Emerging evidence for functional peptides encoded by short open reading frames

Shea J. Andrews and Joseph A. Rothnagel

Abstract | Short open reading frames (sORFs) are a common feature of all genomes, but their coding potential has mostly been disregarded, partly because of the difficulty in determining whether these sequences are translated. Recent innovations in computing, proteomics and high-throughput analyses of translation start sites have begun to address this challenge and have identified hundreds of putative coding sORFs. The translation of some of these has been confirmed, although the contribution of their peptide products to cellular functions remains largely unknown. This Review examines this hitherto overlooked component of the proteome and considers potential roles for sORF-encoded peptides.

Short open reading frames (sORFs). Open reading frames that are usually < 100 codons in length but that can also be longer.

Coding DNA sequence (CDS). An open reading frame (ORF) that encodes a verified protein product. The CDS is typically the first ORF identified and characterized on an mRNA. It defines the end of the 5′ leader and the start of the 3′ trailer sequences.

School of Chemistry and Molecular Biosciences, University of Queensland, St. Lucia, Queensland, 4072, Australia.
Correspondence to J.A.R.
e-mail: j.rothnagel@uq.edu.au
doi:10.1038/nrg3520
Published online 11 February 2014; corrected online 4 March 2014

The proteome varies substantially between different cell types. Diversity in protein expression is derived not only from the straightforward expression of protein-coding genes but also from the use of alternative transcription start sites and processes such as alternative splicing, transcript editing and post-translational modifications. However, recent work has identified an additional component of the proteome: the translation of short open reading frames (sORFs), which has demanded an examination of the potential for sORFs to encode biologically active peptides that have regulatory roles in eukaryotic cells. Although not a topic of this Review, sORFs and their encoded peptides have also been identified in bacteria[1,2].

An open reading frame (ORF) is a potentially translatable sequence that consists of a string of in-frame sense codons beginning with a start codon and ending with a stop codon. A translatable ORF is typically recognized as the coding DNA sequence (CDS) on an mRNA that gives rise to its principle protein product. sORFs are distinguished from all other ORFs by their sizes, but not all sORFs are translated or are indeed translatable. Translatable sORFs have been found within the 5′ leader and 3′ trailer sequences, within or overlapping with the previously known ORF (that is, the CDS) of mRNAs and in various transcripts that were previously thought to be non-coding RNAs (ncRNAs), including long ncRNAs (lncRNAs), intergenic and antisense transcripts. They are potentially translated by leaky scanning and re-initiation (FIG. 1). The sizes of sORFs can range from a theoretical lower limit of two codons (that is, a start codon and a stop codon) to ~100 codons. The smallest translated coding sORF described so far is six codons long[3]. The upper limit is ill-defined, with some studies describing sORFs of 200–250 codons in length, and a consensus has yet to emerge[4–6]. The peptides encoded by these examples blur any distinction between them and the protein products of usually longer ORFs on mRNAs.

Whereas it is straightforward to identify any ORF in a stretch of genomic DNA, it is much more challenging to differentiate coding sORFs from non-coding sORFs. Most *ab initio* gene prediction programs distinguish coding sequences from non-coding ones either by recognizing patterns in genomic sequences that denote features (such as canonical initiation codons, termination sites, splice sites, promoter sequences and polyadenylation signals) or by examining the innate characteristics of the DNA sequence itself (such as codon usage bias, nucleotide composition and in-frame hexamer frequency that might indicate coding potential)[7–9]. However, these gene prediction programs are not well-suited for identifying sORFs with coding potential, as they are designed to assess the coding potential of ORFs that are longer than 100 codons and that are richer in these features[10,11]. Consequently, most prediction programs are not specifically designed to distinguish between coding and non-coding sORFs. Indeed, many gene annotation algorithms will dismiss ORFs of <100 codons as meaningless[12,13]. In addition, these programs use multiple criteria for evaluating the coding potential of an ORF in order to decrease the false-positive identification rate. However, this property also increases the rate of identifying false negatives. This is a particular concern

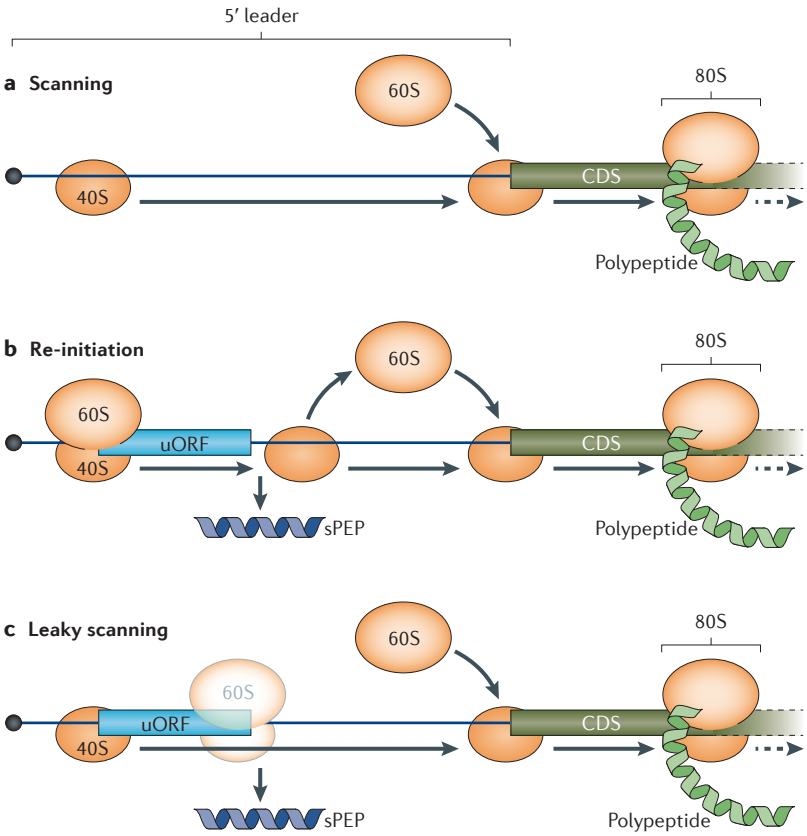

Figure 1 | **Leaky scanning and re-initiation.** **a** | In the standard scanning model of eukaryotic translation, the 40S ribosomal subunit, together with pre-initiation factors (not shown), binds to the 5′ cap of the mRNA and scans along the transcript until the first initiation codon is recognized. The 60S subunit and additional factors (not shown) then combine with the 40S subunit to form the 80S elongation-competent ribosome, which translates the open reading frame (ORF). **b** | Ribosomal re-initiation occurs when a 40S subunit initiates translation at the start codon of a downstream coding DNA sequence (CDS) after completing translation of an upstream ORF (uORF) owing to non-dissociation of the 40S subunit from the mRNA. Translation of the uORF leads to the production of a short peptide encoded by sORFs (sPEP). **c** | In leaky scanning, the 40S ribosomal subunit can either recognize the start codon of an uORF and lead to the translation of a sPEP, or it can scan past the upstream start codon and initiate translation at a downstream start codon. The number of ribosomes initiating at an upstream start codon is determined by the sequences surrounding the start codon. A proportion of ribosomes will initiate at an upstream start codon and translate the uORF, and the rest will ignore the upstream start codon and continue scanning along the transcript to initiate translation at a downstream start codon.

for studies that look for translatable sORFs, as relatively few coding sORFs are likely to be present in the large pool of mostly non-functional sORFs[9,12,13]. As a result, false-negative rejection of potential coding sORFs is a common occurrence[10,14].

Furthermore, many experimental approaches do not adequately cater for the identification of sORFs or their encoded peptides. For example, a random mutagenesis screen is less likely to introduce inactivating mutations in sORFs than in longer coding sequences[15,16]. Short peptides encoded by sORFs (sPEPs) are also less likely to be identified in standard proteomic screens, as these are generally limited to the analysis of proteins that are >10 kDa[17]. Moreover, the average tissue content of sPEPs is estimated to be ~0.1% of protein levels or 10–1,000 molecules per cell[18], and they can often be subject to rapid degradation and/or loss during extraction and purification procedures, which further diminishes the prospects of identifying these peptides[17]. Detection is further impeded by the paucity of information on sORF sequences that is available in the databases used to identify peptides in tandem mass spectrometry outputs. However, recent advances — such as high-throughput identification of translation start sites[19–21], improved sensitivity of proteomic techniques[18], specialized algorithms for identifying putative coding sORFs[22] and better integration of bioinformatic, genomic and proteomic outputs[6,23–25] — have helped to uncover this previously hidden proteomic 'dark matter'.

In this Review, we discuss the various strategies that are available for identifying putative coding sORFs and for distinguishing them from untranslated ones. These include computational and bioinformatic approaches that have been adapted from those used to find larger coding ORFs through the analysis of sequence conservation, codon content and other sequence features, as well as experimental approaches that use transcriptional and translational data. We then provide an overview of the potential for sequences that were previously thought to be non-coding to harbour translatable sORFs, including those present in ncRNAs, and the 5′ leader and 3′ trailer sequences of mRNAs. We appraise the evidence for translation of sORFs and discuss examples of biologically active sPEPs. We also examine the evidence for expression of peptides that originate from the translation of sORFs and discuss their possible functionality and contribution to the proteome.

Identifying putative coding sORFs

The identification of putative coding sORFs has so far relied on three broad strategies that are analogous to those used in conventional gene prediction studies but have been tailored for the identification of sORFs. These are cross-species comparisons of sORF sequences to identify conserved sequences; examination of the codon content and coding features within sORF sequences to differentiate potential coding sORFs from non-coding ones; and analyses of transcriptional and translational experimental data to identify coding sORFs that show evidence of expression.

Computational approaches. The discovery phase involves identifying potential coding sORFs that are distinct from established known coding ORFs. Web-based tools such as sORFfinder[26], HAltORF[27] and uPEPperoni[28] can be used in the initial search phase to locate sORFs that have coding potential. A useful strategy for finding biologically relevant sORFs is to carry out a cross-species comparison to identify conserved sORFs. This step can be combined with a K_a/K_s test to ascertain evidence of purifying, positive or neutral selection[29]. In addition, more stringent tests for conservation can also be used to reduce the likelihood of identifying false positives. For example, these tests can require sequences to be of similar length or located in the

K_a/K_s test
A ratio that compares the number of nonsynonymous substitutions per nonsynonymous site with the number of synonymous substitutions per synonymous site.

same position relative to a previously identified ORF[30], or they can require sequences to be located within the same syntenic region of a comparison genome for intergenic sORFs[31]. Evidence of evolutionary conservation is important, as sORFs that lack cross-species conservation are more likely to be random sequences that do not encode functional peptides[32]. However, non-conserved sORFs should not be dismissed a priori and ought to be retained for further analyses, as species-specific sPEPs may also be biologically relevant. Equally, the search parameters for both genic and intergenic sORFs can be made less stringent by including non-canonical initiation codons. This is an important consideration because ORFs with non-canonical start codons are more common than those initiated from AUG codons[20], and peptides encoded by sORFs with non-AUG start codons have been identified by mass spectrometry[18].

Cross-species comparison approaches can be combined with methods that analyse sequence content and/or gene characteristics to distinguish coding sORFs from non-coding ones. First, sORFs can be sorted by the context of their start codons, as those with an optimal Kozak setting are more likely to be translated[8,33]. Second, the coding potential of sORFs can be determined by analysing features such as nucleotide composition, codon usage and residue bias[34]. Finally, potential coding sORFs can be analysed for the presence of functional domains and/or motifs using algorithms such as Pfam[35] and by grouping sORFs into families of similar sequences[36,37]. These methods are not necessarily mutually exclusive, and a combination of one or more approaches can be used to boost confidence in the identification of bona fide coding sORFs. To this end, sophisticated algorithms, such as Coding Index[10], have been developed specifically for determining the coding potential of sORFs. In addition, gene finder packages that incorporate statistical methods — such as Bayes' estimation, support vector machines and Markov chain models — have also been used; examples of these include CSTMiner[38], CRITICA[39] and Coding Potential Calculator[40].

Experimental evidence. It is necessary to corroborate computational data with biological evidence. For example, the expression of transcripts with sORFs, particularly those from intergenic regions, cannot always be determined *in silico* and needs to be verified. Various methodologies can be used to analyse gene expression, including reverse transcriptase PCR, DNA microarrays, genomic tiling arrays, probing of cDNA or expressed sequence tag (EST) libraries, serial analysis of gene expression (SAGE) and next-generation RNA sequencing (RNA-seq). Although the sensitivity of some of these techniques allows the identification of small and/or rare transcripts, they have limitations such as high background noise, low throughput, prohibitive costs and the requirement for large amounts of RNA[41]. RNA-seq overcomes many of these issues, although transcriptional noise that results from random transcription initiation events may still be problematic[42]. Of course, the usual caveats on transcription data apply with respect to the cell type used and the cell-cycle stage analysed.

Although evidence of transcription of sORF-containing regions is useful, evidence of their translation is essential for determining their contribution to the proteome. A recent advance in this area is ribosome profiling, which uses deep sequencing of ribosome-protected mRNA fragments to provide a genome-wide 'snapshot' of translation[43]. This results in the unbiased identification of regions that are undergoing translation (or at least initiation) at a particular time point, and ribosome profiling can therefore delineate the exact position of all ORFs regardless of whether they are present on protein-coding transcripts[25]. Ribosome profiling also has the additional benefit of identifying non-canonical start codons. However, it should be noted that ribosomal occupancy does not necessarily equate to translation of the following ORF, as it has been shown that start codons are used for the regulation of processes such as attenuation of translation of a downstream ORF and regulation of mRNA availability by inducing nonsense-mediated decay[44]. Alternatively, start codons may be arbitrarily occupied by ribosomes without consequence to the cell or the organism[44]. Therefore, the data provided by ribosomal profiling need to be analysed in combination with the bioinformatic methods outlined above in order to obtain a set of high-confidence potential coding sORFs.

Evidence of sORF translation can also be obtained directly by protein mass spectrometry. Theoretically, any peptide that is present above a threshold level in a cell or tissue fraction can be identified by interrogating its mass spectrum against a database of known or predicted peptides[25,45]. However, currently available databases only contain reference proteomes of experimentally verified or predicted protein sequences and are thus unlikely to contain many sPEP sequences. A high-confidence data set of potential coding sORFs that are obtained from cell-type-specific RNA-seq[18] or ribosome profiling data[25] can greatly assist their identification by mass spectrometry. In contrast to this corroborative approach, proteogenomics offers an unbiased approach for identifying novel and unpredicted sORFs. In this procedure, the mass spectral data is searched against a database that contains the conceptual translation of all six reading frames of the raw genome assembly[45]. Nevertheless, the identification of sPEPs using mass spectrometry is still challenging owing to their small size and low abundance, and particular attention needs to be paid to the peptide isolation and enrichment steps, which are crucial for finding small and/or low-abundance products in cell lysates[25].

sORFs in annotated non-coding transcripts

The intergenic regions of eukaryotic genomes are extensively transcribed[46–49]. Although there is debate in regards to expression levels[50] and function[51], intergenic transcription leads to the expression of a range of RNAs that are mostly assumed to be ncRNAs. Given that sORFs of <100 codons occur frequently by chance[14,52], it is not surprising that sORFs are also present on ncRNAs. As short and intermediate ncRNAs are too small (that is, <200 nucleotides in length) to support

translation[53], it is generally thought that translatable sORFs will only be found on lncRNAs, but this remains to be demonstrated given that ribosome profiling has detected translation initiation sites on smaller transcripts that were previously thought to be non-coding[21]. However, as lncRNAs could potentially account for a substantial proportion of all transcripts[54], they represent a large store of potential protein-coding sequences. A recent study has identified 593,586 intergenic sORFs in *Drosophila melanogaster* by a bioinformatic analysis of non-coding euchromatic DNA[31]. Similarly, an analysis of intergenic[10] and whole-genome[4] sequences in *Arabidopsis thaliana* identified ~600,000 potential intergenic sORFs. When further filters were applied, these numbers dropped significantly to 7,159 (REF. 10) and 33,809 (REF. 4) putative intergenic coding sORFs, respectively. Similar numbers of intergenic sORFs have been found in other species: ~12,000 were found in the hardwood species *Populus deltoides*[6] (that is, cottonwood) and ~41,000 were found in mice[14] (TABLE 1). On average, intergenic sORFs that have high coding potential constitute ~5% of all annotated ORFs for a range of representative eukaryotes in the US National Center for Biotechnology Information (NCBI) RefSeq database[16]. An early bioinformatic analysis of the FANTOM database of mouse transcripts found that 7.3% of intergenic sORFs are likely to encode novel peptides, two-thirds of which are conserved in rats and half of which are conserved in humans[14]. Importantly, most of these intergenic sORFs (94%) are found on expressed transcripts, which indicates that the peptides encoded by these sequences have a high potential to be expressed[14]. A similar observation was made in *A. thaliana*, in which 5% of intergenic sORFs were found to be located in transcribed regions[10].

Short coding sequences in mRNAs

Upstream ORFs. The presence of sORFs within 5′ leader sequences (which are commonly referred to as upstream ORFs (uORFs)) was noted in the first systematic survey of mRNA sequences[33], although an understanding of their coding potential has taken much longer to develop. In general, uORFs modulate ribosome access to a downstream CDS and influence its translation through various mechanisms (BOX 1). uORFs do not share their reading frame with the downstream CDS and are typically shorter. They are ubiquitous and have been identified in most eukaryotes, including vertebrates[55,56], insects[5], fungi[30,57,58] and plants[36,59,60]. They show diverse features such as the number of uORFs in a 5′ leader sequence, their position within the 5′ leader sequence and their length[61]. They are also abundant — 34–48% of mRNAs contain one or more uORFs[55,56,62–66] — although the number of observed uORFs is lower than that expected by chance (the observed/expected ratio is 0.64), which suggests that they are under purifying selection[55,64]. This ratio suggests that the emergence of new uORFs is selected against because of probable deleterious consequences (see BOX 2 for examples of mutations in uORFs) and that retained uORFs have biological roles.

Overlapping and downstream sORFs. Reading frames that overlap but are frameshifted from the generally longer ORF that specifies the CDS have been identified in human and rodent transcripts, and these represent another source of alternatively translated products. A distinction is made between overlapping sORFs and the uORFs that extend past the start codon of the longer, previously characterized ORF (that is, the CDS). Some studies further differentiate overlapping sORFs into additional subtypes: those that lie completely within a known ORF, those that extend from the known ORF into the 3′ trailer sequence and dual-coding transcripts that are generated by alternative splicing[67,68]. Four bioinformatic studies on overlapping ORFs have been published so far[27,67,69,70] (TABLE 1). The earliest study identified 40 overlapping sORF-containing genes using conservative parameters; overlapping sORFs were required to be at least 500 bp (that is, 167 codons) in length and conserved in humans and two other species[69]. By comparison, another study using less stringent parameters identified 1,793 overlapping sORFs, from an analysis of 9,163 human RefSeq sequences, that are >150 bp in length and are conserved in rodents[67]. An examination of the sequence context of their start codons reduced this number to 217 putative overlapping coding sORFs that contain an optimal Kozak ribosome initiation motif.

In contrast to the study of 5′ leader sequences and coding regions, the study of 3′ trailer sequences has attracted little attention with respect to identifying and characterizing downstream sORFs because they were considered not to be translated nor indeed translatable[20]. However, as most 3′ trailer sequences are generally much longer than 5′ leader sequences[56,71], they could be expected to contain more sORFs. Indeed, we found this to be the case: downstream sORFs are ten times more abundant than uORFs in the human and mouse RefSeq database[56].

Translation of sORFs

Short peptides encoded by intergenic sORFs and previously annotated ncRNAs. There is some contention as to whether sORFs on lncRNAs are translatable. Ribosome profiling of mouse embryonic stem cells (mESCs) using a metric based on the density of ribosome profiling reads relative to mRNA expression indicated that nearly half of the lncRNAs analysed contained sORFs that showed evidence of initiation[20]. However, as discussed above, ribosomal occupancy does not necessarily equate to translation, and the latest ribosome profiling studies have tempered the initial predictions that arose from this approach. A recent study that used machine-learning tools to analyse the ribosome profiles of eight early zebrafish developmental stages found that, depending on the data set used, 8–45% of lncRNAs were likely to be translated, and 18–44% of lncRNAs were unlikely to be translated[72]. A re-analysis of the mESC profiling data using the ribosome release score (RRS; that is, a metric based on the release of translating ribosomes after encountering a stop codon) found that most lncRNAs do not

Table 1 | **Studies that have identified putative coding sORFs or coding sORFs**

Location of sORF or data set of sequences analysed	Number of transcripts or sORFs analysed	Approaches						Expression verification		Number of sORFs with coding potential identified	Ref
		Sequence similarity with other species	K_a/K_s analysis	Length or position similarity with other species	Initiation context	Nucleotide composition	Protein domains, motifs and clusters	Transcription analysis	Mass spectrometry		
Humans											
5′ leader	27,660	✓	–	–	–	–	–	–	–	43	55
5′ leader	21,768	✓	✓	–	✓	–	–	–	–	204	56
Overlapping with CDS	14,159	✓	✓	–	–	✓	–	–	–	40	69
Overlapping with CDS	9,163	✓	–	✓	✓	✓	✓	–	–	217	67
Overlapping with CDS	26,009	✓	–	–	✓	–	–	–	–	168	70
Overlapping with CDS	76,000	–	–	–	✓	–	–	–	–	24,547	27
Whole genome*	NA	–	–	–	–	–	–	✓	✓	4	75
Whole genome*	NA	–	–	–	–	–	–	✓	✓	8	76
Whole genome*	NA	–	–	–	–	–	–	✓	✓	90	18
Whole mRNA‡	83,886	–	–	–	–	–	–	–	✓	1,259	24
Mice											
Intergenic	102,801	–	✓	–	–	✓	–	–	–	1,240	14
Drosophila melanogaster											
5′ leader	19,389	✓	✓	–	–	–	–	–	–	44	5
Intergenic	593,586	✓	✓	✓	–	–	–	✓	–	401	31
Arabidopsis thaliana											
5′ leader	34,000	✓	✓	–	–	–	–	–	–	19	36
5′ leader	23,036	✓	✓	✓	–	–	–	–	–	18	22
5′ leader	10,122	✓	✓	–	–	–	–	–	–	18	60
Intergenic	570,948	✓	✓	–	–	✓	–	✓	–	3,241	10
Intergenic	96,358	✓	✓	–	–	✓	–	✓	–	2,302	74
Intergenic	606,285	✓	–	–	–	–	✓	✓	–	1,044	4
Rice											
5′ leader	32,127	✓	–	–	–	–	–	–	–	29§ and 15‖	59
Cottonwood											
Intergenic	12,852	✓	–	–	–	–	✓	✓	✓	611	6
Common bean											
Intergenic	31,576	✓	–	–	–	–	✓	✓	–	776	37
Yeast											
5′ leader	5,542	✓	–	✓	–	–	–	✓	–	15	30
5′ leader	5,602	✓	–	✓	–	–	–	–	–	252	58
5′ leader	2,167	✓	✓	✓	–	–	–	–	–	12	57

CDS, coding DNA sequence; NA, not applicable; sORF, short open reading frame. *Peptides were identified in human cell lines using mass spectrometry, and sequences were then mapped to the US National Center for Biotechnology Information (NCBI) RefSeq database or to in-house RNA and cDNA databases. ‡Reference library consisted of mRNA sequences compiled from GenBank. §Conserved in rice, wheat, barley, maize and sorghum. ‖Conserved in rice and *A. thaliana*.

encode peptides[44]. By comparing RNA-seq and tandem mass spectrometry data from the Encyclopedia of DNA elements (ENCODE) project, a study on the transcriptome of human cells predicted that up to 8% of lncRNAs are translated[73]. However, it should be noted that the mESC study[44] compared classes of transcripts rather than individual transcripts, whereas the human transcriptomic study[73] examined individual transcripts using data from the ENCODE project. There is also supporting proteomic evidence for intergenic sPEP expression in human cells with the identification of 49 peptides encoded by sORFs on lncRNAs, antisense transcripts and unannotated transcripts[18].

A proteomic study in *A. thaliana* found 5,426 short peptides, 905 of which were encoded by genes that had not been previously annotated in the reference databases[23]. Moreover, 155 of these novel peptides mapped to sORFs that were identified in an earlier bioinformatic study that evaluated the coding potential of intergenic sORFs[10]. A follow-up study identified 2,099 coding sORFs that reside on highly expressed transcripts in *A. thaliana*, and 571 of these had identifiable orthologues in other land plants[74]. These authors then carried out a functional assay on 473 intergenic sORFs that matched their selection criteria of high expression and high conservation. Of these sORFs, 49 produced visible phenotypes when overexpressed in transgenic plants, which suggests that the peptides encoded by these sORFs have regulatory roles in plant development[74]. This finding warrants further analyses and biochemical characterization of these sPEPs to determine their functionality. Another study confirmed the expression of 56 novel peptides from *P. deltoides* that had been originally identified using bioinformatic approaches from EST sequences[6].

Box 1 | **The canonical role of uORFs**

A near-universal function of upstream open reading frames (uORFs) is to attenuate translation of their associated downstream coding ORF (that is, the coding DNA sequence (CDS)) by regulating the passage of ribosomes on a 5′ leader sequence. Translation attenuation of the CDS primarily occurs through ribosomal re-initiation and leaky scanning mechanisms[110,111] in which uORFs intercept scanning ribosomes before they reach the CDS[112]. Examples of translational control by uORFs include translation of the mammalian transcripts *THPO* (which encodes thrombopoietin)[113], *CEBPA* (which encodes CCAAT/enhancer binding protein-α) and *CEBPB*[114], and the yeast transcript *GCN4* (REF. 115). Various studies have shown that removal of one or more uORFs from individual 5′ leader sequence increases the level of translation of the CDS[107], examples of which include *ERBB2* (also known as *HER-2*) in monkeys[116] and the *GLI1* (GLI family zinc-finger 1) transcript in humans, monkeys and rodents[117]. Conversely, adding uORFs to a synthetic 5′ leader results in a concomitant reduction in the translation of a reporter transcript with each additional uORF[118]. In addition, an analysis of the expression levels of 11,649 matched mRNA and protein species across a range of mouse tissues and developmental stages indicated that uORFs reduce CDS expression levels by 13–39%[107]. The experimental evidence is supported by the observation that the most abundant proteins are encoded by mRNAs with short 5′ leader sequences, which contain no uORFs, whereas the transcripts of less abundant proteins tend to have longer 5′ leader sequences and more uORFs[62–64,110]. Furthermore, it has long been noted that certain classes of transcripts are more likely to be encumbered by uORFs; transcripts that encode regulatory proteins such as transcription factors contain more uORFs than those encoding structural proteins[65,119].

Short peptides encoded by uORFs. The translation of uORFs was first predicted by the bioinformatic observation that nearly 20% of human uORFs have an initiation codon in an optimal Kozak context and are capable of supporting efficient recognition by scanning ribosomes[56]. This was subsequently confirmed by two high-throughput ribosome profiling studies on human and mouse cell lines. The first study identified 7,936 translation initiation sites upstream of an annotated translation start site, 85% of which were found to be conserved in mice[21]. The second study identified 4,400 translation initiation sites that correspond to uORFs, most of which are conserved in primates and ~60% of which are conserved in mice[19]. Furthermore, the number of translation initiation sites is similar to the number of unique human and mouse uORFs (6,454 and 5,089, respectively) that are found in our bioinformatic analysis of RefSeq transcripts[56]. However, the re-analysed mESC data using the RRS metric indicated that, although ribosomal binding is common in 5′ leader sequences, not all uORFs are translated[44]. Physical evidence that some uORFs are indeed translated has come from proteomic studies on human cell isolates[18,24,25,75,76]. The two early studies[75,76] identified nine novel peptides that originated from seven uORFs, and four of these showed a high degree conservation in mice, which indicates possible functional constraint[75]. Recent studies have corroborated these findings using a combination of more sensitive peptide separation techniques together with enhanced reference databases (which were populated with cell-type-specific sequences using RNA-seq transcriptome or ribosome profiling data). These enhancements resulted in the identification of a further 43 peptides that are derived from the translation of human uORFs[18,24,25].

Short peptides encoded by overlapping sORFs and downstream sORFs. The idea that non-viral transcripts could be translated from more than one reading frame was considered antithetical until the discovery of several peptides that are translated from overlapping sORFs. The early discovery of dual-coding transcripts, such as the tumour antigen *TYRP1* (also known as *TRP*) transcript[77] and the caspase 1 (*CASP1*; also known as *ICE*) transcript[78], underscored the need for improved bioinformatic and experimental approaches that are tailored to their discovery and characterization. An interesting approach has recently been developed to validate the translation of overlapping sORFs on a single transcript using an expression vector that contains two different tags which are out of frame of each other[24,70]. In this approach, the overlapping sORF-containing sequences are cloned upstream of the reporters, transfected into mammalian cell lines and analysed using microscopy and western blots. Proteomic studies have confirmed the expression of 80 short peptides that are encoded by overlapping sORFs[18,24,75,76], and ribosome profiling studies have identified the translation start sites of additional overlapping sORFs[20,68]. The ribosome profiling study estimated that >1% of human protein-coding genes show translation of two reading frames[68].

Box 2 | **Human disease-associated mutations in sPEPs**

The potential of upstream open reading frames (uORFs) to encode biologically active peptides also raises the possibility that mutations in certain short peptides encoded by short open reading frames (sPEPs) could result in a disease phenotype. A comprehensive study identified 509 unique genes in which insertions or deletions within their 5′ leader sequences resulted in the creation or deletion of uORFs; 11 of those are predicted to be disease-causing, and three have been confirmed to cause disease[107]. One such example involves the hair growth-associated (*HR*) transcript, in which mutations within the 5′ leader sequence are associated with the genetic hair disorder Marie Unna hereditary hypotrichosis[108]. The second uORF on the *HR* transcript (which is denoted as *U2HR*) encodes a 34-residue peptide that is essentially invariant in the 15 mammalian sequences analysed. One study[108] found 13 mutations within *U2HR* that result in hair defects and early-onset alopecia. Importantly, four of these are missense mutations that occur at codons 24–28, and all four changes were shown to increase expression of HR–luciferase and HR–GFP reporters. This finding suggests a role for the peptide product of this uORF, rather than (or in addition to) the uORF itself, in regulating the translation of the HR-coding sequence.

There are few reports on the identification and the characterization of short peptides encoded by downstream sORFs. A recent study found that downstream sORFs can be translated after cleavage of the transcript downstream of the CDS, which presumably exposes a ribosomal entry site on the 3′ trailer sequence[71]. Using both capped analysis of gene expression (CAGE) data and cDNA libraries, this study determined that nearly 3.6% of total cDNAs were completely mapped to 3′ trailer sequences, and the 5′ ends of a third of these cDNAs were directly supported by a CAGE cluster[71]. Another study used the CRITICA algorithm[39] and found that 2.9% of 3′ trailer sequence-derived RNAs were likely to encode peptides. The translation of peptides from downstream sORFs is also supported by mass spectrometry analyses of human cell isolates. One study identified three peptides that originated from two downstream sORFs[76], and a second study identified a further six peptides that originated from the 3′ trailer sequence[18]. In contrast to the paucity of peptides found in these studies, a third study identified 45 peptides encoded by downstream sORFs in the HeLa cell line and 14 in a selection of human colon cell lines[24], which indicates broad cell-type-specific differences in expression of these sPEPs. Nevertheless, the expression of sPEPs from 3′ trailer sequences is likely to be rare, and ribosome profiling studies offer little support for translation of peptides from downstream sORFs, as 3′ trailer sequences are found to be almost devoid of ribosomes[20,44,72].

Functionality of sPEPs

sPEPs from intergenic and presumed ncRNAs. Several sPEPs on intergenic regions and ncRNAs (TABLE 2) have been shown to be functional, particularly in plants and insects[79–86]. These sPEPs have diverse regulatory roles, although the mechanisms that underlie their roles are yet to be fully characterized. There are currently two examples of this class of sPEPs for which the mechanism of action is at least partially understood, and both of these are encoded by sORFs within lncRNAs of *D. melanogaster*[86–89]. The first example comes from the sPEPs that originate from four tandem sORFs of near-identical sequence on the *tarsal-less* (*tal*) lncRNA. These sORFs are independently translated to yield peptides of 11 and 32 amino acids in length and have been shown to have a crucial role in embryonic development. Inactivating mutations in *tal* produce an embryonic lethal phenotype that is similar to mutations in the *ovo* (also known as *svb*) gene, which encodes a transcription factor[88,89]. The Tal peptides result in the post-translational modification of Ovo, which promotes proteolytic cleavage and removes the amino-terminal repressor domain of this transcription factor, thereby converting Ovo from a repressor to an activator[89] (FIG. 2). Orthologues of *tal* occur in several other insect species[90], but none have been found in vertebrates, and the mammalian orthologues of Ovo (that is, Ovo-like 1 (OVOL1), OVOL2 and OVOL3) do not seem to be cleaved[91].

The second example comes from two functional sPEPs of 28 and 29 residues that are encoded by a putative ncRNA *pncr003:2L*[86]. Both sPEPs localize to the sarcoendoplasmic reticulum of muscle cells, where they regulate Ca^{2+} transport and thus muscle contraction. Importantly, these sPEPs (which are termed sarcolamban A and B) are conserved in both invertebrates and vertebrates, and they show functional homology to the mammalian peptides sarcolipin and phospholamban[86]. The discovery of translatable sORFs on some lncRNAs suggests that these should be reclassified as mRNAs; or if there is evidence that they also carry out non-coding functions, then they could be classified as bifunctional RNAs. An example of this is the *tal* transcript, which has recently been reclassified as a polycistronic mRNA (see NM_001144577 in the NCBI database).

Short peptides encoded by uORFs. Even though several uORFs have been shown to be translated, only a few have been comprehensively analysed and found to express a biologically active peptide (TABLE 2). Most uORFs act in *cis* and regulate translation of a downstream ORF through conventional mechanisms (BOX 1; FIG. 1), but an interesting subset of these attenuate translation of the downstream ORF in response to environmental signals. The term 'peptoswitch' has recently been coined[92] to describe these regulatory sPEPs, which are activated by binding to small molecules either directly or through an intermediary (FIG. 3). Several peptoswitches have been identified to date in yeast, plants and mammals. The peptoswitch on the yeast *CPA1* transcript downregulates translation of the CDS in response to high arginine levels[93]. In this case, arginine interacts with the 25-residue sPEP that is still attached to its ribosome, which prevents its translocation and initiates nonsense-mediated decay of the transcript[94]. The plant peptoswitches — which are encoded by uORFs on the transcripts of *GBF6* (also known as *BZIP11*, which encodes G-box-binding factor 6), *SAMDC* (which encodes *S*-adenosylmethionine decarboxylase) and *XPL1* (which encodes phosphoethanolamine *N*-methyltransferase 1) — downregulate translation of their CDS in response to increased levels of sucrose[95], polyamine[96] and phosphocholine[97], respectively. These small molecules interact with the nascent

Table 2 | **Functional sPEPs**

Species	Genes or transcripts	Number of residues in sPEP	Notes	Refs
Upstream sPEPs				
Arabidopsis thaliana	*GBF6*	28	Expression of the CDS is modulated by sucrose levels through a conserved sPEP	95
	SAMDC	52	Expression of the CDS is regulated by polyamines binding to the nascent upstream sPEP; orthologous to human *SAMDC*	96
	XPL1	26	Expression of the CDS is regulated by phosphocholine binding to the sPEP	97
Saccharomyces cerevisiae	*CPA1*	25	The sPEP reduces expression of the CDS through ribosomal stalling and blocking translation in response to increased arginine levels	93
Humans	*ASS1*	44	The sPEP regulates expression of *ASS1* in a *trans*-suppressive manner	99
	EPHX1	17 and 26	Expression of *EPHX1* is inhibited by *trans*-acting sPEPs that are encoded by two uORFs through interactions with the translation machinery	100
	HR	34	The sPEP is implicated in the regulation of *HR*; 13 causative mutations of Marie Unna hereditary hypotrichosis have been identified within the second uORF	108
	MKKS	63 and 50	Both sPEPs localize to the mitochondrial membrane and are predicted to function independently of MKKS	101
	NR3C1	93	The sPEP localizes to the cell membrane and regulates expression of the glucocorticoid receptor in a *trans*-acting manner through interaction with unknown cellular factors	98
	SAMDC	6	Expression of the CDS is regulated by polyamines binding to the nascent upstream sPEP; orthologous to *A. thaliana SAMDC*	3
Intergenic sPEPs				
A. thaliana	*PLS*	36	The sPEP is required for correct auxin–cytokinin homeostasis to modulate root growth and leaf vascular patterning	81
	ROT4	53	The sPEP is involved in regulation of leaf shape by reducing cell proliferation in lateral organs	82
Drosophila melanogaster	*Ilp8*	150	The sPEP provides a signal that promotes the delay of metamorphosis in response to conditions that that alter growth in imaginal discs	84,85
	HSPC300	75	The sPEP is a component of the WAVE–SCAR complex and is important in nervous system development for axonogenesis and neuromuscular synapse morphogenesis; *HSPC300* is orthologous to *brk1*	83
	pgc	71	The sPEP is essential for repressing Ser2 phosphorylation in the carboxy-terminal domain of RNA polymerase II in newly formed pole cells (which are the early germline progenitors) and thus has a fundamental role in germ-cell specification	120
	tal	11 and 32	The sORFs encode three peptides of 11 residues and one peptide of 32 residues that are essential for embryonic development and that are required for formation of epithelial architecture; *tal* is orthologous to *Mlpt*	87–89
	SclA and *SclB*	28 and 29	Both sPEPs are involved in the regulation of Ca^{2+} trafficking; alterations result in irregular muscle contractions	86
Maize	*brk1*	84	The sPEP promotes multiple actin-dependent cell polarization events in the developing leaf epidermis; *brk1* is orthologous to *HSPC300*	79
Soybean	*ENOD40-1*	12 and 24	The sPEP binds to nodulin 100 (which is a subunit of sucrose synthase) and is likely to be involved in the control of sucrose use in nitrogen-fixing nodules	80
Tribolium castaneum	*Mlpt*	10, 11, 15 and 23	The sORFs encode four sPEPs with roles in embryonic development, particularly the development of abdominal segments; *Mlpt* is orthologous to *tal*	90
Overlapping sPEPs				
Humans	*TYRP1*	24	The sPEP is co-expressed from the *TYRP1* transcript	77
	CASP1	151	The sPEP is expressed from the intestinal carboxyl esterase gene and is recognized by human leukocyte antigen-B7-restricted renal cell carcinoma-reactive T cell clone	78
	AltPrP	73	The sPEP is co-expressed from the prion protein transcript in brain homogenates, primary neurons and peripheral blood mononuclear cells; it localizes to the mitochondria	102
	AltATXN1	185	The sPEP is co-expressed from the *ATXN1* transcript and is expressed in the cerebellum; it colocalizes and interacts with the ATXN1 protein in the nucleus	103
	AltMRVI1	134	The sPEP colocalizes to the nucleus and interacts with BRCA1	24

AltATXN1, alternative transcript of ataxin-1; AltMRVI1, alternative transcript of murine retrovirus integration site 1 homologue gene; *AltPrP*, alternative transcript of the prion protein gene; *ASS1*, argininosuccinate synthase 1; BRCA1, breast cancer type 1 susceptibility protein; *brk1*, *brick1*; *CASP1*, caspase 1, apoptosis-related cysteine peptidase; CDS, coding DNA sequence; *ENOD40-1*, early nodulin; *EPHX1*, epoxide hydrolase 1; *GBF6*, G-box-binding factor 6; *HR*, hair growth-associated; *Ilp8*, *Insulin-like peptide 8*; *MKKS*, McKusick–Kaufman syndrome; *Mlpt*, mille-pattes; *NR3C1*, nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor); *pgc*, *polar granule component*; *PLS*, *POLARIS*; *ROT4*, *ROTUNDIFOLIA4*; *SAMDC*, S-adenosylmethionine decarboxylase; *SclA*, *Sarcolamban A*; sORF, short open reading frame; sPEP, short peptide encoded by sORF; *tal*, *tarsal-less*; uORF, upstream open reading frame; *TYRP1*, tyrosinase-related protein 1; *XPL1*, phosphoethanolamine *N*-methyltransferase 1.

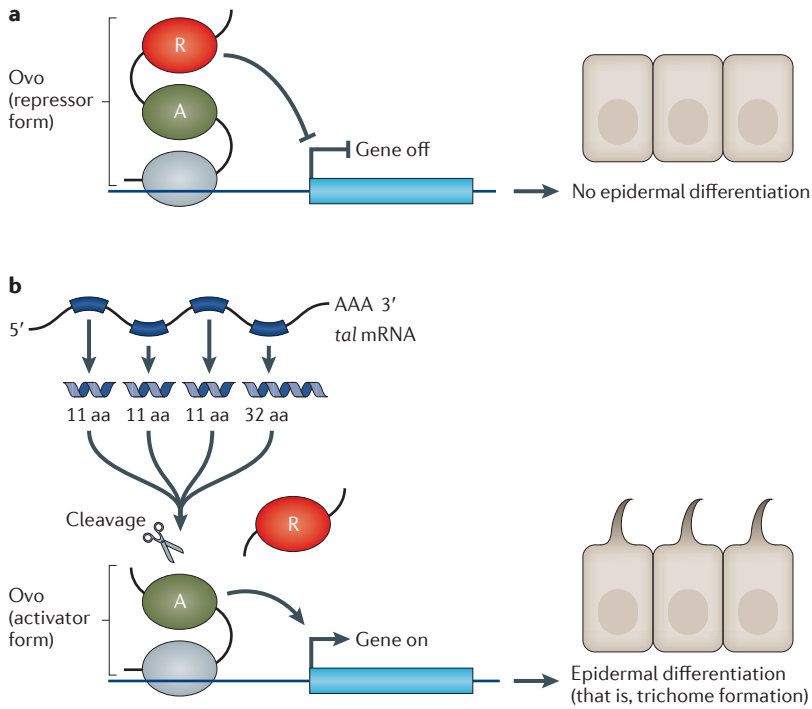

Figure 2 | **The Tal peptides and regulation of Ovo in *Drosophila melanogaster*.** **a** | Full-length Ovo, which is a transcription factor, binds to the promoter of a target gene and represses its transcription. The lack of target-gene expression prevents differentiation of larval epidermal cells, and no trichomes are formed. **b** | The Tal peptides are translated from four tandem short open reading frames on the *tarsal-less* (*tal*) polycistronic transcript that is expressed in epidermal cells. These short peptides promote cleavage of Ovo, which results in the loss of the amino-terminal repressor (R) domain. The truncated Ovo activates target-gene expression and induces trichome formation of larval epidermal cells. A, activating domain; aa, amino acid.

sPEPs to stall ribosomes at the uORF, which prevents their progression and subsequent translation of the downstream ORF. Translation of the CDS of human *SAMDC* is also downregulated by polyamines by a 6-residue sPEP[3], although the uORF has no common sequences with the plant orthologue.

Not all short peptides encoded by uORFs act in *cis*, and some seem to have *trans*-acting capabilities, showing evidence of interaction with cellular components other than the products of their CDSs. Examples of these *trans*-acting sPEPs include those encoded by the mammalian transcripts of the glucocorticoid receptor (*NR3C1*)[98], argininosuccinate synthase 1 (*ASS1*)[99], epoxide hydrolase 1 (*EPHX1*)[100] and the McKusick–Kaufman syndrome (*MKKS*) genes[101]. In the case of the *NR3C1* transcript, removal of the second of five uORFs through mutation of its initiation codon completely inhibited synthesis of the glucocorticoid receptor, whereas removal of any of the other four uORFs had no marked effect on translation of the CDS. Furthermore, reciprocal co-immunoadsorption experiments showed that the peptides encoded by the second uORF do not directly interact with the glucocorticoid receptor but instead bind to other cellular factors, which suggests a *trans*-acting role that has yet to be determined[98].

In the second example of a functional coding uORF, the uORF on the *ASS1* transcript was found to reduce expression of argininosuccinate synthase 1 in both *cis* and *trans* configurations. Overexpression of the ASS1 sPEP in transfected cells resulted in reduced levels of endogenous ASS1. This *trans*-suppression was dependent on both the sequence and the length of the sPEP, although it is not known how this reduction in the translation of the CDS is mediated[99]. Similarly, the two uORFs on an alternative spliced variant of *EPHX1* inhibit translation of the downstream ORF of *EPHX1* by *cis*-suppression of ribosome re-initiation and by the peptides that are encoded by both uORFs. These sPEPs inhibit translation of *EPHX1* in a sequence-specific and dose-dependent *trans*-acting manner[100]. In the final example, all three uORFs on the *MKKS* transcript were found to repress translation of the CDS, most probably by interfering with ribosomal scanning[101]. Intriguingly, this study also found that two of these uORFs encode highly conserved peptides that localize to the mitochondrial membrane (whereas MKKS localizes to the cytoplasm). Although their role in this organelle has yet to be determined, the authors speculate that the cellular functions of the MKKS sPEPs are distinct from that of MKKS. These examples offer a glimpse into the diversity of sPEP activity, but more work is clearly needed to fully understand the mechanisms behind this activity.

Short peptides encoded by overlapping and downstream sORFs. Few short peptides encoded by overlapping sORFs have been identified so far (TABLE 2). Two recently characterized examples are alternative prion protein (AltPrP; also known as APRIO)[102] and alternative ataxin-1 (AltATXN1)[103].

AltPrP is co-expressed from the prion protein transcript (*PRNP*), and AltATXN1 is expressed from the *ATXN1* transcript. Human AltPrP is a 73-residue peptide encoded by an overlapping sORF within the +3 reading frame of *PRNP* and shows extensive conservation and homology in a wide range of mammals. Co-expression of both polypeptides was detected *in vivo* by affinity purification of AltPrP and the prion protein from cell and brain homogenates, and *in vitro* using a translation assay with a haemagglutinin-tagged cDNA construct. Furthermore, AltPrP was observed to localize to the mitochondria, in contrast to the typical localization of the prion protein to the plasma membrane and the Golgi apparatus, and was unregulated in response to endoplasmic reticulum stress and proteasome inhibition.

The second example of a co-expressed sPEP is AltATXN1, which is a 185-amino-acid sPEP encoded by an overlapping sORF within the +3 reading frame of *ATXN1*. Both ATXN1 (which is a 98 kDa nuclear protein of unknown function) and AltATXN1 are co-expressed in the cerebellum and localize to intranuclear inclusions in co-transfected cells. The import of AltATXN1 into the nucleus is dependent on the transcription of *ATXN1*, and the sPEP has been shown to bind to both poly(A)$^+$ RNA and the ATXN1 protein. The functional roles of AltPrP and AltATXN1 remain to be determined.

Transcription activator-like effector nucleases (TALENs). Engineered enzymes that permit precise editing of genomes and that can be used to make specific sequence changes in model organisms such as *Arabidopsis thaliana*, zebrafish and mice.

Microproteins
Negative regulators of multiprotein complexes. In this case, micro refers to the mechanism of action of these proteins rather than to their sizes.

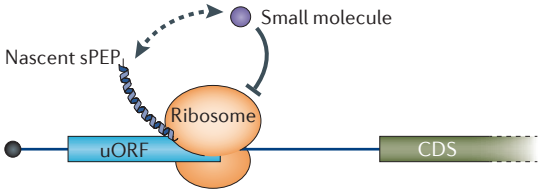

Figure 3 | **Example of a 'peptoswitch'.** Peptoswitches are *cis*-acting regulatory short peptides encoded by short open reading frames (sPEPs) that respond to specific small molecules in the cellular environment. These small molecules interact, either directly or through an intermediary, with the nascent sPEPs (indicated by the dashed arrow) to stall ribosomes at the upstream open reading frame (uORF), which prevents their progression and inhibits translation of the downstream coding DNA sequence (CDS).

One sPEP with regulatory potential has recently been identified by mass spectrometry and is found to be encoded by a sORF within the 3′ trailer sequence of the murine retrovirus integration site 1 homologue (*MRVI1*) gene[24]. The AltMRVI1 sPEP was discovered by re-evaluating data from an earlier yeast two-hybrid screen of proteins that interact with the BRCT domain of the breast cancer type 1 susceptibility protein (BRCA1). The AltMRVI1 sPEP was rejected as a positive hit in the prior study because its ORF was out of frame with the CDS of the *MRVI1* transcript and not recognized as a legitimate protein[24]. AltMRVI1 colocalizes with BRCA1 in the nucleus, and its interaction with BRCA1 was confirmed through co-immunoprecipiation[24], although the function of this interaction and the role of AltMRVI1 are still unknown.

Perspectives

The recent convergence of complementary technologies such as bioinformatics, proteomics and transcriptomics has already resulted in the identification of several hundred putative coding sORFs (TABLE 1) and various sPEPs (TABLE 2). The task now is to ascertain whether these and other sPEPs are functional or whether they are simply unavoidable by-products of sORF *cis*-acting activities. A possible strategy for determining their functions is to overexpress candidate sPEPs in transfected cell lines or whole organisms and monitor changes in phenotypes as exemplified by a recent study in plants[74]. The reverse experiment of removal or inactivation of endogenous sORFs is more challenging because of the difficulty in determining whether any phenotypic changes are due to loss of the sORF or disruption of the transcript in which it lies. New technologies such as transcription activator-like effector nucleases (TALENs)[104] and the clustered regularly interspaced short palindromic repeats (CRISPR)–Cas system[105] are useful in this case, as it may be possible to replace a particular sORF with an unrelated sequence or with a sequence that has specific internal changes in order to examine the consequences of a deficit in the encoded peptide.

It is easy to envisage scenarios in which sPEPs bind to various proteins — such as enzymes, ion channels, growth factors and transcription factors — and regulate their activity by inducing conformational changes, masking functional and regulatory sites, masking nucleic acid-binding sites and/or cofactor-binding sites, or serving as adaptors that potentiate specific interactions. Similarly, they could mimic the binding domains of interacting partners that are involved in multisubunit complexes and regulate their activity through a dominant-negative mechanism, as has been suggested for microproteins[106]. As exemplified by the Tal class of regulators, peptides as small as 11 residues can exert profound biological effects despite their small size, which supports the functionality of sPEPs.

The potential for sPEPs to act as regulators also raises the possibility that sequence variations as a result of mutations or polymorphisms within their respective sORFs could cause disease or contribute to increased risk (BOX 2). A survey of human 5′ leader sequences identified 509 unique genes for which polymorphisms and/or mutations resulted in the creation or deletion of uORFs; 14 of these genes were predicted to be disease-causing, three of which have been confirmed to cause disease[107]. Mutations in uORF-encoded peptides[108] underscore the need for further research on sORFs and their encoded peptides. In addition, it would be prudent for genome-wide association studies to take into account the possible effects of any sequence variations that occur in sORFs and to determine whether mutations within these sequences could be responsible for disease phenotypes or for increased risks.

Although there is now robust evidence for the translation of at least some sORFs, several questions remain to be addressed. For example, how many sORFs are actually translated? How long-lived are their peptide products? Are these peptides incidental by-products of random translation initiation events that merely contribute to cellular 'noise' or are they functional in their own right? If they are functional, then how many of these sPEPs can be shown to be biologically active and what are their functions? An understanding of sPEPs and their modes of action could lead to the development of therapeutic interventions that either mimic their functions or inhibit their activities in a manner that is similar to those already developed for their cytotoxic affects against cancer cells[109]. In conclusion, an appreciation of the role of biologically active peptides that originate from sORFs will offer new avenues of research into eukaryotic regulatory mechanisms.

1. Samayoa, J., Yildiz, F. H. & Karplus, K. Identification of prokaryotic small proteins using a comparative genomic approach. *Bioinformatics* **27**, 1765–1771 (2011).
2. Hobbs, E. C., Fontaine, F., Yin, X. & Storz, G. An expanding universe of small proteins. *Curr. Opin. Microbiol.* **14**, 167–173 (2011).
3. Law, G. L., Raney, A., Heusner, C. & Morris, D. R. Polyamine regulation of ribosome pausing at the upstream open reading frame of *S*-adenosylmethionine decarboxylase. *J. Biol. Chem.* **276**, 38036–38043 (2001).
4. Lease, K. A. & Walker, J. C. The *Arabidopsis* unannotated secreted peptide database, a resource for plant peptidomics. *Plant Physiol.* **142**, 831–838 (2006).
5. Hayden, C. & Bosco, G. Comparative genomic analysis of novel conserved peptide upstream open reading frames in *Drosophila melanogaster* and other dipteran species. *BMC Genomics* **9**, 61 (2008).

6. Yang, X. *et al.* Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Res.* **21**, 634–641 (2011).
7. Sleator, R. D. An overview of the current status of eukaryote gene prediction strategies. *Gene* **461**, 1–4 (2010).
8. Brent, M. R. & Guigó, R. Recent advances in gene structure prediction. *Curr. Opin. Struct. Biol.* **14**, 264–272 (2004).
9. Wang, J. *et al.* Vertebrate gene predictions and the problem of large genes. *Nature Rev. Genet.* **4**, 741–749 (2003).
10. Hanada, K., Zhang, X., Borevitz, J. O., Li, W.-H. & Shiu, S.-H. A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. *Genome Res.* **17**, 632–640 (2007).
11. Cheng, H. *et al.* Small open reading frames: current prediction techniques and future prospect. *Curr. Protein Pept. Sci.* **12**, 503–507 (2011).
12. Basrai, M. A., Hieter, P. & Boeke, J. D. Small open reading frames: beautiful needles in the haystack. *Genome Res.* **7**, 768–771 (1997).
13. Claverie, J.-M. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6**, 1735–1744 (1997).
14. Frith, M. C. *et al.* The abundance of short proteins in the mammalian proteome. *PLoS Genet.* **2**, e52 (2006).
This is the first study to examine the size and nature of the mammalian peptidome.
15. Hashimoto, Y., Kondo, T. & Kageyama, Y. Lilliputians get into the limelight: novel class of small peptide genes in morphogenesis. *Dev. Growth Differ.* **50**, S269–S276 (2008).
16. Kastenmayer, J. P. *et al.* Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res.* **16**, 365–373 (2006).
17. Fälth, M. *et al.* SwePep, a database designed for endogenous peptides and mass spectrometry. *Mol. Cell. Proteom.* **5**, 998–1005 (2006).
18. Slavoff, S. A. *et al.* Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature Chem. Biol.* **9**, 59–64 (2013).
This work builds on previous studies to identify 90 human small proteins using mass spectrometry.
19. Fritsch, C. *et al.* Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.* **22**, 2208–2218 (2012).
20. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011).
21. Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl Acad. Sci.* **109**, E2424–E2432 (2012).
22. Takahashi, H., Takahashi, A., Naito, S. & Onouchi, H. BAIUCAS: a novel BLAST-based algorithm for the identification of upstream open reading frames with conserved amino acid sequences and its application to the *Arabidopsis thaliana* genome. *Bioinformatics* **28**, 2231–2241 (2012).
23. Castellana, N. E. *et al.* Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc. Natl Acad. Sci.* **105**, 21034–21038 (2008).
24. Vanderperre, B. *et al.* Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS ONE* **8**, e70698 (2013).
This proteomic-based study has identified numerous short proteins in several human cell lines and tissues.
25. Menschaert, G. *et al.* 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. Proteom.* **12**, 1780–1790 (2013).
This study shows how ribosome profiling can aid short peptide discovery by mass spectrometry.
26. Hanada, K. *et al.* sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* **26**, 399–400 (2010).
27. Vanderperre, B., Lucier, J. F. & Roucou, X. HAltORF: a database of predicted out-of-frame alternative open reading frames in human. *Database (Oxford)* **2012**, bas025 (2012).
28. Skarshewski, A. *et al.* uPEPperoni: an online tool for upstream open reading frame location and analysis of transcript conservation. *BMC Bioinformatics* http://dx.doi.org/10.1186/1471-2105-15-36 (2014).
29. Hurst, L. D. The K_a/K_s ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18**, 486–487 (2002).
30. Zhang, Z. & Dietrich, F. Identification and characterization of upstream open reading frames (uORF) in the 5′ untranslated regions (UTR) of genes in *Saccharomyces cerevisiae*. *Curr. Genet.* **48**, 77–87 (2005).
31. Ladoukakis, E., Pereira, V., Magny, E., Eyre-Walker, A. & Couso, J. P. Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol.* **12**, R118 (2011).
32. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl Acad. Sci.* **104**, 19428–19433 (2007).
33. Kozak, M. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15**, 8125–8148 (1987).
34. Karlin, S., Campbell, A. M. & Mrázek, J. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* **32**, 185–225 (1998).
35. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).
36. Hayden, C. & Jorgensen, R. Identification of novel conserved peptide uORF homology groups in *Arabidopsis* and rice reveals ancient eukaryotic origin of select groups and preferential association with transcription factor-encoding genes. *BMC Biol.* **5**, 32 (2007).
37. Guillén, G. *et al.* Detailed analysis of putative genes encoding small proteins in legume genomes. *Front. Plant Sci.* **4**, 208 (2013).
38. Castrignanò, T. *et al.* CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison. *Nucleic Acids Res.* **32**, W624–W627 (2004).
39. Badger, J. H. & Olsen, G. J. CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **16**, 512–524 (1999).
40. Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **35**, W345–W349 (2007).
41. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
42. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
43. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
44. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**, 240–251 (2013).
45. Krug, K., Nahnsen, S. & Macek, B. Mass spectrometry at the interface of proteomics and genomics. *Mol. BioSystems* **7**, 284–291 (2011).
46. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
47. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
48. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
49. Clark, M. B. *et al.* The reality of pervasive transcription. *PLoS Biol.* **9**, e1000625 (2011).
50. van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* **8**, e1000371 (2010).
51. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Struct. Mol. Biol.* **14**, 103–105 (2007).
52. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
53. Kageyama, Y., Kondo, T. & Hashimoto, Y. Coding versus non-coding: translatability of short ORFs found in putative non-coding transcripts. *Biochimie* **93**, 1981–1986 (2011).
54. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
55. Iacono, M., Mignone, F. & Pesole, G. uAUG and uORFs in human and rodent 5′untranslated mRNAs. *Gene* **349**, 97–105 (2005).
56. Crowe, M., Wang, X.-Q. & Rothnagel, J. A. Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. *BMC Genomics* **7**, 16 (2006).
57. Neafsey, D. E. & Galagan, J. E. Dual modes of natural selection on upstream open reading frames. *Mol. Biol. Evol.* **24**, 1744–1751 (2007).
58. Cvijovic, M., Dalevi, D., Bilsland, E., Kemp, G. & Sunnerhagen, P. Identification of putative regulatory upstream ORFs in the yeast genome using heuristics and evolutionary conservation. *BMC Bioinformatics* **8**, 295 (2007).
59. Tran, M., Schultz, C. & Baumann, U. Conserved upstream open reading frames in higher plants. *BMC Genomics* **9**, 361 (2008).
60. Vaughn, J. N., Ellingson, S. R., Mignone, F. & von Arnim, A. Known and novel post-transcriptional regulatory sequences are conserved across plant families. *RNA* **18**, 368–384 (2012).
61. Wethmar, K., Smink, J. J. & Leutz, A. Upstream open reading frames: molecular switches in (patho) physiology. *BioEssays* **32**, 885–893 (2010).
62. Pesole, G. *et al.* Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. *Gene* **261**, 85–91 (2000).
63. Suzuki, Y. *et al.* Statistical analysis of the 5′ untranslated region of human mRNA using "oligo-capped" cDNA libraries. *Genomics* **64**, 286–297 (2000).
64. Rogozin, I. B., Kochetov, A. V., Kondrashov, F. A., Koonin, E. V. & Milanesi, L. Presence of ATG triplets in 5′ untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. *Bioinformatics* **17**, 890–900 (2001).
65. Yamashita, R., Suzuki, Y., Nakai, K. & Sugano, S. Small open reading frames in 5′ untranslated regions of mRNAs. *C. R. Biol.* **326**, 987–991 (2003).
66. Chen, C. H., Liao, B. Y. & Chen, F. C. Exploring the selective constraint on the sizes of insertions and deletions in 5′ untranslated regions in mammals. *BMC Evol. Biol.* **11**, 192 (2011).
67. Ribrioux, S., Brungger, A., Baumgarten, B., Seuwen, K. & John, M. R. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. *BMC Genomics* **9**, 122 (2008).
68. Michel, A. M. *et al.* Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.* **22**, 2219–2229 (2012).
69. Chung, W.-Y., Wadhawan, S., Szklarczyk, R., Pond, S. K. & Nekrutenko, A. A. First look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput. Biol.* **3**, e91 (2007).
70. Xu, H. *et al.* Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. *Cell Res.* **20**, 445–457 (2010).
71. Mercer, T. R. *et al.* Expression of distinct RNAs from 3′ untranslated regions. *Nucleic Acids Res.* **39**, 2393–2403 (2011).
72. Chew, G.-L. *et al.* Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. *Development* **140**, 2828–2834 (2013).
73. Bánfai, B. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22**, 1646–1657 (2012).
74. Hanada, K. *et al.* Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc. Natl Acad. Sci.* **110**, 2395–2400 (2013).
This is the first systematic characterization of short open reading frames using transgenic plants.
75. Oyama, M. *et al.* Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Res.* **14**, 2048–2052 (2004).
This is the first study to identify small proteins in human cells using mass spectrometry.
76. Oyama, M. *et al.* Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol. Cell. Proteom.* **6**, 1000–1006 (2007).

77. Wang, R. F., Parkhurst, M. R., Kawakami, Y., Robbins, P. F. & Rosenberg, S. A. Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. *J. Exp. Med.* **183**, 1131–1140 (1996).
78. Ronsin, C. *et al.* A non-AUG-defined alternative open reading frame of the intestinal carboxyl esterase mRNA generates an epitope recognized by renal cell carcinoma-reactive tumor-infiltrating lymphocytes *in situ*. *J. Immunol.* **163**, 483–490 (1999).
79. Frank, M. J. & Smith, L. G. A small, novel protein highly conserved in plants and animals promotes the polarized growth and division of maize leaf epidermal cells. *Curr. Biol.* **12**, 849–853 (2002).
80. Rohrig, H., Schmidt, J., Miklashevichs, E., Schell, J. & John, M. Soybean *ENDO40* encodes two peptides that bind sucrose synthase. *Proc. Natl Acad. Sci.* **99**, 5 (2002).
81. Stuart, A. *et al.* The POLARIS gene of *Arabidopsis* encodes a predicted peptide required for correct root growth and leaf vascular patterning. *Plant Cell* **14**, 16 (2002).
82. Narita, N. N. *et al.* Overexpression of a novel small peptide ROTUNDIFOLIA4 decreases cell proliferation and alters leaf shape in *Arabidopsis thaliana*. *Plant J.* **38**, 699–713 (2004).
83. Abrar, Q. *et al.* HSPC300 and its role in neuronal connectivity. *Neural Dev.* **2**, 18 (2007).
84. Colombani, J., Andersen, D. S. & Léopold, P. Secreted peptide Dilp8 coordinates *Drosophila* tissue growth with developmental timing. *Science* **336**, 582–585 (2012).
85. Garelli, A., Gontijo, A. M., Miguela, V., Caparros, E. & Dominguez, M. Imaginal discs secrete insulin-like peptide 8 to mediate plasticity of growth and maturation. *Science* **336**, 579–582 (2012).
86. Magny, E. G. *et al.* Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* **341**, 1116–1120 (2013).
87. Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A. & Couso, J. P. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.* **5**, e106 (2007).
88. Kondo, T. *et al.* Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biol.* **9**, 660–665 (2007).
89. Kondo, T. *et al.* Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* **329**, 336–339 (2010).
This study identifies the molecular target of the small regulatory peptides encoded by a polycistronic mRNA that was previously thought to be a non-coding transcript.
90. Savard, J., Marques-Souza, H., Aranda, M. & Tautz, D. A segmentation gene in tribolium produces a polycistronic mRNA that codes for multiple conserved peptides. *Cell* **126**, 559–569 (2006).
91. Li, B. *et al. Ovol2*, a mammalian homolog of *Drosophila Ovo*: gene structure, chromosomal mapping, and aberrant expression in blind-sterile mice. *Genomics* **80**, 319–325 (2002).
92. Jorgensen, R. A. & Dorantes-Acosta, A. E. Conserved-peptide upstream open reading frames (CPuORFs) are associated with regulatory genes in angiosperms. *Front. Plant Sci.* **3**, 191 (2012).
93. Werner, M., Feller, A., Messenguy, F. & Piérard, A. The leader peptide of yeast gene *CPA1* is essential for the translational repression of its expression. *Cell* **49**, 805–813 (1987).
94. Gaba, A., Jacobson, A. & Sachs, M. S. Ribosome occupancy of the yeast *CPA1* upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. *Mol. Cell* **20**, 449–460 (2005).
95. Rahmani, F. *et al.* Sucrose control of translation mediated by an upstream open reading frame-encoded peptide. *Plant Physiol.* **150**, 1356–1367 (2009).
96. Hanfrey, C. *et al.* A dual upstream open reading frame-based autoregulatory circuit controlling polyamine-responsive translation. *J. Biol. Chem.* **280**, 39229–39237 (2005).
97. Alatorre-Cobos, F. *et al.* Translational regulation of *Arabidopsis XIPOTL1* is modulated by phosphocholine levels via the phylogenetically conserved upstream open reading frame 30. *J. Exp. Bot.* **63**, 5203–5221 (2012).
98. Diba, F., Watson, C. S. & Gametchu, B. 5′UTR sequences of the glucocorticoid receptor 1A transcript encode a peptide associated with translational regulation of the glucocorticoid receptor. *J. Cell. Biochem.* **81**, 149–161 (2001).
99. Pendleton, L. C., Goodwin, B. L., Solomonson, L. P. & Eichler, D. C. Regulation of endothelial argininosuccinate synthase expression and NO production by an upstream open reading frame. *J. Biol. Chem.* **280**, 24252–24260 (2005).
100. Nguyen, H. L., Yang, X. & Omiecinski, C. J. Expression of a novel mRNA transcript for human microsomal epoxide hydrolase (*EPHX1*) is regulated by short open reading frames within its 5′-untranslated region. *RNA* **19**, 752–766 (2013).
101. Akimoto, C. *et al.* Translational repression of the McKusick–Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. *Biochim. Biophys. Acta.* **1830**, 2728–2738 (2013).
102. Vanderperre, B. *et al.* An overlapping reading frame in the *PRNP* gene encodes a novel polypeptide distinct from the prion protein. *FASEB J.* **25**, 2373–2386 (2011).
103. Bergeron, D. *et al.* An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *J. Biol. Chem.* **288**, 21824–21835 (2013).
104. Joung, J. K. & Sander, J. D. TALENs: a widely applicable technology for targeted genome editing. *Nature Rev. Mol. Cell Biol.* **14**, 49–55 (2013).
105. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nature Methods* **10**, 957–963 (2013).
106. Staudt, A. C. & Wenkel, S. Regulation of protein function by 'microProteins'. *EMBO Rep.* **12**, 35–42 (2011).
107. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl Acad. Sci.* **106**, 7507–7512 (2009).
108. Wen, Y. *et al.* Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nature Genet.* **41**, 228–233 (2009).
This study identified mutations in a highly conserved upstream open reading frame that are associated with genetic hair loss and suggests that an aberrant short peptide may result in disease.
109. Almansour, N. M., Pirogova, E., Coloe, P. J., Cosic, I. & Istivan, T. S. Investigation of cytotoxicity of negative control peptides versus bioactive peptides on skin cancer and normal cells: a comparative study. *Future Med. Chem.* **4**, 1553–1565 (2012).
110. Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**, 283–292 (1986).
111. Kozak, M. Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Mol. Cell. Biol.* **7**, 3438–3445 (1987).
112. Morris, D. R. & Geballe, A. P. Upstream open reading frames as regulators of mRNA translation. *Mol. Cell. Biol.* **20**, 8635–8642 (2000).
113. Ghilardi, N., Wiestner, A. & Skoda, R. C. Thrombopoietin production is inhibited by a translational mechanism. *Blood* **92**, 4023–4030 (1998).
114. Calkhoven, C. F., Müller, C. & Leutz, A. Translational control of C/EBPα and C/EBPβ isoform expression. *Genes Dev.* **14**, 1920–1932 (2000).
115. Hinnebusch, A. G. Translational regulation of yeast *GCN4*. *J. Biol. Chem.* **272**, 21661–21664 (1997).
116. Child, S. J., Miller, M. K. & Geballe, A. P. Translational control by an upstream open reading frame in the *HER-2/neu* transcript. *J. Biol. Chem.* **274**, 24335–24341 (1999).
117. Wang, X.-Q. & Rothnagel, J. A. Post-transcriptional regulation of the *GLI1* oncogene by the expression of alternative 5′ untranslated regions. *J. Biol. Chem.* **276**, 1311–1316 (2001).
118. Wang, X. Q. & Rothnagel, J. A. 5′-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. *Nucleic Acids Res.* **32**, 1382–1391 (2004).
119. Kozak, M. An analysis of vertebrate mRNA sequences: intimations of translational control. *J. Cell Biol.* **115**, 887–903 (1991).
120. Hanyu-Nakamura, K., Sonobe-Nojima, H., Tanigawa, A. & Lasko, P. *Drosophila* Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. *Nature* **451**, 730–733 (2008).

Acknowledgement
This work was supported by a grant to J.A.R. from the Australian National Health and Medical Research Council (ID631551).

Competing interests statement
The authors declare no competing interests.