to designate individual exons as constitutive or alternative but the relationship between distant exons cannot be judged, since these are never represented on the same sequencing fragment (**D**; upper).

With the emergence of technologies for long-read sequencing it is now possible to read full-length isoforms as single molecules, negating the challenges posed by transcript assembly [7,8]. Single-molecule techniques have been used to resolve complex, organized alternative splicing events, such as mutually exclusive or inclusive relationships between distant exons (**D**; lower) [8]. However, these techniques are currently expensive, meaning depth remains a constraint and rare transcripts may fall below the limits of sampling.

Box 2: Averages lie - transcriptomic insights from single-cell RNA-Seq



Until recently, all RNA-Seq experiments were performed on bulk tissue or cell samples. These experiments have been instrumental in advancing our understanding of mammalian transcription and have generated hugely valuable resources, such as the ENCODE [32] and Genotype-Tissue Expression (GTEx) [115] transcription catalogs.

However, the analysis of a homogenized tissue can be likened to the analysis of a smoothie; measurements of gene expression or alternative splicing report a single population average among pooled cells, ignoring heterogeneity between cells (**A**; left). For instance, a measurement from bulk tissue cannot discriminate between the possibilities that *Gene X* is expressed at a moderate level in all cells or, alternatively, that *Gene X* is expressed in just a subset of cells but at a high level. Likewise, an observed increase in the expression of *Gene X* could be attributed either to a uniform increase in its transcription across all cells or, alternatively, an increase in the number of cells expressing *Gene X*.

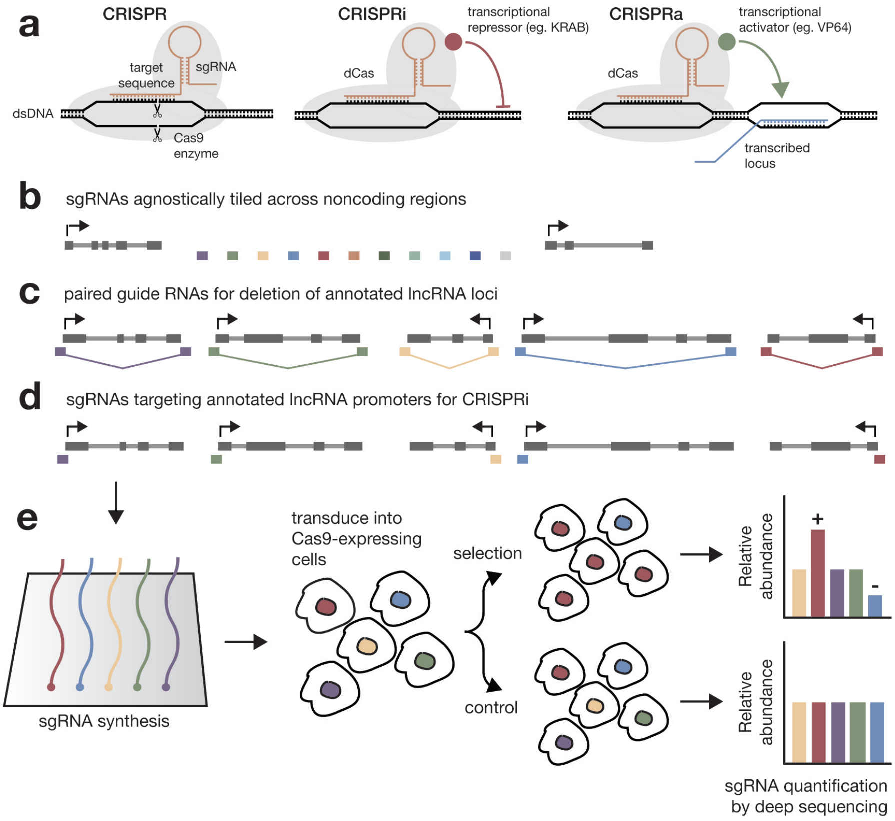
Single-cell RNA-Seq, by contrast, provides information about the population structure of gene expression within a sample and can simultaneously measure the proportion of cells expressing *Gene X* and the magnitude of its expression in each (**A**; right panel). Single-cell RNA-Seq commonly reveals bimodality in gene expression and alternative isoform usage within cell populations, which is overlooked in pooled experiments. For instance, Shalek *et al* used single-cell RNA-Seq to resolve bimodal responses in the expression of key immune genes among mouse dendritic cells stimulated with lipopolysaccharide [75].

Heterogeneity in gene expression within a sample might be biologically relevant. Most obviously, it might reflect diversity of cell-types or indicate that unresolved subtypes are present in a seemingly homogenous population (as speculated in the study just mentioned [75]). These might be intermixed (e.g. in blood), spatially organized (e.g. among cortical layers) or temporally organized (e.g. among differentiating cells) (**B**). Partnering single-cell RNA-Seq with modern histology and microscopy provides powerful insight into the physical and transcriptomic architecture of complex tissues [116,117].

Long before the advent of single-cell RNA-Seq, highly precise spatial organization of lncRNAs had been observed by in *situ* hybridization, which showed that the expression of many lncRNAs is restricted to individual brain regions, structures or cell types (**C**) [15].

Evidence from single-cell RNA-Seq supports this hypothesis. Profiling individual human neocortical cells, Liu *et al* recently found lncRNAs to be expressed at comparable levels to mRNAs in individual cells but expressed in fewer cells overall. This explained parallel measurements from pooled cells, in which mRNAs were considerably more abundant (**D**) [16].

Box 3: High-throughput forward genetics for lncRNA characterization


a
CRISPR
target sequence
sgRNA
dsDNA
Cas9 enzyme
CRISPRi
dCas
transcriptional repressor (eg. KRAB)
CRISPRa
dCas
transcriptional activator (eg. VP64)
transcribed locus
b
sgRNAs agnostically tiled across noncoding regions
c
paired guide RNAs for deletion of annotated lncRNA loci
d
sgRNAs targeting annotated lncRNA promoters for CRISPRi
e
sgRNA synthesis
transduce into Cas9-expressing cells
selection
control
Relative abundance
+
-
Relative abundance
sgRNA quantification by deep sequencing

Until recently there has been no robust, scalable strategy for the genetic interrogation of lncRNA functionality. Traditional mutagenic screens generate frame-shift mutations to knockout protein-coding genes and are therefore inappropriate for lncRNAs. Likewise, RNA interference is plagued by off-target activity (exacerbated in instances of low target stoichiometry), incomplete knockdown and the difficulty of targeting nuclear-localized, highly alternatively spliced lncRNA transcripts [18,91].

Therefore, it is noteworthy that several laboratories have recently developed high-throughput phenotypic screens that utilize the CRISPR-Cas9 system, rather than mutagens or si/shRNAs, to interrogate putative functional elements in the genome [18].

The Cas9 nuclease is delivered to a specific genomic location by a single guide RNA (sgRNA), based on the latter's complementarity to the target (**A**). By introducing a large library of different sgRNAs to a pool of cells expressing Cas9, with different cells taking up different sgRNAs, many genomic sites may be independently targeted, in parallel (**B-D**). Cells are cultured and may be subject particular selective conditions. The frequency of sgRNA markers in the pool can then be measured using deep sequencing, revealing biases in cell survival/proliferation specific to individual sgRNAs. Most sgRNAs should not change in relative frequency, however deleterious sgRNAs will be relatively depleted and those that have a positive influence enriched. In this fashion, transcripts or genomic elements whose perturbation has functional consequences relevant to the selection paradigm may be identified.

CRISPR-Cas9 introduces double-stranded DNA breaks at precise genomic locations, often generating small indels at these sites (**A**; left). This is effective for the perturbation of functional protein-coding genes or noncoding elements (e.g. enhancers), which have been identified agnostically by tiling sgRNAs across noncoding regions (**B**) [118].

However, this approach is not well suited to lncRNAs, since the small indels generated by Cas9 rarely have a strong effect on their function. One strategy [96] that can be used to overcome this is to use paired guide RNAs (pgRNAs) to induce large genomic deletions between the two targeted sites, thereby removing entire lncRNA domains and/or loci (**C**). An alternative [97] is to use CRISPR interference (CRISPRi), wherein a catalytically inactive Cas9 enzyme (dCas) is fused to a transcriptional repressor (e.g. KRAB) in order to inhibit gene expression at genomic target sites (**A**; right). The initial success of these approaches implies widespread lncRNA functionality and that the design of elegant, well-directed paradigms for phenotypic screening will lead to further success.

Recently, pooled CRISPR screens have been combined with single-cell RNA-seq, directly linking sgRNA expression to transcriptome responses in thousands of individual cells and, thereby, enabling more subtle, context specific effects to be polled [119–121].

References

1 Liu, G. *et al.* (2013) A meta-analysis of the genomic and transcriptomic composition of complex life. *Cell Cycle* 12, 2061–2072

2 Morris, K. V and Mattick, J.S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet.* 15, 423–437

3 Cheng, J. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308, 1149–1154

4 Carninci, P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563

5 Wang, Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63

6 Mercer, T.R. *et al.* (2011) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* 30, 99–104

7 Sharon, D. *et al.* (2013) A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31, 1009–1014

8 Tilgner, H. *et al.* (2015) Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* 33, 736–742

9 Kolodziejczyk, A.A. *et al.* (2015) The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620

10 Stephen, S. *et al.* (2008) Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol. Biol. Evol.* 25, 402–408

11 Iyer, M.K. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* 47, 199–208

12 Ulitsky, I. and Bartel, D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell* 154, 26–46

13 Ulitsky, I. (2016) Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* 17, 601–614

14 Pheasant, M. and Mattick, J.S. (2007) Raising the estimate of functional human sequences. *Genome Res.* 17, 1245–1253

15 Mercer, T.R. *et al.* (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. U.S.A.* 105, 716–721

16 Liu, S.J. *et al.* (2016) Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol.* 17, 67

17 Deveson, I.W. *et al.* High-resolution chromosome 21 transcriptome analysis reveals distinct architecture of protein-coding and noncoding RNA. [pre-print submission in preparation]

18 Shalem, O. *et al.* (2015) High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.* 16, 299–311

19 McFadden, E.J. and Hargrove, A.E. (2016) Biochemical methods to investigate lncRNA and the influence of lncRNA:protein complexes on chromatin. *Biochemistry* 55, 1615–1630

20 Kapranov, P. *et al.* (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916–919

21 Rinn, J.L. *et al.* (2003) The transcriptional activity of human Chromosome 22. *Genes Dev.* 17, 529–540

22 Kapranov, P. *et al.* (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* 15, 987–997

23 Okazaki, Y. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573

24 Katayama, S. *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science* 309, 1564–1566

25 St Laurent, G. *et al.* (2012) Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. *BMC Genomics* 13, 504

26 van Bakel, H. *et al.* (2010) Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* 8, e1000371-21

27 Jiang, L. *et al.* (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* 21, 1543–1551

28 Hardwick, S.A. *et al.* (2016) Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Meth.* 13, 792–798

29 Clark, M.B. *et al.* (2011) The reality of pervasive transcription. *PLoS Biol.* 9, e1000625-6

30 Djebali, S. *et al.* (2012) Landscape of transcription in human cells. *Nature* 489, 101–108

31 Birney, E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the

ENCODE pilot project. *Nature* 447, 799–816

32 ENCODE (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74

33 FANTOM (2014) A promoter-level mammalian expression atlas. *Nature* 507, 462–470

34 Yue, F. *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515, 355–364

35 Brannan, C.I. *et al.* (1990) The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.* 10, 28–36

36 Cabili, M.N. *et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927

37 Derrien, T. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789

38 Quek, X.C. *et al.* (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* 43, D168-73

39 Hon, C.-C. *et al.* (2017) An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*

40 Quinn, J.J. and Chang, H.Y. (2016) Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* 17, 47–62

41 Mattick, J.S. (2009) The genetic signatures of noncoding RNAs. *PLoS Genet.* 5, e1000459

42 Lagarde, J. *et al.* (2016) Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). *Nat. Commun.* 7, 12339

43 Furuno, M. *et al.* (2006) Clusters of internally primed transcripts reveal novel long noncoding RNAs. *PLoS Genet.* 2, e37

44 Dieci, G. *et al.* (2007) The expanding RNA polymerase III transcriptome. *Trends Genet.* 23, 614–622

45 Necsulea, A. *et al.* (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505, 635–640

46 Mele, M. *et al.* (2017) Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res.* 27, 27–37

47 Guttman, M. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227

48 Alam, T. *et al.* (2014) Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes. *PLoS One* 9, e109443

49 Clark, M.B. *et al.* (2012) Genome-wide analysis of long noncoding RNA stability. *Genome Res.* 22, 885–898

50 Mukherjee, N. *et al.* (2017) Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat. Struct. Mol. Biol.* 24, 86–96

51 Kowalczyk, M.S. *et al.* (2012) Molecular biology: RNA discrimination. *Nature* 482, 310–311

52 West, J.A. *et al.* (2014) The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol. Cell* 55, 791–802

53 Washietl, S. *et al.* (2014) Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* 24, 616–628

54 Cabili, M.N. *et al.* (2015) Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* 16, 20

55 Chen, L.-L. (2016) Linking long noncoding RNA localization and function. *Trends Biochem. Sci.* 41, 761–772

56 Hutchinson, J.N. *et al.* (2007) A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* 8, 39

57 Sunwoo, H. *et al.* (2009) MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.* 19, 347–359

58 Penny, G.D. *et al.* (1996) Requirement for Xist in X chromosome inactivation. *Nature* 379, 131–137

59 Rinn, J.L. *et al.* (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311–1323

60 Dinger, M.E. *et al.* (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* 18, 1433–1445

61 Askarian-Amiri, M.E. *et al.* (2011) SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. *RNA* 17, 878–891

62 Johnson, R. *et al.* (2009) Regulation of neural macroRNAs by the transcriptional repressor REST. *RNA* 15, 85–96

63 Mercer, T.R. *et al.* (2010) Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. *BMC Neurosci.* 11, 14

64 Ng, S.-Y. *et al.* (2012) Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J.* 31, 522–533

65 Yan, L. *et al.* (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1131–1139

66 Berget, S.M. *et al.* (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. USA.* 74, 3171–3175

67 Croft, L. *et al.* (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat. Genet.* 24, 340–341

68 Wang, E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476

69 Pan, Q. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415

70 Barbosa-Morais, N.L. *et al.* (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science.* 338, 1587–1593

71 Merkin, J. *et al.* (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science.* 338, 1593–1599

72 Merkin, J.J. *et al.* (2015) Origins and impacts of new mammalian exons. *Cell Rep.* 10, 1992–2005

73 Gonzalez-Porta, M. *et al.* (2013) Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* 14, R70

74 Tilgner, H. *et al.* (2014) Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. USA.* 111, 9869–9874

75 Shalek, A.K. *et al.* (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498, 236–240

76 Marinov, G.K. *et al.* (2014) From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.* 24, 496–510

77 Welch, J.D. *et al.* (2016) Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Res.* 44, e73

78 Zhang, X. *et al.* (2016) Cell-type-specific alternative splicing governs cell fate in the developing cerebral cortex. *Cell* 166, 1147–1162.e15

79 Karlsson, K. and Linnarsson, S. (2017) Single-cell mRNA isoform diversity in the mouse brain. *BMC Genomics* 18, 126

80 Braunschweig, U. *et al.* (2014) Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* 24, 1774–1786

81 Wong, J.J.-L. *et al.* (2016) Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology. *Bioessays* 38, 41–49

82 Boutz, P.L. *et al.* (2015) Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev.* 29, 63–80

83 Ashe, H.L. *et al.* (1997) Intergenic transcription and transinduction of the human beta-globin locus. *Genes Dev.* 11, 2494–2509

84 Williamson, L. *et al.* (2017) UV irradiation induces a noncoding RNA that functionally opposes the protein encoded by the same gene. *Cell* 168, 843–855

85 Clark, M.B. *et al.* (2015) Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat. Meth.* 12, 339–342

86 Kornblihtt, A.R. *et al.* (2013) Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* 14, 153–165

87 Tilgner, H. *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, 1616–1625

88 Mercer, T.R. *et al.* (2011) Expression of distinct RNAs from 3' untranslated regions. *Nucleic Acids Res.* 39, 2393–2403

89 Flynn, R.A. and Chang, H.Y. (2014) Long Noncoding RNAs in Cell-Fate Programming and Reprogramming. *Cell Stem Cell* 14, 752–761

90 Duret, L. *et al.* (2006) The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene.

Science 312, 1653–1655

91 Goff, L.A. and Rinn, J.L. (2015) Linking RNA biology to lncRNAs. *Genome Res.* 25, 1456–1465

92 Li, L. and Chang, H.Y. (2014) Physiological roles of long noncoding RNAs: insight from knockout mice. *Trends Cell Biol.* 24, 594–602

93 Sauvageau, M. *et al.* (2013) Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* 2, e01749

94 Mattick, J.S. (2013) Probing the phenomics of noncoding RNA. *Elife* 2, e01968

95 Lewejohann, L. *et al.* (2004) Role of a neuronal small non-messenger RNA: behavioural alterations in BC1 RNA-deleted mice. *Behav. Brain Res.* 154, 273–289

96 Zhu, S. *et al.* (2016) Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nat. Biotechnol.* 34, 1279–1286

97 Liu, S.J. *et al.* (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science.* 355,

98 Engreitz, J.M. *et al.* (2016) Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* 539, 452–455

99 Smola, M.J. *et al.* (2016) SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proc. Natl. Acad. Sci. U.S.A.* 113, 10322–10327

100 Sarma, K. *et al.* (2014) ATRX directs binding of PRC2 to Xist RNA and Polycomb targets. *Cell* 159, 869–883

101 Chu, C. *et al.* (2015) Systematic discovery of Xist RNA binding proteins. *Cell* 161, 404–416

102 McHugh, C.A. *et al.* (2015) The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* 521, 232–236

103 Engreitz, J.M. *et al.* (2013) The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science.* 341, 1237973

104 Simon, M.D. *et al.* (2013) High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature* 504, 465–469

105 Mercer, T.R. and Mattick, J.S. (2013) Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* 20, 300–307

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

107 Wang, K.C. *et al.* (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472, 120–124

108 Tsai, M.-C. *et al.* (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329, 689–693

109 Di Ruscio, A. *et al.* (2014) DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature* 503, 371–376

110 Arab, K. *et al.* (2014) Long noncoding RNA TARID directs demethylation and activation of the tumor suppressor TCF21 via GADD45A. *Mol. Cell* 55, 604–614

111 Hasegawa, Y. *et al.* (2010) The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev. Cell* 19, 469–476

112 Lai, F. *et al.* (2013) Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* 494, 497–501

113 Mattick, J.S. *et al.* (2010) A global view of genomic information--moving beyond the gene and the master regulator. *Trends Genet.* 26, 21–28

114 Mattick, J.S. (2010) RNA as the substrate for epigenome-environment interactions: RNA guidance of epigenetic processes and the expansion of RNA editing in animals underpins development, phenotypic plasticity, learning, and cognition. *Bioessays* 32, 548–552

115 Mele, M. *et al.* (2015) Human genomics. The human transcriptome across tissues and individuals. *Science* 348, 660–665

116 Stahl, P.L. *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82

117 Marques, S. *et al.* (2016) Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* 352, 1326–1329

118 Sanjana, N.E. *et al.* (2016) High-resolution interrogation of functional elements in the noncoding genome. *Science* 353, 1545–1549

119 Adamson, B. *et al.* (2016) A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein pesponse. *Cell* 167, 1867–1882.e21

120 Jaitin, D.A. *et al.* (2016) Dissecting immune circuits by linking CRISPR-pooled Screens with single-cell RNA-seq. *Cell* 167, 1883–1896.e15

121 Datlinger, P. *et al.* (2017) Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Meth.* 14, 297–301

Trends Box

The mammalian transcriptome is hugely diverse, thanks to pervasive transcription and alternative splicing. Advances in RNA-Seq (e.g. targeted, single-molecule and single-cell techniques) continue to shape our understanding.

LncRNA diversity remains under-appreciated, and dynamics of expression, splicing and functional roles poorly characterized.

High-resolution and single-cell studies show lncRNAs are not lowly expressed but expressed with heightened spatiotemporal precision. LncRNAs are also enriched for splicing, with noncoding exons near-universally alternative.

Emergence of high-throughput forward-genetic screens utilizing CRISPR-Cas9 targeted genome manipulation and precise, scalable methods for resolving RNA structure and RNA-protein interactions accelerate lncRNA characterization.

Precise, dynamic expression and complex splicing fit with central role of lncRNAs in mammalian developmental program.

Outstanding Questions

The true breadth and complexity of the mammalian transcriptome remains unrealized. Targeted RNA-Seq experiments continue to unearth widespread transcription below the limit of detection for RNA-Seq. Single-molecule technologies reveal complex, organized patterns of alternative splicing that cannot be resolved with short reads. A genome-wide single-molecule transcriptome survey at saturating depth has yet to be achieved, even for a single tissue or cell-line.

Although in relative infancy, single-cell RNA-Seq has shown instances of organized, biologically relevant heterogeneity of gene expression between, and even within, cell populations. Further experiments are necessary to resolve the distribution of alternative isoforms between cells. Are alternative isoforms distributed uniformly or in an organized fashion, potentially reflecting biological differences between cells? This question is especially pertinent for lncRNAs, which exhibit highly cell-specific expression and enriched alternative splicing.

High-throughput forward genetic screens utilizing the CRISPR-Cas9 system have yielded initial success in identifying functional lncRNAs. However, selection has been limited to relatively crude (cell growth) phenotypes. The challenge now is to design elegant, well-directed paradigms for phenotypic screening that are appropriate for the subtle and highly context-specific roles enacted by many lncRNAs.

The ultimate challenge for developmental genetics is to understand not just how cell identity is defined but how cells are organized into precisely structured organs and tissues. It appears increasingly that the developmental program in complex organisms is epigenetically orchestrated and guided by regulatory RNAs. To reconcile the characteristic aspects of lncRNA biology (precise expression, promiscuous alternative splicing) with their proposed role in this program remains a key challenge for the community.

Glossary

LncRNA: Long noncoding RNA; i.e. an RNA molecule longer than ~200 nucleotides that does not contain a substantial open reading frame.

RNA-Seq: Massively parallel sequencing of cDNA molecules (typically fragmented into short sequencing fragments) derived by reverse transcription from RNA transcripts.

Targeted RNA-Seq; aka CaptureSeq: Sequencing of cDNA molecules (as per RNA-Seq) selected by hybridization to specific oligonucleotide baits. Allows enriched coverage of specific genes or genomic regions.

Forward genetics: An approach used to identify genes responsible for a particular phenotype of an organism (as opposed to reverse genetics, which studies the phenotype of an organism following disruption of a known gene).

Ultraconserved element: A region of DNA that is identical in multiple different species. Some ultraconserved elements have been found to be transcriptionally active.

Adaptive radiation: An evolutionary process in which organisms diversify rapidly from an ancestral species into a multitude of new forms.

CRISPR: Clustered regularly interspaced short palindromic repeats. A genetic element found in prokaryotes, which forms the basis of a recent genome engineering technology (CRISPR-Cas9) that enables permanent modification of genes *in vivo*.

Tiling array: A subtype of microarray chips, which probe intensively for sequences known to exist in a contiguous region.

Gene deserts: Genomic regions thought to be transcriptionally silent.

Morphogens: Signaling molecules that control cell fate specification in developing tissues.

Paraspeckles: Relatively recently discovered sub-nuclear bodies that are formed by the interaction of the lncRNA *NEAT1* and various proteins.

Transcript assembly: The process of reconstructing full-length RNA transcripts from short sequencing reads.

Intron retention: A mode of alternative splicing in which a sequence that is normally intronic is retained in the mature mRNA transcript.

Nonsense mediated decay: A surveillance pathway that exists in eukaryotes whose function is to reduce errors in gene expression by eliminating mRNA transcripts that contain premature stop codons.

Branch point: A genetic element involved in splicing located near the 3' end of the intron and just upstream of the poly-pyrimidine tract.

Ectopic expression: Abnormal gene expression in a cell type or developmental stage.

MiTranscriptome (http://mitranscriptome.org/): A catalog of human long poly-adenylated RNA transcripts derived from computational analysis of high-throughput RNA sequencing (RNA-Seq) data from over 6,500 samples spanning diverse cancer and tissue types.

Gencode catalog (https://www.gencodegenes.org/): The reference human genome annotation for The ENCODE Project.