Review

CellPress

The Landscape of long noncoding RNA classification

Georges St. Laurent[2,3], Claes Wahlestedt[4], and Philipp Kapranov[1,2]

[1] Institute of Genomics, School of Biomedical Sciences, Huaqiao Univerisity, 668 Jimei Road, Xiamen, China 361021
[2] St. Laurent Institute, 317 New Boston St., Suite 201, Woburn, MA 01801 USA
[3] Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, 185 Meeting Street, Providence, RI 02912, USA
[4] Center for Therapeutic Innovation and Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, 1501 NW 10th Ave, Miami, FL 33136 USA

Advances in the depth and quality of transcriptome sequencing have revealed many new classes of long noncoding RNAs (lncRNAs). lncRNA classification has mushroomed to accommodate these new findings, even though the real dimensions and complexity of the noncoding transcriptome remain unknown. Although evidence of functionality of specific lncRNAs continues to accumulate, conflicting, confusing, and overlapping terminology has fostered ambiguity and lack of clarity in the field in general. The lack of fundamental conceptual unambiguous classification framework results in a number of challenges in the annotation and interpretation of noncoding transcriptome data. It also might undermine integration of the new genomic methods and datasets in an effort to unravel the function of lncRNA. Here, we review existing lncRNA classifications, nomenclature, and terminology. Then, we describe the conceptual guidelines that have emerged for their classification and functional annotation based on expanding and more comprehensive use of large systems biology-based datasets.

The ncRNA universe

The classic view of the transcriptome landscape and its mRNA-centric paradigm for transcript annotation has undergone a fundamental change [1,2]. The ENCODE project (see Glossary) estimates that (mostly noncoding) transcripts cover 62–75% of our genome [3], and contribute greatly to the overall estimate of 80% potentially functional sequence in our DNA [4]. Similarly, RNAseq studies show that transcripts from these noncoding regions dominate the population of nonribosomal, nonmitochondrial RNAs in a human cell [5]. ncRNAs have emerged as a major source of biomarkers [6–10], targets for therapeutics [8,11], and potential explanations for the function of noncoding variants from genome-wide association studies (GWAS) [12]. Constant expansion of the evidence for broad functionality of ncRNAs [13], in processes ranging from heritable epigenetic change [14] to species-specific changes in cognition [15], may finally answer the long-standing question of the role of ncDNA in eukaryotic biology [16,17].

Although there has been an emphasis on the annotation and classification of ncRNAs with properties similar to those of protein-coding mRNAs, the vast majority of genomic space used for RNA production remains underexplored. Moreover, although some ncRNAs share properties with coding mRNAs, such as splicing and polyadenylation [18,19], sequence conservation [18,19], and export to the cytosol [20], many others do not, highlighting the differences in the functionalities of coding versus ncRNAs [5,11,13,21–24].

The state of noncoding annotation is still in its early days, but the field now has sufficient perspective to establish a

Glossary

5′-cap: an altered nucleotide present at 5′ ends of a eukaryotic RNA and vital for its functioning.
ENCODE project: the Encyclopedia of DNA Elements; a public research consortium launched in September 2003 by the National Human Genome Research Institute. The goal of the project is to identify all functional elements in the human genome sequence.
Endogenous retrovirus (ERV): a genomic element that was traced back to a retrovirus integrated into an ancestral genome and since retained. ERV sequences comprise ~8% of the human genome.
Expressed sequence tag (EST): a relatively short and typically partial sequence of a longer RNA molecule.
FANTOM consortium: an international research consortium established by scientists at RIKEN, Japan in 2000, initially to assign functional annotations to the full-length cDNAs collected during the Mouse Encyclopedia Project. FANTOM has since developed and expanded over time to encompass different fields of transcriptome analysis.
Genomic bin approach: an approach designed to detect differentially expressed regions of the genome in the regions where no annotation is available.
Long tandem repeat (LTR): identical pieces of DNA found at the ends of retroviruses and critical for viral life cycle. LTRs contain elements required for viral gene expression. LTRs of ERVs often retain these elements and thus can initiate or control expression of host transcripts.
Paraspeckle: a subcellular compartment that could be identified in nuclear interchromatin space.
Polycomb repressive complex 2 (PRC2): a multi-protein complex that reversibly modifies chromatin structure and silences target genes.
Tiling microarray: a microarray design (typically oligonucleotide-based) where probes interrogate an entire genomic region of interest at regular intervals agnostic of genomic annotations. This design differs from other microarrays that target only specific genomic features of interest, like exons of known genes.

Corresponding authors: Wahlestedt, C. (cwahlestedt@med.miami.edu); Kapranov, P. (philippk08@hotmail.com).
Keywords: long non-coding RNA; annotation of long non-coding RNAs; classification of long non-coding RNAs; function of long non-coding RNAs; transcriptome; systems biology; lncRNA; lincRNA; vlincRNA.

0168-9525/
 http://dx.doi.org/10.1016/j.tig.2015.03.007

logical conceptual framework for the classification of the universe of transcripts that emanate from noncoding genomic regions. New methodologies for integrated classification have accompanied a massive expansion of global transcriptome datasets, particularly from genomics consortia such as FANTOM [25], ENCODE [3], and GTEx [26]. Methods for grouping RNA sequencing (RNA-seq) (Box 1) reads into a single transcribed region have improved rapidly [27–30]. Progress has also been made in machine-learning approaches aimed at the integration and biological interpretation of diverse datasets [29,31,32]. All this has culminated in widely used sets of annotations of lncRNAs, such as the one provided by the GENCODE consortium [33], and others [31,34–37].

Here, we review existing lncRNA classes and then describe the conceptual guidelines that have emerged for their classification and functional annotation based on the expanding and more efficient use of large systems-biology-based datasets. This framework endeavours to guide researchers in the classification of ncRNAs and interpretation of next-generation sequencing (NGS) data, especially in noncoding portions of the genome.

Box 1. Overview of high-throughput technologies used to detect and quantify ncRNAs

RNA sequencing (RNA-seq): currently, one of the most commonly used procedures in transcriptome profiling. Typically, RNA is converted into cDNA using random hexamers followed by massive random sequencing of the resulting cDNAs using NGS technologies. As a result, millions of short sequence tags can be generated per experiment. Subsequent mapping of the tags reveals the genomic position encoding the RNA and its relative mass in the cell. The procedure is suitable for various aspects of transcriptome research: RNA mapping, quantitation, alternative splicing analysis etc.

mRNA sequencing (mRNA-seq): RNA-seq on polyA+ fraction of RNA, often synonymous with RNA-seq.

Direct RNA sequencing (DRS): sequencing of native RNA, without library preparation including cDNA conversion step [143], has been successfully used to sequence native polyA+ and identify alternative polyadenylation sites. DRS is particularly useful in applications where artifacts of reverse transcription are undesirable, such as precise strand of origin determination, and in applications that deal with minute amounts of nucleic acids such as single-cell applications. Theoretically, it can provide multiple tags per molecule, however so far it has been used in applications that provide a single tag per molecule at the polyadenylation site.

Cap assisted gene expression (CAGE): a transcriptome profiling procedure that targets RNAs with a 5′-cap [144]. CAGE generates short (typically 27 nucleotides) sequence tags from 5′ ends of such RNAs, with one tag per RNA molecule. It enables accurate mapping of 5′ ends this subset of RNAs.

Serial analysis of gene expression (SAGE): targets polyadenylated messages and generates a single internal (typically close to the 3′ end) tag per RNA molecule [145].

Paired-end tag (PET): also targets polyadenylated RNAs and generates a tag that combines information on 5′ and 3′ ends of the same RNA molecule [146].

Rapid amplification of cDNA ends (RACE): an 'outward' PCR-based method designed to identify sequences connected to a given region, which can be used in conjunction with NGS or microarrays, for deep transcriptome profiling of a specific locus [44].

Targeted RNA sequencing: selection of RNAs from a locus of interest using tiling microarrays followed by RNA-seq to achieve the same goal [45].

GRO-seq: A typical RNA profiling experiment measures steady-state levels of RNA. By contrast, GRO-seq [134] combines nuclear run-on experiments and NGS analysis to provide information on transcription competent RNA polymerase complexes.

Criteria and features of existing classes and categories of lncRNAs

The classification of the great majority of lncRNAs relies on the empirical attributes originally used to detect them (Table 1, Figure 1). This reflects their short history relative to protein-coding genes, and provides a convenient basis by which to classify these uncharacterized RNA species.

Classification based on transcript length

The length estimate of ncRNAs serves as the most commonly used attribute for their classification. Typically, a threshold of 200 bases separates long from short ncRNAs [38,39] (Table 1). Often, our knowledge is limited to sequence reads mapped to a 'region of transcription', and even with improvements in NGS read length [40], this will probably continue for the foreseeable future.

Building transcribed regions based on RNA-seq profiling of total RNA (rather than the polyA+ fraction, see below) led to the discovery that intergenic space encodes thousands of very long intergenic ncRNAs (vlincRNAs), whose primary transcripts can range in length from 50 kb to 1 Mb [28,29,41]. Spanning at least 10% of the human genome [5,29], vlincRNAs have been implicated in important biological processes such as pluripotency [29], cancer [28,29], apoptosis [29], cell cycle progression [28,42], and cellular senescence [41].

Classification based on association with annotated protein-coding genes

This commonly used attribute (Table 1, Figure 1) serves as the foundation of the GENCODE classification of lncRNAs [33]. It underlies the logical challenge of overlapping noncoding and coding transcripts at a given locus – called 'transcriptional forests' by the FANTOM consortium [43]. Targeted methods (Box 1) based on rapid amplification of cDNA ends (RACE) experiments [44] and RNA-seq [45] indicate that transcriptional forests constitute a general feature of the human genome. A prominent category of ncRNAs has emerged from these transcriptional forests composed of sense ncRNAs that overlap coding mRNAs on the same strand and share some sequence with the latter, yet do not encode proteins [44,46–48]. This category includes unspliced sense partially intronic RNAs (PINs) [49], and spliced transcripts that combine exons from coding and noncoding regions of a gene [47,48]. GENCODE recognizes the existence of such spliced lncRNAs in their 'sense overlapping' biotype [33].

The PIN and sense overlapping categories allow for overlap between lncRNAs and exons of a protein-coding gene. However, a protein-coding gene can produce lncRNAs found exclusively in its introns, known as totally intronic RNAs (TINs) [49] (Table 1, Figure 1). TINs make up the majority (~70%) of all non-coding (non-rRNA) nuclear-encoded RNA and 40–50% of all cellular (non-rRNA) RNAs by mass, as established by single-molecule sequencing [50]. Evidence that large numbers of introns encode stand-alone RNAs originally came from microarray expression profiling

Table 1. Different known classes of lncRNAs

Category	Abbreviation	Refs	Specific examples
Classification based on transcript length			
Long noncoding RNA	lncRNA	[38,39]	
Long-intergenic noncoding RNA; large intervening noncoding RNA, long-intervening noncoding RNA	lincRNA	[18]	ANRIL [117], H19 [147], HOTAIR [18], HOTTIP [148], lincRNA-p21 [149], XIST [150], Paupar [151]
Very long intergenic noncoding RNA	vlincRNA	[29]	HELLP transcript [42], Vlinc_21, vlinc_185, vlinc_377, vlinc_500 [29]
macroRNA		[28,152]	Airn, Gtl2lt, KCNQOT1, Lncat, Nespas (reviewed in [152]), STAiR1 [28]
Promoter-associated long RNA	PALR	[38]	
Classification based on association with annotated protein-coding genes			
Intronic ncRNA; stable intronic sequence RNA; totally intronic RNA, partially intronic RNA	sisRNA, TIN, PIN	[49,50,54] additional references in the text	
Circular intronic RNAs	ciRNAs	[55]	
Sense ncRNA		[44]	
Natural antisense ncRNA	asRNA, NAT	[57]	BACE1-AS [153], aHIF [154], Tsix [155]
Mirror antisense		[44,67]	Globin antisense [67]
Exonic circular RNAs	ecircRNAs	[62]	cANRIL [118]
Chimeric RNAs, trans-spliced RNAs, exon juxtaposition		[44,63–65]	
Stand-alone ncRNAs made from 3′UTRs	uaRNA	[60]	
Chromatin-interlinking RNA	ciRNA	[68]	
Transcription start site-associated RNAs	TSSa-RNAs	[156]	
Classification based on association with other DNA elements of known function			
Enhancer-associated RNA	eRNA	[157]	
Promoter-associated long RNA	PALR	[38]	
Upstream antisense RNA	uaRNA	[158]	
PROMoter uPstream Transcript	PROMPT	[89]	
Telomeric repeat-containing RNA	TERRA	[159]	
Classification based on protein-coding RNA resemblance			
mRNA-like noncoding RNAs	mlncRNAs	[18]	
Long-intergenic noncoding RNA; large intervening noncoding RNA, long-intervening noncoding RNA	lincRNA	[18]	ANRIL [117], H19 [147], HOTAIR [18], HOTTIP [148], lincRNA-p21 [149], XIST [150]
Classification based on association with repeats			
C0T-1 repeat RNA		[160]	
Long interspersed nuclear element	LINE1/2	[161]	
Transcribed endogenous retroviruses		[81]	
Expressed Satellite Repeats		[162]	
Non-coding RNA driven by promoters within repeats	vlincRNAs, NASTs	[29,76]	Vlinc_21, vlinc_185, vlinc_377, vlinc_500 [29]
Polypurine-repeat-containing RNA	GRC-RNA	[163]	
Transcribed pseudogenes		[83]	PTENP1 and KRASP1 [86]
Classification based on association with a biochemical pathway or stability			
Nrd1-unterminated transcript	NUT	[164]	
miRNA primary transcripts		[165]	H19 [166]
piRNA primary transcripts		[167]	
Cryptic unstable transcript	CUT	[88]	
PROMoter uPstream Transcript	PROMPT	[89]	
Xrn1-sensitive unstable transcript	XUT	[91]	
Stable Uncharacterized Transcript, Stable Unannotated Transcript	SUT	[92]	
Classification based on sequence and structure conservation			
Transcribed-ultraconserved regions	T-UCR	[95]	UCR106 [95]
Hypoxia-induced noncoding ultraconserved transcript	HINCUT	[100]	
Long-intergenic noncoding RNA; large intervening noncoding RNA, long-intervening noncoding RNA	lincRNA	[18]	HOTAIR [18], HOTTIP [148]
RNA-Z regions		[97]	
EvoFold regions		[98]	

Table 1 (*Continued*)

Category	Abbreviation	Refs	Specific examples
Classification based on expression in different biological states			
Long stress-induced noncoding transcript	LSINCT	[101]	
Hypoxia-induced noncoding ultraconserved transcript	HINCUT	[100]	
Non-Annotated Stem Transcript	NAST	[76]	
Classification based on association with subcellular structures			
Chromatin-associated RNA	CAR	[102]	
Chromatin-interlinking RNA	ciRNA	[68]	
Nuclear bodies associated RNAs		[168]	
PRC2 associated RNAs		[19,103]	
Classification based on function			
Long noncoding RNAs with enhancer-like function; ncRNA-activating	ncRNA-a	[108]	ncRNA-a7 [108]
miRNA primary transcripts		[165]	H19 [166]
piRNA primary transcripts		[167]	
Competing endogenous RNA	ceRNA	[109]	PTENP1 and KRASP1 [86]

[49,51] and *in silico* analysis of expressed sequence tag (EST) databases [49,52]. Even genomes as compact as those of human viruses can encode functional intronic RNAs [53]. Large numbers of stand-alone intronic RNAs recently found in *Xenopus* oocytes [54] and mice [50] support the conclusion that introns encode functional ncRNAs on a global scale. Some of these transcripts likely represent circular intronic ncRNAs (ciRNAs) (produced from introns that escape debranching) that can accumulate in cells and regulate expression of their parent genes [55] (Figure 1). Overlap on the opposite DNA strand from their associated protein-coding gene represents another frequently used attribute of lncRNAs. These natural antisense transcripts (NATs) (Figure 1) occur in 50–70% of all protein-coding genes [56,57].

ncRNAs can also be composed solely of sequences of exons of protein-coding mRNAs (Figure 1, Table 1). For example, transcript cleavage followed by post-transcriptional 5′-cap addition [58,59] can result in production of stand-alone ncRNAs from various parts of mRNAs [58], notably 3′ untranslated regions (UTRs) [60]. In fact, the type 0 variant of the cap structure may associate with the post-transcriptionally capped 5′ ends [61]. Additional cellular processes could produce this type of ncRNA, such as the reverse splicing implicated in the production of circular exonic RNAs [62], trans-splicing leading to production of chimeric RNAs [63,64], exon juxtaposition [65], and presumed RNA copying, leading to production of 'mirror antisense' transcripts [44,66,67]. Finally, RNAs whose sequences have features of bona fide coding transcripts may have other roles as revealed by the class of chromatin-interlinking RNAs (ciRNAs). These RNAs participate in maintaining interphase chromatin configuration and mostly include spliced transcripts with long 3′ UTRs [68].

Classification based on association with other DNA elements of known function

Notable classes of such RNAs include enhancer- and promoter-associated long RNAs (Table 1, Figure 1). These long RNAs are involved in linking the dynamics of nuclear architecture, chromatin signalling plasticity, and transcriptional regulation [69]. Interestingly, enhancers that give rise to RNA species have greater likelihood of functionality in reporter assays than those that do not [70], arguing for a functional, rather than spurious, link between RNA and this type of genomic element.

Classification based on mRNA resemblance

As mentioned previously, research has focussed on ncRNAs with a spliced structure, conserved sequence, and a polyA tail [18,19,34,35,71] (Figure 1). In fact, lncRNAs annotated by GENCODE – even those solely confined to intronic sequences – represent primarily spliced transcripts [33]. These features were used to identify thousands of transcripts in mice [18] and humans [19], called long intervening ncRNAs (lincRNAs) [18]. This approach has revealed many important functional lncRNAs, such as HOTAIR, which mediates gene silencing by facilitating localization of the epigenetic repressor Polycomb repressive complex (PRC)2 to its target sequence [72]. Expression analysis of ~10 000 human lincRNAs across 1300 tumour samples using microarrays revealed hundreds of noncoding transcripts potentially driving four different cancer types [73]. Numerous other studies have implicated lincRNAs in human development and disease [74]. As an indication of their functionality, expression analysis of 11 tetrapod species found 2508 lincRNAs expressed in at least three species and originating >90 million years ago [71].

Classification based on association with repeats

About half of the human genome consists of repeats of various categories, and many ncRNA-encoding genomic regions overlap these elements (Figure 1). Promoters within repeats drive expression of many ncRNAs [75], especially in pluripotent [29,76,77] and cancerous cells [29]. Promoters within the long tandem repeats (LTRs) of endogenous retroviruses specifically associate with ncRNAs from several classes of nonannotated stem transcripts (NASTs) [76] in pluripotent stem cells, including lincRNAs [77] and vlincRNAs [29] (Table 1). In addition, LTR-driven vlincRNAs identify common regulatory architectures between stem and cancer cells [29]; an interesting reminder of ideas from the stem cell theory of cancer [78].

Individual copies of repeats are expressed from their own promoters and contribute to the ncRNA transcriptome. For example, RNA polymerase (Pol) III transcribes

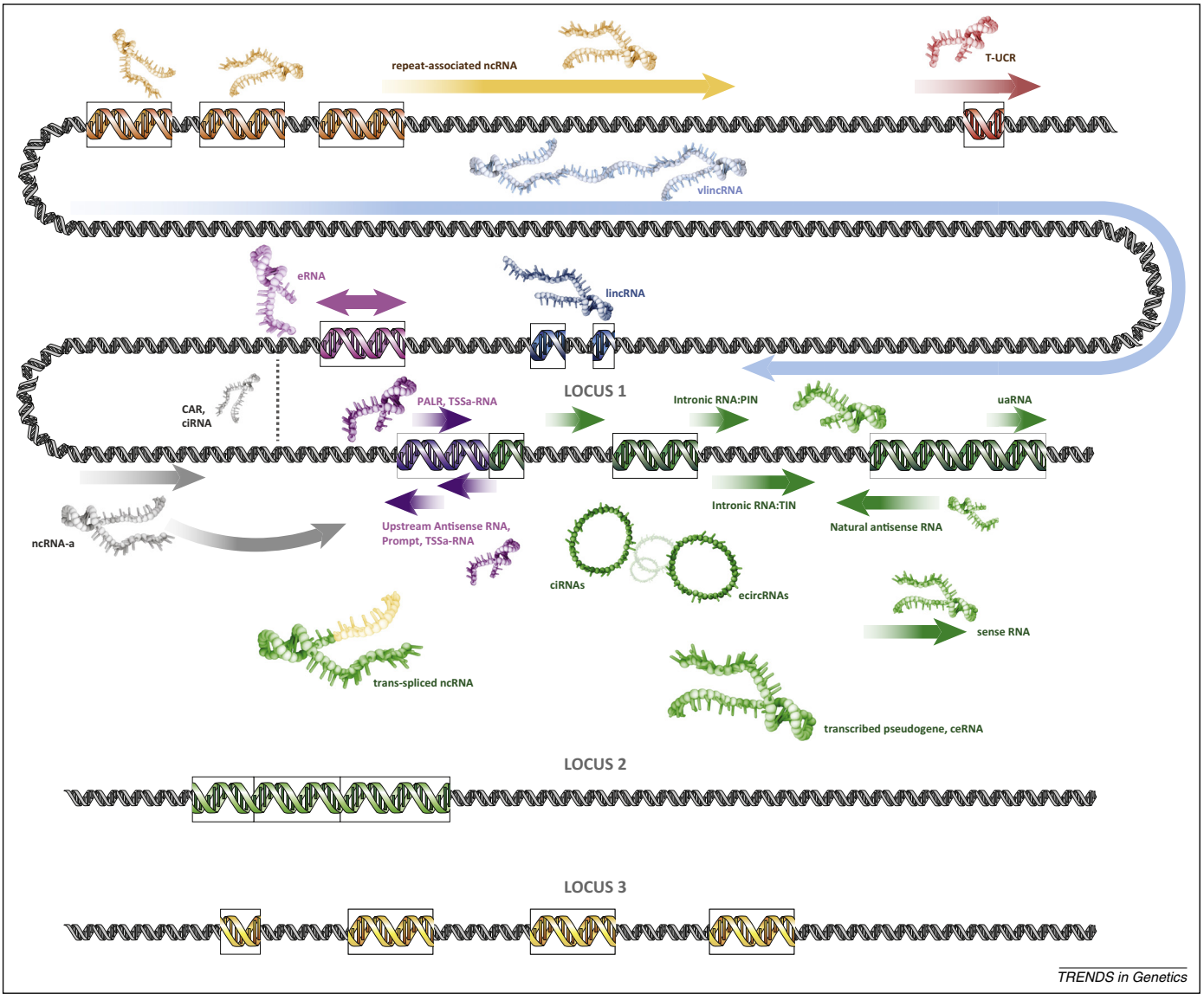

Figure 1. Schematic diagram illustrating various classes of ncRNAs. Three hypothetical loci are shown. Protein coding exons are shown as green (locus 1) or yellow boxes (locus 3). Locus 2 signifies a pseudogene of locus 1. Regulatory regions of locus 1 are shown in purple (promoter) and magenta (enhancer). Repeats are denoted by brown boxes. Lines with arrows represent ncRNAs. The role depicted here for CARs and ciRNAs in stabilising a chromatin loop is hypothetical. Abbreviations: CAR, chromatin-associated RNA; ceRNA, competing endogenous RNA; ciRNA, chromatin-interlinking RNA (grey) or circular intronic RNA (green); eRNA, enhancer-associated RNA; ecircRNA, exonic circular RNA; lincRNA, long intervening non-coding RNA; ncRNA, noncoding RNA; ncRNA-a, activating ncRNAs; PALR, promoter-associated long RNA; PIN, partially intronic RNA; TIN, totally intronic RNA; TSSa-RNA, transcription-start-site-associated RNA; T-UCR, transcribed ultraconserved regions; uaRNA, 3′ UTR-derived RNAs; vlincRNA, very long intergenic ncRNA.

noncoding repeated elements, such as Alu, B1, and B2, that can bind to RNA Pol II and affect its activity in response to stress [79]. Long interspersed nuclear elements (LINEs) comprise 20% of the genome and express mostly noncoding transcripts due to 3′ truncation and accumulated mutations [80]. Similarly, expression of noncoding endogenous retroviruses (ERVs) is a well-documented phenomenon [81]. Finally, examination of transcripts containing repeat copies continues to reveal additional regulatory functions for these molecules, as exemplified by Alu-mediated intermolecular interaction of coding and ncRNAs in *trans* described recently [82].

Transcripts from a specific subset of repeated sequences – noncoding copies of protein-coding genes or mRNAs (pseudogenes) [83] – have gained prominence upon realisation that they can function in various ways [84], including binding and titrating regulatory molecules that normally interact with the functional copies [85,86]. Moreover, pseudogenes can be transcribed from the opposite strand thus producing transcripts capable of intermolecular interaction with the productive copy [87] or its promoter [85].

Classification based on a biochemical pathway or stability

Classification of ncRNA based on their association with substrate pools of different RNA degradation pathways and enzymes has recently gained popularity. Inhibition of components of the exosome (RRP6, RRP40, and RRP44) or nonsense-mediated decay (XRN1) has revealed populations of ncRNAs not easily observed in wild type cells [88–92] (Table 1). This approach also provides information

about the pathways of their metabolism. The latter is another attribute used for classification of ncRNAs, as exemplified by XUTs (Xrn1-sensitive unstable transcripts) [91] (Table 1). Most of the pathways analysed so far in this classification involve RNA degradation, and these lncRNAs overlap with classes of NATs [91] and promoter-associated RNAs [89,90,92].

Classification based on sequence or structure conservation

Sequence conservation, though highly informative in predicting functional protein-coding mRNAs, remains a metric with controversial merit in the noncoding space. Its absence – typical for lncRNAs [93] – does not universally imply lack of functionality [22,24]. Still, many ultra-conserved regions (UCRs) – sequences of DNA 100% conserved in humans, rats, and mice – map to the noncoding space of the genome [94]. A large number of UCRs are transcribed as ncRNAs, and some are associated with malignant states [95]. As secondary RNA structure plays a crucial role in ncRNA function [24,96], a number of bioinformatics approaches such as RNA-Z [97] and EvoFold [98] leverage structure rather than sequence conservation to predict ncRNA-encoding regions (Table 1) [99].

Classification based on biological states

A number of cancer-associated transcribed UCRs (T-UCRs) encoding ncRNAs were induced by hypoxia and thus further subclassified as hypoxia-induced noncoding ultraconserved transcripts (HINCUTs) [100]. They serve as an example of another attribute used for ncRNA classification: induction after treatment with a stimulus or association with a certain biological state. Another example is long stress-induced noncoding transcripts (LSINCTs) [101].

Classification based on subcellular localisation

RNA localisation can provide important clues to its function. ncRNAs tend to be enriched in the nucleus [38,56], which suggests their involvement in the temporal–spatial regulation of nuclear architecture. For example, chromatin-associated RNAs (CARs) – both intronic and intergenic – form an integral component of chromatin, with the potential to regulate the expression of nearby genes [102] (Figure 1). The ENCODE consortium performed extensive profiling of three subnuclear compartments (chromatin, nucleolus, and nucleoplasm) revealing their RNA compositions [3]. Within the nucleus, ncRNA association with, and targeting of, the gene-silencing PRC2 complex led to the identification of thousands of PRC2-associated ncRNAs in mouse embryonic stem cells [103] and human cell lines [19]. ncRNAs form components of other nuclear subcompartments such as paraspeckles, the nucleolus, and the nuclear matrix [104]. Presumably, additional classes of ncRNAs associated with these and other compartments likely await discovery. Interestingly, some lncRNAs localise to the cytosol [105] and actually associate with ribosomes. Even the small mitochondrial genome encodes lncRNAs [106], underscoring the variety of different processes in which these transcripts could participate (see below).

Classification based on function

lncRNAs can participate in a plethora of different cellular processes: chromatin remodelling, regulation of transcription and translation, RNA stability, scaffolding, and innate immunity. We discuss here only examples of functions used for classification, and direct the reader to other reviews focusing on lncRNA molecular mechanisms [6,7,13,24,39,107].

Activating ncRNAs (ncRNA-a), which have enhancer-like properties, represents an example of classification based on function (Table 1). This class is distinguished from enhancer RNAs (eRNAs) [108] by positively regulating nearby genes (Figure 1). One notable member of this class, designated ncRNA-a7, regulates the Snai1 transcription factor. Depleting ncRNA-a7 leads to major phenotypic changes at both cellular and molecular levels [108]. The category of ncRNA-a will probably continue to grow, as the accumulation of high-quality expression datasets identify more lncRNAs that positively correlate with nearby genes (St. Laurent *et al.*, unpublished).

Another example is competing endogenous RNAs (ceRNAs) [109]. They share sequence similarity with protein-coding transcripts and function by competing for regulatory molecules [109]. Any ncRNA sharing a sequence with another (coding or noncoding) RNA could potentially be a ceRNA, such as transcribed pseudogenes, which represent important ceRNAs [86] (Figure 1). Conceivably, the ceRNAs could form part of a complex regulatory matrix driven by differential affinity among many contextually associated RNA molecules [110].

Some lncRNAs serve as precursors for shorter functional RNAs as exemplified by primary transcripts for mi- and piwi-interacting RNAs (piRNAs) (Table 1). In fact, the long and short cleavage products could have distinct functions, as demonstrated by short noncoding tRNA-like molecule produced during maturation of metastasis associated lung adenocarcinoma transcript 1 (MALAT1) lncRNA [111]. The ENCODE consortium estimates that ~6% of all annotated coding and noncoding transcripts overlap with short RNAs [3]. A recent report suggests that an 18-nucleotide short RNA produced from a coding mRNA regulates translation [112]. Also, ncRNAs derived from 3′ ends of mRNAs associate with Argonaute proteins, suggesting they represent novel regulatory molecules [90]. Short RNAs derived from protein-coding transcripts can also mediate transgenerational silencing of the parent gene [14]. Conceivably, cleavage could also generate functional lncRNAs from a longer precursor ncRNA, where both the precursor and the product could have different functions.

Finally, we note that not every lncRNA transcript functions solely as a noncoding element. Peptide sequencing data revealed presence of 250 novel mouse peptides encoded by presumed lncRNAs [113]. The full extent of the novel mammalian proteome encoded by lncRNAs is not yet clear however. Although many lncRNAs appear to associate with ribosomes [105,114], this frequently does not result in protein synthesis [115,116], but instead could reveal lncRNA regulation of translation [105]. Nonetheless, the protein-coding potential is currently used as one metric in lncRNA definition [37].

Challenges of current lncRNA classification

As described above, the existing classifications of lncRNAs rest on their descriptive and distinctive properties: from their size, to their localization, to their function. For example, the GENCODE system, as one of the few practical and up-to-date classifications available, also classifies lncRNAs into antisense RNA or lincRNA, in addition to intron-associated biotypes [33]. Although logical principles guide these classifications, they have inherited a number of unavoidable shortcomings. First, the existing classes capture a small fraction of lncRNAs present in the cell as illustrated in Figure 2. The various lists of lncRNAs annotated based on resemblance to protein-coding mRNAs account for only 0.05–1.12% of cellular RNA (Figure 2), while functional intronic RNAs could constitute as much as 16% [50]. Second, the overlap between multiple existing annotations of lncRNAs derived by different groups is small [33]. Third, the descriptions of the classes in the current annotation schemes can be vague. For example, a lncRNA could initiate at an enhancer element or initiate a large distance away and merely overlap it, yet currently they would both be classified as eRNAs. Fourth, the existing classes are not mutually exclusive. Thus, a lincRNA could theoretically also classify as an eRNA, and an LSINCT, and a CAR, and a T-UCF. For example, ANRIL is a lncRNA [117], a NAT [117], and a circular RNA [118]. This point is particularly problematic, as few datasets cover all of these characteristics and therefore many ncRNAs are not comprehensively assessed. Fifth, they lack systematisation: following the current schemes, in the future one might expect hundreds of overlapping classes of ncRNAs as new knowledge is incorporated into the classification. There are already at least 50 associations with a multitude of biological or biochemical processes (Table 1). Sixth, an attribute used in the current classifications may decline over time in relevance or utility. Considering the growing role of trans regulation by ncRNAs via intermolecular interactions [42,82,119], the fact that a lncRNA associates with an enhancer, promoter, or intron, or is antisense to a known gene may not reflect the actual function of that ncRNA. Instead, the latter could function by interacting with transcripts derived from elsewhere in a genome.

The consolidated conceptual framework of lncRNA classification

The concepts driving lncRNA classification have begun to benefit from recent developments in the annotation of noncoding transcripts and the dramatically improved techniques for measuring them. Below, we review the conceptual components that provide the basis for these ongoing improvements (Figure 3).

Tier 1: mapping the longest unprocessed transcript

The fact that the existing lists of lncRNAs miss a large fraction of ncRNA mass (Figure 2) argues that the annotation process has to start at a higher level. The first logical step in this effort is mapping the longest non-coding transcript (Figure 3).

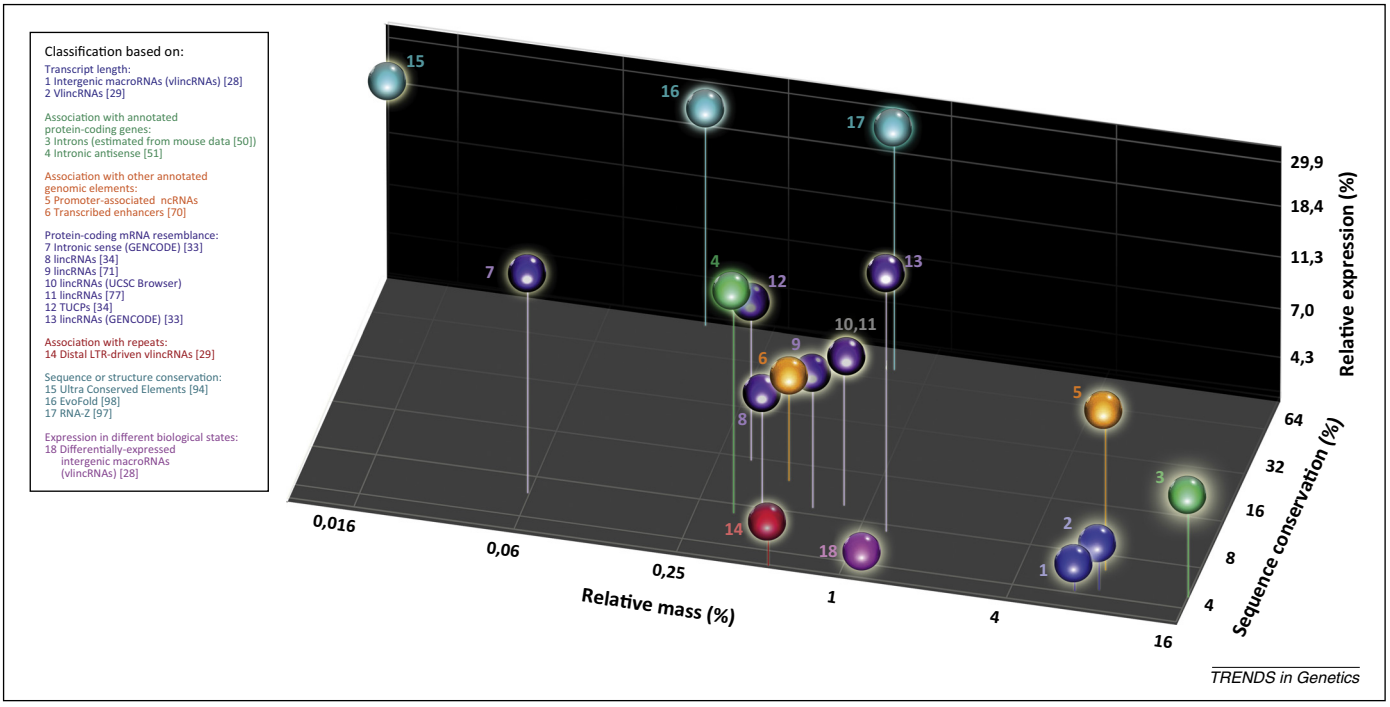

Figure 2. Properties of different published lists of human transcripts representing various classes of ncRNAs. Sequence conservation was defined by the conserved elements from the Vertebrate MULTIZ Alignment & Conservation (100 Species) database from the University of California Santa Cruz (UCSC) Genome Browser [169]. Relative conservation represents the fraction of conserved bases relative to the total lengths for each list of ncRNAs. Relative mass and expression levels represent averages of several malignant and normal tissues profiled using single-molecule RNA-seq analysis [5,29]. Only non-ribosomal RNA reads uniquely mapping to the nuclear genome were considered. Relative mass represents proportion of reads mapping to a particular genomic element relative to all reads. The relative expression is the relative mass divided by the total length of each list and normalized to the relative expression of coding exons (defined by UCSC Genes). Promoter-associated RNAs were defined by the regions 3 kb upstream of annotated start sites of UCSC Genes. Given the lack of a comprehensive list of standalone human intronic RNAs, we extrapolated the relative mass of those based on mouse data [50]. The GENCODE annotations [33] are based on v19. Abbreviations: lincRNA, long intervening non-coding RNA; LTR; long tandem repeat; ncRNA, noncoding RNA; TUCP, transcript of uncertain coding potential; vlincRNA, very long intergenic non-coding RNA.

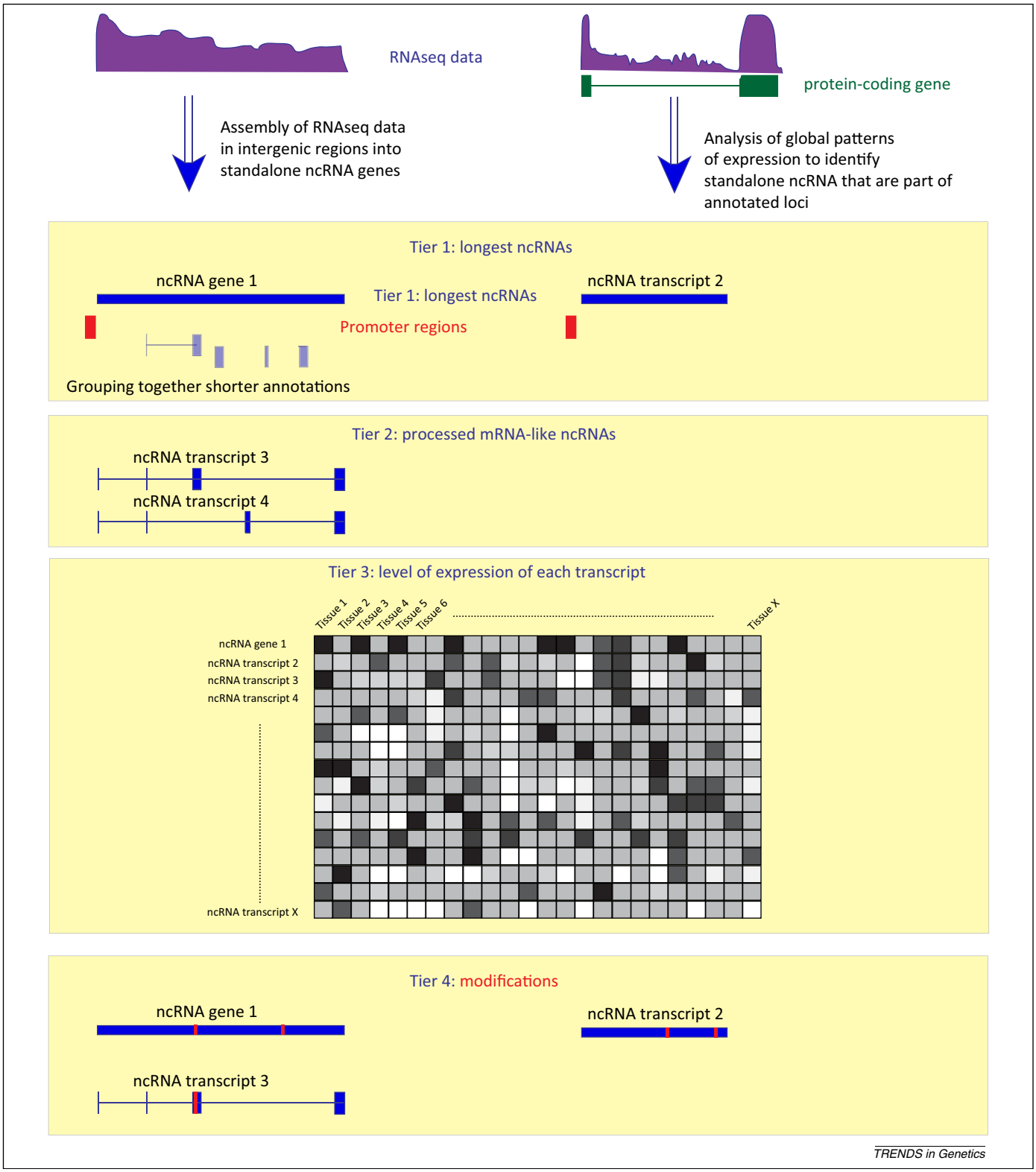

Figure 3. Outline of the consolidated conceptual framework of ncRNA classification. Highly accurate empirical RNA-seq data drives both annotation and quantification of the longest ncRNA (Tier 1) and of processed ncRNA species (Tier 2) across the entire genome. The quantitation data serves as the basis for the combined global matrix of knowledge of expression of each (coding and noncoding) RNA gene and transcript across multiple biological sources (Tier 3). This information provides the input for the functional annotation of non-coding transcripts using systems biology approaches. Mapping of RNA modifications provides the final layer of knowledge in this scheme. Abbreviations: ncRNA, noncoding RNA.

Subdividing the intergenic space into standalone ncRNA loci (genes) has obvious benefits. First, it allows for consolidation of disparate and often incomplete ncRNAs represented by ESTs, lincRNAs, and mRNAs, into a single locus. As an illustration, the clinically important 8q24 region upstream of the *MYC* gene contains a number of different lncRNA elements [12] (Figure 4). Given the distances that separate them, it may not be obvious that

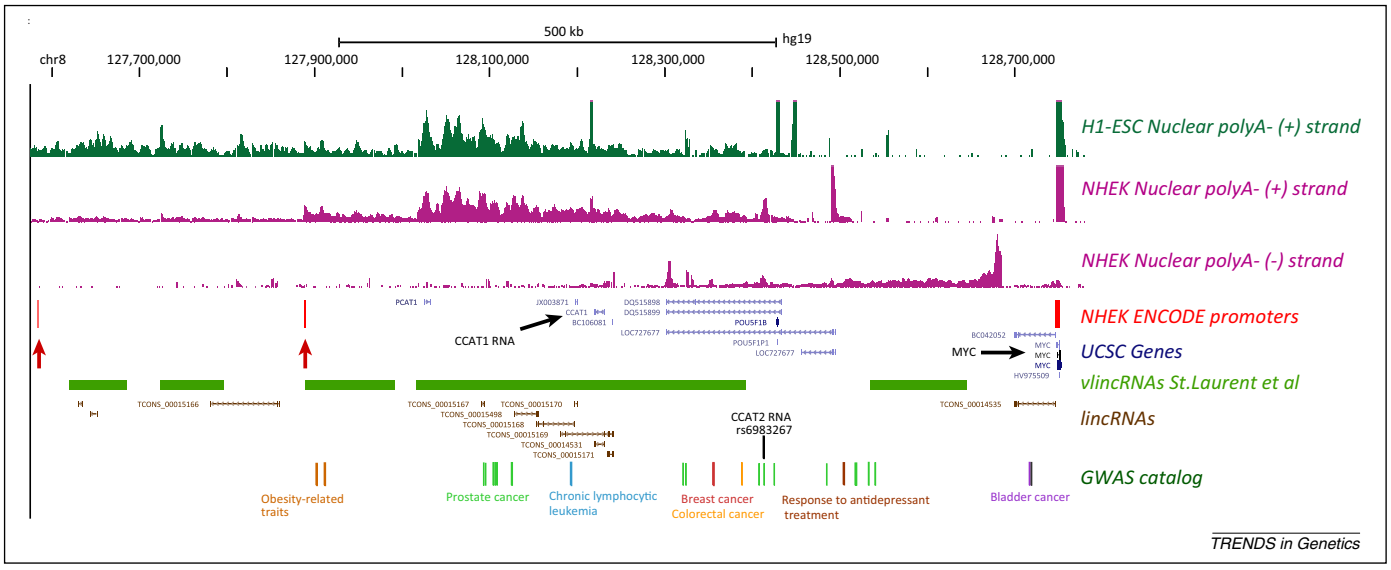

Figure 4. A genomic view of the 8q24 region upstream of the human *MYC* gene. This clinically important locus containing many GWAS hits associated with several cancers represents an example of a genomic region that could clearly benefit from the new annotation scheme. The RNA-seq analysis reveals fairly strong signal on both strands covering most of this >1 Mbp region. Yet, the known lncRNA annotations represent only a small fraction of this locus and judging by the distribution of the RNA-seq signal and known promoters, are likely part of much larger transcript units (for example vlincRNAs shown on the figure). Transcriptome RNA-seq data are represented by the polyA- nuclear RNA from normal epidermal keratinocytes (NHEK) and embryonic stem cells (H1) generated by the ENCODE consortium [3]. In addition, vlincRNAs [29], promoters [32], and disease-associated variants from GWASs [170] are shown. Abbreviations: GWAS, genome-wide association study; lncRNA, long noncodingRNA; vlincRNA, very long intergenic noncoding RNA. Reproduced, with permission, from [12].

these annotations are part of the same transcript, yet the RNA-seq signal clearly groups them together into one locus associated with a specific regulatory region (Figure 4). Second, such a grouping would allow experiments to focus on the locus rather than its many different genomic elements, allowing for seamless integration of the data from independent experiments. Third, it would clarify the issue of RNA association with different genomic features, for example enhancers, by showing whether the transcripts originate from these DNA elements, or merely overlap them. Overall, the longest transcripts would serve as scaffolds to bring together all the disparate annotations into gene-like structures with their own specific transcription regulatory regions, helping to resolve the problem of overlap. In this case, promoter information and CAGE tag data (Box 1) [25] would help in both assessing the quality of the map and understanding the regulation of such genes.

A lncRNA may not always be produced from its own dedicated promoter, as exemplified by circular intronic RNAs [55]. Such stand-alone functional intronic RNAs would however have certain features, such as low correlation with other exons or introns of the same gene, relatively high expression levels with low variance, and occasionally, differential expression in a biological time course [50]. These properties can now be measured by highly quantitative analysis of RNA levels across multiple diverse samples. Thus, defining standalone transcripts would require an additional dimension – quantitative measurement to allow for analysis of coexpression with multiple neighbouring transcripts.

Tier 2: defining processed transcripts

The transcriptional forest concept implies that multiple RNA species share the same genomic space, either transcribed independently or derived by processing of longer precursors [38,43,120]. Mapping sites of polyadenylation [121,122] provides additional information on completeness of such maps. Application of highly sensitive methods targeted to specific regions using RACE [47] or capture-sequencing [45] would increase the discovery of processed species. Multiple levels of processing exist, such as A to I editing [123,124] and others [125] each under their own regulatory control.

Tier 3: the additional dimension of expression levels

In the past, genomic coordinates alone determined genomic annotation. However, in the case of overlapping transcripts it cannot predict which isoforms likely function in a particular tissue. Thus, progress of our understanding of the complexities of the transcriptome (Box 1) argues for an additional dimension – expression of each RNA produced at a given locus (Figure 3). The pioneering efforts by the FANTOM Consortium [25] make this undertaking possible.

Tier 4: RNA modifications

A map of all (>100) RNA modifications [125] constitutes the final layer of annotation (Figure 3). These patterns could represent an information rich source for distinguishing RNA molecules, thus assisting in their classification. So far, technological limitations prevent us from efficient genome-wide mapping of most RNA modifications, and assaying technically accessible modifications is fraught with pitfalls [124], such as false discovery due to technological noise [126]. Hopefully, existing [127] and emerging technologies [128] will enable progress in cataloguing RNA modifications.

From consolidated conceptual framework to function

The first key of the new framework is consolidation, achieved by grouping disparate lncRNA transcripts into genes or standalone Tier 1 transcripts. The second aspect

moves from phenomenological description of lncRNAs to their genomic coordinates, which parallels the evolution of the concept of the gene [129]. The third aspect uses empirical data to unravel the layers of overlapping transcripts, as exemplified by the study of intronic RNAs in mice [50].

The fourth aspect assigns functional weight to a lncRNA by integrating information from diverse high throughput methods [130]. Among others, these methods include cross-linking immunoprecipitation (CLIP)-seq [131] for detection and measurement of RNA – protein interactions, selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE)-seq [132] for analysis of RNA secondary structure, chromatin isolation by RNA purification (ChIRP)-seq [133] for measurement of RNA–chromatin interactions, and global run-on (GRO)-seq (Box 1) to measure transcription [134]. This multifaceted approach combines independent sources of evidence for the functionality of an RNA molecule, underscoring the complexities of lncRNA involvement in the flow of biological information [24,135]. Fortunately, new machine-learning methods have emerged to identify and decipher complex patterns in the data, yielding probabilistic evaluation of ncRNA function across large populations of transcripts [71,136].

The evolution of the conceptual framework of ncRNA classification described above provides a roadmap for the analysis of an RNA-seq experiment and its integration into a broader knowledge base of high-throughput multidimensional information. The availability of a common set of genomic coordinates for various stages of ncRNA processing (Figure 3, Tiers 1 and 2) provides the key resource that enables the integration of data from multiple experiments. Tiers 3 and 4 assist in refinement of the classification by separating overlapping transcripts that have different patterns of gene expression and modification.

For effective data integration, systems biology approaches require expression datasets that cover large numbers of biological sources with extraordinary precision [137]. Small yet biologically important effects [138] can be lost in technological noise [139]. Similarly, loss of ncRNAs and transcriptome complexity can occur during library preparation [140] and RNA isolation [141]. Processing of NGS data also presents a number of challenges. For example, algorithms building transcripts (Tier 1) should account for regions of the genome that have poor alignability due to repetitive regions [142]. NGS reads unassigned after these steps can then serve as input into *ab initio* algorithms such as the genomic binning approach [50,139] to detect differentially expressed regions [27–30]. Without a doubt, cycles of iterations consisting of annotation, expression measurement, and addition of new transcripts and transcribed regions from global RNA measurement experiments will illuminate the puzzle of pervasive transcription.

Concluding remarks

Assigning functions to the mass of lncRNAs produced in the cell requires novel thinking and approaches. Many of the classic reductionist methods that worked well for coding genes have proven less useful to the challenges of deciphering the elaborate populations of transcripts generated by pervasive ncRNA transcription. Instead, global, systems-biology and genomics-driven approaches have emerged, which rely on an integrative framework of annotation and classification. This framework increases emphasis on the quality of genome-wide RNA measurements to allow for the ready integration of data from multiple types of experiments. It facilitates the development of improved tools for the integration of the highly multi-dimensional data from these experiments into the classification framework, thereby revealing associations between both coding and non-coding transcripts. Finally, it supports the rational and structured selection of subsets of these predictions for biological follow-up using reductionist methods.

Online links

FANTOM Consortium: http://fantom.gsc.riken.jp/
ENCODE Consortium: http://www.genome.gov/encode/
St. Laurent Institute: http://www.stlaurentinstitute.com/
Database of RNA modifications: http://mods.rna.albany.edu/
NIH Roadmap Epigenomics project: http://www.roadmapepigenomics.org/

Acknowledgements

We wish to thank Maxim Ri, Denis Antonets and Dmitry Shtokalo for help with the bioinformatics analysis and Mark Mazaitis and Anna Miminoshvili and for expert assistance with the figure preparations. Studies on long noncoding RNAs in the Wahlestedt laboratory are currently supported by the US National Institute of Health awards DA035592, MH084880 and NS071674.

References

1 Kapranov, P. *et al.* (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916–919
2 Okazaki, Y. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573
3 Djebali, S. *et al.* (2012) Landscape of transcription in human cells. *Nature* 489, 101–108
4 Bernstein, B.E. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74
5 Kapranov, P. *et al.* (2010) The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol.* 8, 149
6 Clark, B.S. and Blackshaw, S. (2014) Long non-coding RNA-dependent transcriptional regulation in neuronal development and disease. *Front. Genet.* 5, 164
7 Gibb, E.A. *et al.* (2011) The functional role of long non-coding RNA in human carcinomas. *Mol. Cancer* 10, 38
8 Qureshi, I.A. and Mehler, M.F. (2013) Long non-coding RNAs: novel targets for nervous system disease diagnosis and therapy. *Neurotherapeutics* 10, 632–646
9 Reis, E.M. and Verjovski-Almeida, S. (2012) Perspectives of long non-coding RNAs in cancer diagnostics. *Front. Genet.* 3, 32
10 Vergara, I.A. *et al.* (2012) Genomic "dark matter" in prostate cancer: exploring the clinical utility of ncRNA as biomarkers. *Front. Genet.* 3, 23
11 Wahlestedt, C. (2013) Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nat. Rev. Drug Discov.* 12, 433–446
12 St Laurent, G. *et al.* (2014) Dark matter RNA illuminates the puzzle of genome-wide association studies. *BMC Med.* 12, 97
13 Clark, M.B. *et al.* (2013) The dark matter rises: the expanding world of regulatory RNAs. *Essays Biochem.* 54, 1–16
14 Liebers, R. *et al.* (2014) Epigenetic regulation by heritable RNA. *PLoS Genet.* 10, e1004296
15 Smalheiser, N.R. (2014) The RNA-centred view of the synapse: non-coding RNAs and synaptic plasticity. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 369

16 Penman, S. (1991) If genes just make proteins and our proteins are the same, then why are we so different? *J. Cell. Biochem.* 47, 95–98
17 Zuckerkandl, E. (1981) A general function of noncoding polynucleotide sequences. Mass binding of transconformational proteins. *Mol. Biol. Rep.* 7, 149–158
18 Guttman, M. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227
19 Khalil, A.M. *et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11667–11672
20 Zhang, K. *et al.* (2014) The ways of action of long non-coding RNAs in cytoplasm and nucleus. *Gene* 547, 1–9
21 Kapranov, P. and St Laurent, G. (2012) Dark matter RNA: existence, function, and controversy. *Front. Genet.* 3, 60
22 Mattick, J.S. (2009) The genetic signatures of noncoding RNAs. *PLoS Genet.* 5, e1000459
23 Morris, K.V. and Mattick, J.S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet.* 15, 423–437
24 St Laurent, G., III and Wahlestedt, C. (2007) Noncoding RNAs: couplers of analog and digital information in nervous system function? *Trends Neurosci.* 30, 612–621
25 FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014) A promoter-level mammalian expression atlas. *Nature* 507, 462–470
26 Lonsdale, J. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–585
27 Guttman, M. *et al.* (2010) *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510
28 Hackermuller, J. *et al.* (2014) Cell cycle, oncogenic and tumor suppressor pathways regulate numerous long and macro non-protein coding RNAs. *Genome Biol.* 15, R48
29 St Laurent, G. *et al.* (2013) VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. *Genome Biol.* 14, R73
30 Trapnell, C. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578
31 Boley, N. *et al.* (2014) Genome-guided transcript assembly by integrative analysis of RNA sequence data. *Nat. Biotechnol.* 32, 341–346
32 Ernst, J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49
33 Harrow, J. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774
34 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
35 Jia, H. *et al.* (2010) Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA* 16, 1478–1487
36 Amaral, P.P. *et al.* (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.* 39, D146–D151
37 Volders, P.J. *et al.* (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.* 43, D174–D180
38 Kapranov, P. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488
39 Wang, K.C. and Chang, H.Y. (2011) Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914
40 Sharon, D. *et al.* (2013) A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31, 1009–1014
41 Lazorthes, S. *et al.* (2015) A vlincRNA participates in senescence maintenance by relieving H2AZ-mediated repression at the INK4 locus. *Nat. Commun.* 6, 5971
42 van Dijk, M. *et al.* (2012) HELLP babies link a novel lincRNA to the trophoblast cell cycle. *J. Clin. Invest.* 122, 4003–4011
43 Carninci, P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563
44 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
45 Mercer, T.R. *et al.* (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* 30, 99–104
46 Denoeud, F. *et al.* (2007) Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* 17, 746–759
47 Djebali, S. *et al.* (2008) Efficient targeted transcript discovery via array-based normalization of RACE libraries. *Nat. Methods* 5, 629–635
48 Makrythanasis, P. *et al.* (2009) Variation in novel exons (RACEfrags) of the MECP2 gene in Rett syndrome patients and controls. *Hum. Mutat.* 30, E866–E879
49 Nakaya, H.I. *et al.* (2007) Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol.* 8, R43
50 St Laurent, G., III *et al.* (2012) Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. *BMC Genomics* 13, 504
51 Fachel, A.A. *et al.* (2013) Expression analysis and in silico characterization of intronic long noncoding RNAs in renal cell carcinoma: emerging functional associations. *Mol. Cancer* 12, 140
52 Engelhardt, J. and Stadler, P.F. (2012) Hidden treasures in unspliced EST data. *Theory Biosci.* 131, 49–57
53 Moss, W.N. and Steitz, J.A. (2013) Genome-wide analyses of Epstein–Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA. *BMC Genomics* 14, 543
54 Gardner, E.J. *et al.* (2012) Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of *Xenopus tropicalis*. *Genes Dev.* 26, 2550–2559
55 Zhang, Y. *et al.* (2013) Circular intronic long noncoding RNAs. *Mol. Cell* 51, 792–806
56 Cheng, J. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308, 1149–1154
57 Katayama, S. *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science* 309, 1564–1566
58 Affymetrix ENCODE Transcriptome Project; Cold Spring Harbor Laboratory ENCODE Transcriptome Project (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* 457, 1028–1032
59 Otsuka, Y. *et al.* (2009) Identification of a cytoplasmic complex that adds a cap onto 5′-monophosphate RNA. *Mol. Cell. Biol.* 29, 2155–2167
60 Mercer, T.R. *et al.* (2011) Expression of distinct RNAs from 3′ untranslated regions. *Nucleic Acids Res.* 39, 2393–2403
61 Abdelhamid, R.F. *et al.* (2014) Multiplicity of 5′ cap structures present on short RNAs. *PLoS ONE* 9, e102895
62 Jeck, W.R. *et al.* (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19, 141–157
63 Djebali, S. *et al.* (2012) Evidence for transcript networks composed of chimeric RNAs in human cells. *PLoS ONE* 7, e28213
64 Finta, C. and Zaphiropoulos, P.G. (2002) Intergenic mRNA molecules resulting from trans-splicing. *J. Biol. Chem.* 277, 5882–5890
65 Zaphiropoulos, P.G. (1999) RNA molecules containing exons originating from different members of the cytochrome P450 2C gene subfamily (CYP2C) in human epidermis and liver. *Nucleic Acids Res.* 27, 2585–2590
66 Kapranov, P. *et al.* (2010) New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism. *Nature* 466, 642–646
67 Volloch, V. *et al.* (1996) Antisense globin RNA in mouse erythroid tissues: structure, origin, and possible function. *Proc. Natl. Acad. Sci. U.S.A.* 93, 2476–2481
68 Caudron-Herger, M. *et al.* (2011) Coding RNAs with a non-coding function: maintenance of open chromatin structure. *Nucleus* 2, 410–424
69 Rinn, J.L. and Chang, H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81, 145–166
70 Andersson, R. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461
71 Necsulea, A. *et al.* (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505, 635–640
72 Gupta, R.A. *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076

73 Du, Z. *et al.* (2013) Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* 20, 908–913
74 Harries, L.W. (2012) Long non-coding RNAs and human disease. *Biochem. Soc. Trans.* 40, 902–906
75 Faulkner, G.J. *et al.* (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* 41, 563–571
76 Fort, A. *et al.* (2014) Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.* 46, 558–566
77 Kelley, D.R. and Rinn, J.L. (2012) Transposable elements reveal a stem cell specific class of long noncoding RNAs. *Genome Biol.* 13, R107
78 Sell, S. (2010) On the stem cell origin of cancer. *Am. J. Pathol.* 176, 2584–2594
79 Mariner, P.D. *et al.* (2008) Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell* 29, 499–509
80 Belancio, V.P. *et al.* (2009) LINE dancing in the human genome: transposable elements and disease. *Genome Med.* 1, 97
81 Flockerzi, A. *et al.* (2008) Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project. *BMC Genomics* 9, 354
82 Wang, J. *et al.* (2013) Control of myogenesis by rodent SINE-containing lncRNAs. *Genes Dev.* 27, 793–804
83 Zheng, D. *et al.* (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res.* 17, 839–851
84 Pink, R.C. *et al.* (2011) Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA* 17, 792–798
85 Johnsson, P. *et al.* (2013) A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nat. Struct. Mol. Biol.* 20, 440–446
86 Poliseno, L. *et al.* (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465, 1033–1038
87 Kerin, T. *et al.* (2012) A noncoding RNA antisense to moesin at 5p14.1 in autism. *Sci. Transl. Med.* 4, 128ra140
88 Arigo, J.T. *et al.* (2006) Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. *Mol. Cell* 23, 841–851
89 Ntini, E. *et al.* (2013) Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat. Struct. Mol. Biol.* 20, 923–928
90 Valen, E. *et al.* (2011) Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nat. Struct. Mol. Biol.* 18, 1075–1082
91 van Dijk, E.L. *et al.* (2011) XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* 475, 114–117
92 Xu, Z. *et al.* (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature* 457, 1033–1037
93 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
94 Bejerano, G. *et al.* (2004) Ultraconserved elements in the human genome. *Science* 304, 1321–1325
95 Hudson, R.S. *et al.* (2013) Transcription signatures encoded by ultraconserved genomic regions in human prostate cancer. *Mol. Cancer* 12, 13
96 Mauger, D.M. *et al.* (2013) The genetic code as expressed through relationships between mRNA structure and protein function. *FEBS Lett.* 587, 1180–1188
97 Washietl, S. *et al.* (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2454–2459
98 Pedersen, J.S. *et al.* (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* 2, e33
99 Washietl, S. *et al.* (2007) Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.* 17, 852–864
100 Ferdin, J. *et al.* (2013) HINCUTs in cancer: hypoxia-induced noncoding ultraconserved transcripts. *Cell Death Differ.* 20, 1675–1687
101 Silva, J.M. *et al.* (2010) Identification of long stress-induced non-coding transcripts that have altered expression in cancer. *Genomics* 95, 355–362
102 Mondal, T. *et al.* (2010) Characterization of the RNA content of chromatin. *Genome Res.* 20, 899–907
103 Zhao, J. *et al.* (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* 40, 939–953
104 Bergmann, J.H. and Spector, D.L. (2014) Long non-coding RNAs: modulators of nuclear structure and function. *Curr. Opin. Cell Biol.* 26, 10–18
105 van Heesch, S. *et al.* (2014) Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol.* 15, R6
106 Rackham, O. *et al.* (2011) Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. *RNA* 17, 2085–2093
107 Kung, J.T. *et al.* (2013) Long noncoding RNAs: past, present, and future. *Genetics* 193, 651–669
108 Orom, U.A. *et al.* (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell* 143, 46–58
109 Tay, Y. *et al.* (2014) The multilayered complexity of ceRNA crosstalk and competition. *Nature* 505, 344–352
110 St Laurent, G. *et al.* (2012) Dark matter RNA: an intelligent scaffold for the dynamic regulation of the nuclear information landscape. *Front. Genet.* 3, 57
111 Wilusz, J.E. *et al.* (2008) 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* 135, 919–932
112 Pircher, A. *et al.* (2014) An mRNA-derived noncoding RNA targets and regulates the ribosome. *Mol. Cell* 54, 147–155
113 Prabakaran, S. *et al.* (2014) Quantitative profiling of peptides from RNAs classified as noncoding. *Nat. Commun.* 5, 5429
114 Ingolia, N.T. *et al.* (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802
115 Banfai, B. *et al.* (2012) Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* 22, 1646–1657
116 Guttman, M. *et al.* (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154, 240–251
117 Pasmant, E. *et al.* (2011) ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* 25, 444–448
118 Burd, C.E. *et al.* (2010) Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet.* 6, e1001233
119 Holdt, L.M. *et al.* (2013) Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLoS Genet.* 9, e1003588
120 Kapranov, P. *et al.* (2007) Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* 8, 413–423
121 Jan, C.H. *et al.* (2011) Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature* 469, 97–101
122 Ozsolak, F. *et al.* (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 143, 1018–1029
123 Peng, Z. *et al.* (2012) Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat. Biotechnol.* 30, 253–260
124 St Laurent, G. *et al.* (2013) Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in Drosophila. *Nat. Struct. Mol. Biol.* 20, 1333–1339
125 Grosjean, H. (2009) Nucleic acids are not boring long polymers of only four types of nucleotides:a guided tour. In *DNA and RNA Modification Enzymes: Structure, Mechanism, Function and Evolution* (7th edn) (Grosjean, H., ed.), pp. 1–18, Landes Bioscience
126 Kleinman, C.L. and Majewski, J. (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 335, 1302 author reply 1302
127 Flusberg, B.A. *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461–465
128 Barhoumi, A. and Halas, N.J. (2011) Detecting chemically modified DNA bases using surface-enhanced raman spectroscopy. *J. Phys. Chem. Lett.* 2, 3118–3123
129 Gerstein, M.B. *et al.* (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res.* 17, 669–681

130 Mudge, J.M. *et al.* (2013) Functional transcriptomics in the post-ENCODE era. *Genome Res.* 23, 1961–1973
131 Zhang, C. and Darnell, R.B. (2011) Mapping *in vivo* protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.* 29, 607–614
132 Siegfried, N.A. *et al.* (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* 11, 959–965
133 Chu, C. *et al.* (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell* 44, 667–678
134 Hah, N. *et al.* (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* 145, 622–634
135 Mattick, J.S. (2007) A new paradigm for developmental biology. *J. Exp. Biol.* 210, 1526–1547
136 Dozmorov, M.G. *et al.* (2013) Systematic classification of non-coding RNAs by epigenomic similarity. *BMC Bioinform.* 14 (Suppl 14), S2
137 Wan, Y.W. *et al.* (2014) On the reproducibility of TCGA ovarian cancer microRNA profiles. *PLoS ONE* 9, e87782
138 St Laurent, G. *et al.* (2013) On the importance of small changes in RNA expression. *Methods* 63, 18–24
139 Raz, T. *et al.* (2011) Protocol dependence of sequencing-based gene expression measurements. *PLoS ONE* 6, e19287
140 Fu, G.K. *et al.* (2014) Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proc. Natl. Acad. Sci. U.S.A.* 111, 1891–1896
141 Sultan, M. *et al.* (2014) Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* 15, 675
142 Derrien, T. *et al.* (2012) Fast computation and applications of genome mappability. *PLoS ONE* 7, e30377
143 Ozsolak, F. *et al.* (2009) Direct RNA sequencing. *Nature* 461, 814–818
144 Kodzius, R. *et al.* (2006) CAGE: cap analysis of gene expression. *Nat. Methods* 3, 211–222
145 Philippe, N. *et al.* (2014) Combining DGE and RNA-sequencing data to identify new polyA+ non-coding transcripts in the human genome. *Nucleic Acids Res.* 42, 2820–2832
146 Fullwood, M.J. *et al.* (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.* 19, 521–532
147 Pachnis, V. *et al.* (1984) Locus unlinked to alpha-fetoprotein under the control of the murine raf and Rif genes. *Proc. Natl. Acad. Sci. U.S.A.* 81, 5523–5527
148 Wang, K.C. *et al.* (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472, 120–124
149 Huarte, M. *et al.* (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 142, 409–419
150 Brown, C.J. *et al.* (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* 71, 527–542
151 Vance, K.W. *et al.* (2014) The long non-coding RNA Paupar regulates the expression of both local and distal genes. *EMBO J.* 33, 296–311
152 Koerner, M.V. *et al.* (2009) The function of non-coding RNAs in genomic imprinting. *Development* 136, 1771–1783
153 Faghihi, M.A. *et al.* (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med.* 14, 723–730
154 Thrash-Bingham, C.A. and Tartof, K.D. (1999) aHIF: a natural antisense transcript overexpressed in human renal cancer and during hypoxia. *J. Natl. Cancer Inst.* 91, 143–151
155 Lee, J.T. *et al.* (1999) Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat. Genet.* 21, 400–404
156 Seila, A.C. *et al.* (2008) Divergent transcription from active promoters. *Science* 322, 1849–1851
157 Kim, T.K. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465, 182–187
158 Flynn, R.A. *et al.* (2011) Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10460–10465
159 Luke, B. and Lingner, J. (2009) TERRA: telomeric repeat-containing RNA. *EMBO J.* 28, 2503–2510
160 Hall, L.L. *et al.* (2014) Stable C0T-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes. *Cell* 156, 907–919
161 Rangwala, S.H. *et al.* (2009) Many LINE1 elements contribute to the transcriptome of human somatic cells. *Genome Biol.* 10, R100
162 Ting, D.T. *et al.* (2011) Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* 331, 593–596
163 Zheng, R. *et al.* (2010) Polypurine-repeat-containing RNAs: a novel class of long non-coding RNA in mammalian cells. *J. Cell Sci.* 123, 3734–3744
164 Schulz, D. *et al.* (2013) Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell* 155, 1075–1087
165 Saini, H.K. *et al.* (2008) Annotation of mammalian primary microRNAs. *BMC Genomics* 9, 564
166 Cai, X. and Cullen, B.R. (2007) The imprinted H19 noncoding RNA is a primary microRNA precursor. *RNA* 13, 313–316
167 Li, X.Z. *et al.* (2013) Defining piRNA primary transcripts. *Cell Cycle* 12, 1657–1658
168 Mao, Y.S. *et al.* (2011) Biogenesis and function of nuclear bodies. *Trends Genet.* 27, 295–306
169 Blanchette, M. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14, 708–715
170 Hindorff, L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367