Trends in Genetics

CellPress
REVIEWS

Review

Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation

Jorge Ruiz-Orera[1] and M. Mar Albà[1,2,*]

The translatome can be defined as the sum of the RNA sequences that are translated into proteins in the cell by the ribosomal machinery. Until recently, it was generally assumed that the translatome was essentially restricted to evolutionary conserved proteins encoded by the set of annotated protein-coding genes. However, it has become increasingly clear that it also includes small regulatory open reading frames (ORFs), functional micropeptides, *de novo* proteins, and the pervasive translation of likely nonfunctional proteins. Many of these ORFs have been discovered thanks to the development of ribosome profiling, a technique to sequence ribosome-protected RNA fragments. To fully capture the diversity of translated ORFs, we propose a comprehensive classification that includes the new types of translated ORFs in addition to standard proteins.

Highlights

Ribosome profiling sequencing techniques have provided evidence that many small ORFs are translated outside annotated coding sequences.

The functional proteome includes long conserved proteins, alternative isoforms, small proteins hidden in long noncoding RNAs, and recently evolved *de novo* proteins.

The translation of upstream small ORFs can regulate expression of the main coding sequence.

The pervasive translation of the transcriptome generates abundant raw material for the formation of functional proteins *de novo*.

Ribosome Profiling Reveals a Complex and Pervasive Translatome

In recent years we have made much progress in identifying the full set of sequences that are translated in the cell, or the **translatome** (see Glossary). We have learnt that, in addition to sequences encoding classical long proteins, there are many small **open reading frames (ORFs)** that are engaged by the translation machinery, resulting in the translation of small proteins or peptides. Some of these ORFs encode functional products, others are regulatory, and others are recent evolutionary events that may not yet have a function. Here, we provide an overview of the different types of translated ORFs and propose a classification that takes into account their evolutionary properties.

A recently developed RNA sequencing technique known as **ribosome profiling** has greatly contributed to the progress made in the past decade [1–6]. In contrast to **RNA sequencing (RNA-Seq)**, which targets complete RNA sequences, ribosome profiling is specific for ribosome-protected RNA fragments. The technique is very sensitive, and allows the detection of small proteins that may go undetected by proteomics-based methods. The sequencing reads provide a snapshot of the regions that are being translated in the cell. In addition, as the ribosome scans the coding sequences one codon at a time, the reads that correspond to actively translated regions show a characteristic three-nucleotide periodicity. Ribosome profiling has a single nucleotide resolution, because it is possible to identify the precise location of the **peptidyl-site (P-site)** of each sequencing read. The three-nucleotide periodicity of the reads can be used to predict *bona fide* translated sequences and discard spurious signals in the nontranslated transcriptome fraction (Figure 1). A number of computational approaches use three-nucleotide periodicity to predict translated ORFs in the transcriptome (Box 1) [7].

[1]Evolutionary Genomics Group, Research Programme in Biomedical Informatics, Hospital del Mar Research Institute, Universitat Pompeu Fabra, Barcelona, Spain
[2]Catalan Institution for Research and Advanced Studies, Barcelona, Spain

*Correspondence:
malba@imim.es (M.M. Albà).

 https://doi.org/10.1016/j.tig.2018.12.003
© 2018 Elsevier Ltd. All rights reserved.

CrossMark

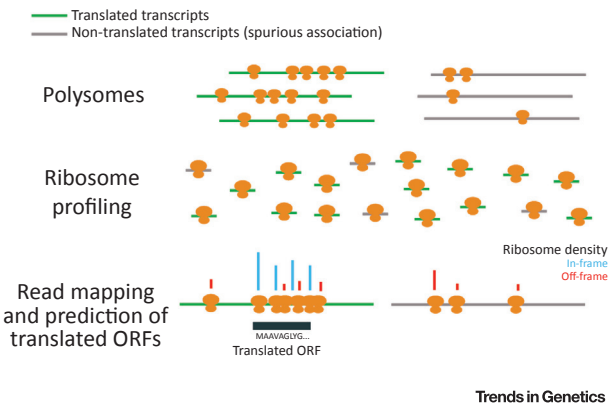

Figure 1. Workflow to Identify Actively Translated ORFs. Ribosome-protected RNA fragments are sequenced and mapped to all annotated transcripts in the species genome. On the basis of the three-nucleotide periodicity of the reads, and additional features, different software can predict actively translated ORFs. The prediction is based on a high fraction of reads in the correct frame (in-frame, blue colour) when compared to alternative ones (off-frame, red colour). This step is performed with single-nucleotide resolution after computing the read P-site per each read length. Abbreviations: ORF, open reading frame.

Ribosome profiling has confirmed the translation of thousands of annotated protein-coding sequences in different species, including many alternative protein isoforms [8]. Translation activity has also been observed in thousands of ORFs located upstream of annotated coding sequences (**upstream ORFs** or **uORFs**) [9,10] or in alternative frames of known coding sequences [11]. Finally, many transcripts thought to be noncoding have been reported to

Box 1. Programmes to Identify Actively Translated Sequences from ribosome profiling Data

ORFscore [30]: first method that analysed the three-nucleotide periodicity of translating ribosomes across ORFs. It requires the previous calculation of P-site positions per read length.

RibORF [14]: uses a support vector machine classifier that defines active translation of predicted ORF sequences based on three-nucleotide periodicity and uniformity across codons by calculating the percentage of maximum entropy values. It requires the previous calculation of P-site positions per each read length.

RiboTaper [15]: uses a multitaper spectral analysis method to remove noise and analyze three-nucleotide periodicity. It requires the previous calculation of P-site positions per read length.

RiboHMM [16]: builds a probabilistic model that takes into account the abundance of ribosome-protected fragments as well as their periodicity.

SPECtre [80]: analytical tool that models the overall periodicity of ribosomal occupancy using a classifier based on spectral coherence.

RB-BP [21]: an unsupervised Bayesian approach to predict translated ORFs from ribosome profiles. It incorporates and propagates uncertainty in the prediction process. Besides, there is automatic Bayesian selection of read lengths and ribosome P-site offsets.

PRICE [81]: infers actively translated codons using maximum likelihood for ribosome profiling reads. Based on the inferred codons, translated ORF candidates are identified using hypothesis tests based on the generalized binomial distribution.

RiboCode [20]: quantitative analysis of three-nucleotide periodicity for the *de novo* annotation of the full translatome. It calculates if the distribution of reads in a single frame is statistically significant when compared to the other two alternative frames. It requires the previous calculation of P-site positions per each read length.

Glossary

***De novo* ORF:** new ORF generated in a sequence.
***De novo* protein:** new protein generated from a previously non-coding region of the genome.
Genetic drift: process by which mutations change in frequency in the population over time by chance alone; some mutations with no beneficial effects become fixed and contribute to the divergence between species.
Long noncoding RNA (lncRNA): transcript longer than a given pre-established threshold (typically 200 nucleotides) with an experimentally validated noncoding function, or presumed not to encode any protein.
Micropeptide: functional small protein usually shorter than 100 aa.
Nonsense-mediated decay: surveillance pathway that eliminates translated transcripts with premature stop codons.
Open reading frame (ORF): sequence that can encode a protein. It starts with an initiation codon (often ATG for methionine) and ends in a stop codon.
Peptidyl-site (P-site): site for peptidyl transfer, which is the second binding site for transfer RNA in the ribosome.
Proteome: entire set of proteins in a cell, tissue, or organism.
Purifying selection: removal of deleterious mutations by natural selection; also referred to as negative selection.
Ribosome profiling: high-throughput sequencing of ribosome-protected RNA fragments, or ribosome profiling; it provides information on the parts of a transcript that are being translated.
Ribosome stalling: mechanisms by which a ribosome is paused while translating an ORF, affecting the translation of downstream ORFs.
RNA sequencing (RNA-Seq): high-throughput sequencing of the RNA fraction of the cell, tissue, or organism.
Translation initiation context: sequence surrounding the translation initiation codon, it influences the efficiency of translation.
Translation reinitiation: mechanism by which a ribosome that translates an ORF can restart translation at a downstream ORF.

show clear signatures of translation in different species, such as humans [12–16], mice [2,17], *Drosophila* [13], zebrafish [18], yeast [1], and plants [19]. Some of the predictions have been subsequently confirmed by mass spectrometry proteomics data [17,20,21].

Translatome: sum of translated ORFs.
Upstream ORF (uORF): ORF within the 5′ UTR of a protein-coding transcript.

In summary, ribosome-profiling experiments have revealed that translation is more extensive than initially thought. A significant fraction of the translatome maps to untranslated regions (UTRs) and sequences previously considered to be noncoding; careful identification and classification of the translated sequences is required to fully understand how the genes are regulated and translated.

Evolution Rules: A New Classification of Translated ORFs

Conservation of ORFs across species can help identifying functional small proteins, or **micropeptides**, that have been preserved by natural selection [22]. The set of small functional proteins also includes proteins with unique sequences that have recently emerged *de novo* from previously noncoding sequences [23]. Translated ORFs, however, do not need to encode functional proteins. In mouse, a large fraction of the peptides derived from species-specific ORFs appear to evolve neutrally [17]. In the case of regulatory ORFs, translation of the ORF can affect the stability of the transcript, or the translation of the main ORF, irrespective of the sequence of the encoded product [24]. Proteins that are small and poorly conserved across species are likely to be missed by automatic annotation pipelines; thus, the use of genome-wide experimental methods (ribosome profiling, mass-spectrometry-based proteomics) is key to their identification.

The identification of homologues of a new candidate protein can provide strong evidence of its functionality. However, the power of sequence similarity searches depends on the size of the protein. As a result, homologues of small proteins may be missed even if the sequence is similar in other species [22]. Highly divergent or small homologous ORFs can be identified using genomic alignments, but in this case it is important to ensure that the similarity is significant and not simply the result of a short divergence time between the species. We also have to consider that some ORFs might be located in sequences that are conserved for reasons other than the preservation of the protein, such as the presence of regulatory elements [25]. The genomic alignments can also help confirm cases in which the ORF is truly species or lineage specific [23].

The presence of **purifying selection** in translated sequences can be assessed using the ratio between the number of nonsynonymous and synonymous substitutions (dN/dS or Ka/Ks) in homologous sequence alignments. Under no selection at the protein sequence level, we expect the ratio to be approximately 1, whereas a ratio that is significantly lower than 1 is consistent with purifying selection [26]. In sequences that are not conserved across species it is possible to use single nucleotide variants instead of substitutions to test for signatures of selection [17]. In the case of ORFs located in genomic regions that can be aligned across multiple species, phylogenetic codon substitution frequencies (PhyloCSFs) can help to discriminate between coding and noncoding regions [27]. The UCSC database also scores the conservation level of small genomic elements (PhastCons) or nucleotides (PhyloP) in different genomes by analysing multiple species alignments [28]. These scores offer a possible alternative to the ratio between nonsynonymous to synonymous substitution to detect putative functional micropeptides [29–31]. They can also be useful to search for selection signatures not directly related to the protein, such as the need to maintain a given ORF **translation initiation context** to ensure proper levels of translation.

To account for the diversity of translated ORFs, we propose an ORF classification that takes into account sequence and/or positional conservation across species as well as the presence

or absence of purifying selection signatures (Figure 2, Key Figure). This classification encompasses long conserved ORFs but also small ORFs with potentially different roles in regulation and evolutionary innovation. The different types of translated ORFs will be described in detail in the next sections.

The Missing Part: (Very) Small Proteins

The number of annotated proteins varies extensively depending on the species and the annotation pipeline. Most of the annotated **proteomes** are almost exclusively composed of long conserved proteins (Figure 2, Class F1). This is not surprising, as both a minimum sequence length (typically 100 amino acids) and species conservation are usually required for the automatic annotation of coding sequences. This ensures a low number of false positives but many small proteins remain unannotated [22,32].

A large fraction of the transcriptome is composed of transcripts without any predicted long ORF. These transcripts are usually classified as **long noncoding RNAs** (**lncRNAs**; Box 2). However, some of them have been subsequently found to contain micropeptides (Table 1; Figure 2, Class F2). The functions of these micropeptides are very heterogeneous, although several of them have been shown to play roles in muscle function (e.g., myoregulin [33], myomixer/myomerger [34,35], MOXI/mitoregulin [36,37]). Promising candidates are identified combining different approaches, including the prediction of translation features by ribosome profiling or proteomics and the identification of sequence conservation across species [38–40].

An interesting case is the developmental gene *Mlpt/Tal/Pri* found in different insect species. The gene was initially believed to be noncoding but it was later discovered that it encoded several functional peptides [41–44]. These peptides, produced from a polycistronic transcript, direct the proteolytic cleavage of the transcription factor Shavenbaby (*Svb*), converting it from a repressor to an activator. Another example is the gene *MIEF1*, which translates a protein involved in the regulation of mitochondrial fission but also contains another ORF translating a 70-amino-acid (aa) micropeptide. The latter peptide is conserved across vertebrates (MIEF1-MP) and regulates mitochondrial translation by binding to the mitoribosome [45].

Many gene isoforms encode putative small proteins, but few of them have been characterised. One recently studied example is the human *C7orf49* gene [46]. This gene encodes a 157 aa polypeptide that modulates retrovirus infection, but also an alternative isoform which is only 69 aa long and is involved in DNA repair. The smaller isoform was discovered by a proteomics approach directed towards the identification of small nonannotated proteins, highlighting the importance of using novel strategies for the detection of the small proteome.

Functional proteins may also be encoded by transcripts known to have noncoding regulatory roles. One striking example is a protein recently described to be encoded by the telomerase RNA component, which appears to regulate crosstalk between autophagy and apoptosis [47]. A reverse example is the *oskar* mRNA, which translates a protein that is required for the formation of the germline in *Drosophila* but also has a regulatory role in early oogenesis through its 3′ UTR [48]. These interesting cases point to the possibility that many other transcripts are bifunctional, and that the current dichotomy between coding and noncoding needs to be reconsidered.

When the Act of Translation, and Not the Protein Product, Is What Matters

Translation has traditionally been associated with the production of a functional protein or peptide. However, some translated ORFs can exert a regulatory effect on the same or

Key Figure

Classification of Translated ORFs Based on Evolutionary Properties

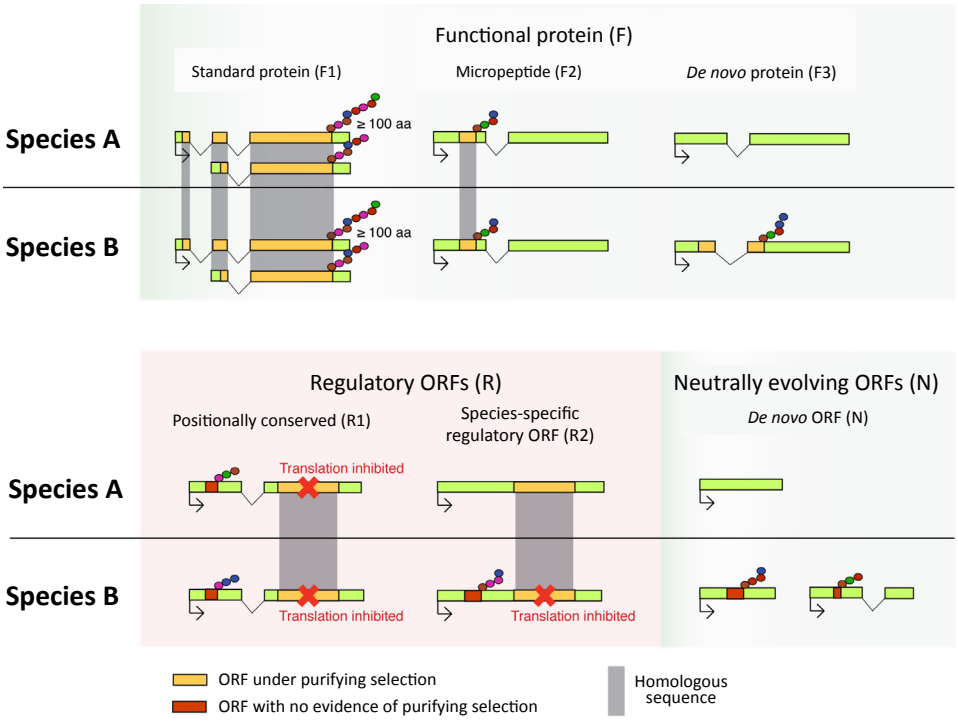

Figure 2. We propose different classes of translated ORFs based on functionality and conservation between two (or more) species. Sequences that show signatures of purifying selection are likely to be functional proteins (Class F). While most proteins are ≥100 aa (Class F1), the number of functional smaller proteins, or micropeptides (Class F2), is growing. Finally, each species contains new functional proteins that have evolved *de novo* (Class F3). Other ORFs are regulatory (Class R) and the sequences are poorly or no constrained at the sequence level because, in most cases, the peptide itself is not functional. These ORFs are often located upstream of a main ORF encoding a functional protein (uORFs). The act of translation of uORFs can regulate the stability or translation of the main ORF. In the figure, the translation of the represented regulatory ORFs inhibits the translation of the main ORF. Although the product of these ORFs is usually not conserved across species, the relative position of the ORF may be constrained (positional homologues, class R1). Some of these uORFs, however, are species specific (class R2). Finally, each species and lineage contains a large number of transcripts containing *de novo* ORFs that are translated but evolve neutrally (Class N). For simplicity, in the diagram each gene only contains one ORF class but different classes may be translated from the same transcript. Abbreviations: aa, amino acids; ORF, open reading frame; uORF, upstream ORF.

neighbouring genes. Often, the relative position of the ORF in the transcript needs to be conserved, but the actual protein sequence is irrelevant (Figure 2, Class R1).

uORFs are the main class of regulatory small ORFs. It has been observed that regulation through uORFs is conserved across vertebrates for dozens of genes [24,49] and in many cases translation starts from non-AUG start codons [2]. Most uORFs do not appear to be constrained at the sequence level. For example, a genome-wide study in *Drosophila* has found that 93.5%

Box 2. The Vast World of lncRNAs

Over the past decade, numerous studies have reported the expression of a large number of lncRNAs in human and mouse cells by using large-scale cDNA sequencing and next-generation sequencing techniques [82]. Perhaps the most iconic of these efforts was the ENCODE project, which described a large number of unannotated noncoding transcripts in humans [83]. The genes formed complex transcriptional units that could include both coding and noncoding isoforms. Other transcripts were part of loci that appeared to be entirely noncoding.

In general, lncRNAs do not contain long ORFs, are polyadenylated, and short (they often have just two exons). Although some have well characterized functions in the nucleus, in general they tend to accumulate in the cytoplasm [84]. With the exception of some lncRNAs with known regulatory functions, lncRNAs are expressed at lower levels than protein-coding RNAs are, and are more prominent in certain tissues, such as testes [85]. Despite being annotated as noncoding, many lncRNAs contain short ORFs; there is now strong evidence that some of these ORFs can be translated [12–16].

The number of annotated lncRNAs has dramatically increased over recent years and, in mammals, it nearly equals that of protein-coding genes. This is a conservative estimate, as many additional lncRNAs can be detected by *de novo* transcript assembly of RNA-Seq reads. Variations in the number of estimated lncRNAs may arise from differences in the data source, sequencing strategies, or computational methods. lncRNAs are a common feature of eukaryotic genomes; *Drosophila*, with a more compact genome than mammals, contains around 3000 different lncRNAs [86].

of the uORFs with ribosome profiling translation signatures have PhyloCSF scores that are below those observed for functional coding sequences [50]. Therefore, only a small subset of the uORFs is likely to encode peptides that have a function *per se*.

The presence of an ORF in a UTR often produces a repressive effect on the translation and/or transcription of the main coding sequence (reviewed in [9]). As a result, uORFs are often depleted from regions in the proximity of coding sequences [24]. Several metabolic or neurological diseases in humans are associated with mutations that generate new uORFs and repress the expression of the main protein [9]. In *Drosophila* it has been found that mutations in

Table 1. Examples of Identified Functional Micropeptides in Transcripts Previously Annotated as Noncoding

Name	Length (aa)	Conservation	Function	Refs
Toddler/APELA	54 aa	Human and mouse	Embryonic signal that promotes cell movement via Apelin receptors.	[87]
Sarcolamban	28 or 29 aa	Human and flies	Regulation of cardiac calcium uptake.	[88]
ENOD40	12–24 aa	Plants	Binds to the sucrose synthase. Bifunctional RNA.	[89]
Myomixer/Myomerger	84 aa	Human and mouse	Controls cell fusion and muscle formation.	[34,35]
SPAAR	75 aa	Human and mouse	Regulation of mTORC1 and muscle regeneration.	[90]
NoBody	68 aa	Human and mouse	Interacts with the mRNA decapping complex.	[91]
DWORF	34 aa	Human and mouse	Activates the SERCA pump by physical interaction.	[92]
Myoregulin	46 aa	Human and mouse	Interacts directly with SERCA and impedes Ca^{2+} uptake into the SR.	[33]
CASIMO1	83 aa	Human and mouse	Controls cell proliferation and interacts with squalene epoxidase, modulating lipid droplet formation.	[93]
MOXI/Mitoregulin	56 aa	Human and mouse	Enhancer of mitochondrial β-oxidation.	[36,37]
Pint	87 aa	Human and mouse	Inhibits transcriptional elongation of multiple oncogenes.	[94]
Bouncer	125 aa	Zebrafish	Mediates sperm–egg binding and is crucial for fertilization.	[95]
Mlpt-Tal-Pri	10–32 aa	Insects	Required for embryonic and imaginal development	[41–44]

uORF-translated sequences have low derived allele frequencies, as expected if some of them are deleterious [50].

The repression of the translation of the main protein by uORFs may occur by several mechanisms, including **ribosome stalling**, inhibition of **translation reinitiation,** or uORF induced **nonsense-mediated decay (NMD)** [51]. In the *Arabidopsis bZIP11* gene, a peptide produced by a uORF can sense sucrose abundance, and promote ribosome stalling at the uORF stop codon at high sucrose concentrations, inhibiting expression of the main gene product [52]. Another example of a uORF-regulated gene is *GCN4*, which encodes a transcription factor that activates the response to amino acid starvation in baker's yeast. The 5′ UTR of *GCN4* contains several uORFs that inhibit translation reinitiation, preventing expression of the main protein product. Under stress conditions, however, the repressive uORFs are not efficiently translated, and the GCN4 protein can be produced [53].

Even if only a small fraction of regulatory ORFs in UTRs are constrained at the sequence level, features related to translational regulation can be the targets of selection. This includes the uORF relative position and length, which can affect the ribosome reinitiation efficiency, the presence of secondary structures, and the translation initiation sequence context. Actually, different studies support the idea that uORFs contain initiation contexts that prevent their efficient translation except in specific conditions. For example, the *GGP* gene in plants contains a highly conserved uORF starting with ACG that is selectively translated under high ascorbate conditions, when the translation of GDP-l-galactose phosphorylase (GGP) protein is not required. Mutating ACG to AUG leads to malfunctioning of this regulatory mechanism, and GGP is no longer translated in normal conditions [54]. In this case, the use of a nonoptimal start codon for the uORF protects the main coding sequence. In other cases, the inhibition of translation is achieved by ORFs that partially overlap the main translated sequence. For instance, the translation of overlapping ORFs in human and yeast *STN1* gene strongly reduces translation and also induces NMD of the transcript, affecting telomere function [55].

Translated ORFs in lncRNAs have sometimes been compared to uORFs in protein-coding transcripts [18]. However, the peptides translated from lncRNAs tend to be longer than those translated from uORFs [17]. A similar regulatory function is difficult to imagine in lncRNAs, as there is no downstream ORF encoding a functional protein. As it occurs in aberrant protein-coding transcripts, some ORFs can trigger NMD in lncRNAs; in yeast, up to 16.8% of unannotated transcripts increased their abundance $\geq$2-fold in the absence of NMD [56], suggesting that they are normally degraded by this mechanism. Other ORFs in lncRNAs have been hypothesized to protect the transcript from the translation of downstream elements that might produce harmful peptides or trigger transcript decay. For instance, suppression of an upstream ORF in the mouse lncRNA *Jpx* increases the rate of translation downstream of the uORF, suggesting that it has a regulatory function [57].

Some ORFs have been found to enhance RNA stability and/or translation. Many antisense transcripts can physically interact with their protein-coding pairs to increase the translation of the main protein [58]. The analysis of ribosome-profiling data from *Arabidopsis* identified many antisense transcripts containing actively translated ORFs. Five of them were upregulated during phosphorus starvation and their expression was positively correlated with expression of the corresponding sense mRNAs, consistent with such a mechanism [19].

Finally, plant precursors of miRNAs and siRNAs also contain translated ORFs [19,59]. Peptides produced in miRNA precursors have been defined as miRNA-encoded peptides (miPEPs) and

belong to ancient miRNA families that are conserved across all flowering plants. It has been reported that these peptides enhance the accumulation of the mature miRNAs by increasing the transcription of the precursor miRNAs [59].

De novo ORFs and Evolutionary Innovation

The analysis of ribosome profiling has also uncovered a large number of translated sequences that are species or lineage specific [17,60,61]. The putative new translated proteins have unique sequences and may potentially contribute to novel functions. Proteins that have presumably emerged from initially noncoding regions of the genome are said to have originated *de novo* [62–64]. A study in yeast termed such newly translated ORFs protogenes, to indicate that they have the potential to become new functional genes [61]. The resulting proteins are usually shorter than 100 aa and show little or no signatures of purifying selection (Figure 2, Class N) [17,61]. Although selection signatures may be more difficult to identify in proteins that have only recently become functional than in well-established older proteins, the data suggest that many new proteins are likely to be nonfunctional [17].

The existence of many species-specific small proteins or peptides can probably be explained by the pervasive nature of translation, affecting a large part of the transcriptome [12–14]. Furthermore, the genome is exposed to a fast transcriptional turnover, as inferred from the differences in the genomic regions that are transcribed across closely related mouse taxa [65]. These high transcriptional dynamics, together with the continuous accumulation of ORF-enabling and -disabling mutations (affecting start and stop codons), is expected to result in a high rate of gain or loss of translatable ORFs over time. Comparison of RNA-Seq and ribosome profiling data for human and mouse brain samples [66] provides many examples of such translation volatility (Figure 3). One example is a mouse-specific uORF in the 5′ UTR of the conserved human and mouse gene *Elfn2* (Figure 3A); the second example is a putative *de novo* gene from mice, overlapping the conserved *Kif3c* gene and expressing a new peptide (Figure 3B).

The translation of nonfunctional new peptides may seem paradoxical. Why should the cell spend energy in translating something that is not useful? First, we have to consider that species-specific transcripts are usually expressed at low levels [13,67,68], which greatly reduces the cost of producing them. Besides, theoretical estimates indicate that, in eukaryotes, the cost of transcribing and translating a gene, especially if expressed at low or moderate levels, is probably too small to be counteracted by negative selection [69]. Provided the peptides are not toxic, translation-enabling mutations may simply fix in the population by **genetic drift**. If the peptides provide an advantage the fixation process can be much more rapid. The relative influence of drift and selection depends on the species under consideration [70]. In humans, most neutrally evolving lineage-specific transcripts appear to be fixed in the population, which is consistent with drift [71]. The new proteins can generate novel antigens that need to be recognised as self by the adaptive immune system, otherwise the cells expressing them may be attacked and destroyed. *De novo* proteins that are expressed in more than one tissue are enriched in the thymus and spleen, which contain specialised cells that regulate the distinction between self and non-self [72].

Over evolutionary time, some of the newly generated peptides may provide a selective advantage, becoming a ***de novo* protein** that continues to evolve under purifying selection (Figure 2, Class F3) [23]. The gain of a useful function may occur by mutations in the ORF itself, but also by changes in other molecules, or in the environment, turning a dispensable peptide into something useful once the context has changed. It is worth noting that random peptides

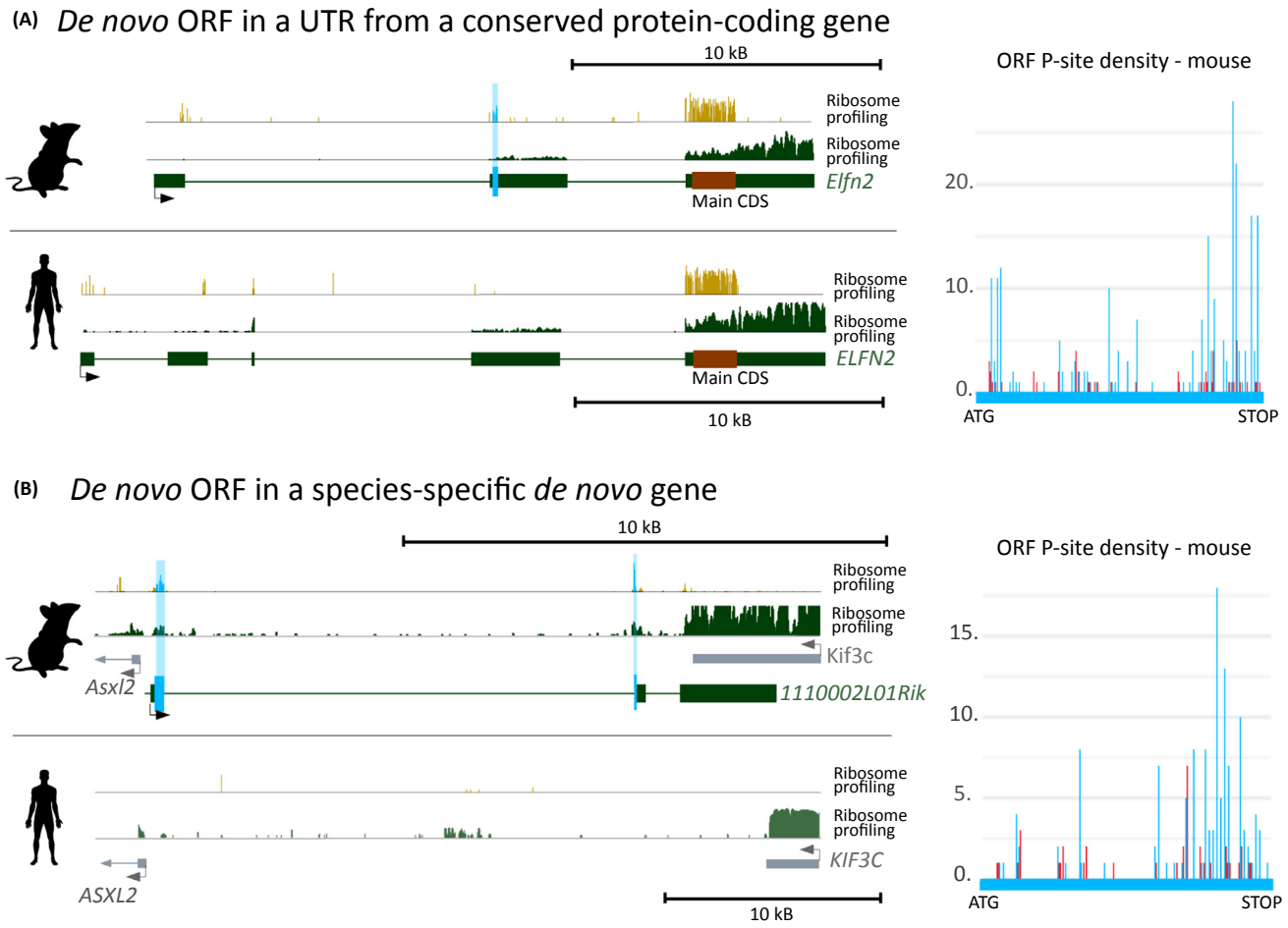

Figure 3. Evidence of Translated *de novo* ORFs in Ribosome-Profiling Experiments. Ribosome profiling and RNA-Seq sequencing data from mouse and human brain tissues were mapped to the corresponding genome sequences. Syntenic genomic regions that include conserved protein-coding genes are shown (*ELFN2* and *KIF3C*). ORFs identified as translated by RibORF, and the corresponding ribosome profiling reads, are indicated in blue. Moreover, two diagrams representing the density of ribosome profiling P-sites per ORF position are shown; blue peaks correspond to inframe reads, read peaks correspond to off-frame reads. (A) An upstream ORF in the gene *Elfn2* is detected in mice but not in humans. (B) Divergent transcription from the promoter of the *Actg1* gene results in a new transcript in mouse (1110002L01Rik) that encodes a putative peptide; no similar transcript is observed in human. Abbreviations: ORF, open reading frame; P-site, peptidyl-site; RNA-Seq, RNA sequencing.

can have biochemical activities; for example, selection for ATP-binding activity in a library of randomly generated 80 aa polypeptides successfully identified several candidates capable of binding to ATP [73]. In *Escherichia coli*, selection for stress tolerance in a random library of 20 aa peptides identified a specific peptide that conferred resistance to nickel [74]. In summary, ***de novo* ORFs** provide abundant material for the birth of new functional proteins.

There are many examples of proteins likely to have evolved *de novo* [23]. The gene *BSC4* in *Saccharomyces cerevisiae* was one of the first reported cases [63]. This gene, encoding a 132 aa protein, may help yeast to survive periods of starvation. The syntenic genomic regions of the closely related species *Saccharomyces paradoxus* and *Saccharomyces mikatae* are also transcribed, but they lack the capacity to encode a similar protein, suggesting that the ancestral gene may have been a noncoding RNA. Another *de novo* gene in *S. cerevisiae*, *MDF1*, confers

a selective advantage in rich medium by promoting the rapid consumption of glucose [75]. The putative *de novo* gene *QQS* in *Arabidopsis* modulates the production of protein versus starch in plants [76]. Finally, several *de novo* proteins in mammals are likely to be involved in the immune response or cancer [23,77].

Recently evolved *de novo* proteins are small and poorly conserved across species; therefore, many of them are likely to be missing from the current gene annotations, or be wrongly annotated as noncoding loci. The use of experimental methods to identify translated ORFs, including ribosome profiling and proteomics, should help close this gap and improve our understanding of the impact of *de novo* proteins in evolutionary innovation. Regulatory uORFs can also be considered to have arisen *de novo* from previously noncoding sequences (Figure 2, Class R2), making *de novo* ORF emergence a common evolutionary process. Although some newly arisen uORFs may be deleterious, others show evidence of positive selection [50]. It seems clear that the continuous emergence of novel uORFs provides new opportunities for fine-tuning regulation of gene expression through the control of translation.

Concluding Remarks and Future Perspectives

New studies over the past decade have revealed the existence of multiple translated ORFs hidden within the transcriptome, revealing an additional level of complexity in the translatome that was previously unappreciated [1,4,22]. Translated ORFs can correspond to functional proteins, but also be of a regulatory nature. The high sensitivity of ribosome profiling has also opened a window to the detection of many translated ORFs that have not yet been purged by selection or drift [17].

Regulatory ORFs are usually small and located in regions generally believed to be noncoding, such as 5′ UTRs [10,24]. They can modulate expression of other proteins or RNAs *in trans*, by the action of the encoded peptide, or *in cis*, by directly interfering with the expression of a downstream ORF. The number of transcripts with potentially translated uORFs is large, and future studies will be required to identify those with the highest impact in translational regulation.

The pervasive translation of the transcriptome facilitates the birth of new functional proteins *de novo* [23]. Initially, many of the translated proteins are likely to be nonfunctional but, over time, a small fraction of them may provide a selective advantage and be maintained by selection. By definition, randomly occurring ORFs tend to be small. This implies that recently generated *de novo* proteins will be short. These proteins may become longer over time by the fusion of different proteins or protein domains, or by the co-option of adjacent noncoding exonic sequences into a coding function [78,79].

The development of new tools and methodologies to identify translated sequences, and the refinement of the existing ones, will help to complete the characterization of the translatome for different species (see Outstanding Questions). With the information gained from these methods it should be possible to develop improved algorithms to predict whether a small ORF will be translated or not. Very small proteins (<24 aa) are still challenging to identify even with current methods. We also need to find better ways to predict the functionality of the translated ORFs on a one-to-one basis.

Finally, the discovery of thousands of translated small ORFs opens the question of how these ORFs should be annotated in the databases. Ribosome profiling is a powerful tool to identify translation events but translation is not necessarily equivalent to protein accumulation, and some proteins could be rapidly degraded after being produced. In annotated protein-coding

Outstanding Questions

What is the real extent of the translatome? How can we incorporate the described plethora of small translated ORFs into current gene annotations?

Ribosome-profiling applications are still in their infancy. What can we expect from future applications of this technology to different cell types and disease states? How many small proteins await discovery?

How can we identify recently evolved proteins that are functional? Selection tests based on amino acid substitutions may not be sensitive enough to identify signatures of selection in these proteins.

How many upstream ORFs are regulatory? How can we differentiate between measurable effects and actual fitness-affecting activities?

Many translation events are species specific. How many of these events are likely to be maintained over time? What percentage of the functional proteins have arisen *de novo*?

genes the label 'protein evidence' (yes or no) is used to indicate whether there is evidence of the existence of the protein by proteomics or other methods. A similar tag named 'translation evidence' could be used for translation events identified by ribosome-profiling-based methods. Additionally, the new ORFs should be classified according to their relatively localisation in the transcript and their level of evolutionary conservation. Given the large number of species- or lineage-specific ORF translation events that are detected, we argue that an evolutionary perspective is key to integrating and interpreting the new data to come.

Acknowledgements

We would like to acknowledge three anonymous reviewers for their useful comments. The work was funded by grant BFU2015-65235-P from Ministerio de Economía e Innovación (Spanish Government) co-funded by FEDER (EU), and by grants 2014SGR1121 and 2017SGR01020 from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, Generalitat de Catalunya).

References

1. Ingolia, N.T. *et al.* (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–223
2. 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
3. Andreev, D.E. *et al.* (2017) Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res.* 45, 513–526
4. Brar, G.A. and Weissman, J.S. (2015) Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* 16, 651–664
5. Aspden, J.L. *et al.* (2014) Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *Elife* 3, e03528
6. Ingolia, N.T. (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* 15, 205–213
7. Calviello, L. and Ohler, U. (2017) Beyond read-counts: Ribo-seq data analysis to understand the functions of the transcriptome. *Trends Genet.* 33, 728–744
8. Weatheritt, R.J. *et al.* (2016) The ribosome-engaged landscape of alternative splicing. *Nat. Struct. Mol. Biol.* 23, 1117–1123
9. Barbosa, C. *et al.* (2013) Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet.* 9, e1003529
10. Cabrera-Quio, L.E. *et al.* (2016) Decoding sORF translation –from small proteins to gene regulation. *RNA Biol.* 13, 1051–1059
11. Michel, A.M. *et al.* (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.* 22, 2219–2229
12. Ingolia, N.T. *et al.* (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* 8, 1365–1379
13. Ruiz-Orera, J. *et al.* (2014) Long non-coding RNAs as a source of new peptides. *Elife* 3, 1–24
14. Ji, Z. *et al.* (2015) Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* 4, e08890
15. Calviello, L. *et al.* (2016) Detecting actively translated open reading frames in ribosome profiling data. *Nat. Meth.* 13, 165–170
16. Raj, A. *et al.* (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* 5, e13328
17. Ruiz-Orera, J. *et al.* (2018) Translation of neutrally evolving peptides provides a basis for *de novo* gene evolution. *Nat. Ecol. Evol.* 2, 890–896
18. Chew, G.-L. *et al.* (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. *Development* 140, 2828–2834
19. Bazin, J. *et al.* (2017) Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. *Proc. Natl. Acad. Sci. U. S. A.* 114, E10018–E10027
20. Xiao, Z. *et al.* (2018) *De novo* annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Res.* 46, e61
21. Malone, B. *et al.* (2017) Bayesian prediction of RNA translation from ribosome profiling. *Nucleic Acids Res.* 45, 2960–2972
22. Couso, J.-P. and Patraquim, P. (2017) Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* 18, 575–589
23. McLysaght, A. and Hurst, L.D. (2016) Open questions in the study of *de novo* genes: what, how and why. *Nat. Rev. Genet.* 17, 567–578
24. Chew, G.-L. *et al.* (2016) Conservation of uORF repressiveness and sequence features in mouse human and zebrafish. *Nat. Commun.* 7, 11663
25. Tautz, D. (2009) Polycistronic peptide coding genes in eukaryotes – how widespread are they? *Brief. Funct. Genomic. Proteomic* 8, 68–74
26. Hurst, L.D. (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18, 486
27. Lin, M.F. *et al.* (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27, i275–i282
28. Pollard, K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20, 110–121
29. Aspden, J.L. *et al.* (2014) Extensive translation of small ORFs revealed by Poly-Ribo-Seq. *Elife* 3, e03528
30. Bazzini, A.A. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* 33, 981–993
31. Mackowiak, S.D. *et al.* (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* 16, 1–21
32. Basrai, M.A. *et al.* (1997) Small open reading frames: beautiful needles in the haystack. *Genome Res.* 7, 768–771
33. Anderson, D.M. *et al.* (2015) A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* 160, 595–606
34. Bi, P. *et al.* (2017) Control of muscle formation by the fusogenic micropeptide myomixer. *Science* 356, 323–327
35. Quinn, M.E. *et al.* (2017) Myomerger induces fusion of non-fusogenic cells and is required for skeletal muscle development. *Nat. Commun.* 8, 15665
36. Makarewich, C.A. *et al.* (2018) MOXI is a mitochondrial micropeptide that enhances fatty acid β-oxidation. *Cell Rep.* 23, 3701–3709

37. Stein, C.S. *et al.* (2018) Mitoregulin: a lncRNA-encoded microprotein that supports mitochondrial supercomplexes and respiratory efficiency. *Cell Rep.* 23, 3710–3720.e8
38. Crappé, J. *et al.* (2014) Little things make big things happen: a summary of micropeptide encoding genes. *EuPA Open Proteomics* 3, 128–137
39. Makarewich, C.A. and Olson, E.N. (2017) Mining for micropeptides. *Trends Cell. Biol.* 27, 685–696
40. Fields, A.P. *et al.* (2015) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell.* 60, 816–827
41. Kondo, T. *et al.* (2007) Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat. Cell Biol.* 9, 660–665
42. Zanet, J. *et al.* (2015) Pri sORF peptides induce selective proteasome-mediated protein processing. *Science* 349, 1356–1358
43. Savard, J. *et al.* (2006) A segmentation gene in tribolium produces a polycistronic mRNA that codes for multiple conserved peptides. *Cell* 126, 559–569
44. Kondo, T. *et al.* (2010) Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* 329, 336–339
45. Rathore, A. *et al.* (2018) MIEF1 microprotein regulates mitochondrial translation. *Biochemistry* 57, 5564–5575
46. Slavoff a, S. *et al.* (2014) A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J. Biol. Chem.* 289, 10950–10957
47. Rubtsova, M. *et al.* (2018) Protein encoded in human telomerase RNA is involved in cell protective pathways. *Nucleic Acids Res.* 46, 8966–8977
48. Jenny, A. *et al.* (2006) A translation-independent role of oskar RNA in early *Drosophila* oogenesis. *Development* 133, 2827–2833
49. Johnstone, T.G. *et al.* (2016) Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* 35, 706–723
50. Zhang, H. *et al.* (2018) Genome-wide maps of ribosomal occupancy provide insights into adaptive evolution and regulatory roles of uORFs during *Drosophila* development. *PLoS Biol.* 16, e2003903
51. Hellens, R.P. *et al.* (2016) The emerging world of small ORFs. *Trends Plant Sci.* 21, 317–328
52. Yamashita, Y. *et al.* (2017) Sucrose sensing through nascent peptide-meditated ribosome stalling at the stop codon of *Arabidopsis* bZIP11 uORF2. *FEBS Lett.* 591, 1266–1277
53. Hinnebusch, A.G. (2005) Translational regulation of GCN4 and the general amino acid control of yeast. *Annu. Rev. Microbiol.* 59, 407–450
54. Laing, W.A. *et al.* (2015) An upstream open reading frame is essential for feedback regulation of ascorbate biosynthesis in *Arabidopsis*. *Plant Cell* 27, 772–786
55. Torrance, V. and Lydall, D. (2018) Overlapping open reading frames strongly reduce human and yeast STN1 gene expression and affect telomere function. *PLoS Genet.* 14, e1007523
56. Smith, J.E. *et al.* (2014) Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. *Cell Rep.* 7, 1858–1866
57. Hezroni, H. *et al.* (2017) A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biol.* 18, 162
58. Carrieri, C. *et al.* (2012) Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* 491, 454–457
59. Lauressergues, D. *et al.* (2015) Primary transcripts of microRNAs encode regulatory peptides. *Nature* 520, 90
60. Wilson, B.A. and Masel, J. (2011) Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.* 3, 1245–1252
61. Carvunis, A.-R. *et al.* (2012) Proto-genes and *de novo* gene birth. *Nature* 487, 370–374
62. Begun, D.J. *et al.* (2007) Evidence for *de novo* evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* clade. *Genetics* 176, 1131–1137
63. Cai, J. *et al.* (2008) De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* 179, 487–496
64. Toll-Riera, M. *et al.* (2009) Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* 26, 603–612
65. Neme, R. and Tautz, D. (2016) Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence. *Elife* 5, e09977
66. Gonzalez, C. *et al.* (2014) Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J. Neurosci.* 34, 10924–10936
67. 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
68. 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
69. Lynch, M. and Marinov, G.K. (2015) The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci. U. S. A.* 112, 15690–15695
70. Haerty, W. and Ponting, C.P. (2013) Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol.* 14, R49
71. Ruiz-Orera, J. *et al.* (2015) Origins of *de novo* genes in human and chimpanzee. *PLoS Genet.* 11, e1005721
72. Bekpen, C. *et al.* (2018) Dealing with the adaptive immune system during *de novo* evolution of genes from intergenic sequences. *BMC Evol. Biol.* 18, 121
73. Keefe a, D. and Szostak, J.W. (2001) Functional proteins from a random-sequence library. *Nature* 410, 715–718
74. Stepanov, V.G. and Fox, G.E. (2007) Stress-driven *in vivo* selection of a functional mini-gene from a randomized DNA library expressing combinatorial peptides in *Escherichia coli*. *Mol. Biol. Evol.* 24, 1480–1491
75. Li, D. *et al.* (2010) A *de novo* originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.* 20, 408–420
76. Li, L. *et al.* (2015) QQS orphan gene regulates carbon and nitrogen partitioning across species via NF–YC interactions. *Proc. Natl. Acad. Sci. U. S. A.* 112, 14734–14739
77. Luis Villanueva-Cañas, J. *et al.* (2017) New genes and functional innovation in mammals. *Genome Biol. Evol.* 9, 1886–1900
78. Toll-Riera, M. and Albà, M.M. (2013) Emergence of novel domains in proteins. *BMC Evol. Biol.* 13, 47
79. Bornberg-Bauer, E. *et al.* (2015) Emergence of *de novo* proteins from 'dark genomic matter' by 'grow slow and moult'. *Biochem. Soc. Trans.* 43, 867–873
80. Chun, S.Y. *et al.* (2016) SPECtre: a spectral coherence-based classifier of actively translated transcripts from ribosome profiling sequence data. *BMC Bioinformatics* 17, 482
81. Erhard, F. *et al.* (2018) Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods* 15, 363
82. Okazaki, Y. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573
83. Djebali, S. *et al.* (2012) Landscape of transcription in human cells. *Nature* 489, 101–108
84. van Heesch, S. *et al.* (2014) Extensive localization of long non-coding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol.* 15, R6
85. Washietl, S. *et al.* (2014) Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* 24, 616–628
86. Brown, J.B. *et al.* (2014) Diversity and dynamics of the *Drosophila* transcriptome. *Nature* 512, 393–399
87. Pauli, A. *et al.* (2014) Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* 343, 1248636
88. Magny, E.G. *et al.* (2013) Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* 341, 1116–1120

89. Rohrig, H. *et al.* (2002) Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proc. Natl. Acad. Sci. U. S. A.* 99, 1915–1920
90. Matsumoto, A. *et al.* (2016) mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* 541, 228
91. D'Lima, N.G. *et al.* (2017) A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* 13, 174–180
92. Nelson, B.R. *et al.* (2016) A peptide encoded by a transcript annotated as long noncoding RNA enhances CERCA activity in muscle. *Science* 351, 271–275
93. Polycarpou-Schwarz, M. *et al.* (2018) The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene* 37, 4750–4768
94. Zhang, M. *et al.* (2018) A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nat. Commun.* 9, 4475
95. Herberg, S. *et al.* (2018) The Ly6/uPAR protein Bouncer is necessary and sufficient for species-specific fertilization. *Science* 361, 1029–1033