Biochimica et Biophysica Acta xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagrm

Review

Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs

Gali Housman, Igor Ulitsky *

Department of Biological Regulation, Weizmann Institute of Science, Rehovot 76100, Israel

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form 18 June 2015
Accepted 19 July 2015
Available online xxxx

Keywords:
Long noncoding RNAs
Translation
Computational genomics

ABSTRACT

Long noncoding RNAs (lncRNAs) are a diverse class of RNAs with increasingly appreciated functions in vertebrates, yet much of their biology remains poorly understood. In particular, it is unclear to what extent the current catalog of over 10,000 annotated lncRNAs is indeed devoid of genes coding for proteins. Here we review the available computational and experimental schemes for distinguishing between coding and noncoding transcripts and assess the conclusions from their recent genome-wide applications. We conclude that the model most consistent with the available data is that a large number of mammalian lncRNAs undergo translation, but only a very small minority of such translation events results in stable and functional peptides. The outcomes of the majority of the translation events and their potential biological purposes remain an intriguing topic for future investigation.

1. Introduction — the gray area between coding and noncoding transcripts

Genome-wide surveys of transcription in mammalians [1–8] and more recently in other vertebrates [9–14] have shown that vertebrate genomes are pervasively transcribed and produce a large variety of processed transcripts, many of which do not overlap canonical genes. These include long noncoding RNAs (lncRNAs), which closely resemble mRNAs in that they possess an m^7Gpppn cap at the 5′ end, a poly(A) tail at the 3′ end, and in most cases undergo splicing [15]. The expression of many lncRNAs has been shown to be altered in a wide range of human diseases, including cancer [16], which promotes the interest in understanding their roles. A variety of functional studies in different species have shown that lncRNAs play important functions in numerous key cellular pathways, including regulation of gene expression, progression through the cell cycle, and establishment of cell identity during embryonic development [17–20]. While it remains unclear how most of these functions are carried out, it is almost certain that different lncRNAs employ radically different modes of action, some of which take place predominantly in the nucleus, and others in the cytoplasm [15]. As the biological mechanism of action of the vast majority of lncRNAs remains unknown, a lurking question is whether some of them actually encode short functional peptides, and so are misannotated and in fact function as mRNAs.

How can one distinguish between coding and noncoding transcripts among long RNAs? lncRNAs rarely contain highly structured regions, making tools for predicting classical structured noncoding RNAs of limited utility. Further, vertebrate lncRNAs are on average ~1000 nt long, and so by chance they are expected to contain multiple translatable open reading frames (ORFs) with at least 50 (and sometimes even more than 100) in-frame codons enclosed within an AUG codon and a stop codon [21]. By mere chance, many of the AUG start codons in such ORFs will also be found in favorable Kozak contexts that will promote translation initiation. Therefore, according to textbook knowledge of translation, many annotated lncRNA transcripts should in fact be coding. The options that we have to consider for each candidate lncRNA are therefore as follows: (i) it codes for a functional protein and is therefore misannotated as a lncRNA; (ii) it employs a mechanism that allows it to avoid productive translation; and (iii) it is translated, and translation does not result in functional peptides, and so is either tolerated by the cell as "waste" or has other roles. We note that while the third group is translated, we believe they are nevertheless "noncoding" RNAs, as they do not function as templates encoding functional proteins. As we will discuss here, the relative fraction of currently known lncRNAs that belong to each of the three groups is controversial, with some studies suggesting that hundreds of human lncRNAs are coding for uncharacterized proteins [22,23] (and are therefore not lncRNAs),

* Corresponding author.
E-mail address: igor.ulitsky@weizmann.ac.il (I. Ulitsky).

http://dx.doi.org/10.1016/j.bbagrm.2015.07.017
1874-9399/

Table 1
Recent methods for distinguishing between coding and noncoding transcripts. 'Features and methodology' refers to the types of features and means for their combination used by each method or tool: "ORF" refers to the length (absolute or relative) of ORFs in the transcript; "composition" refers to features describing frequencies of nucleotides, amino acids or their combinations in the entire transcript or in specific ORFs; "similarity" refers to features describing the degree of similarity between the transcripts, or their predicted ORFs, to known protein sequences; "domains" refers to the potential of the transcript to encode a protein domain found in a protein domain database, such as Pfam [31]; "evolution" refers to the substitution patterns of nucleotides in the transcript, evaluated using whole-genome alignments; "conservation" refers to the use of general sequence conservation, e.g., computed using PhastCons [109]; "RNA structure" refers to features describing different characteristics of secondary structures predicted to form by the transcript; "SVM" refers to the use of a supporting vector machine to combine the features in a classifier framework.

Name	Input	Features and methodology used	URL
PLAR[a] [14]	RNA-seq data or transcriptome assembly	ORF; composition; similarity; domains; evolution	Implementation: http://webhome.weizmann.ac.il/home/igoru/PLAR
iSeeRNA [27]	Transcriptome assembly	ORF; composition; conservation; SVM	Web server and implementation: http://www.myogenesisdb.org/iSeeRNA
CPC [29]	RNA sequences, protein database	ORF; composition; similarity; SVM	Web server and implementation: http://cpc.cbi.pku.edu.cn/
CONC [30]	Sequences	ORF; composition; similarity; SVM	N/A
RNAcode[99]	Sequence alignments	Evolution	Implementation: http://wash.github.io/rnacode/
PhyloCSF[b] [100]	Sequence alignments	Evolution	Implementation: https://github.com/mlin/PhyloCSF/wiki
Re et al. [101]	Alignments	Evolution	N/A
HMMER	Translated protein sequences, HMM models (e.g., from Pfam)	Domains	Implementation: http://hmmer.janelia.org/
PORTRAIT [102]	RNA sequences	ORF; composition; protein product features	Server and implementation: http://bioinformatics.cenargen.embrapa.br/portrait
CPAT [26]	RNA sequences	ORF; composition; linear regression	Implementation: http://code.google.com/p/cpat/ Webserver: http://lilab.research.bcm.edu/cpat/index.php
CNCI [103]	RNA sequences	ORF; composition; SVM	Implementation: http://www.bioinfo.org/software/cnci
PLEK [104]	RNA sequences	Composition; SVM	Implementation: http://sourceforge.net/projects/plek/files/
Wang et al. [105]	RNA sequences	Composition; RNA structure; SVM	N/A
CNCTDISCRIMINATOR [106]	RNA-seq data	ORF; composition; RNA structure; expression level	Implementation: http://biomecis.uta.edu/~ashis/res/cnctdiscriminator/suppl
Ulveling et al. [107]	RNA sequences	Composition	N/A
HLRF [108]	RNA sequences or genomic sequences	Composition; RNA structure; logistic regression and random forrest	http://ncrna-pred.com/HLRF.htm

[a] Uses CPC, HMMER, RNAcode and combines their results.
[b] Parameters currently available only for whole-genome alignments for mammals, flies, mosquitos and yeast.

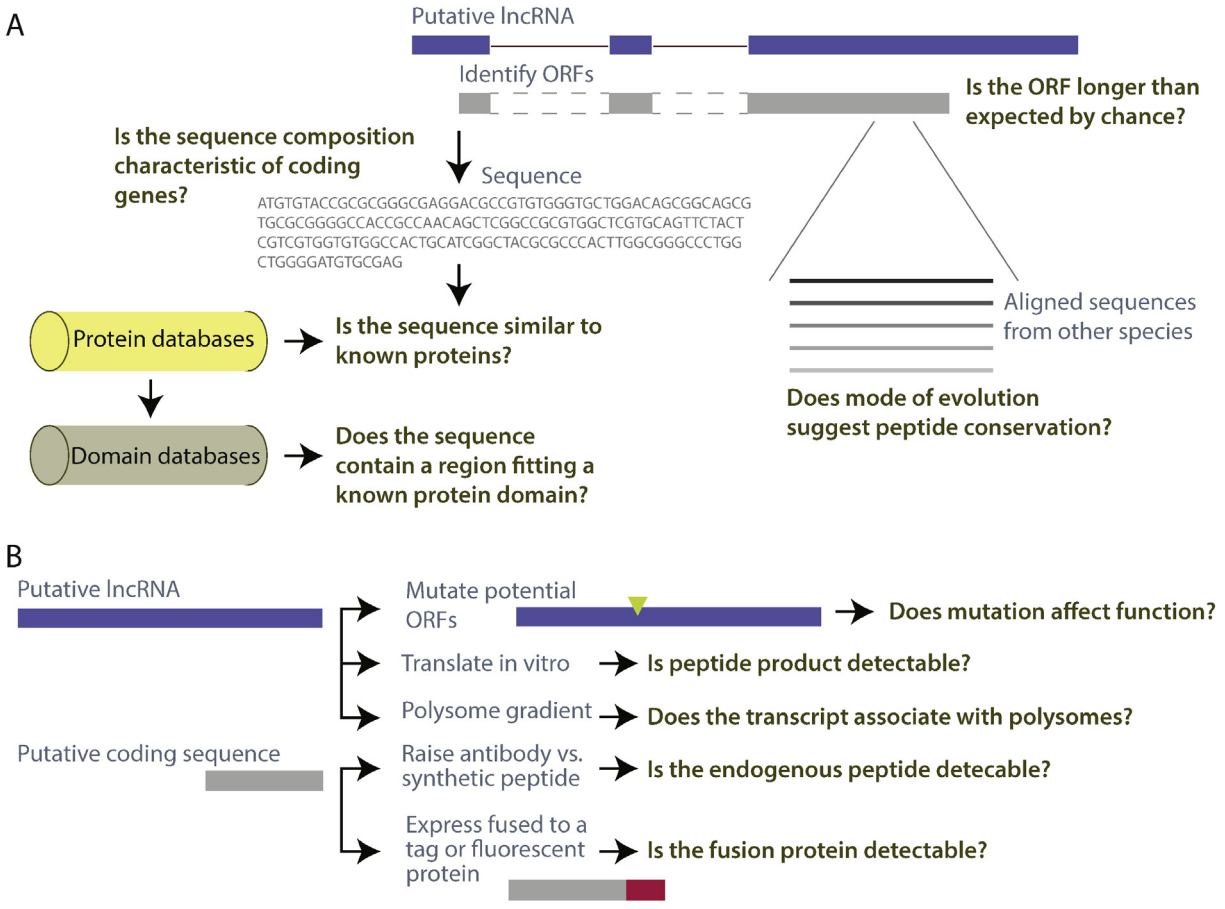

Fig. 1. Methods for distinguishing between protein-coding and noncoding transcripts. A scheme of the common computational (A) and experimental (B) approaches for evaluating the protein-coding potential of a specific putative lncRNA transcript.

while others concluding that only very few are in fact protein-coding [24,25]. We propose that most current evidence points to the prevalence of the third option — lncRNAs are being pervasively translated, but products of their translation are very unstable or nonfunctional.

2. Computational methods for distinguishing between protein-coding and lncRNA genes

Different computational schemes can be used to assess the sequence or the evolution of an uncharacterized transcript and predict whether it is likely to encode a protein. As most features that can be used for such classification have limited discriminatory power, methods usually rely on a combination of diverse features. An overview of recently introduced tools for distinguishing between coding and noncoding transcripts is found in Table 1. We note that most approaches are best suited for those lncRNAs that do not overlap other genes at all (long intervening noncoding RNAs or lincRNAs), and approaches that use genome alignments (see below) are particularly unsuitable when the lncRNA overlaps the coding sequence of a protein-coding gene on either the sense or the antisense strand. The following groups of features have

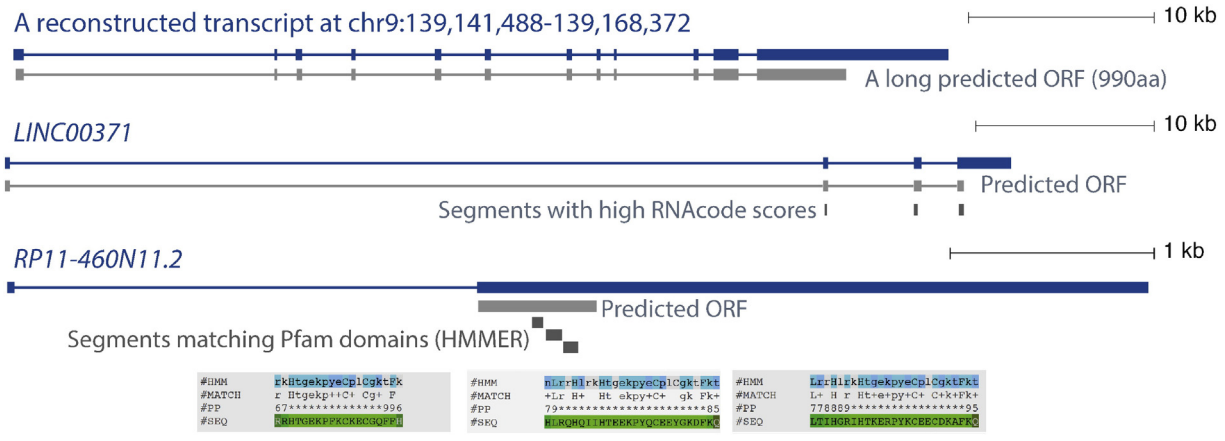

Fig. 2. Examples of putative human lncRNA transcripts that do not pass one or more of the commonly used filters for detecting protein-coding genes. The first transcript is reconstructed from human RNA-seq data by Cufflinks [14] and contains a long ORF. The second is presently annotated as a lincRNA but overlaps several domains predicted to be coding by RNAcode [99]. The third encodes an ORF that is predicted by HMMer to encode zinc-finger domains annotated in the Pfam database.

Please cite this article as: G. Housman, I. Ulitsky, Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of lon..., Biochim. Biophys. Acta (2015), http://dx.doi.org/10.1016/j.bbagrm.2015.07.017

been proposed by different studies as inputs for classification schemes distinguishing between coding and noncoding transcripts (Figs. 1–2):

ORF length — coding regions tend to be much longer than expected by chance [21], and so the presence of a long ORF (e.g., >300 nt long, encoding a protein with >100 amino acids) can serve as an indicator of the coding potential of a sequence. Since the likelihood of seeing a long ORF increases with transcript length, tools such as CPAT [26] and iSeeRNA [27] also examine the length of the longest ORF as a fraction of the entire transcript sequence. It should be noted that by itself, ORF length has limited predictive ability — a transcript of 2 Kb is expected to have an ORF of ~200 nt by chance, and an ORF of 300 nt is only one standard deviation longer than expected by chance [21]. Indeed, well characterized human lncRNAs including *Xist*, *Meg3*, *Hotair*, *Kcnq1ot1*, and *H19* all have ORFs of >100 codons [21] (we note that *H19* was also predicted by some to be protein-coding [28]).

Nucleotide, codon or short word frequencies — nucleotide frequencies in ORFs encoding functional proteins are dictated by non-random codon usage, and so the spectrum of nucleotide or word frequencies in entire transcripts or predicted ORFs can be used as indicators of coding potential. As shown in Table 1, many of the recent tools for distinguishing between coding and noncoding transcripts rely mostly on this group of features, in part because they are easier and faster to compute than features based on sequence conservation or similarity. Some tools, such as CPC [29] and CONC [30], look at triplet composition, and at the properties of the amino acids encoded by them, whereas others, such as CPAT [26], also look at frequencies of individual bases at each possible frame, and hexamer frequencies in the entire sequence.

Substitution patterns — protein-coding sequences evolve under selective pressures to preserve specific amino acids or amino acid types at defined positions and to maintain open reading frames. The presence of such pressures can be measured by inspection of multiple sequence alignments: by comparing substitution frequencies in different positions within a reading frame, and by testing whether insertions and deletions (indels) are depleted, and the reading frame is preferentially preserved when indels do occur. Features derived from sequence alignments are particularly powerful for detecting transcripts that encode conserved peptides, even very short ones that are difficult to recognize using other features. Naturally, such features are less useful for detecting peptides that only recently became functional, or for annotating lncRNAs in species where related genome sequences or whole genome alignments are not available.

Presence of sequences encoding known functional domains — protein-coding genes typically contain common protein domains and the probabilistic models describing those domains have been collected in dedicated databases, such as Pfam [31] which contains Hidden Markov Model (HMM) representations of both well-characterized and putative domains [32]. Tools such as HMMER (http://hmmer.janelia.org/) can then be used to examine possible products of transcripts in all three frames and compute the likelihood that the transcript encodes a common domain, which in turn provides substantial evidence that the transcript is protein-coding.

Similarity to known proteins — coding regions are likely to bear sequence similarities to entries in known protein databases. This feature is superficially simpler to implement than some of the other features, but it is important to carefully choose and filter the protein sequence database that the putative lncRNA is compared to, as some databases, including both GenBank and Ensembl, frequently contain "hypothetical protein" sequences or models with no experimental support. It is also worth noting that true noncoding transcripts that have recently adopted sequence from a coding transcript (e.g., from a pseudogene) may contain elements that score highly as potential functional domains or as similar to other proteins, but those elements will typically not reside in a coherent open reading frame.

Other feature groups — additional features have been proposed as suitable for classification, but those have more limited support in known bona fide lncRNAs. For instance, it is well appreciated that lncRNAs evolve much faster than protein coding genes, and so iSeeRNA [27] uses general sequence conservation as a criterion for distinguishing between coding and noncoding RNAs. However this feature likely introduces a bias against the rare conserved lncRNAs [9,13]. Other tools (Table 1) also use the presence of structured elements (quantified as minimum free energy of the predicted fold) as features, but there is no conclusive evidence that lncRNAs are more structured than mRNAs [15,33,34]. In fact there are computational predictions that lncRNAs may be less structured than mRNAs as a group [35], and experimental data suggesting lncRNAs are slightly more structured than mRNAs, but still much less structured than canonical short or intermediate-size ncRNAs [36].

Most of the studies extracted some of the features described above from collections of "positive" and "negative" samples (known protein-coding sequences and some set of "confidently noncoding" RNAs) and then trained a classification algorithm, such as a support vector machine (SVM) or a logistic regression [37]. For instance, the widely used Coding Potential Calculator (CPC) [29], uses the length of the ORF in a transcript, its codon frequencies, and the BLASTP-computed similarity scores between transcript ORFs and a known protein database as inputs to an SVM [37] that outputs a binary prediction (coding/noncoding) along with a confidence score. As very few long RNAs have been rigorously shown to be bona fide noncoding, the most challenging aspect of such studies is the construction of the set of "negative" samples — sequences that are known with high confidence to be noncoding. The use of different datasets across studies makes it difficult to compare and evaluate the reported performance.

The underlying assumption behind all the criteria currently used is that short recently evolved yet functional proteins are relatively rare. As we will describe below, current mass spectrometry data support the notion that such proteins are rare, but the extent of their prevalence remains the pivotal question in this controversial topic.

Complementing the tools for detecting transcripts that are likely to be noncoding are methods for detection of high-confidence translated small ORFs (sORFs) (reviewed in [38]), such as of sORFinder (http://evolver.psc.riken.jp/) [39], HAltORF (http://www.roucoulab.com/haltorf/) [40] and uPEPperoni (http://upep-scmb.biosci.uq.edu.au/) [41].

3. Experimental approaches for testing whether individual transcripts are translated into proteins

In pursuing the function of a specific lncRNA, even if it passes the computational filters described in the previous section, it is desirable to test experimentally whether it is indeed noncoding (Fig. 1B). Since most transcripts have only few ORFs that are relatively long or contain potentially conserved amino acids, it is feasible to use several alternative methods to test if any of those are translated into detectable peptides. The best option, in our opinion, is to test whether the functionality of the transcript is preserved when the ORFs are perturbed, e.g. by introducing frameshift-inducing mutations [9,42]. This approach is only applicable when the function of the RNA is known and when the transcript sequence can be manipulated (e.g., when the relevant

Please cite this article as: G. Housman, I. Ulitsky, Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of lon..., Biochim. Biophys. Acta (2015), http://dx.doi.org/10.1016/j.bbagrm.2015.07.017

phenotype results from transcript over-expression [42] or if a custom-made transcript sequence can be used in a rescue experiment [9]).

Each of the other available methods is associated with caveats. One can test if the transcript yields peptides when translated in vitro [43–45], but a sequence may be translated in vitro but not in vivo and vice versa. Predicted peptides can also be synthesized and used to produce antibodies that can then be used to detect the putative peptide by various methods such as immunohistochemistry or western blot, though those methods may have limited sensitivity [46]. Another option is to fuse the predicted ORF with a C-terminal tag such as FLAG or GFP that can then be used for detection using Western blotting or microscopy [45–49]. The problem here is that fusion of a peptide to such a tag can turn a very unstable peptide into a part of a stable longer protein.

3.1. Monitoring translation with Ribo-seq and polysome profiling

A combination of the approaches listed above can lead to quite conclusive evidence about the coding potential of specific transcripts. However, the holistic question of the extent to which the currently annotated lncRNAs are "noncoding" requires globally applicable methods.

Translation of transcripts can be monitored by looking at the global association of RNAs with ribosomes at a given time. Two somewhat complementary methods have been developed — ribosome profiling (Ribo-seq) and polysomal fractionation. Ribo-seq [50] allows one to take a snapshot of RNA regions that are associated with translating ribosomes. It combines classical approaches for obtaining "ribosome protected fragments" (RPFs) — footprints of actively translating ribosomes [51] — with the advances made in deep sequencing in the past decade, in order to obtain a global map of the positions within eukaryotic RNAs occupied by 80S ribosomes. Ribo-seq gives indication of what regions are translated, but not of the fraction of the copies of a transcript that are actively translated, whereas polysomal fractionation can address the latter question. With polysomal fractionation, one separates transcripts according to the number of ribosomes associated with them using a sucrose gradient, and then high-throughput methods such as microarrays or RNA-seq are employed to identify and count the transcripts in each fraction. Unlike Ribo-seq, when using polysome profiling, it is unknown which region of the transcript is actively translated. Lack of association with polysomes has been used as a criterion for classifying transcripts as noncoding [52,53], but it is important to keep in mind that association with polysomes does not necessarily imply that the protein products are functional, and RNAs found in some cellular granules may migrate with heavy polysomes, regardless of their state of translation [54]. Indeed, transcripts considered bona fide lncRNAs, such as HULC, associate with polysomes [46].

Application of Ribo-seq in different systems, including mice [22], zebrafish [55], plants [56], and yeast [57], has shown that a substantial portion of expressed lncRNAs are associated with translating ribosomes. For instance, Ingolia et al. [22] found that the majority of annotated lncRNAs in mouse ES cells have regions associated with ribosomes with efficiencies closely resembling those of coding sequences and substantially higher than those observed for mRNA 3′ UTRs. The initial interpretation of this finding was that many lncRNAs are misannotated and encode, perhaps polycistronically, short peptides [22].

Subsequent studies challenged this notion. Guttman et al. [25] and Chew et al. [58] highlighted other characteristics of these data that distinguish lncRNAs from known coding sequences, such as a sharp decrease in ribosome occupancy downstream to a stop codon seen in canonical coding sequences [25], but much less so in lncRNAs. This criterion, by itself or in combination with other footprint- or ORF-based features, can be used to effectively distinguish lncRNAs from canonical coding sequences, though not between lncRNAs and 5′ UTRs [25,58].

Another important drawback raised was that some of the putative RPFs might result from RNA protection by ribonucleoprotein complexes other than translating 80S ribosomes, since some footprints were found on canonical ncRNAs that are certainly not translated such as the RNA components of RNAse P and telomerase, and since the lengths of these footprints were different than those of RPFs from canonical coding sequences [25,59]. The most recent reanalysis of the existing data along with new techniques introduced by Ingolia et al. [59] suggest that while such non-ribosomal footprints are indeed present in typical Ribo-seq libraries, they can be effectively removed to further enrich for reads corresponding to bona fide 80S footprints. One method that allows this distinction is FLOSS [59], a metric that ranks ORFs according to similarity of their RPF length distribution to the length distribution of RPFs mapping to known coding sequences. However, while FLOSS scores effectively removed footprints on mitochondrial genes and classical ncRNAs, the vast majority of the RPFs observed on lncRNAs and 5′ UTRs were still retained, reinforcing the observation that those regions were indeed actively translated. Additional experiments, including direct pulldown of ribosomal proteins added strong support to this conclusion.

The current notion is therefore that many annotated lncRNAs undergo active translation, but that this translation resembles that observed in 5′ UTRs. Several recent studies combined Ribo-seq with other computational and experimental methods in order to further increase confidence in detecting translation events:

- Drug treatments: different drugs target translation at specific stages, and conducting Ribo-seq experiments after such treatments can yield a better understanding of the RPF origins, or be used for particular goals. For example, the use of harringtonine [22] or puromycine [60] uncovered unannotated translation initiation sites and the prevalence of non-AUG start codons. Depletion of RPFs from an ORF following treatment with Pateamine A, an eIF4A inhibitor, has been used to test if ORFs are indeed actively translated [61].
- Polysomal fractionation: RNA-seq in subcellular fractions of human cells showed that lncRNAs are present in different cellular compartments — nucleus, free cytosolic, and ribosome-associated fractions, with most being enriched in the free cytosolic [62] and in the light polysomal fraction [63]. Ribosome profiling following polysomal fractionation (Poly-Ribo-seq), allows enrichment of transcripts that are translated by only a few ribosomes and therefore likely contain small ORFs (smORFs). Aspden et al. [64] used Poly-Ribo-seq in search of novel smORFs in flies, but generally found that smORFs in ncRNAs had profiles different than those of known protein coding genes encoding long or short ORFs. The novel smORFs had a lower translational efficiency, similar to UTRs and their products could rarely be validated with mass-spectrometry or with FLAG tag signal. Notably, the 'dwarf' smORFs in lncRNAs were shorter (~20 aa) than known smORFs in flies.
- Machine learning methods for filtering footprints: as noted above with FLOSS, one can use metrics to classify the Ribo-seq reads into those that represent translation and those that do not. Such metrics can classify by read density similar to the density in coding regions [65], the match of the RPF to the reading frame [55], and nucleotide composition and conservation [66].

Importantly, these studies showed association of many lncRNAs with translating ribosomes and presumably the formation of translation products, but as discussed below, translation does not warrant production of a functional peptide. Indeed, for the most part, there is no current evidence indicating that more than a handful of the peptides encoded by lncRNAs that pass all the computational filters described above are stable, functional or conserved. For example, Bazzini et al. [55] used a computational approach to predict 303 coding smORFs in transcripts considered non-coding, 71% of which are under 100 aa long. However, only six novel peptide products of these smORFs, all in lncRNAs or UTRs, could be detected by mass spectrometry (see below).

Please cite this article as: G. Housman, I. Ulitsky, Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of lon..., Biochim. Biophys. Acta (2015), http://dx.doi.org/10.1016/j.bbagrm.2015.07.017

3.2. Mass spectrometry for detection of peptides from novel genes — lack of evidence or lack of power?

The next question is to what extent the peptide products of the translation events on lncRNAs accumulate in cells at detectable levels? Mass spectrometry (MS) is currently the leading proteomic platform for peptide detection and a multitude of recent studies used different flavors of MS for studying peptides expressed in human, mouse and zebrafish tissues [23,24,47,55,67–74]. A priori, it might be difficult to detect potential peptides originating from translation of annotated lncRNAs using classical MS design due to at least three issues: (i) experimental biases against detection of peptides smaller than 10 kD; (ii) paucity of potential peptides coming from unannotated transcripts in databases used for spectra search; and (iii) the inherent limited sensitivity of MS which leads to bias against detection of lowly expressed genes. Accordingly, recent studies have proposed improved MS methods, including a peptidomics approach [47,75], combined fractional diagonal chromatography (COFRADIC) [72,74,76,77], and other improvements [67,73]; constructed potential peptide databases using RNA-seq or Ribo-seq-based transcriptome reconstructions [47,67,72,78]; and combined multiple MS experiments to increase proteome coverage depth [23,68].

Despite these advances that have increased the a priori chances of observing lncRNA translation products, and specific focus of recent MS-based surveys on identifying such spectra (which may have even led to some degree of ascertainment bias), the most recent attempts found evidence of peptide products traceable to a very limited number of lncRNA genes:

- As part of the ENCODE project, Banfai et al. [24] analyzed shotgun MS/MS data from K562 and GM12878 cell lines, and after filtering unannotated protein-coding genes found peptides mapping to only two lncRNAs, each of which was supported by just one peptide. When compared with peptide detection rates of mRNAs with expression levels matching those of lncRNAs, lncRNAs were 13- to 20-fold depleted for detected translation given their expression levels. Importantly, this study did not detect a general bias against detection of peptides coming from short ORFs [24].
- Slavoff et al. [47] used a peptidomics method and identified just eight peptides <50 aa derived from lncRNAs, and for those eight, the evidence that they come from mature peptides is limited [79].
- Kim et al. [68] identified nine annotated noncoding peptide-producing transcripts in a very deep proteomic survey (16 million MS/MS spectra). The same group reported peptides from 34 lncRNA candidates in zebrafish [71], where a higher number of mis-annotations is expected due to less mature genome sequence and annotation.
- Bazzini et al. [55] used a dedicated MS approach for identifying protein-products of zebrafish small ORFs predicted based on conservation and Ribo-seq (see above), but found peptides from products of just six new ORFs in annotated lncRNAs.
- Prabakaran et al. [78] reported 250 novel peptides coming from unannotated regions, but only 25 of those mapped to regions outside the boundaries of protein-coding genes, and only three of those had support from Ribo-seq data. When we mapped the 250 peptides from this study against a collection of lncRNAs defined by PLAR [14] and passing stringent filters, we found only two matches, and both mapped to probable pseudogenes.
- Another study specifically designed to detect peptides coming from unannotated genes or lncRNAs did not detect any evidence of translation from lncRNAs [69].
- Finally, several other recent studies did not report any peptides derived from lncRNAs, despite the use of a Ribo-seq-based peptide database [74,77].

Notably, products derived from translation events in 5′ UTRs were also very rarely detected, suggesting similarities between translation outcomes in lncRNAs and 5′ UTRs [55,72].

The studies described above appear to contrast a recent study by Wilhelm et al. [23] that reported 430 peptides from 404 lncRNAs detected by MS. This unexpectedly high number suggested that many lncRNAs may indeed be translated and produce detectable peptides, that were somehow missed by others. However, our analysis of the data suggests that these large numbers result from promiscuous search parameters when matching spectra to potential transcripts. In fact, when we used BLASTN to map the 430 peptides from Wilhelm et al. [23] to lncRNA exons in our recent collection of >10,000 human lncRNA genes passing stringent filters [14], we found only five hits with BLAST E-value of $<10^{-3}$, and all of those mapped to pseudogenic regions. Our analysis is concordant with a recent re-analysis of these data by Valencia et al. [80] that concluded that the peptide identifications reported by Wilhelm et al. [23] likely contain many false positives due to inclusion of low-quality spectra and relaxed peptide detection thresholds.

We conclude that evidence from about a dozen recent studies suggests that translation products originating from ORFs in lncRNAs, even those that appear translated in Ribo-seq data, are essentially invisible in MS data. While it is possible this is due to limitations of MS in analyzing peptides shorter than 50 aa [55], we propose that a more likely explanation is that such peptides are very unstable and therefore do not accumulate to consequential levels in mammalian cells.

4. Discussion

4.1. Best practices for annotating a transcriptome

As described above, the toolbox for distinguishing between coding and noncoding transcripts is substantial and rapidly growing, and many of the tools are available both as web servers and as stand-alone tools (listed in Table 1). Some of these tools (typically those that rely only on sequence composition features) are relatively simple to use, while others require dedicated input processing. In our experience each of the tools has limitations when used in isolation, and so better accuracy is achieved by combining multiple methods, as we have recently done as part of our Pipeline for lncRNA annotation from RNA-seq data (PLAR [14]). It also remains very important to manually inspect transcripts of particular interest (i.e., positive hits from a functional screen) even if they are predicted as coding, and test whether the evidence of protein-coding potential is not a result of an artifact, and indeed converges on a specific ORF that is likely to be coding. For example, sequences that are not repeat-masked will sometimes contain regions resembling protein domains, and regions that are very highly conserved will sometimes be called coding simply because their alignments are not informative enough for tools inspecting substitution patterns.

It should also be noted that, in our experience, when studying a well-annotated genome, such as human or mouse, only few RNA-seq reconstructed transcripts that are both classified as coding by the above-mentioned tools and do not overlap known genes actually correspond to likely bona fide unannotated proteins. Rather, the majority of such transcripts are usually parts of pseudogenes, which are sometimes difficult to identify [81]. It is also notable that some recently described lncRNAs with specific functions fail one or more of the commonly used filters. For example, TINCR [82] contains a region predicted by RNAcode to be coding and the region covering its 456 nt ORF is predicted to encode a peptide that matches an uncharacterized Pfam-B domain. Another lncRNA, lnc-DC, is a human ortholog of the mouse protein-coding gene 110000G20Rik which encodes for a Wdnm1-like protein (Wfdc21) that is highly conserved in mammals [83], and therefore would not pass protein-similarity filters.

Please cite this article as: G. Housman, I. Ulitsky, Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of lon..., Biochim. Biophys. Acta (2015), http://dx.doi.org/10.1016/j.bbagrm.2015.07.017

4.2. Functionality of pervasive translation of lncRNAs?

The question of the functionality of translation in lncRNAs can be reduced to the question of the extent to which peptides <50 aa, after their release from the ribosome, can remain stable and functional in vertebrate cells. Inspection of proteins with currently known functions that are annotated in Ensembl reveals only 16 that are <50 aa. Some of these are clearly independently translated proteins, such as Ost4, a 37 aa, 3.4 kD membrane protein conserved all the way from human to yeast. Another micropeptide, MLN, recently discovered by Eric Olson and colleagues [45], is encoded by a 46 aa ORF in a transcript previously annotated as a lncRNA. Such short proteins can therefore be stable (perhaps due to protection from degradation by the membrane, or by the SERCA protein in the case of MLN) and functional in vertebrates, but their relative scarcity indicates that they may be the exception rather than the rule. We suggest that the vast majority of other short-peptide translation events formed as part of "pervasive translation" [59,84], which appear very common based on Ribo-seq data, are largely nonfunctional products that are tolerated in cells because they are usually rapidly degraded by abundant cytosolic endopeptidases, aminopeptidases, and other machineries active in cells [85] and therefore have limited impact. Systematic evidence for stability of short (<50 aa) peptides in cells would be invaluable for addressing this question, but no such resources are presently available, and existing anecdotal evidence suggests that such peptides are generally very unstable [85–88].

When the cost of production of such peptides is considered, it is important to keep in mind that the vast majority of such translation events occur in 5′ UTRs of coding genes rather than in lncRNAs. Using Ribo-seq data from U2OS cells [89] we find that a similar fraction of all mapped RNA-seq reads (and by proxy nucleotides in polyadenylated transcripts) are found in 5′ UTRs and lncRNAs, but 5′ UTRs account for 10-times more Ribo-seq reads (and by proxy occupy 10-times more ribosomes) (Fig. 3). When comparing to coding regions and using the Ribo-seq reads as a proxy for the number of ribosomes actively translating lncRNAs and mRNAs, there are over 1000-times more ribosomes associated with mRNAs than with lncRNAs. Similar results are obtained in other cell types (I.U., unpublished data). The total fitness cost of all the translation events from all lncRNAs is thus small when compared to the cost of mRNA translation.

A different and interesting way to view pervasive translation of lncRNAs is that these genes can serve as a platform for de novo protein evolution [65,90]. Characteristics of a majority of lncRNAs, in particular, their relative novelty and lineage specificity, suggest that these lncRNAs could be precursors for new proteins. Furthermore, new proteins are expected to be very short and under weak evolutionary constraints, fitting the translated smORFs identified in lncRNAs.

4.3. Regulatory roles of translation in noncoding transcripts

The fate of the translation products arising from lncRNAs is part of a bigger question of translation of "untranslated regions", and among those most prominently of 5′ UTRs. Since mRNAs as a group are ~100-fold more abundant than lncRNAs in cells, and most 5′ UTRs have at least one uORF that undergoes canonical translation, 5′ UTRs are much more abundant templates for translation of small ORFs than lncRNAs, yet the products of this translation are also almost entirely absent from MS data (as detailed above) and the ORFs that are being translated are only very rarely conserved in sequence. Very few ORFs in 5′ UTRs are conserved between human and mouse, and even in those that are conserved, selection on the encoded peptide sequences is mostly weak [91].

It is not known how much of the translation observed in uORFs is functional and how much is tolerated noise of the translational machinery, but some of these events regulate the translation of the main (almost always longest) ORF in the mRNA [92,93]. We can envision parallel "regulatory" roles for translation of ORFs in lncRNAs. Specific possible regulatory roles include regulation of stability of the lncRNA (e.g., by modulating its degradation by nonsense mediated decay [94,95] or other pathways), regulation of its localization in the cell, or protection of parts of the lncRNA sequence from scanning ribosomes that are potent helicases. Translation-facilitated degradation of long noncoding RNAs can serve as an elegant mechanism for regulating the accumulation of snoRNA precursors, which is further modulated during stress [96–98] — and so translation can certainly induce a regulatory mechanism in long noncoding RNAs. We argue that such mechanisms are probably much more prevalent than currently appreciated and should serve as a fruitful direction of further research into the biological consequences of pervasive translation of lncRNA genes.

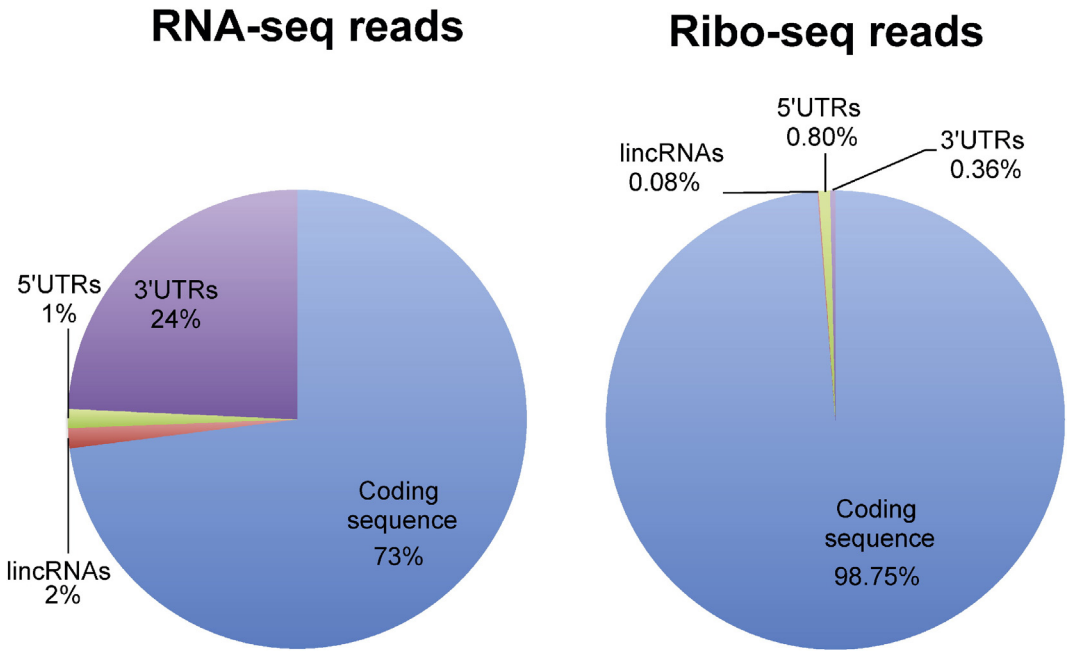

Fig. 3. Distribution of RNA-seq and Ribo-seq reads. RNA-seq and Ribo-seq data from human osteosarcoma U2OS cells [89] were mapped to the human transcriptome, and only reads mapping to mRNAs or lncRNAs were considered. Plotted is the fraction of reads mapping to the indicated features.

Please cite this article as: G. Housman, I. Ulitsky, Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of lon..., Biochim. Biophys. Acta (2015), http://dx.doi.org/10.1016/j.bbagrm.2015.07.017

Transparency Document

The Transparency document associated with this article can be found, in the online version.

Acknowledgments

We thank Rory Johnson, Yoav Lubelsky, Lisha Qiu Jin Lim, Noa Gil, Hadas Hezroni and Neta Degani for useful discussions and comments on the manuscript. I.U. is an incumbent of the Sygnet Career Development Chair for Bioinformatics and recipient of an Alon Fellowship. Work in the Ulitsky lab is supported by grants to I.U. from the Israeli Science Foundation (1242/14 and 1984/14), the I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation (grant no. 1796/12), the Minerva Foundation, the Fritz-Thyssen Foundation, the Rising Tide foundation and by a research grant from The Abramson Family Center for Young Scientists.

References

[1] A.M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B.E. Bernstein, A. van Oudenaarden, A. Regev, E.S. Lander, J.L. Rinn, Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 11667–11672.

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

[3] M. Guttman, I. Amit, M. Garber, C. French, M.F. Lin, D. Feldser, M. Huarte, O. Zuk, B.W. Carey, J.P. Cassady, M.N. Cabili, R. Jaenisch, T.S. Mikkelsen, T. Jacks, N. Hacohen, B.E. Bernstein, M. Kellis, A. Regev, J.L. Rinn, E.S. Lander, Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals, Nature 458 (2009) 223–227.

[4] T. Ravasi, H. Suzuki, K.C. Pang, S. Katayama, M. Furuno, R. Okunishi, S. Fukuda, K. Ru, M.C. Frith, M.M. Gongora, S.M. Grimmond, D.A. Hume, Y. Hayashizaki, J.S. Mattick, Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome, Genome Res. 16 (2006) 11–19.

[5] Y. Okazaki, M. Furuno, T. Kasukawa, J. Adachi, H. Bono, S. Kondo, I. Nikaido, N. Osato, R. Saito, H. Suzuki, I. Yamanaka, H. Kiyosawa, K. Yagi, Y. Tomaru, Y. Hasegawa, A. Nogami, C. Schonbach, T. Gojobori, R. Baldarelli, D.P. Hill, C. Bult, D.A. Hume, J. Quackenbush, L.M. Schriml, A. Kanapin, H. Matsuda, S. Batalov, K.W. Beisel, J.A. Blake, D. Bradt, V. Brusic, C. Chothia, L.E. Corbani, S. Cousins, E. Dalla, T.A. Dragani, C.F. Fletcher, A. Forrest, K.S. Frazer, T. Gaasterland, M. Gariboldi, C. Gissi, A. Godzik, J. Gough, S. Grimmond, S. Gustincich, N. Hirokawa, I.J. Jackson, E.D. Jarvis, A. Kanai, H. Kawaji, Y. Kawasawa, R.M. Kedzierski, B.L. King, A. Konagaya, I.V. Kurochkin, Y. Lee, B. Lenhard, P.A. Lyons, D.R. Maglott, L. Maltais, L. Marchionni, L. McKenzie, H. Miki, T. Nagashima, K. Numata, T. Okido, W.J. Pavan, G. Pertea, G. Pesole, N. Petrovsky, R. Pillai, J.U. Pontius, D. Qi, S. Ramachandran, T. Ravasi, J.C. Reed, D.J. Reed, J. Reid, B.Z. Ring, M. Ringwald, A. Sandelin, C. Schneider, C.A. Semple, M. Setou, K. Shimada, R. Sultana, Y. Takenaka, M.S. Taylor, R.D. Teasdale, M. Tomita, R. Verardo, L. Wagner, C. Wahlestedt, Y. Wang, Y. Watanabe, C. Wells, L.G. Wilming, A. Wynshaw-Boris, M. Yanagisawa, I. Yang, L. Yang, Z. Yuan, M. Zavolan, Y. Zhu, A. Zimmer, P. Carninci, N. Hayatsu, T. Hirozane-Kishikawa, H. Konno, M. Nakamura, N. Sakazume, K. Sato, T. Shiraki, K. Waki, J. Kawai, K. Aizawa, T. Arakawa, S. Fukuda, A. Hara, W. Hashizume, K. Imotani, Y. Ishii, M. Itoh, I. Kagawa, A. Miyazaki, K. Sakai, D. Sasaki, K. Shibata, A. Shinagawa, A. Yasunishi, M. Yoshino, R. Waterston, E.S. Lander, J. Rogers, E. Birney, Y. Hayashizaki, Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs, Nature 420 (2002) 563–573.

[6] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M.C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V.B. Bajic, S.E. Brenner, S. Batalov, A.R. Forrest, M. Zavolan, M.J. Davis, L.G. Wilming, V. Aidinis, J.E. Allen, A. Ambesi-Impiombato, R. Apweiler, R.N. Aturaliya, T.L. Bailey, M. Bansal, L. Baxter, K.W. Beisel, T. Bersano, H. Bono, A.M. Chalk, K.P. Chiu, V. Choudhary, A. Christoffels, D.R. Clutterbuck, M.L. Crowe, E. Dalla, B.P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C.F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T.R. Gingeras, T. Gojobori, R.E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T.K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S.P. Krishnan, A. Kruger, S.K. Kummerfeld, I.V. Kurochkin, L.F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K.C. Pang, W.J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J.F. Reid, B.Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S.L. Salzberg, A. Sandelin, C. Schneider, C. Schonbach, K. Sekiguchi, C.A. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S.L. Tan, S. Tang, M.S. Taylor, J. Tegner, S.A. Teichmann, H.R. Ueda, E. van Nimwegen, R. Verardo, C.L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S.M. Grimmond, R.D. Teasdale, E.T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J.S. Mattick, D.A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, Y. Hayashizaki, The transcriptional landscape of the mammalian genome, Science 309 (2005) 1559–1563.

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

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

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

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

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

[12] S. Washietl, M. Kellis, M. Garber, Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals, Genome Res. 24 (2014) 616–628.

[13] A. Necsulea, M. Soumillon, M. Warnefors, A. Liechti, T. Daish, U. Zeller, J.C. Baker, F. Grutzner, H. Kaessmann, The evolution of lncRNA repertoires and expression patterns in tetrapods, Nature 505 (2014) 635–640.

[14] H. Hezroni, D. Koppstein, M.G. Schwartz, A. Avrutin, D.P. Bartel, I. Ulitsky, Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species, Cell Rep. 11 (7) (2015) 1110–1122.

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

[16] T. Gutschner, S. Diederichs, The hallmarks of cancer: a long non-coding RNA point of view, RNA Biol. 9 (2012) 703–719.

[17] A. Fatica, I. Bozzoni, Long non-coding RNAs: new players in cell differentiation and development, Nat. Rev. Genet. 15 (2014) 7–21.

[18] M. Sauvageau, L.A. Goff, S. Lodato, B. Bonev, A.F. Groff, C. Gerhardinger, D.B. Sanchez-Gomez, E. Hacisuleyman, E. Li, M. Spence, S.C. Liapis, W. Mallard, M. Morse, M.R. Swerdel, M.F. D'Ecclessis, J.C. Moore, V. Lai, G. Gong, G.D. Yancopoulos, D. Frendewey, M. Kellis, R.P. Hart, D.M. Valenzuela, P. Arlotta, J.L. Rinn, Multiple knockout mouse models reveal lincRNAs are required for life and brain development, eLife 2 (2013) e01749.

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

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

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

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

[23] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas Gholami, M. Lieberenz, M.M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, B. Kuster, Mass-spectrometry-based draft of the human proteome, Nature 509 (2014) 582–587.

[24] B. Banfai, H. Jia, J. Khatun, E. Wood, B. Risk, W.E. Gundling Jr., A. Kundaje, H.P. Gunawardena, Y. Yu, L. Xie, K. Krajewski, B.D. Strahl, X. Chen, P. Bickel, M.C. Giddings, J.B. Brown, L. Lipovich, Long noncoding RNAs are rarely translated in two human cell lines, Genome Res. 22 (2012) 1646–1657.

[25] M. Guttman, P. Russell, N.T. Ingolia, J.S. Weissman, E.S. Lander, Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins, Cell 154 (2013) 240–251.

[26] L. Wang, H.J. Park, S. Dasari, S. Wang, J.P. Kocher, W. Li, CPAT: coding-potential assessment tool using an alignment-free logistic regression model, Nucleic Acids Res. 41 (2013) e74.

[27] K. Sun, X. Chen, P. Jiang, X. Song, H. Wang, H. Sun, iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data, BMC Genomics 14 (Suppl. 2) (2013) S7.

[28] D.K. Gascoigne, S.W. Cheetham, P.B. Cattenoz, M.B. Clark, P.P. Amaral, R.J. Taft, D. Wilhelm, M.E. Dinger, J.S. Mattick, Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes, Bioinformatics 28 (2012) 3042–3050.

[29] L. Kong, Y. Zhang, Z.Q. Ye, X.Q. Liu, S.Q. Zhao, L. Wei, G. Gao, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine, Nucleic Acids Res. 35 (2007) W345–W349.

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

[31] R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L. Sonnhammer, J. Tate, M. Punta, Pfam: the protein families database, Nucleic Acids Res. 42 (2014) D222–D230.

Please cite this article as: G. Housman, I. Ulitsky, Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of lon..., Biochim. Biophys. Acta (2015), http://dx.doi.org/10.1016/j.bbagrm.2015.07.017

[32] S.R. Eddy, Profile hidden Markov models, Bioinformatics 14 (1998) 755–763.
[33] E. Rivas, S.R. Eddy, Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs, Bioinformatics 16 (2000) 583–605.
[34] D. Managadze, I.B. Rogozin, D. Chernikova, S.A. Shabalina, E.V. Koonin, Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs, Genome Biol. Evol. 3 (2011) 1390–1404.
[35] J. Yang, J. Zhang, Human long noncoding RNAs are substantially less folded than messenger RNAs, Mol. Biol. Evol. 32 (4) (2014) 970–977.
[36] D. Incarnato, F. Neri, F. Anselmi, S. Oliviero, Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome, Genome Biol. 15 (2014) 491.
[37] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition Elsevier Science, 2005.
[38] S.J. Andrews, J.A. Rothnagel, Emerging evidence for functional peptides encoded by short open reading frames, Nat. Rev. Genet. 15 (2014) 193–204.
[39] K. Hanada, K. Akiyama, T. Sakurai, T. Toyoda, K. Shinozaki, S.H. Shiu, sORF finder: a program package to identify small open reading frames with high coding potential, Bioinformatics 26 (2010) 399–400.
[40] B. Vanderperre, J.F. Lucier, X. Roucou, HAltORF: a database of predicted out-of-frame alternative open reading frames in human, Database: The Journal of Biological Databases and Curation2012 (bas025).
[41] A. Skarshewski, M. Stanton-Cook, T. Huber, S. Al Mansoori, R. Smith, S.A. Beatson, J.A. Rothnagel, uPEPperoni: an online tool for upstream open reading frame location and analysis of transcript conservation, BMC Bioinforma. 15 (2014) 36.
[42] W. Hu, B. Yuan, J. Flygare, H.F. Lodish, Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation, Genes Dev. 25 (2011) 2573–2578.
[43] R.B. Lanz, N.J. McKenna, S.A. Onate, U. Albrecht, J. Wong, S.Y. Tsai, M.J. Tsai, B.W. O'Malley, A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex, Cell 97 (1999) 17–27.
[44] M.I. Galindo, J.I. Pueyo, S. Fouix, S.A. Bishop, J.P. Couso, Peptides encoded by short ORFs control development and define a new eukaryotic gene family, PLoS Biol. 5 (2007) e106.
[45] D.M. Anderson, K.M. Anderson, C.L. Chang, C.A. Makarewich, B.R. Nelson, J.R. McAnally, P. Kasaragod, J.M. Shelton, J. Liou, R. Bassel-Duby, E.N. Olson, A micropeptide encoded by a putative long noncoding RNA regulates muscle performance, Cell 160 (2015) 595–606.
[46] K. Panzitt, M.M. Tschernatsch, C. Guelly, T. Moustafa, M. Stradner, H.M. Strohmaier, C.R. Buck, H. Denk, R. Schroeder, M. Trauner, K. Zatloukal, Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA, Gastroenterology 132 (2007) 330–342.
[47] S.A. Slavoff, A.J. Mitchell, A.G. Schwaid, M.N. Cabili, J. Ma, J.Z. Levin, A.D. Karger, B.A. Budnik, J.L. Rinn, A. Saghatelian, Peptidomic discovery of short open reading frame-encoded peptides in human cells, Nat. Chem. Biol. 9 (2013) 59–64.
[48] A. Pauli, M.L. Norris, E. Valen, G.L. Chew, J.A. Gagnon, S. Zimmerman, A. Mitchell, J. Ma, J. Dubrulle, D. Reyon, S.Q. Tsai, J.K. Joung, A. Saghatelian, A.F. Schier, Toddler: an embryonic signal that promotes cell movement via Apelin receptors, Science 343 (2014) 1248636.
[49] P. Wang, Y. Xue, Y. Han, L. Lin, C. Wu, S. Xu, Z. Jiang, J. Xu, Q. Liu, X. Cao, The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation, Science 344 (2014) 310–313.
[50] N.T. Ingolia, S. Ghaemmaghami, J.R. Newman, J.S. Weissman, Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling, Science 324 (2009) 218–223.
[51] J.A. Steitz, Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA, Nature 224 (1969) 957–964.
[52] N. Brockdorff, A. Ashworth, G.F. Kay, V.M. McCabe, D.P. Norris, P.J. Cooper, S. Swift, S. Rastan, The product of the mouse *Xist* gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus, Cell 71 (1992) 515–526.
[53] S. Carpenter, D. Aiello, M.K. Atianand, E.P. Ricci, P. Gandhi, L.L. Hall, M. Byron, B. Monks, M. Henry-Bezy, J.B. Lawrence, L.A. O'Neill, M.J. Moore, D.R. Caffrey, K.A. Fitzgerald, A long noncoding RNA mediates both activation and repression of immune response genes, Science 341 (2013) 789–792.
[54] A.M. Krichevsky, K.S. Kosik, Neuronal RNA granules: a link between RNA localization and stimulation-dependent translation, Neuron 32 (2001) 683–696.
[55] A.A. Bazzini, T.G. Johnstone, R. Christiano, S.D. Mackowiak, B. Obermayer, E.S. Fleming, C.E. Vejnar, M.T. Lee, N. Rajewsky, T.C. Walther, A.J. Giraldez, Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation, EMBO J. 33 (2014) 981–993.
[56] P. Juntawong, T. Girke, J. Bazin, J. Bailey-Serres, Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) E203–E212.
[57] J.E. Smith, J.R. Alvarez-Dominguez, N. Kline, N.J. Huynh, S. Geisler, W. Hu, J. Coller, K.E. Baker, Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*, Cell Rep. 7 (2014) 1858–1866.
[58] G.L. Chew, A. Pauli, J.L. Rinn, A. Regev, A.F. Schier, E. Valen, Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs, Development 140 (2013) 2828–2834.
[59] N.T. Ingolia, G.A. Brar, N. Stern-Ginossar, M.S. Harris, G.J. Talhouarne, S.E. Jackson, M.R. Wills, J.S. Weissman, Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes, Cell Rep. 8 (2014) 1365–1379.
[60] C. Fritsch, A. Herrmann, M. Nothnagel, K. Szafranski, K. Huse, F. Schumann, S. Schreiber, M. Platzer, M. Krawczak, J. Hampe, M. Brosch, Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting, Genome Res. 22 (2012) 2208–2218.
[61] N. Stern-Ginossar, B. Weisburd, A. Michalski, V.T. Le, M.Y. Hein, S.X. Huang, M. Ma, B. Shen, S.B. Qian, H. Hengel, M. Mann, N.T. Ingolia, J.S. Weissman, Decoding human cytomegalovirus, Science 338 (2012) 1088–1093.
[62] S. van Heesch, M. van Iterson, J. Jacobi, S. Boymans, P.B. Essers, E. de Bruijn, W. Hao, A.W. MacInnes, E. Cuppen, M. Simonis, Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes, Genome Biol. 15 (2014) R6.
[63] J. Carlevaro-Fita, A. Rahim, R. Guigo, L. Vardy, R. Johnson, Widespread Localisation of Long Noncoding RNAs to Ribosomes: Distinguishing Features and Evidence for Regulatory Roles, 2015.
[64] J.L. Aspden, Y.C. Eyre-Walker, R.J. Phillips, U. Amin, M.A. Mumtaz, M. Brocard, J.P. Couso, Extensive translation of small open reading frames revealed by poly-Ribo-seq, eLife 3 (2014) e03528.
[65] B.A. Wilson, J. Masel, Putatively noncoding transcripts show extensive association with ribosomes, Genome Biol. Evol. 3 (2011) 1245–1252.
[66] J. Crappe, W. Van Criekinge, G. Trooskens, E. Hayakawa, W. Luyten, G. Baggerman, G. Menschaert, Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs, BMC Genomics 14 (2013) 648.
[67] J. Ma, C.C. Ward, I. Jungreis, S.A. Slavoff, A.G. Schwaid, J. Neveu, B.A. Budnik, M. Kellis, A. Saghatelian, Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue, J. Proteome Res. 13 (2014) 1757–1765.
[68] M.S. Kim, S.M. Pinto, D. Getnet, R.S. Nirujogi, S.S. Manda, R. Chaerkady, A.K. Madugundu, D.S. Kelkar, R. Isserlin, S. Jain, J.K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N.A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L.D. Selvan, A.H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S.K. Sreenivasamurthy, A. Marimuthu, G.J. Sathe, S. Chavan, K.K. Datta, Y. Subbannayya, A. Sahu, S.D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K.R. Murthy, N. Syed, R. Goel, A.A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T.C. Huang, J. Zhong, X. Wu, P.G. Shaw, D. Freed, M.S. Zahari, K.K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C.J. Mitchell, S.K. Shankar, P. Satishchandra, J.T. Schroeder, R. Sirdeshmukh, A. Maitra, S.D. Leach, C.G. Drake, M.K. Halushka, T.S. Prasad, R.H. Hruban, C.L. Kerr, G.D. Bader, C.A. Iacobuzio-Donahue, H. Gowda, A. Pandey, A draft map of the human proteome, Nature 509 (2014) 575–581.
[69] H. Sun, C. Chen, M. Shi, D. Wang, M. Liu, D. Li, P. Yang, Y. Li, L. Xie, Integration of mass spectrometry and RNA-seq data to confirm human ab initio predicted genes and lncRNAs, Proteomics 14 (2014) 2760–2768.
[70] J. Crappe, E. Ndah, A. Koch, S. Steyaert, D. Gawron, S. De Keulenaer, E. De Meester, T. De Meyer, W. Van Criekinge, P. Van Damme, G. Menschaert, PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration, Nucleic Acids Res. 43 (5) (2014) e29.
[71] D.S. Kelkar, E. Provost, R. Chaerkady, B. Muthusamy, S.S. Manda, T. Subbannayya, L.D. Selvan, C.H. Wang, K.K. Datta, S. Woo, S.B. Dwivedi, S. Renuse, D. Getnet, T.C. Huang, M.S. Kim, S.M. Pinto, C.J. Mitchell, A.K. Madugundu, P. Kumar, J. Sharma, J. Advani, G. Dey, L. Balakrishnan, N. Syed, V. Nanjappa, Y. Subbannayya, R. Goel, T.S. Prasad, V. Bafna, R. Sirdeshmukh, H. Gowda, C. Wang, S.D. Leach, A. Pandey, Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis, Mol. Cell Proteomics 13 (2014) 3184–3198.
[72] G. Menschaert, W. Van Criekinge, T. Notelaers, A. Koch, J. Crappe, K. Gevaert, P. Van Damme, Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events, Mol. Cell Proteomics 12 (2013) 1780–1790.
[73] A.G. Schwaid, D.A. Shannon, J. Ma, S.A. Slavoff, J.Z. Levin, E. Weerapana, A. Saghatelian, Chemoproteomic discovery of cysteine-containing human short open reading frames, J. Am. Chem. Soc. 135 (2013) 16750–16753.
[74] A. Koch, D. Gawron, S. Steyaert, E. Ndah, J. Crappe, S. De Keulenaer, E. De Meester, M. Ma, B. Shen, K. Gevaert, W. Van Criekinge, P. Van Damme, G. Menschaert, A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites, Proteomics 14 (2014) 2688–2698.
[75] A.D. Tinoco, D.M. Tagore, A. Saghatelian, Expanding the dipeptidyl peptidase 4-regulated peptidome via an optimized peptidomics platform, J. Am. Chem. Soc. 132 (2010) 3819–3830.
[76] A. Staes, P. Van Damme, K. Helsens, H. Demol, J. Vandekerckhove, K. Gevaert, Improved recovery of proteome-informative, protein N-terminal peptides by combined fractional diagonal chromatography (COFRADIC), Proteomics 8 (2008) 1362–1370.
[77] P. Van Damme, D. Gawron, W. Van Criekinge, G. Menschaert, N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men, Mol. Cell Proteomics 13 (2014) 1245–1261.
[78] S. Prabakaran, M. Hemberg, R. Chauhan, D. Winter, R.Y. Tweedie-Cullen, C. Dittrich, E. Hong, J. Gunawardena, H. Steen, G. Kreiman, J.A. Steen, Quantitative profiling of peptides from RNAs classified as noncoding, Nat. Commun. 5 (2014) 5429.
[79] J. Crappé, W. Van Criekinge, G. Menschaert, Little things make big things happen: a summary of micropeptide encoding genes, EuPA Open Proteomics 3 (2014) 128–137.
[80] I. Ezkurdia, J. Vazquez, A. Valencia, M. Tress, Analyzing the first drafts of the human proteome, J. Proteome Res. 13 (8) (2014) 3854–3855.
[81] B. Pei, C. Sisu, A. Frankish, C. Howald, L. Habegger, X.J. Mu, R. Harte, S. Balasubramanian, A. Tanzer, M. Diekhans, A. Reymond, T.J. Hubbard, J. Harrow, M.B. Gerstein, The GENCODE pseudogene resource, Genome Biol. 13 (2012) R51.
[82] M. Kretz, Z. Siprashvili, C. Chu, D.E. Webster, A. Zehnder, K. Qu, C.S. Lee, R.J. Flockhart, A.F. Groff, J. Chow, D. Johnston, G.E. Kim, R.C. Spitale, R.A. Flynn, G.X. Zheng, S. Aiyer, A. Raj, J.L. Rinn, H.Y. Chang, P.A. Khavari, Control of somatic tissue differentiation by the long non-coding RNA *TINCR*, Nature 493 (2013) 231–235.

Please cite this article as: G. Housman, I. Ulitsky, Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of lon..., Biochim. Biophys. Acta (2015), http://dx.doi.org/10.1016/j.bbagrm.2015.07.017

[83] J.M. Dijkstra, K.T. Ballingall, Non-human lnc-DC orthologs encode Wdnm1-like protein, F1000Research 3 (2014) 160.
[84] A. Pauli, E. Valen, A.F. Schier, Identifying (non-)coding RNAs and small peptides: challenges and opportunities, BioEssays 37 (2015) 103–112.
[85] T. Saric, C.I. Graef, A.L. Goldberg, Pathway for degradation of peptides generated by proteasomes: a key role for thimet oligopeptidase and other metallopeptidases, J. Biol. Chem. 279 (2004) 46723–46732.
[86] S. Baboo, P.R. Cook, "Dark matter" worlds of unstable RNA and protein, Nucleus 5 (2014) 281–286.
[87] P.B. Hackett, R.B. Petersen, C.H. Hensel, F. Albericio, S.I. Gunderson, A.C. Palmenberg, G. Barany, Synthesis in vitro of a seven amino acid peptide encoded in the leader RNA of Rous sarcoma virus, J. Mol. Biol. 190 (1986) 45–57.
[88] A. Chen, Y.F. Kao, C.M. Brown, Translation of the first upstream ORF in the hepatitis B virus pregenomic RNA modulates translation at the core and polymerase initiation codons, Nucleic Acids Res. 33 (2005) 1169–1181.
[89] S.W. Eichhorn, H. Guo, S.E. McGeary, R.A. Rodriguez-Mias, C. Shin, D. Baek, S.H. Hsu, K. Ghoshal, J. Villen, D.P. Bartel, mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues, Mol. Cell 56 (2014) 104–115.
[90] J. Ruiz-Orera, X. Messeguer, J.A. Subirana, M.M. Alba, Long non-coding RNAs as a source of new peptides, eLife 3 (2014) e03523.
[91] M.L. Crowe, X.Q. Wang, J.A. Rothnagel, Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides, BMC Genomics 7 (2006) 16.
[92] K. Wethmar, J.J. Smink, A. Leutz, Upstream open reading frames: molecular switches in (patho)physiology, BioEssays 32 (2010) 885–893.
[93] S.E. Calvo, D.J. Pagliarini, V.K. Mootha, Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 7507–7512.
[94] J.T. Mendell, N.A. Sharifi, J.L. Meyers, F. Martinez-Murillo, H.C. Dietz, Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise, Nat. Genet. 36 (2004) 1073–1078.
[95] A. Gaba, A. Jacobson, M.S. Sachs, Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay, Mol. Cell 20 (2005) 449–460.
[96] C.M. Smith, J.A. Steitz, Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5′-terminal oligopyrimidine gene family reveals common features of snoRNA host genes, Mol. Cell. Biol. 18 (1998) 6897–6909.
[97] S. Lykke-Andersen, Y. Chen, B.R. Ardal, B. Lilje, J. Waage, A. Sandelin, T.H. Jensen, Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes, Genes Dev. 28 (2014) 2498–2517.
[98] G. Dieci, M. Preti, B. Montanini, Eukaryotic snoRNAs: a paradigm for gene expression flexibility, Genomics 94 (2009) 83–88.
[99] S. Washietl, S. Findeiss, S.A. Muller, S. Kalkhof, M. von Bergen, I.L. Hofacker, P.F. Stadler, N. Goldman, RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data, RNA 17 (2011) 578–594.
[100] M.F. Lin, I. Jungreis, M. Kellis, PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions, Bioinformatics 27 (2011) i275–i282.
[101] M. Re, G. Pesole, D.S. Horner, Accurate discrimination of conserved coding and non-coding regions through multiple indicators of evolutionary dynamics, BMC Bioinforma. 10 (2009) 282.
[102] R.T. Arrial, R.C. Togawa, M. Brigido Mde, Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis, BMC Bioinforma. 10 (2009) 239.
[103] L. Sun, H. Luo, D. Bu, G. Zhao, K. Yu, C. Zhang, Y. Liu, R. Chen, Y. Zhao, Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts, Nucleic Acids Res. 41 (2013) e166.
[104] A. Li, J. Zhang, Z. Zhou, PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme, BMC Bioinforma. 15 (2014) 311.
[105] Y. Wang, Y. Li, Q. Wang, Y. Lv, S. Wang, X. Chen, X. Yu, W. Jiang, X. Li, Computational identification of human long intergenic non-coding RNAs using a GA-SVM algorithm, Gene 533 (2014) 94–99.
[106] A.K. Biswas, B. Zhang, X. Wu, J.X. Gao, CNCTDiscriminator: coding and noncoding transcript discriminator — an excursion through hypothesis learning and ensemble learning approaches, J. Bioinforma. Comput. Biol. 11 (2013) 1342002.
[107] D. Ulveling, M.E. Dinger, C. Francastel, F. Hube, Identification of a dinucleotide signature that discriminates coding from non-coding long RNAs, Front. Genet. 5 (2014) 316.
[108] S. Lertampaiporn, C. Thammarongtham, C. Nukoolkit, B. Kaewkamnerdpong, M. Ruengjitchatchawalya, Identification of non-coding RNAs with a new composite feature in the hybrid random forest ensemble algorithm, Nucleic Acids Res. 42 (2014) e93.
[109] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, G.M. Weinstock, R.K. Wilson, R.A. Gibbs, W.J. Kent, W. Miller, D. Haussler, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes, Genome Res. 15 (2005) 1034–1050.

Please cite this article as: G. Housman, I. Ulitsky, Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of lon..., Biochim. Biophys. Acta (2015), http://dx.doi.org/10.1016/j.bbagrm.2015.07.017