New peptides under the s(ORF)ace of the genome

Article (Accepted Version)


This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/66151/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk
New peptides under the s(ORF)ace of the genome

J.I. Pueyo†, E.G. Magny†, and J.P. Couso*  

1: Brighton and Sussex Medical School, University of Sussex, United Kingdom  
†: These authors contributed equally to this work  
*: correspondence to j.p.couso@sussex.ac.uk (J.P. Couso)

Keywords: smORFs, peptides, Ribosome Profiling, peptidomics, LncRNAs, uORFs.

Abstract

Hundreds of previously unidentified functional small peptides could exist in most genomes, but these sequences have been generally overlooked. The discovery of genes encoding small peptides with important functions in different organisms, has ignited the interest in these sequences, and led to an increasing amount of effort towards their identification.

Here, we review the advances, both, computational, and biochemical, that are leading the way in the discovery of putatively functional smORFs, as well as the functional studies that have been carried out as a consequence of these searches. The evidence suggests that smORFs form a substantial part of our genomes, and that their encoded peptides could have important functions in a variety of cellular functions.
Introduction

Deciphering the genetic information encoded in a genome is one of the main challenges in Biology. A constant improvement of sequencing and bioinformatics techniques has greatly advanced our understanding of this information but has also revealed the extent of its complexity. The difficulties associated with accurately predicting and annotating Small Open Reading Frame genes (smORFs) perfectly illustrate this complexity and the challenges it poses.

In the genome of most organisms there are hundreds of thousands of putative smORFs, consisting of a start-codon followed by in-frame codons and ending with a stop-codon [1-2]. Distinguishing translated and functional smORFs among this overwhelming and mostly spurious pool of sequences represents a major issue, which is particularly difficult to resolve since standard computational algorithms to identify coding sequences are generally not suited for small sequences [3-5]. Initially, short coding sequences (<100aa) were excluded from genome annotation pipelines [6], with the assumption that the majority of coding genes would code for larger proteins [7]. However, genes encoding small peptides have been identified in several organisms [8], like the *tarsal-less/polished rice/millepattes* gene, which codes for 11 aa-long peptides with important developmental functions in arthropods [9-12]. Such examples have led to the realisation that previously uncharacterised protein-coding smORFs with promising biological functions could exist in most genomes, and an increasing amount of effort has been directed towards their identification.

Here we will focus on the advances, both computational and biochemical that have been used to identify smORFs, and will present some of the different examples of smORFs which have been functionally characterised as a consequence of these studies.
Altogether, there is evidence suggesting that smORFs form a substantial part of our genomes and that their encoded peptides could be involved in a variety of cellular functions. Their characterisation could therefore lead to discoveries with important implications in cell biology and human health.

**Systematic searches for putative coding smORFs using computational approaches.**

Initial genome-wide searches for putative functional smORFs were conducted by bioinformatics methods designed to overcome the limitations of standard gene annotation algorithms. Generally, these methods were based on the analysis of sequence-composition frequencies (Figure 1A; see sORFfinder and CRITICA in BOX1), and/or on the evaluation of:

a) the conservation of candidate smORF sequences in related species using pair-wise alignment-based tools (Figure 1B; BLAST [13]), and b) of their purifying selection (conservation of the aa relative to nt sequence) [14]. These initial studies identified several hundreds, and even thousands of putatively functional novel smORFs in the genomes of yeast, plants, flies, and mice [15-19], generally representing about 3-5% of the annotated genes in these organisms (Table 1).

In order to identify conserved coding sequences, more recent methods based on multiple sequence alignments incorporate phylogenetic distances and a model of nucleotide substitution rates, in the case of PhastCons [20], or a model of codon substitution frequencies, in the case of phyloCSF [21], both built upon known coding and no-coding sequences (see Box 1). As shown below, these methods have sometimes been used
together with experimental methods in order to validate, or strengthen, the functionality of the smORFs identified as translated.

Ribosome Profiling: a biochemical approach for genome-wide translation assessment of smORFs

Next generation RNA sequencing (RNA-seq) has allowed to identify entire transcriptomes [22] and led to the unexpected realisation that a much higher than anticipated portion of the genome is transcribed (up to 85% in mammals [23] and 75% in flies [24]). A large proportion of these transcripts lack a “long” ORF of more than 100 aa, and have therefore been considered as long non-coding RNAs (LncRNAs), even though they otherwise resemble canonical mRNAs, having similar lengths, being transcribed by RNA-polymerase II, capped, poly-adenylated, and most even accumulating in the cytoplasm [25]. Although several LncRNAs have a well-established non-coding function [26], for the vast majority this remains unknown, making it plausible that some LncRNAs actually encode smORFs.

A method known as ribosome profiling (or Ribo-seq; Figure 1C) [27], consisting in sequencing nuclease-protected mRNA fragments (or footprints) bound by translating ribosomes (stabilized with an elongation inhibitor like cyclohexamide (CHX), allows to quantitatively and qualitatively measure the translation of these transcriptomes [28].

Different ribosome profiling studies, in a wide variety of species [29-40] have found that translation occurs in an almost pervasive fashion, detecting ribosome footprints in LncRNAs, in the untranslated regions (UTRs) of annotated transcripts, either upstream (uORFs) or
downstream (dORFs) of the CDS, and even overlapping the CDS of canonical mRNAs, with the vast majority of these corresponding to short ORFs (Table 1, and Table S1).

However there is some ambiguity with this method, since a ribosome bound fragment (RBF) read does not always strictly equate to an actively translated RNA fragment; a fragment of similar size could be obtained by a scanning ribosome, or other RNA-binding proteins [28]. Ribo-seq studies therefore employ different experimental or computational strategies to identify more accurately actively translated regions, involving the use of different metrics, such as RBF coverage, translation efficiency (TE: the ratio of RBFs / total mRNA reads), ribosomal release score (RRS), or codon phasing (see BOX1, [2]). Translation inhibitors, such as harringtonine (HR), which generates a pile-up of RBFs at the start codon, have also been used to identify translation initiation sites in actively translated ORFs [30]. Some of these studies have focused exclusively on the identification of translated smORFs in fruit flies [32], zebrafish [31], yeast [36], and mice [41].

In *Drosophila*, Aspden et al.[32] incorporated polysomal fractionation before ribo-seq to isolate cytoplasmic RNAs bound by 2-6 ribosomes and therefore actively translated, rather than those being scanned by single non-productive ribosomes or other RNA-binding proteins, this also enriched for RNAs encoding short ORFs (6 ribosomes being the maximum number that could fit in a 300 nt ORF). Using stringent RFB density and coverage thresholds, they corroborated the translation of 83% of the annotated smORFs transcribed in *Drosophila S2* cells (228 out of 274), and found 2,708 and 313 novel translated smORFs in 5’UTRs and LncRNAs (Figure 2). Annotated smORFs, were found to be longer (~80 aa median) and with similar levels of “functionality” as canonical coding genes (conservation, aa usage and secondary structures) whereas the smORFs detected in 5’UTRs and LncRNAs,
were shorter (~20aa media length), and lacked the functional signatures observed in longer smORFs. However, some of these 5’UTR and LncRNA smORFs could be detected in epitope tagging experiments, displaying a similar sub-cellular localisation as canonical proteins, suggesting that some of them may encode functional peptides.

Bazzini et al. [31] performed Ribo-seq in zebrafish embryos, using ORFscore (See Box 1), a method which quantifies the 3-codon periodicity of the distribution of RBFs relative to the predicted ORF (phasing), a feature consistent with those ORFs being actively translated. Using this method, they validated the translation of 302 (52%) previously annotated smORFs, and identified 190 novel smORFs in previously uncharacterised transcripts and LncRNAs, as well as 311 uORFs and 93 ORFs in 3’UTRs (Figure 2; Table 1). In parallel, 63 novel smORFs were found using a conservation-based computational pipeline (micPDP) (see Box 1) in a catalogue of non-coding transcripts, 23 of them were also deemed translated by Ribo-seq, representing a pool of peptides highly likely to be translated and functional in zebrafish.

In yeast, 1,088 previously uncharacterised transcripts associated with poly-ribosomes (supporting their translation) were found [36]. Ribo-seq identified 185 of these as having sufficient footprint coverage, and TE scores to support smORF translation. Furthermore, 61 out of 80 transcripts from this pool, showed a codon triplet phasing bias to a single frame, suggesting their translation. Finally, 39 of these translated smORFs also showed varying extents of conservation among divergent yeast species, implying that they could be functional (Table 1).

Crappe et al.[41], combined a computational smORF search with Ribo-seq data to identify potentially coding smORFs in the mouse genome. A systematic search for potential coding
smORFs, conserved across mammalian species, was performed using sORFinder [16] and PhastCons (Box 1). Subsequently, a Support Vector Machine (SVM) learning algorithm, trained with sets of putatively non-coding and coding sequences, was used to classify the predicted smORFs, leading to the identification of some 40,000 smORFs with high coding probability in intergenic regions and LncRNAs. Independently, they re-analysed available ribosome profiling data from a mouse embryonic cell line [30], to identify translated smORFs (passing a coverage threshold, and showing a pile-up at their start codon when treated with HR). They identified 528 intergenic smORFs and 226 smORFs in LncRNAs, of these 401 and 89, respectively, were also found in the computational pipeline, representing a pool of smORFs likely to encode functional peptides (Table 1).

This study highlights the discrepancy in numbers that can exist between computational predictions and experimental detection. Part of this discrepancy could be explained by a possible high false positive rate in the bioinformatic pipeline, which could be due, for example, to the presence of conserved elements such as transposons, pseudogenes, and simple repeats [42]. It could also be explained by the fact that computational pipelines can search whole genomes for putative smORFs, whereas only the smORFs within transcripts expressed above a certain threshold in specific cells or tissues studied will be tested in Riboseq (or HPLC-MS) experiments.

Lee et al. [33] used a similar method, treating human and mouse cell lines with lactimidomycin (LTM), another initiation phase inhibitor. In this study they identified 227 annotated Human smORFs as translated (out of 694 annotated smORFs in ENSEMBL), as well as 288 ORFs in LncRNAs and 1,194 uORFs (most of them <100 aa long) (Figure 2; Table 1).
Altogether these studies show that thousands of smORFs are translated in eukaryotic genomes, with a substantial portion showing conservation and coding potential features, suggesting that a large repertoire of functional, yet uncharacterized peptides could exist in these organisms.

**Detection of smORF peptides by mass spectrometry**

The high-performance Liquid chromatography Mass-spectrometry (HPLC-MS) proteomics approach [43] has also been adapted to identify small peptides, mainly, by modifying the protocols for data analysis: instead of comparing candidate peptide spectrum matches (PSMs) to databases of annotated proteins, these are compared to databases generated *de novo*, based on all the possible translations of a given transcriptome (Figure 1D).

Furthermore, standard proteomics require protein sequences to be supported by multiple PSMs. Because smORFs are too short to fit more than one PSM, this single PSM is usually required to pass the most stringent criteria in order to be unambiguously assigned to that smORF, potentially leading to a higher rate of false negatives.

Slavoff *et al.*[44] developed a peptidomics strategy, also applying specific experimental optimizations: inhibiting proteolysis, arguing that the proteolytic fragments of canonical proteins greatly increases the complexity of the peptidome and deteriorates the signal to noise ratio when it comes to identifying short peptides (themselves more susceptible to protease degradation), and using electrostatic-repulsion hydrophilic interaction chromatography (ERLIC) prior to HPLC-MS. They identified 86 novel peptides in human cells: 33 of them mapping to alternative CDS’ in annotated transcripts (corresponding to uORFs,
dORFs, and smORFs overlapping annotated CDS’), 8 mapping to LncRNAs, and 49 of them mapping to previously un-annotated transcripts (Table 1).

This method was tested against other workflows [45], leading to two important observations: first, the use of ERLIC fractionation greatly increases the number of peptides detected (~10 fold) and second, there is an important lack of overlap between the peptides identified by different workflows, and even by different technical repeats, highlighting the stochastic nature of this technique, and the requirement of several repeats to achieve an optimal sampling saturation of the peptidome. In total, they analysed 3 different cell lines and a tumor sample, and identified a total of 311 short peptides, of which 237 are novel, with ~80% of them mapping to previously unannotated transcripts (Table 1), and the rest to alternative CDS within annotated transcripts with a similar distribution, in UTRs and overlapping CDS’, as found by Slavoff et al.[44].

Another study, identified 1,259 novel peptides [46], by matching the spectra of different HPLC-MS data-sets (16 in total, covering a range of different human samples) to a custom database of predicted ORFs within annotated transcripts (mapping to UTRs and overlapping CDS’), suggesting that the translation of these “alternative” smORFs could be a wide-spread phenomenon. Interestingly, the majority of these peptides were identified in plasma and serum samples (1,118 / 1,259), implying that they could be secreted, although the reason or mechanism leading to this remains unknown (Table S1). Again, given the stochastic nature of this technique, this seemingly high number of identified novel peptides could be explained, in part, to the large number of samples analysed in this study.

Some of the Ribo-seq-based studies covered above have used HPLC-MS in order to validate their results (Table 1). In general, previously annotated smORFs tend to be more abundantly
detected by HPLC-MS than uORFs or LncRNA smORFs; in Aspden et al.[32] and Bazzini et al.[31] detected almost a third of the 228 and 302 annotated translated smORFs, but Aspden et al.[32] failed to identify any peptide from LncRNAs or uORFs, and Bazzini et al. [31] only identified 3 and 17, respectively. Similarly, only a handful of peptides corresponding to uORFs and LncRNAs have been detected by HPLC-MS in studies that detected hundreds by Ribo-seq in humans (Figure 2; Table 1). These results clearly highlight a difference of sensitivity between these methods, and could also be in agreement with the segregation of these smORFs into two different functional classes, as observed by Aspden et al.[32], with smaller uORF and LncRNA ORFs showing lower conservation features and being less likely to be translated into functional peptides than longer annotated smORFs. These results could also be explained by the stochastic nature of the peptidomics method observed by Slavoff et al.[44], with the peptides from LncRNAs or uORFs being generally smaller, and therefore probably less stable, less abundant, and having lower chances of being detected. In that sense, detection by peptidomics could be considered as a convincing proof of translation, and as an indication of probable function, but the opposite is not necessarily true (Figure 2). Also, it is important to point out that these studies did not use the extensively optimized protocols (with proteolysis free conditions, ERLIC fractionation, and multiple technical repeats), which may have improved the detection of these smaller peptides.

**Computational and biochemical strategies unravel novel smORF peptide functions**

Although these computational and biochemical approaches have identified hundreds of translated and conserved smORFs, previous systematic functional studies (based on random
mutagenesis) in different organisms, have failed to find them. This disparity could be explained by the lower probability of mutagens to target a small ORF in comparison to larger canonical ones. In addition, and as exemplified by several of the examples covered below, these small peptides may act as regulators of cellular processes requiring a very specific and in-depth phenotypic analysis, in order to detect mutants. As a result, only a handful of smORFs, found serendipitously, had been functionally characterised prior to these extensive smORF searches [8].

However, these genome-wide smORF searches have aided the functional characterisation of smORFs by identification of putative candidates. Following their bioinformatic prediction, some studies have carried out high-throughput smORF functional screens in yeast, [15] and in plants [47], and found dozens of functional smORFs, with several being essential (Table 1).

Other studies have focused on a more in-depth characterisation of specific smORFs. One example is the Sarcolamban (Scl) gene [48], previously annotated as a non-coding gene [49], but identified by a bioinformatics approach as a potential functional smORF in Drosophila (Table 1;[17]). Scl encodes for two 28 and 29aa transmembrane related-micropeptides which act as inhibitors of SERCA calcium pump and regulate heart muscle contraction (Figure 2A;[48]). Importantly, these peptides appear to be functional homologues to the vertebrate Sarcolipin and Phospholamban peptides, thereby uncovering an ancestral family of smORFs conserved from insects to humans [48]. More recently, another member of this family, Myoregulin (46 aa) [50], and a novel small peptide, DWORF (34 aa) [51] with an antagonistic function (since it enhances the activity of SERCA), have both been identified in mice from previously non-coding annotated transcripts. Toghether with Sln and Pln, (and Scl
in flies) these peptides contribute to the smORF-based regulatory repertoire that regulates calcium dynamics and seemingly participate in conferring different muscles with specific contractility properties [50].

Another example is the toddler/apela gene, which was identified in a ribo-seq-based search for novel signalling peptides in zebrafish (Table 1;[52]). The apela gene encodes a secreted 58 aa peptide, that binds to the Apelin receptor and promotes cell mobility during gastrulation [52]. This novel peptide also shows a great extent of conservation across vertebrates.

Similarly, the Drosophila hemotin (hemo) gene was identified as a putative functional smORF by a computational study [17], and subsequently, its translation supported by ribosome profiling and proteomics studies [32, 53](Table 1). hemo is expressed in hemocytes (Drosophila macrophages) where it regulates endosomal maturation, and phagocytosis, by inhibiting the activity of phosphotidylinositol kinases through an interaction with 14-3-3z (Figure 2B;[54]). Interestingly, this regulatory mechanism also appears to have been conserved across evolution as the vertebrate Stannin (Snn), a factor involved in organometallic cytotoxicity [55], is the functional Hemo homologue in flies and mouse macrophages [54].

In humans, the MRI-2 peptide, which was shown to stimulate double-strand break repair through a direct interaction with the DNA end binding protein Ku (Figure 2C;[56]), was functionally characterised because it appeared as translated in a HPLC-MS screen for novel short peptides (Table 1;[44]).
These examples highlight the contribution of these bioinformatic and experimental approaches in the identification of functional smORFs. They also strengthen our view about the complexity and biological relevance of these peptides, which can regulate a diversity of cellular processes, with their function being conserved, in some cases, across vast evolutionary distances. Overall, their study can certainly have important implications in cell biology, and even in human medical research[57].

**Concluding remarks and future perspectives**

In this review, we have shown that there is extensive evidence supporting the translation of substantial numbers of smORFs in a variety of organisms. This evidence is likely to increase as new methods and metrics are developed to analyse ribo-seq data more robustly in order to identify *bona fide* translated regions (BOX 2). For example, Ingolia et al.[58] recently developed a metric, based on assessing the distribution of RPF lengths (Fragment Length Organisation Similarity Score or FLOSS) which can accurately distinguish between reads protected by the translation-engaged 80s ribosomal conformation, from reads obtained from the protection of other non-translating ribosomal conformations (40s and 60s), or other RNA-binding proteins. Other groups have used classification algorithms, such as the random forest-based Translated ORF classifier (TOC) [29, 52], the logistic regression-based ORF-rater [59], or the SVM-based RibORF [60], which integrate different Ribo-seq metrics, and their profiles on known coding and non-coding regions, to identify translated ORFs. All these studies support the translation of hundreds of novel small peptides encoded in transcripts previously thought to be non-coding and as uORFs, in vertebrates.
It remains challenging, however, to identify which among this ever-growing set of Ribo-seq-supported translated smORFs encode functional peptides, from those representing “translational noise”, or acting as translation-dependent regulatory sequences. Abundant evidence supports the role of uORFs as translational regulators, exerting this function through their engagement of ribosomes [61-63], inferring this is the main, canonical function of uORFs. Similarly, it has been suggested that smORFs within LncRNAs, or overlapping annotated coding mRNAs, could function mainly as regulators of transcript stability, by engaging the non-sense mediated decay (NMD) pathway [64-65]. Nonetheless, as shown above, several smORF-encoded peptides with important functions have been identified in previously non-coding RNAs proving that these sequences can certainly encode functional peptides [10, 48, 50-52, 54]. There are even examples of canonical non-coding RNAs, such as pri-miRNAs [66] and ribosomal RNAs [67-69], encoding biologically active peptides with well characterised functions. Similarly some uORFs have been shown to exert their regulatory function through their encoded peptides, with this regulation depending on their aa sequence [70], and being able to occur in trans [71-74].

To assess the functional potential of smORFs, some studies have used an integrative approach to take advantage of the extensive RNA-seq, Ribo-seq, and HPLC-MS datasets available, to assess the translation, conservation, and coding potential of smORFs in several organisms. Mackoviak et al.[75] identified, computationally, a total of 2,002 novel putatively functional smORFs in 5 different organisms, based on their conservation patterns (obtained, briefly, with an SVM-based classifier, taking into account ORF conservation in multiple alignments, and PhyloCSF and PhastCons scores). These peptides map mostly to UTRs and LncRNAs, show little homology to known proteins, and are shorter than annotated smORFs,
also having different aa sequence properties [75]. Interestingly these smORFs have Ribo-seq ORFscore values that are higher than non-coding controls, but lower than annotated smORFs. Similarly, Ruiz-Orera et al. [76] found that, in several species, smORFs in IncRNAs have intermediate Ribo-seq and conservation features, which resemble those of newly evolved peptides. These results are, overall, reminiscent to those of Aspden et al. [32] in Drosophila, reinforcing the idea of functionally distinct classes of smORFs.

Although these studies provide valuable information regarding the functional potential of smORFs, they remain elusive about their specific functions. These systematic smORF searches have the ultimate aim of advancing genome annotations, which ultimately entails the attribution of specific functions to these newly detected smORFs, and this functional characterisation certainly poses the next challenge towards which an increased amount of efforts should be directed. Advances in gene editing technologies such as CRIPSR, which allow to relatively quickly generate specific mutants in most organisms [77], and the development of more sensitive phenotypical screens and biochemical assays to accurately assess peptide functions [57], will help to start filling this void of functional information in the genome.


60. Ji, Z., et al., Many IncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife, 2015. 4.
69. Paharkova, V., et al., Rat Humanin is encoded and translated in mitochondria and is localized to the mitochondrial compartment where it regulates ROS production. Mol Cell Endocrinol, 2015. 413: p. 96-100.


FIGURE 1

Bioinformatic approaches

A  Sequence Composition Analysis

B  Conservation and Homology

Biochemical approaches

C  Ribosome Profiling

D  Mass Spectrometry
Figure 1. Bioinformatic and biochemical approaches for the prediction of putative function smORFs.

Bioinformatic approaches: A- Nucleotide composition analyses of primary smORF sequences (tarsal-less 1A ORF; yellow), such as codon composition or hexamere nucleotide frequencies, are able to determine their coding potential, since the nucleotide frequencies of functional protein-coding ORFs are not random, due to a biased codon usage. B- Functional protein-coding sequences are under evolutionary constrains. Identification of smORF in closely related species allows to assess whether nucleotide changes are constrained to maintain the aa sequence (Ka/Ks). Furthermore, phylogenetic analyses of smORF homologues predict conserved motifs, or protein domains, which can be further used to identify distant homologues, as shown by the phylogenetic tree of Sarcolamban family.

Biochemical approaches: C- Ribosome profiling is based on sequencing of nuclease protected-ribosome bound RNA fragments (footprints), and allows a qualitative and quantitative genome-wide assessment of translation. Separation of polysomal fractions (red rectangle), enables the isolation of actively translated smORF transcripts (Poly-Ribo-Seq), and in combination with Ribo-seq, has detected translated smORFs. D- Mass spectrometry (MS) has detected smORF-encoded products from a digested protein sample by matching experimental spectra to predicted spectra from a reference/custom protein-database.
FIGURE 2
**Figure 2. Different classes of smORFs detected by Ribo-seq and HPLC-MS in Humans, Zebrafish and fruit flies.**

**A-** Venn diagrams representing the number of smORFs detected by Ribo-seq (blue) or HPLC-MS (pink), relative to the total number of transcripts encoding each class of smORF (yellow) in humans[33, 44], zebrafish [31] and fruit flies [32]. In these organisms, HPLC-MS detects very few peptides from LncRNAs and uORFs (0%-0.3%), compared to annotated smORFs (12-33%), whereas Ribo-seq still detects a substantial amount of LncRNA smORFs and uORFs (3-30%, compared to 30-80% annotated smORFs), highlighting the difference in sensitivity between these techniques. The number of transcribed uORFs (*) was inferred from the number of transcripts with uORFs identified in other studies, for humans [61] and for zebrafish [29]; the number of peptides identified in humans by HPLC-MS (†) were obtained from Mackowiak et al.[75]

**B-** The higher detection rates of annotated smORFs by HPLC-MS could be due to their higher levels of expression, and larger (and more stable) peptides, which also correlate with their closer resemblance to canonical proteins, in terms of functional signatures (protein domain content, conservation). Although these observations imply that annotated smORFs represent a functionally distinct class from LncRNA smORFs and uORFs, the identification of a growing number of biologically active peptides encoded in previously non-coding RNAs and uORFs (italics) proves that their functionality should not be systematically discarded.
**Figure 3. Cellular functions of conserved smORF micropeptides.**

**A-** Muscle performance depends on intracellular levels of Ca\(^{2+}\) regulated by the Ryanodine receptors (RyR) and Sarcoendoplasmic reticulum (SER) calcium ATPase (SERCA) pump. A conserved family of smORF peptides bind SERCA inhibiting its activity. Their members, Sarcolamban (Scl) in *Drosophila* and Sarcolipin (Sln) and Phospholamban (Pln) and Myoregulin (Mln) in vertebrates, display specific expression patterns. In addition, a new vertebrate smORF, DWORF, activates SERCA by competitively displacing SERCA inhibitors.

**B-** The *Hemotin (Hemo)/Stannin (Snn)* family is necessary for regulation of phagocytosis in *Drosophila* and mouse macrophages. Trafficking of phagocytised particles depends on the phosphorylation states of phosphatidyl-inositol (PI). At early endosomes, PI is phosphorylated into PI(3)P by the PI3Kinase (PI3K68D). At late endosomes PI(3)P is phosphorylated into PI(3,5)P2, which leads to lysosome fusion (acidification) and degradation of cargo. Vesicle trafficking can be reversed by PI(3,5)P2 dephosphorylation by Myotubulurin phosphatases (Mtm). Therefore, maturation of phagocytized particles correlates positively with acidification and negatively with PI(3)P. The 88aa-Hemo/Snn peptides inhibit a 14-3-3ζ-mediated Pi368Dkinase activation.

**C-** In humans, the MRI-2 peptide is involved in the non-homologous end joining (NHEJ) double-strand break (DSB) DNA repair. This 69aa-long peptide is recruited to the nucleus upon DSBs induction, where it binds Ku70/Ku80 heterodimers, and stimulates DNA ligation through NHEJ.
Table 1. The Identification of putative functional smORFs using computational and experimental approaches, has led to their functional characterisation.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Method/Metrics</th>
<th>Identified smORFs:</th>
<th>Comments:</th>
<th>Ref.</th>
<th>Functional outcomes:</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>Conservation, transcription, optimal codon adaptation index</td>
<td>588</td>
<td>Putative functional novel smORFs.</td>
<td>19</td>
<td>122/140 smORFs with evidence of transcription, or translation, or conservation produce growth defects in different conditions.</td>
<td>15</td>
</tr>
<tr>
<td>Drosophila</td>
<td>Conservation, purifying selection, syntenic and transcription</td>
<td>401</td>
<td>Putative functional novel smORFs.</td>
<td>17</td>
<td>porocolumbo: regulates SERCA mediated-calcium uptake and heart rhythmicity in Drosophila.</td>
<td>48</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>Hexamer sequence-composition frequencies, conservation, transcription and purifying selection</td>
<td>3,241</td>
<td>Putative functional novel smORFs.</td>
<td>16</td>
<td>49/472 conserved and transcribed smORFs produce over-expression morphological phenotypes.</td>
<td>47</td>
</tr>
<tr>
<td>Mouse</td>
<td>Hexamer sequence composition (CRITICAL), and purifying selection</td>
<td>1,240</td>
<td>Putatively functional smORFs, of which 495 lack similarity to known proteins.</td>
<td>18</td>
<td>Epitope tag of 14/15 show translation and specific subcellular localization in HeLa cells.</td>
<td>18</td>
</tr>
<tr>
<td>Drosophila s2 cells</td>
<td>RBF density and coverage</td>
<td>228</td>
<td>Out of 274 annotated smORFs.</td>
<td>32</td>
<td>Epitope tag of 7 lncRNA smORFs and 3 uORFs show translation and specific subcellular localization in s2 cells.</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>815</td>
<td>Out of 915 possible smORFs in 125 lincRNAs transcribed in 52 cells.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2,708</td>
<td>Out of 9,069 possible uORFs within 52 cell transcripts.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhastCons</td>
<td></td>
<td>212</td>
<td>Pass a threshold that distinguishes intergenic from coding regions at 10% FDR (out of 228 annotated, translated smORFs).</td>
<td>22</td>
<td>hromatin: regulates endocytic maturation and phagocytosis in Drosophila hemocytes.</td>
<td>54</td>
</tr>
<tr>
<td>Custom database</td>
<td></td>
<td>99</td>
<td>Out of 228 annotated smORFs, 60 found in this study, plus 59 from PeptideAtlas.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zebrafish embryos</td>
<td>Enhanced Translated ORF classifier (TOC).</td>
<td>395</td>
<td>Novel coding genes (using chow et al ribo-seq data).</td>
<td>52</td>
<td>Toddler: promotes cell migration during gastrulation by activating Epidermal growth factor receptor signalling.</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
<td>smORFs. (12 secreted and 15 TMM).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human cells</td>
<td>Proteasome inhibition, and ELMIC fractionation Custom database</td>
<td>86</td>
<td>Novel peptides, 49 mapping to previously un-annotated transcripts, 8 mapping to lincRNAs, and 37 to 3’UTRs, 5’UTRs, and overlapping annotated CDSs.</td>
<td>44</td>
<td>MRR1-2: stimulates double-strand break repair through a direct interaction with the DNA end binding protein Ku.</td>
<td>56</td>
</tr>
</tbody>
</table>

* smORF searches with functional outcomes

<table>
<thead>
<tr>
<th>Computational</th>
<th>Ribo-seq</th>
<th>Mass-spectrometry</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organisms</strong></td>
<td><strong>Method/Metrics</strong></td>
<td><strong>Identified smORFs:</strong></td>
</tr>
<tr>
<td>Yeast</td>
<td>Conservation, transcription, optimal codon adaptation index</td>
<td>588</td>
</tr>
<tr>
<td>Drosophila</td>
<td>Conservation, purifying selection, syntenic and transcription</td>
<td>401</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>Hexamer sequence-composition frequencies, conservation, transcription and purifying selection</td>
<td>3,241</td>
</tr>
<tr>
<td>Mouse</td>
<td>Hexamer sequence composition (CRITICAL), and purifying selection</td>
<td>1,240</td>
</tr>
<tr>
<td>Drosophila s2 cells</td>
<td>RBF density and coverage</td>
<td>228</td>
</tr>
<tr>
<td></td>
<td></td>
<td>815</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2,708</td>
</tr>
<tr>
<td>PhastCons</td>
<td></td>
<td>212</td>
</tr>
<tr>
<td>Custom database</td>
<td></td>
<td>99</td>
</tr>
<tr>
<td>Zebrafish embryos</td>
<td>Enhanced Translated ORF classifier (TOC).</td>
<td>395</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
</tr>
<tr>
<td>Human cells</td>
<td>Proteasome inhibition, and ELMIC fractionation Custom database</td>
<td>86</td>
</tr>
</tbody>
</table>

**sORF finder**: bioinformatic package to identify smORFs with high confident coding potential based on their similarity in nucleotide composition to bona fide coding genes by hidden Markov model. Potential coding sORFs are further tested for functionality by searching homologues and evolutionary constrains [16].

**Coding Region Identification Tool Invoking Comparative Analysis (CRITICA)**: gene prediction algorithm, which intergrates a purifying selection analysis of pair-wise aligned homologous regions into a hexamere sequence composition-analysis [18].

**PhastCons**: program that predicts conserved elements in multiple alignment sequences. It is based on a statistical hidden Markov phylogenetic model (phylo-HMM) that takes into account the probability of nucleotide substitutions at each site in a genome and how this probability changes from one site to the next [20].

**PhyloCSF**: comparative sequence method that analyses multiple alignments of nucleotide sequence using statistical comparison of phylogenetic codon models to ascertain the likelihood to be a conserved protein coding sequence [21].

**Micropeptide detection pipeline (micPDP)**: method that evaluates the existence of purifying selection on aa sequence from codon nucleotide changes. This pipeline filters candidate alignments according to coverage and reading frame conservation and then PhyloCSF method is applied to assess their coding potential from codon substitutions in genome-wide multi-alignments [31].

**ORFscore**: translation-dependent metric that exploits the 3-nt step movement of translating ribosomes across the transcript. Therefore, the Ribo-seq reads in coding ORFs tend to show a tri-nucleotide codon periodicity on the frame of translation (phasing)[31]. This method requires a strict cut-off for the size of analysed RBFs (only precise size reads, usually 28-29nt, are used), which could lead to a significant loss in read density.

**Ribosome Release Score (RRS)**: metric based on the releasing ability of ribosomes from the translating RNA after they encounter a stop codon. RRS is defined as the ratio between the total number of Ribo-seq reads in the ORF and the total number Ribo-seq reads in the subsequent 3’UTR, normalized respectively to the total length of their regions divided by the normalized number of RNA-seq reads in each region computed in the same fashion [78].

**Fragment length organisation similarity score (FLOSS)**: this method relies on the difference of the fragment size distribution of the ribo-seq footprints in coding genes and non-coding RNAs. This metric scores the coding potential of ORFs according to their similarity of the length of protected footprints of known coding genes [58].
BOX-2. Evaluation of coding potential and translation of smORFs by computer learning classifiers.

Coding Potential Calculator (CPC): bioinformatics tool that scores six sequences features to distinguish coding vs non-coding ORFs, three relate to the quality of the longest ORF (ORF size, Coverage, integrity) whereas the other three are based on sequence conservation using BLASTX (number of hits, quality of the hits, frame distribution of hits) that are incorporated in a Support Vector learning machine classifier. [79-80].

Translated ORF Classifier (TOC): a Ribo-seq classifier based on a random Forest model that assess the coding potential of each ORF within a transcript based on 4 metrics: Translation Efficiency (ratio of the Ribo-seq reads/RNA-seq read within the ORF: Level of translation), Inside vs Outside (coverage inside ORF/coverage outside ORF; coverage is number of nucleotides having Ribo-seq reads/total number of nucleotides), Fraction Length (fraction of the transcript covered by ORF) and Disengagement score (DS) assess the release efficiency of the ribosome after a stop codon which is a characteristic of ribosome translating coding ORFs by measuring the Ribo-seq reads in the ORF/Ribo-seq reads downstream. [29]. Pauli et al.[52] improved TOC classifier by adding a “cover” metric (number of nucleotides of the ORF covered by Ribo-seq reads).

ORF Regression Algorithm for Translational Evaluation of RPFs (ORF-RATER): this metric is able to identify and quantify translation in ORFs from Ribo-seq data by comparing the patterns of ribosome occupancy (initiation and termination peaks and elongation phase) to that of coding ORFs. ORF-RATER uses a linear regression model that allows the integration of multiple lines of evidence and evaluates each ORF according to the nearby context [59].

RibORF Classifier: a Ribo-seq Support Vector Machine classifier that defines active translation of ORFs according to the evaluation of phasing by using 5’ footprint off-set distances to the ribosome A-site, from canonical proteins, to identify 3nt periodicity, and uniformity of footprint distribution across codons by calculating the percentage of maximum entropy values [60].
Supplementary Table 1: Number smORFs identified using computational, Ribo-seq and proteomics approaches in different organisms.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Organism</th>
<th>Method/Metrics</th>
<th>Identified smORFs:</th>
<th>Comments:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kusar et al. (2003)</td>
<td>Yeast</td>
<td>Conservation, transcription, optimal codon adaptation index</td>
<td>588</td>
<td>Putative functional novel smORFs.</td>
</tr>
<tr>
<td>Reib et al. (2006)</td>
<td>Mouse</td>
<td>Hexamere sequence composition (CRITICA), and purifying selection</td>
<td>1,240</td>
<td>Putatively functional smORFs, of which 495 lack similarity to known proteins.</td>
</tr>
<tr>
<td>Asplin et al. (2014)</td>
<td>Drosophila S2 cells</td>
<td>RBF density and coverage</td>
<td>228</td>
<td>Out of 274 annotated smORFs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>313</td>
<td>Out of 918 possible smORFs in 125 lncRNAs transcribed in S2 cells.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2,708</td>
<td>Out of 9,069 possible uORFs within S2 cell transcripts.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PhastCons</td>
<td>212</td>
<td>Pass a threshold that distinguishes intergenic from coding regions at 10% FDR (out of 228 annotated, translated smORFs).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Custom database</td>
<td>99</td>
<td>Out of 228 annotated smORFs, 60 found in this study, plus 39 from PeptideAtlas.</td>
</tr>
<tr>
<td>Bazzini et al. (2014)</td>
<td>Zebrafish embryos</td>
<td>RBF coverage and phasing (ORFScore)</td>
<td>302</td>
<td>Out of 581 annotated smORFs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>190</td>
<td>Out of 2540 possible smORFs in un-annotated transcripts and lncRNAs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>311</td>
<td>gORFs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Custom database</td>
<td>98</td>
<td>Previously annotated smORFs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>micPDP</td>
<td>6</td>
<td>smORFs in un-annotated transcripts and lncRNAs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>17</td>
<td>gORFs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>89</td>
<td>smORFs (12 secreted and 15 TMHHM).</td>
</tr>
<tr>
<td>Grappe et al. (2013)</td>
<td>Mouse E14 mESCs cells</td>
<td>sORFFinder + SVM</td>
<td>1,100</td>
<td>smORFs in lncRNAs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>27, 371</td>
<td>Intergenic smORFs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RBF Coverage, and start codon pile-up with Harringtonine treated samples</td>
<td>23,127</td>
<td>Other smORFs (intronic, overlapping coding exons).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>528</td>
<td>smORFs in lncRNAs (401 pass sORFFinder + SVM filter).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>228</td>
<td>Intergenic smORFs (89 pass sORFFinder + SVM filter)</td>
</tr>
<tr>
<td>Lee et al. (2012)</td>
<td>Human HEK93 cells</td>
<td>Global initiation site sequencing (GTI-seq) with LTM treated samples</td>
<td>288</td>
<td>smORFs (median size 18 aa) out of 5,763 ncRNA transcripts (Refseq).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6,729</td>
<td>gORFs in 3352 genes (mostly smORFs).</td>
</tr>
<tr>
<td>Ma et al. (2013)</td>
<td>Human (cells / tissues)</td>
<td>Protease inhibition, and ELRIC fractionation Custom database</td>
<td>237</td>
<td>Novel peptides, 80% mapping to previously un-annotated transcripts and 20% to 3'UTRs, 5'UTRs, and overlapping annotated CDSS.</td>
</tr>
<tr>
<td>Vanderperre et al. (2013)</td>
<td>Human (cells / tissues)</td>
<td>56 HPLC-MS datasets from peptide Atlas matched to custom database</td>
<td>1,256</td>
<td>Novel peptides within annotated transcripts, 3'UTRs, 5'UTRs, and overlapping annotated CDSS.</td>
</tr>
<tr>
<td>Sieroff et al. (2013)</td>
<td>Human cells</td>
<td>Protease inhibition, and ELRIC fractionation Custom database</td>
<td>86</td>
<td>Novel peptides, 49 mapping to previously un-annotated transcripts. 8 mapping to LncRNAs, and 37 to 3'UTRs, 5'UTRs, and overlapping annotated CDSS.</td>
</tr>
</tbody>
</table>