

Sussex Research

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

Jose Pueyo-Marques, Emile G Magny, Juan Pablo Couso

Publication date

05-08-2015

Licence

This work is made available under the **Copyright not evaluated** licence and should only be used in accordance with that licence. For more information on the specific terms, consult the repository record for this item.

Document Version

Accepted version

Citation for this work (American Psychological Association 7th edition)

Pueyo-Marques, J., Magny, E. G., & Couso, J. P. (2015). *New peptides under the s(ORF)ace of the genome* (Version 1). University of Sussex. <https://hdl.handle.net/10779/uos.23440169.v1>

Published in

Trends in Biochemical Sciences

Link to external publisher version

<https://doi.org/10.1016/j.tibs.2016.05.003>

Copyright and reuse:

This work was downloaded from Sussex Research Open (SRO). This document is made available in line with publisher policy and may differ from the published version. Please cite the published version where possible. Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners unless otherwise stated. For more information on this work, SRO or to report an issue, you can contact the repository administrators at sro@sussex.ac.uk. Discover more of the University's research at <https://sussex.figshare.com/>

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

J.I. Pueyo^{1†}, E.G. Magny^{1†}, and J.P. Couso^{1*}

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, perhaps thousands 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 small Open Reading Frame genes (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.

Identifying functional smORFs, a major challenge for genome annotations.

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 putatively translated 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 because standard computational algorithms to identify coding sequences are generally not suited for small sequences [3-5]. Initially, short coding sequences (<100 amino acids (aa)) 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], such as the polycistronic *tarsal-less* 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, which have been used to identify smORFs, and will present some of the different examples of smORFs that 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 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 sORFinder and CRITICA in Box 1), 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 nucleotide (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 (Figure 2).

In order to identify conserved coding sequences, more recent methods based on multiple sequence alignments incorporate phylogenetic distances and a model of nt substitution rates, in the case of PhastCons [20], or a model of codon substitution frequencies, in the case of phyloCSF [21]. Both are built upon known coding and non-coding sequences (see

Box 1). As shown later in this Review, 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 the identification of entire transcriptomes [22] and has 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 a similar length, being transcribed by RNA-polymerase II, capped, and 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] allows the quantitative and qualitative measurement of the translation of these transcriptomes. This method consists of sequencing nuclease-protected mRNA fragments (or footprints) bound by translating ribosomes stabilized with an elongation inhibitor such as cycloheximide (CHX) [28].

Different ribosome profiling studies, in a wide variety of species [29-40], have found that translation occurs in an almost pervasive fashion. Ribosome footprints are detected in

LncRNAs, in the untranslated regions (UTRs) of annotated transcripts either upstream (uORFs) or downstream (dORFs) of the coding DNA sequence (CDS), and even overlapping the CDS of canonical mRNAs, with the vast majority of these corresponding to short ORFs (Figures 2 and S2).

However there is some ambiguity with this method, since a ribosome protected fragment (RPF) 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 RPF coverage, translation efficiency (TE: the ratio of RPFs / total mRNA reads), ribosomal release score (RRS), or codon phasing (see Box 1, [2]). Translation inhibitors, such as harringtonine (HR), which generate a pile-up of RPFs at the start codon, have also been used to identify translation initiation sites in actively translated ORFs [30]. Unexpectedly, studies using this approach [27,30,33,38,41], as well as some peptidomics studies [42-43], have shown that a considerable amount of translation, including that of many novel smORFs, initiates from non-canonical start codons, which increases the complexity of the proteome, and highlights the importance of biochemical detection methods to obtain comprehensive translation profiles. The sensitivity offered by Ribo-Seq, has allowed several studies to use this technique to identify translated smORFs in yeast [36], fruit flies [32], zebrafish [31], mice [44], and humans [33].

In *Drosophila*, Aspden *et al.* [32] incorporated polysomal fractionation before Ribo-Seq to isolate cytoplasmic RNAs bound by 2-6 ribosomes. This selected for those RNAs being actively translated, rather than those being scanned by single non-productive ribosomes or

other RNA-binding proteins, and also enriched for RNAs encoding short ORFs (6 ribosomes being the maximum number that could fit in a 300 nt ORF). Using stringent RPF 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, respectively (Figure 3). Annotated smORFs were found to be ~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 sub-cellular localizations similar to those of canonical proteins, suggesting that some of them may encode functional peptides.

In zebrafish embryos, Ribo-Seq profiles were analysed using ORFscore (See Box 2) [31], a method that quantifies the 3-codon periodicity of the distribution of RPFs 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 dORFs (Figures 2 and 3). 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 were found to associate with poly-ribosomes (supporting their translation) [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 (Figure 2).

In another study, human and mouse cell lines, were treated with lactimidomycin (LTM), another initiation phase inhibitor, prior to Ribo-Seq, in order to globally identify translation initiation sites [33]. 227 annotated Human smORFs were identified as translated (out of 694 of annotated smORFs in ENSEMBL), as well as 288 ORFs in LncRNAs and 1,194 uORFs (most of them <100 aa long) (Figures 2 and 3).

Altogether these studies show that thousands of smORFs are translated in eukaryotic genomes. A substantial number of smORFs exhibit 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 tandem Mass-spectrometry (HPLC-MS/MS) proteomics approach has also been adapted to identify novel small peptides, by modifying, mainly, the protocols for data analysis. Here, 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). Custom databases greatly increase the peptide spectra search space. This could potentially lead to higher rates of false positives, particularly since post-translational

modifications, which have been shown to account for a substantial portion of unassigned HPLC-MS/MS spectra [45] , are not always taken into account, and could therefore lead to miss-identification of peptides.

One study identified 1,259 novel peptides by matching the spectra of 16 different HPLC-MS/MS data-sets from different human samples, to such a custom database, which included all possible alternative ORFs (mapping to UTRs and overlapping CDS') in the human transcriptome [46] (Figure S2). This study suggests 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. 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 analyzed in this study.

Another important consideration, is that standard proteomics require protein sequences to be supported by multiple PSMs. smORFs are often too short to fit more than one PSM; this single PSM should therefore be required to pass the most stringent criteria in order to be unambiguously assigned to that smORF. This higher stringency will reduce the rate of false positives, but may compromise the detection of *bona-fide* smORFs peptides. Slavoff *et al.* [42] developed a peptidomics strategy, taking into account these analytical considerations, while also applying specific experimental optimizations. First, they inhibited proteolysis, arguing that the proteolytic fragments of canonical proteins greatly increase the complexity of the peptidome and deteriorate the signal to noise ratio when it comes to identifying short peptides; second, they used electrostatic-repulsion hydrophilic interaction chromatography (ERLIC) prior to HPLC-MS/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 (Figure 2).

This method was tested against other workflows [43], 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. This latter finding highlights 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 (Figure S2). The rest map to alternative CDS within annotated transcripts with a similar distribution (in UTRs and overlapping CDS') as found by Slavoff *et al.* [42].

Identification of smORFs by multiple approaches

In an attempt to more reliably identify translated smORFs, different studies have combined different computational and biochemical methods.

In one study, a computational smORF search was carried out in order to identify potentially coding smORFs in the mouse genome [44], which were then compared to an available ribosome profiling dataset from a mouse cell line [30]. Putatively coding smORFs, conserved across mammalian species, were recovered 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 28, 471 smORFs with high coding probability in intergenic regions and LncRNAs. In parallel, they re-analysed the Ribo-Seq dataset, and identified 528 intergenic smORFs and 226 smORFs in LncRNAs, passing a coverage threshold and showing a pile-up at their start codon when treated with HR; of these, 401 and 89, respectively, were also found in the computational pipeline, representing a pool of smORFs likely to encode functional peptides (Figure S2).

This study highlights the discrepancy in numbers that can exist between computational predictions and experimental detections. Part of this discrepancy could be explained on one hand by a possible high false positive rate in the bioinformatic pipeline, which could be due to, for example, to the presence of conserved elements such as transposons, pseudogenes, and simple repeats [47]. These false positive rates vary greatly among studies, depending on the stringency of filters applied in the computational pipelines. On the other hand, it could also be partly explained by the fact that computational pipelines 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 experimental approaches.

Some of the Ribo-Seq-based studies covered above have used HPLC-MS/MS in order to validate their results (Figures 2 and S2). In general, previously annotated smORFs tend to be more abundantly detected by HPLC-MS/MS than uORFs or LncRNA smORFs; Aspden *et al.*[32] and Bazzini *et al.*[31] detected almost a third of the 228 and 302 annotated translated smORFs, respectively, but Aspden *et al.*[32] failed to identify any peptide from LncRNAs or uORFs, and Bazzini *et al.* [31] only identified 3 and 17 peptides, respectively. Similarly, only a handful of peptides corresponding to uORFs and LncRNAs have been detected by HPLC-MS/MS in studies that detected hundreds by Ribo-Seq in humans [48-49]

(Figures 2, S2 and 3). These results clearly highlight a difference of sensitivity between these methods. They are also in agreement with the stochastic nature of peptidomics, observed by Slavoff *et al.*[42], and with the shorter size and lower translation efficiency of LncRNA smORFs and uORFs observed by Aspden *et al.*[32]; the peptides from LncRNAs or uORFs being generally smaller and probably less stable, and overall less abundant, have lower chances of being detected. In that sense, detection by peptidomics could be considered as a convincing proof of translation, and may indicate that the detected peptide is probably functional, provided that this detection is not the result of “translational noise” or a false positive; the absence of detection by peptidomics, however, should not be used to discard functionality (Figure 3). It is also 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 rates of these smaller peptides.

Other studies have taken advantage of the extensive RNA-Seq, Ribo-Seq, and HPLC-MS/MS datasets available, to assess the translation, conservation, and coding potential of smORFs in several organisms. Mackoviak *et al.* [49] identified 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. 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.*[50] found that, in several species, smORFs in LncRNAs

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.

Computational and biochemical strategies lead to novel functional smORF peptides

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, these small peptides may act as regulators of cellular processes requiring a very specific and in-depth analysis in order to detect their mutant phenotype. As a result, only a handful of smORFs, found serendipitously, had been functionally characterised prior to these extensive smORF searches [8].

These genome-wide smORF searches have aided the functional characterisation of smORFs by identifying candidates for functional analysis. Following their bioinformatic predictions, some studies have carried out high-throughput functional screens in yeast [15] and in plants [51], and found dozens of functional smORFs, with several being essential (Figure 2).

Other studies have focused on a more in-depth characterisation of specific smORFs, like that of *Sarcolamban (Scl)* in *Drosophila* [52], a gene previously annotated as non-coding [53], but identified, by a bioinformatics approach, to encode potentially functional smORFs (Figure 2;[17]). *Scl* encodes two 28 and 29aa related transmembrane peptides that act as inhibitors of the sarco-endoplasmic reticulum Ca^{2+} ATPase, and regulate heart muscle contraction (Figure 4A;[52]). Importantly, these peptides appear to be functional

homologues to the vertebrate Sarcolipin (Sln) and Phospholamban (Pln) peptides, thereby uncovering an ancestral family of smORFs conserved from insects to humans [52]. More recently, another member of this family, Myoregulin (46 aa) [54], and a novel small peptide with an antagonistic function, called DWORF (34 aa) [55], have both been identified in mice from transcripts previously annotated as non-coding.

Another example is the *toddler/apela* gene, initially characterised in vertebrates [56], and identified through a Ribo-Seq-based search for novel signalling peptides in zebrafish (Figure 2;[57]). The *apela* gene encodes a secreted 58 aa peptide that binds to the Apelin receptor and promotes cell mobility during gastrulation [57]. 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 its translation subsequently supported by ribosome profiling and proteomics studies [32,58] (Figure 2). *hemo* is expressed in hemocytes (*Drosophila* macrophages) where it regulates endosomal maturation and phagocytosis by inhibiting the activity of phosphatidylinositol kinases through an interaction with 14-3-3z (Figure 4B;[59]). Interestingly, the vertebrate Stannin (Snn) peptide, a factor involved in organometallic cytotoxicity [60], was identified as the functional homologue of Hemotin in vertebrates [59], showing that this regulatory mechanism is also conserved across evolution.

In humans, the 69 aa long MRI-2 peptide, was shown to stimulate double-strand break repair through a direct interaction with the DNA end binding protein *Ku* (Figure 4C;[61]). This peptide was functionally characterized as a direct result of a HPLC-MS/MS screen for novel short peptides (Figure 2; [42]) which detected it as translated in K562 cells.

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 and are functionally conserved, in some cases, across vast evolutionary distances. Overall, their study can certainly have important implications in cell biology, and in human medical research [62].

Concluding remarks and future perspectives

In this Review, we have shown 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, like, for example, the FLOSS and RiboTaper (See Box 2) methods: the former, based on the assessment of the organization of RPF lengths, specifically identifies the reads protected by translation-engaged ribosomes [41]; the latter applies a spectral theory-based analytical method to identify the ribosome foot-print profiles across given transcripts that follow a tri-nucleotide periodicity and represent translated regions [48]. Other groups have used classification algorithms, such as the random forest-based Translated ORF classifier (TOC) [29,57], the logistic regression-based ORF-rater [63], or the SVM-based RibORF [64], which integrate different Ribo-Seq metrics, and their profiles on known coding and non-coding regions, to identify translated ORFs. All of these studies support the translation of hundreds of novel small peptides in vertebrates, encoded in transcripts previously thought to be non-coding or encoded as uORFs.

It remains challenging, however, to distinguish 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 through their engagement of ribosomes [65-67], which has been inferred to be their main function. 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 [36,68-69]. Nonetheless, as shown in this Review, several smORF-encoded peptides with important functions have been identified in previously non-coding RNAs, proving that these sequences can certainly encode functional peptides [10,52,54-55,57,59]. There are even examples of canonical non-coding RNAs, such as pri-miRNAs in plants [70] and ribosomal RNAs in mammals [71-73], encoding biologically active peptides with well characterised functions. Similarly, several uORFs show conservation features reminiscent of coding proteins in mice and humans [74], and some uORFs have been shown to exert their regulatory function through their encoded peptides, with this regulation depending on their aa sequence [75], and being able to occur in *trans* [76-79].

Systematic smORF searches have the ultimate aim of advancing genome annotations, which entails the attribution of specific functions to these newly detected smORFs. Although these studies, whether based on computational predictions, or experimental detection, by Ribo-Seq or Mass-spec or both, provide valuable information regarding the functional potential of smORFs, they remain elusive about their specific functions. This functional characterisation certainly poses the next challenge towards which an increased amount of effort should be directed.

Some general functional characteristics, like the segregation of smORFs into distinct functional classes, have been proposed [32], and appear to be supported by the results of

different studies [49-50], but these also remain to be tested through detailed phenotypic analysis. Advances in gene editing technologies such as CRIPSR, which allow one to relatively quickly generate specific mutants in most organisms [80], and the development of more sensitive phenotypical screens and biochemical assays to accurately assess peptide functions [62], will help to start filling this void of functional information in the genome.

References:

1. Basrai MA, Hieter P, Boeke JD (1997) Small Open Reading Frames: Beautiful Needles in the Haystack. *Genome Research* 7: 768-771.
2. Mumtaz MA, Couso JP (2015) Ribosomal profiling adds new coding sequences to the proteome. *Biochem Soc Trans* 43: 1271-1276.
3. Wang J, Li S, Zhang Y, Zheng H, Xu Z, et al. (2003) Vertebrate gene predictions and the problem of large genes. *Nat Rev Genet* 4: 741-749.
4. Cheng H, Chan WS, Li Z, Wang D, Liu S, et al. (2011) Small Open Reading Frames: Current Prediction Techniques and Future Prospect. *Curr Protein Pept Sci*.
5. Sleator RD (2010) An overview of the current status of eukaryote gene prediction strategies. *Gene* 461: 1-4.
6. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, et al. (1996) Life with 6000 genes. *Science* 274: 546, 563-547.
7. Dujon B, Alexandraki D, Andre B, Ansorge W, Baladron V, et al. (1994) Complete DNA sequence of yeast chromosome XI. *Nature* 369: 371-378.
8. Andrews SJ, Rothnagel JA (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* 15: 193-204.
9. Savard J, Marques-Souza H, Aranda M, Tautz D (2006) A segmentation gene in *Tribolium* produces a polycistronic mRNA that codes for multiple conserved peptides. *Cell* 126: 559-569.
10. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP (2007) Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biology* 5: 1052-1062.
11. Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, et al. (2007) Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biology* 9: 660-U687.
12. Zanet J, Benrabah E, Li T, Pelissier-Monier A, Chanut-Delalande H, et al. (2015) Pri sORF peptides induce selective proteasome-mediated protein processing. *Science* 349: 1356-1358.
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal Of Molecular Biology* 215: 403-410.
14. Nekrutenko A, Makova KD, Li WH (2002) The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res* 12: 198-202.
15. Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au W-C, et al. (2006) Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res* 16: 365-373.
16. Hanada K, Zhang X, Borevitz JO, Li WH, Shiu SH (2007) 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 Research* 17: 632-640.
17. Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP (2011) Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol* 12: R118.
18. Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, et al. (2006) The abundance of short proteins in the mammalian proteome. *Plos Genetics* 2: 515-528.
19. Kessler MM, Zeng Q, Hogan S, Cook R, Morales AJ, et al. (2003) Systematic Discovery of New Genes in the *Saccharomyces cerevisiae* Genome. *Genome Res* 13: 264-271.
20. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15: 1034-1050.
21. Lin MF, Jungreis I, Kellis M (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27: i275-282.
22. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10: 57-63.

23. Hangauer MJ, Vaughn IW, McManus MT (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* 9: e1003569.
24. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, et al. (2011) The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471: 473-479.
25. van Heesch S, van Iterson M, Jacobi J, Boymans S, Essers PB, et al. (2014) Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol* 15: R6.
26. Kung JT, Colognori D, Lee JT (2013) Long noncoding RNAs: past, present, and future. *Genetics* 193: 651-669.
27. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324: 218-223.
28. Ingolia NT (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* 15: 205-213.
29. Chew GL, Pauli A, Rinn JL, Regev A, Schier AF, et al. (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* 140: 2828-2834.
30. Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of Mammalian proteomes. *Cell* 147: 789-802.
31. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, et al. (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J*. 2014/04/08 ed. pp. embj.201488411.
32. Aspden JL, Eyre-Walker YC, Philips RJ, Brocard M, Amin U, et al. (2014) Extensive translation of small ORFs revealed by polysomal ribo-Seq *eLife* 3: e03528.
33. Lee S, Liu B, Lee S, Huang SX, Shen B, et al. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A* 109: E2424-2432.
34. Duncan CD, Mata J (2014) The translational landscape of fission-yeast meiosis and sporulation. *Nat Struct Mol Biol* 21: 641-647.
35. Vasquez JJ, Hon CC, Vanselow JT, Schlosser A, Siegel TN (2014) Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages. *Nucleic Acids Res* 42: 3623-3637.
36. Smith JE, Alvarez-Dominguez JR, Kline N, Huynh NJ, Geisler S, et al. (2014) Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. *Cell Rep* 7: 1858-1866.
37. Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, et al. (2012) High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 335: 552-557.
38. Stern-Ginossar N, Weisburd B, Michalski A, Le VT, Hein MY, et al. (2012) Decoding human cytomegalovirus. *Science* 338: 1088-1093.
39. Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS (2013) Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *eLife* 2: e01179.
40. Juntawong P, Girke T, Bazin J, Bailey-Serres J (2014) Translational dynamics revealed by genome-wide profiling of ribosome footprints in *Arabidopsis*. *Proc Natl Acad Sci U S A* 111: E203-212.
41. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, et al. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* 8: 1365-1379.
42. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, et al. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* 9: 59-64.
43. Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, et al. (2014) Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res* 13: 1757-1765.

44. Crappe J, Van Crielinge W, Trooskens G, Hayakawa E, Luyten W, et al. (2013) Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics* 14: 648.
45. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, et al. (2015) A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* 33: 743-749.
46. Vanderperre B, Lucier JF, Bissonnette C, Motard J, Tremblay G, et al. (2013) Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS One* 8: e70698.
47. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* 7: e1002384.
48. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, et al. (2016) Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* 13: 165-170.
49. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, et al. (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol* 16: 179.
50. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM (2014) Long non-coding RNAs as a source of new peptides. *Elife* 3: e03523.
51. Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, et al. (2013) Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc Natl Acad Sci U S A* 110: 2395-2400.
52. Magny EG, Pueyo JI, Pearl FM, Cespedes MA, Niven JE, et al. (2013) Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* 341: 1116-1120.
53. Tupy JL, Bailey AM, Dailey G, Evans-Holm M, Siebel CW, et al. (2005) Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America* 102: 5495-5500.
54. Anderson Douglas M, Anderson Kelly M, Chang C-L, Makarewich Catherine A, Nelson Benjamin R, et al. (2015) A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* 160: 595-606.
55. Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, et al. (2016) Muscle physiology. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* 351: 271-275.
56. Chng SC, Ho L, Tian J, Reversade B (2013) ELABELA: a hormone essential for heart development signals via the apelin receptor. *Dev Cell* 27: 672-680.
57. Pauli A, Norris ML, Valen E, Chew GL, Gagnon JA, et al. (2014) Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* 343: 1248636.
58. Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, et al. (2007) A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat Biotechnol* 25: 576-583.
59. Pueyo JI, Magny EG, Sampson CJ, Amin U, Evans IR, et al. (2016) Hemotin, a regulator of phagocytosis encoded by a small ORF and conserved across metazoans. *PloS Biology* 14: e1002395.
60. Billingsley ML, Yun J, Reese BE, Davidson CE, Buck-Koehntop BA, et al. (2006) Functional and structural properties of stannin: roles in cellular growth, selective toxicity, and mitochondrial responses to injury. *J Cell Biochem* 98: 243-250.
61. Slavoff SA, Heo J, Budnik BA, Hanakahi LA, Saghatelian A (2014) A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J Biol Chem* 289: 10950-10957.
62. Saghatelian A, Couso JP (2015) Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat Chem Biol* 11: 909-916.

63. Fields AP, Rodriguez EH, Jovanovic M, Stern-Ginossar N, Haas BJ, et al. (2015) A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol Cell* 60: 816-827.
64. Ji Z, Song R, Regev A, Struhl K (2015) Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* 4.
65. Calvo SE, Pagliarini DJ, Mootha VK (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A* 106: 7507-7512.
66. Wang XQ, Rothnagel JA (2004) 5'-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. *Nucleic Acids Res* 32: 1382-1391.
67. Morris DR, Geballe AP (2000) Upstream open reading frames as regulators of mRNA translation. *Mol Cell Biol* 20: 8635-8642.
68. Mendell JT, Sharifi NA, Meyers JL, Martinez-Murillo F, Dietz HC (2004) Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat Genet* 36: 1073-1078.
69. Tani H, Torimura M, Akimitsu N (2013) The RNA Degradation Pathway Regulates the Function of GAS5 a Non-Coding RNA in Mammalian Cells. *PLoS ONE* 8: e55684.
70. Laressergues D, Couzigou JM, Clemente HS, Martinez Y, Dunand C, et al. (2015) Primary transcripts of microRNAs encode regulatory peptides. *Nature* 520: 90-93.
71. Guo B, Zhai D, Cabezas E, Welsh K, Nouraini S, et al. (2003) Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature* 423: 456-461.
72. Lee C, Zeng J, Drew BG, Sallam T, Martin-Montalvo A, et al. (2015) The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. *Cell Metab* 21: 443-454.
73. Paharkova V, Alvarez G, Nakamura H, Cohen P, Lee KW (2015) Rat Humanin is encoded and translated in mitochondria and is localized to the mitochondrial compartment where it regulates ROS production. *Mol Cell Endocrinol* 413: 96-100.
74. Crowe ML, Wang XQ, Rothnagel JA (2006) Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. *BMC Genomics* 7: 16.
75. Ebina I, Takemoto-Tsutsumi M, Watanabe S, Koyama H, Endo Y, et al. (2015) Identification of novel *Arabidopsis thaliana* upstream open reading frames that control expression of the main coding sequences in a peptide sequence-dependent manner. *Nucleic Acids Research* 43: 1562-1576.
76. Akimoto C, Sakashita E, Kasashima K, Kuroiwa K, Tominaga K, et al. (2013) 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.
77. Nguyen HL, Yang X, Omiecinski CJ (2013) 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.
78. Pendleton LC, Goodwin BL, Solomonson LP, Eichler DC (2005) Regulation of endothelial argininosuccinate synthase expression and NO production by an upstream open reading frame. *J Biol Chem* 280: 24252-24260.
79. Diba F, Watson CS, Gametchu B (2001) 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.
80. Sander JD, Joung JK (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol* 32: 347-355.

81. Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, et al. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research* 35: W345-W349.
82. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* 147: 1537-1550.
83. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154: 240-251.

Bioinformatic approaches

A Sequence Composition Analysis

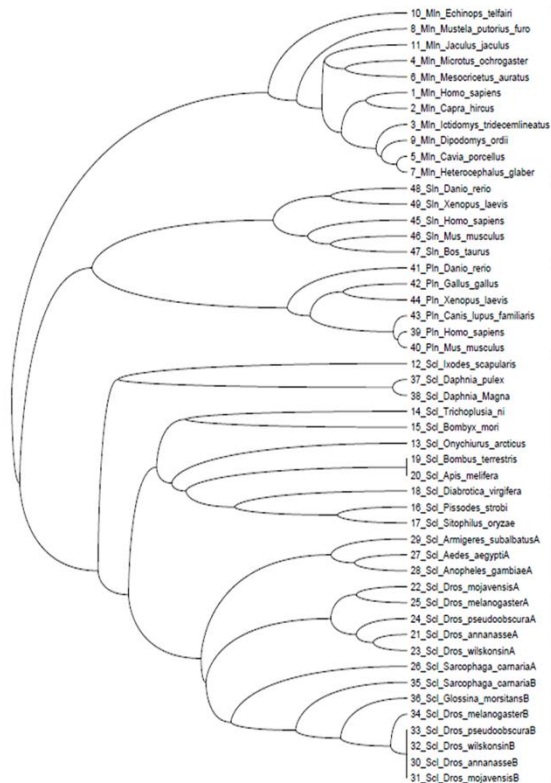
```

aaaaagccagctcggtttgtcattcaagtatttttggtcaatacacggcatacgaatg
K K P A R F C H S S I F G Q Y T A Y E
gcagcctacttggatcccaactggccagctactaaagaagctacacgacagcagaaga
A A Y L D P T G Q Y * R S Y T T T A R H
cgtaatcgtagacctcttttagaaaatccaataaatcacagatcttcgcatggccgcct
R N R R P L L E N P I N H R S S P W P P
atctggatcccaactggctcagctactgaagttggagcaagcagaagcagcaatattt
I W I P L V S T E V G A S K Q K Q Q Y F

```

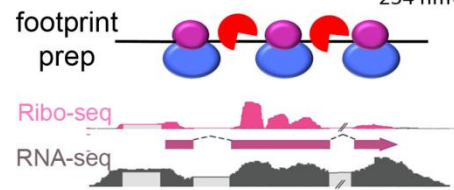
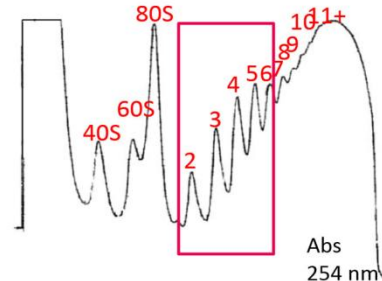
1A
11 aa

B Conservation and Homology



Biochemical approaches

C Ribosome Profiling



D Mass Spectrometry

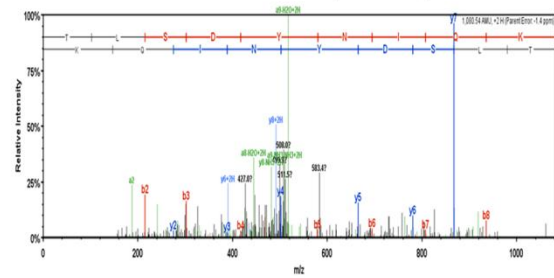
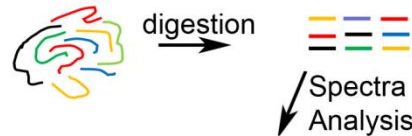


FIGURE 1

Figure 1. Bioinformatic and biochemical approaches for the prediction of putatively functional smORFs.

Bioinformatic approaches (A, B): **A-** Nucleotide composition analyses of primary smORF sequences (tarsal-less 1A ORF; yellow), such as codon composition or hexamer 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 constraints. Identification of smORFs in closely related species allows one to assess whether nucleotide changes are constrained to maintain the aa sequence (K_a/K_s). 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,D): **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 a subset of polysomal fractions (red rectangle) enables the isolation of actively translated smORF transcripts (Poly-Ribo-Seq); in combination with Ribo-Seq, this 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 or custom protein-database.

Figure 2. The Identification of putative functional smORFs using computational and experimental approaches has led to their functional characterisation.

smORF searches with functional outcomes						
Organism:	Method/Metrics:	Identified smORFs:	Comments:	Ref.	Functional outcomes:	Ref.
Yeast	Conservation, transcription, optimal codon adaptation index	588	Out of 140,000 smORFs, at least 18 codons long identified in the <i>S. cerevisiae</i> genome.	19	22/140 smORFs with evidence of transcription, or translation, or conservation produce growth defects in different conditions.	15
Drosophila	Conservation, purifying selection, synteny and transcription	401	Out of 556,554 smORFs, at least 10 aa long, identified in non-exonic euchromatic regions of the <i>D. melanogaster</i> genome.	17	<i>sarcolamban</i> : regulates SERCA mediated-calcium uptake and heart rhythmicity in <i>Drosophila</i> .	52
Arabidopsis	Hexamer sequence-composition frequencies, conservation, transcription and purifying selection	3,241	Out of 570, 948, at least 30 aa long, identified in the intergenic sequences of the <i>A. thaliana</i> genome.	16	49/473 conserved and transcribed smORFs produce over-expression morphological phenotypes.	51
Drosophila S2 cells	RPF density and coverage	228	Out of 274 annotated smORFs.	32	<i>hemotin</i> : regulates endocytic maturation and phagocytosis in <i>Drosophila</i> hemocytes.	59
		313	Out of 918 possible smORFs in 125 lncRNAs transcribed in S2 cells.			
		2,708	Out of 9,069 possible uORFs within S2 cell transcripts.			
	PhastCons	212	Pass a threshold that distinguishes intergenic from coding regions at 10% FDR (out of 228 annotated, translated smORFs).			
	Custom database	99	Out of 228 annotated smORFs, 60 found in this study, plus 39 from PeptideAtlas.			
Zebrafish embryos	Enhanced Translated ORF Classifier (TOC).	399	Novel coding genes (using chew et al ribo-seq data).	57	<i>Toddler</i> : promotes cell migration during gastrulation by activating Apelin receptor signalling.	57
		89	smORFs (12 secreted and 15 TMHMM).			
Human cells	Protease inhibition, and ELRIC fractionation Custom database	86	Novel peptides, 49 mapping to previously un-annotated transcripts. 8 mapping to lncRNAs, and 37 to 3'UTRs, 5'UTRs, and overlapping annotated CDSs.	42	<i>MRI-2</i> : stimulates double-strand break repair through a direct interaction with the DNA end binding protein Ku.	61

Computational
Ribo-Seq
Mass-spectrometry

FIGURE 2

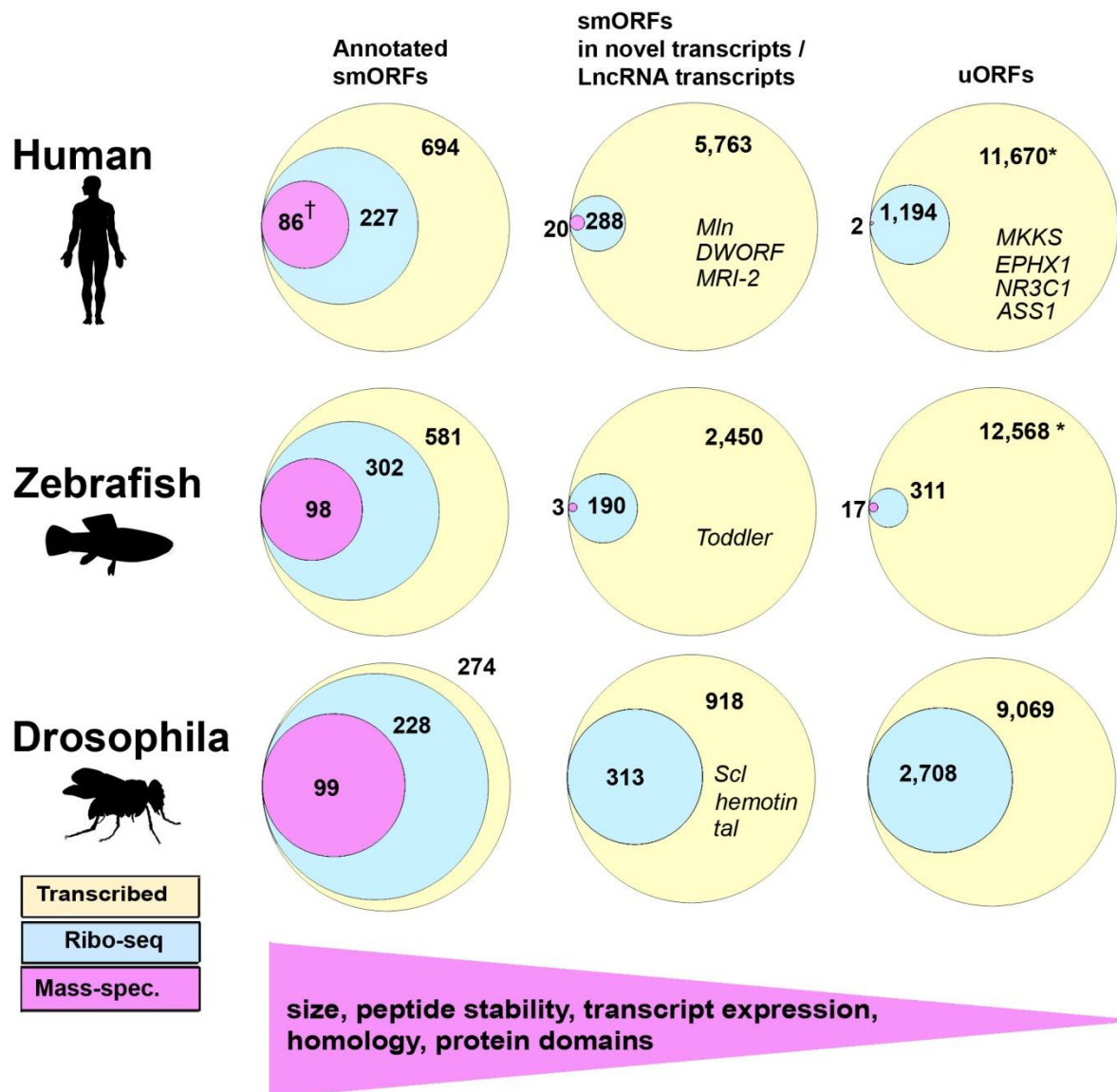


FIGURE 3

Figure 3. Different classes of smORFs detected by Ribo-Seq and HPLC-MS/MS in Humans, Zebrafish and fruit flies.

Venn diagrams representing the number of smORFs detected by Ribo-Seq (blue) or HPLC-MS/MS (Mass spec, pink), relative to the total number of transcripts encoding each class of smORF (yellow) in humans[33,42], zebrafish [31]and fruit flies [32]. In these organisms, HPLC-MS/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 [65] and for zebrafish [29]; the number of peptides identified in humans by HPLC-MS/MS (†) were obtained from Mackowiak et al.[49]. The higher detection rates of annotated smORFs by HPLC-MS/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.

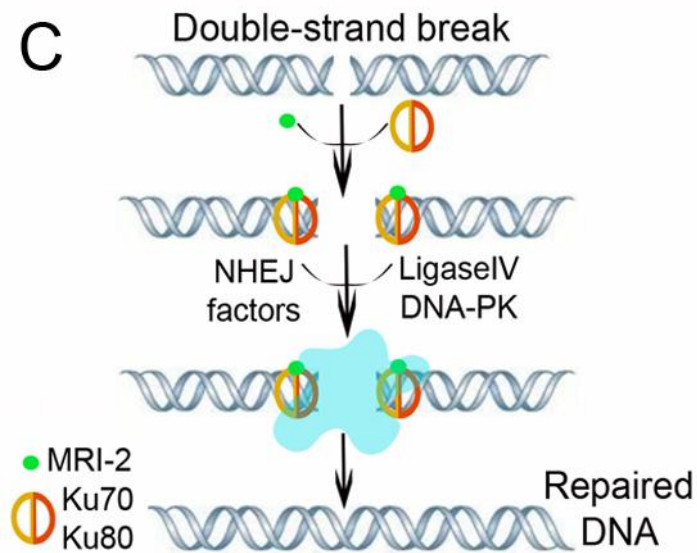
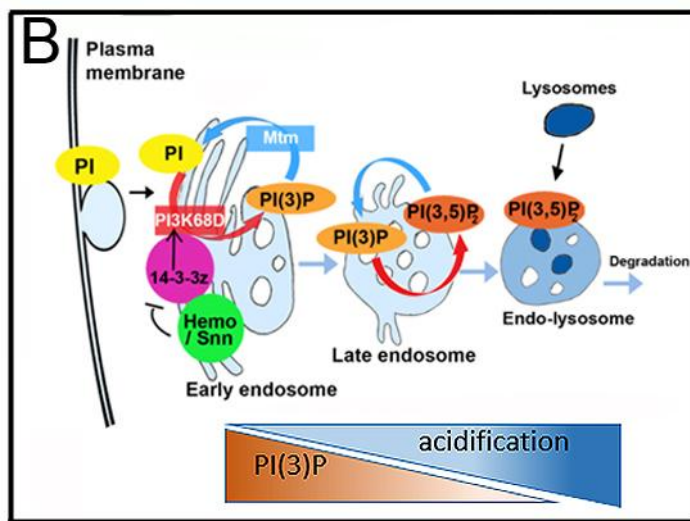
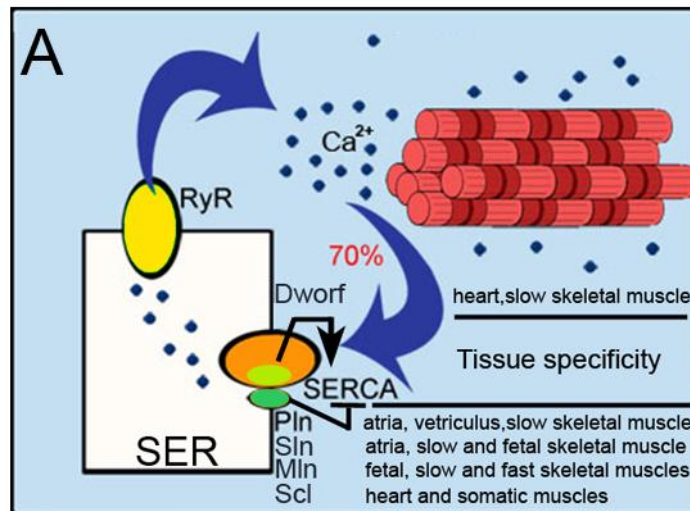


FIGURE 4

Figure 4. Cellular functions of conserved smORF micropeptides.

A- Muscle performance depends on intracellular levels of Ca^{2+} regulated by the Ryanodine receptors (RyR) and Sarco-endoplasmic reticulum (SER) Ca^{2+} ATPase (SERCA) pump. A conserved family of smORF peptides bind SERCA inhibiting its activity. Their members, Sarcolamban (Scl) in *Drosophila*, and Sarcolipin (Sln), Phospholamban (Pln) and Myoregulin (Mln) in vertebrates, display specific expression patterns. In addition, a new vertebrate smORF peptide, DWORF, activates SERCA by competitively displacing SERCA inhibitors. These peptides contribute to the smORF-based regulatory repertoire that regulates calcium dynamics and, because of their tissue-specificity, they seemingly participate in conferring different muscles with specific contractility properties [54].

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). The 88aa-Hemo/Snn peptides inhibit a 14-3-3 ζ -mediated Pi368Dkinase activation. 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.

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 DSB induction, where it binds Ku70/Ku80 heterodimers and stimulates DNA ligation through NHEJ.

Supplementary Figure 2: Number smORFs identified using computational, Ribo-seq and proteomics approaches in different organisms.

smORF searches:				
Organism:	Method/Metrics:	Identified smORFs:	Comments:	Ref.
Mouse	Hexamer sequence composition (CRITICA), and purifying selection	1,240	Computationally predicted smORFs, of which 495 lack similarity to known proteins.	18
Zebrafish embryos	RPF coverage and phasing (ORFscore)	302	Out of 581 annotated smORFs.	31
		190	Out of 2540 possible smORFs in un-annotated transcripts and LncRNAs.	
		311	uORFs	
	Custom database	98	Previously annotated smORFs.	
		6	smORFs in un-annotated transcripts and LncRNAs.	
		17	uORFs	
	micPDP	63	Out of 15,674 ORFs in transcripts without annotated CDS. (23/63 also found by ribo-seq).	
Mouse E14 mESCs cells	sORFfinder +SVM	1,100	smORFs in ncRNAs	44
		27, 371	Intergenic smORFs	
		23,127	Other smORFs (intronic, overlapping coding exons).	
	RPF Coverage, and start codon pile up with Harringtonine treated samples (using data from RF).	528	smORFs in ncRNAs (401 pass sORFfinder +SVM filter).	
		226	intergenic smORFs (89 pass sORFfinder +SVM filter)	
Human HEK293 cells	Global initiation site sequencing (GTI-seq) with LTM treated samples	288	smORFs (median size 18 aa) out of 5763 ncRNA transcripts (Refseq).	33
		6,729	uORFs in 3352 genes (mostly smORFs).	
Human	SVM (using ORF conservation in multiple alignments, PhyloCSF and phastCons scores), and overlap filters used to predict novel smORFs.	831	119 uORFs and 357 LncRNAs	49
Zebra fish		211	14 uORFs and 97 LncRNAs	
Fly		194	50 uORFs and 52 LncRNAs	
Human	predicted smORFs passing ORFscore threshold of 6 in publicly available Ribo-Seq datasets.	45	12 uORFs and 28 LncRNAs	
Zebra fish		50	4 uORFs and 28 LncRNAs	
Human	predicted smORFs detected in publicly available HPLC-MS/MS datasets.	34	1 uORF and 21 LncRNAs	
Zebra fish		2	1 uORF	
Fly		3	1 uORF and 2 LncRNAs	
Human HEK293 cells	Ribotaper: multitaper spectral analysis of RPF distributions to identify periodic footprint profiles	656	annotated genes with translated uORFs	48
		504	translated ORFs in ncRNAs: 79 LncRNAs, 114 antisense genes	
	Using HEK293 HPLC-MS/MS dataset (ref), and custom dataset built with ribotaper ORFs	12	uORFs	
		14	smORFs in ncRNAs: 8 LncRNAs, 6 antisense genes	
Human (cells / tissues)	Protease inhibition, and ELRIC fractionation Custom database	237	Novel peptides, 80% mapping to previously un-annotated transcripts and 20% to 3'UTRs, 5'UTRs, and overlapping annotated CDSs.	43
Human (cells / tissues)	16 HPLC-MS/MS datasets from peptide Atlas matched to custom database	1,256	Novel peptides within annotated transcripts, 3'UTRs, 5'UTRs, and overlapping annotated CDSs.	46

Computational
Ribo-Seq
Mass-spectrometry

BOX-1 Assessment of smORF-coding potential based on sequence and conservation methods.

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 smORFs are further tested for functionality by searching homologues and evolutionary constraints [16].

Coding Region Identification Tool Invoking Comparative Analysis (CRITICA): gene prediction algorithm that integrates 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 the PhyloCSF method is applied to assess their coding potential from codon substitutions in genome-wide multi-alignments [31].

Coding Potential Calculator (CPC): bioinformatics tool that scores six sequence features to distinguish coding vs non-coding ORFs. Three of the features 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. [81-82].

BOX-2. Evaluation of smORF-coding potential and translation by Ribosomal profiling methods.

ORFscore: translation-dependent metric that exploits the 3nt step movement of translating ribosomes across the transcript. Therefore, the Ribo-Seq reads in coding ORFs tend to show a tri-nucleotide periodicity on the frame of translation (phasing)[31]. This method requires a restricted sample of RPFs, with sizes matching the more abundant average ribosomal footprint, usually 28-29nt.

Ribosome Release Score (RRS): metric 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 [83].

Fragment length organisation similarity score (FLOSS): this method relies on the difference of RPF length distribution between coding genes and non-coding RNAs. This metric scores the coding potential of ORFs according to the similarity between their RPF length distribution, and that of known coding genes [41].

Translated ORF Classifier (TOC): Random Forest classifier that assesses the ORF-coding potential within a transcript according to 4 metrics: Translation Efficiency (ratio of the Ribo-Seq reads/RNA-Seq read within the ORF), Inside vs Outside (coverage inside ORF/coverage outside ORF; coverage: nucleotides having Ribo-Seq reads/total number of nucleotides), Fraction Length (fraction of the transcript covered by ORF) and Disengagement score (DS) (assesses the efficiency of ribosomal release after a stop codon). [29]. Pauli *et al.*[57] improved the TOC by adding a “coverage” metric.

ORF Regression Algorithm for Translational Evaluation of RPFs (ORF-RATER): this metric quantifies the translation of 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 [63].

RibORF Classifier: a Support Vector Machine classifier that defines active translation of ORFs based on the evaluation of phasing parameters obtained from canonical proteins. This method identifies 3-nt periodicity, and uniformity of footprint distribution across codons by calculating the percentage of maximum entropy values [64].

RiboTaper: Similar to ORFscore but uses a multitaper spectral analysis method to obtain 3nt periodicity from raw Ribo-Seq read data, which is typically noisy. This allows to calculate framing patterns using reads of varied lengths, provided that the P-site position is determined for each length [48].

