Classification and function of small open reading frames

Juan Pablo Couso, Pedro Patraquim

Publication date
09-06-2023

Published in
Classification and Function of small Open Reading Frames

Link to external publisher version
https://doi.org/10.1038/nrm.2017.58

Classification and function of small open-reading frames

Juan-Pablo Couso[1,2]* and Pedro Patraquim[2]

[1]Centro Andaluz de Biologia del Desarrollo, CSIC-UPO, Sevilla, Spain and [2]Brighton and Sussex Medical School, University of Sussex, Brighton, United Kingdom.

*Author for correspondence: jpcou@upo.es

Abstract

Small open-reading frames (smORFs or sORFs) of 100 codons or less are usually - if arbitrarily - excluded from canonical proteome annotations. Despite this, the genomes of a wide range of metazoans, including humans, contain hundreds of smORFs, some of which fulfil key physiological functions. Recently, ribosomal profiling has been employed to show that the transcriptome of the model organism *Drosophila melanogaster* contains thousands of smORFs of different classes actively undergoing translation which produces peptides of mostly unknown function. Here we present a comprehensive analysis of the smORF repertoire in flies, mice and humans. We propose the existence of several classes of smORFs with different functions, from inert DNA sequences to transcribed and translated cis-regulators of translation, and finally to expression of functional peptides with a propensity to act as regulators of canonical membrane-associated proteins, or as components of ancestral protein complexes in the cytoplasm. We suggest that the different smORF classes could represent steps during the evolution of novel peptide and protein sequences. Our analysis introduces a distinction between different peptide-coding classes in animal genomes, and highlights the role of *Drosophila melanogaster* as a model organism for the study of small peptide biology in the context of development, physiology and human disease.

Introduction

The encoding of genetic information in DNA is one of the great discoveries of our times, as it allowed the physical characterisation of genes, which had been previously defined as abstract units of function and inheritance. It was followed by the discovery of the expression of the encoded genetic information into a “messenger” RNA (mRNA) and its decoding into

proteins, the 'Central Dogma' of molecular biology. Hence, a shift in the concept of the gene into a physical nucleotide sequence took place. Initially, gene sequences were identified as containing Open Reading Frames (**ORFs**) potentially translatable into proteins. More recently, the full molecular complexity of genes has been exposed, culminating in the ENCODE project findings, and the updated concept of a gene to include regulatory regions and transcripts[1]. Excitingly, the 'Central Dogma' has been challenged by the discovery of a high number of genes producing mRNA-like RNAs apparently not translated into proteins, called long non-coding RNAs (**lncRNAs**). These genes and their products have, in a short time, revolutionised our understanding of gene regulation and RNA metabolism[2, 3].

There is another class of genetic elements that also challenge the understanding our genomes' coding potential: putatively functional small Open Reading Frames (**smORFs**) of 10 to 100 codons[4]. Hundreds of thousands to millions of smORF sequences are found in eukaryotic genomes[5-7], and thousands can be mapped to transcripts, in many cases, to putative lncRNAs[8, 9]. It is as if we have a genome within our genome; a hidden genome about which we know very little. smORFs have been deemed non-coding on the basis of: **a)** their short length, which defeats standard methods for computational analysis; **b)** little experimental corroboration of their function, and **c)** convenience, since their very high numbers present a challenge for annotation and curation. As a consequence, functional smORFs are often not annotated because they are not experimentally corroborated, and not corroborated because they are not annotated, a difficulty which is rarely (and serendipitously) surpassed.

As it is the case for canonical protein-coding ORFs, we rely on computational and experimental evidence to distinguish between functional and inert smORFs[4]. For computational evidence, a fundamental tool is sequence similarity, showing 1) conservation of putative coding sequence across species, and thus, across time, indicating a selective value and hence function; and 2) similarity with proteins and protein domains having an experimentally corroborated function, hence suggesting a similar function for the smORF. However, true conservation and homology of smORFs is difficult to establish due to two fundamental problems: short sequences accrue lower quantitative conservation scores (that is, the sum of scores from each amino-acid) than longer canonical proteins, whereas reciprocally, the probability for short sequences obtaining such 'low conservation score' by chance is higher. For example, BLAST penalises the identification of protein sequences below 80 amino acids, and fails below 20[6]. Thus, it is difficult to identify functional smORFs

based on computational information alone. Given their high number, and the expectation that most short ORFs in the genome are not functional (see below), arbitrary cut-offs for minimal ORF length of 50 or 100 amino acids are used in genome annotation, discarding those ORFs below these sizes that do not have clear experimental evidence of function.

Obtaining experimental evidence for smORF function is also difficult. Standard biochemical methods for protein isolation fail to detect peptides below 10kD, which escape a typical gel or filter, and can be masked by degradation peptides from larger proteins. Genetics also encounters problems, as short sequences such as smORFs offer a small target for random mutagenesis screens and other gene-discovery protocols, while the huge number of smORFs in the genome (see below) makes unpractical a systematic directed mutagenesis program. In the unlikely event of a smORF mutation being isolated, it is often assigned to adjacent canonical genes, since most smORFs are not annotated.

smORFs are finally receiving attention and breaking out of this impasse. There is a growing realisation that hundreds, if not thousands, of smORFs are translated[8, 10, 11]; and that smORF-encoded peptides (SEPs) can have important functions and be widely conserved across metazoans[12, 13]; (reviewed in [14-16]). However, the full repertoire of smORF peptide functions is still not known, nor the genomic and evolutionary roles of smORF sequences. Attempts have been made to experimentally characterise smORFs at a genomic level in yeast, bacteria and plants showing that hundreds can produce a phenotype[5, 17, 18] and refuting the classical view that smORFs are non-functional and thus irrelevant. In metazoans, anecdotal experimental evidence has also accumulated to support these conclusions, although still far from a full genomic sweep. There are SEPs[16, 19] (sometimes referred as 'micropeptides'[20, 21]) or annotated as having biological activity as antibacterial peptides[22], cell signals[23], cytoskeletal regulators[24], and other regulators of canonical proteins[13, 25]. These functions can be essential for animal life, but only a small minority (a few hundreds) of the putative smORFs in each genome have a suspected function (inferred by homology)[26], and even fewer (tens) have experimentally corroborated function[14]. These functional smORFs tend to be longer (~80 amino acids), and thus are more amenable to standard homology searches, as well as biochemical and genetic analyses. However, 90% of smORFs are much shorter (~20 amino acids) and thus not suited to these studies, yet can display sequence conservation and translation evidence similar to ‘functional’ smORFs[6, 8]. Further, even these shorter smORFs can have crucial developmental and physiological functions and homologues across vast

evolutionary distances, including humans[12, 13]. Therefore, there could be many more yet uncharacterised smORFs having biomedically relevant functions, but until now we could not identify such smORFs, nor predict what their functions may be.

Current experimental evidence shows that only about 1.2% of short ORFs in animal genomes are transcribed, and of those only about a third appear translated (*see below*). However, their numbers are such that functional smORFs could be producing tens of thousands of yet uncharacterised peptides in each animal species. Even if only a fraction of these peptides would have biological activity, we could be missing hundreds of peptides that could shed light on many aspects of biology and medicine. The challenge is then how to identify the bioactive smORFs and their peptides, or the 'beautiful needles in a haystack'[4].

Here we present an emerging scenario arising from our own new data, plus a re-analysis of previous data in the literature. Although there is evidence for pervasive usage of non-AUG start codons and translation of overlapping ORFs[14, 27, 28], we focus on the population of non-overlapping ORFs with canonical AUG start codons in the reference genomes of three metazoans (fruit flies, mice and humans). We propose that sufficient information has accrued to approach smORFs not as an undersize discard bin, but as a group of novel molecular actors with specific characteristics, evolutionary origin and biological functions at both the RNA (non-coding) and peptide (coding) levels. We present a classification of animal smORFs based on characteristics of their sequence and the structure of their RNAs and encoded peptides, a classification that interestingly provides predictions on the function of not yet fully characterised smORFs. Finally, we show evidence indicating that a) novel smORFs can randomly and continuously appear in animal genomes, and b) that different smORF classes represent steps in the evolution of new canonical proteins.

Classification of smORFs

Drosophila melanogaster translated smORFs have been identified at the genomic level using a combination of ribosomal profiling, peptide tagging and bioinformatic analysis[8]. Two main groups of translated smORFs were identified, according to their profiling metrics and bioinformatic characteristics, with one such group enriched in peptides allocated to cell membranes and organelles. This was important, because it seemed to identify smORFs with

high chances of being functional. Such a classification could direct research to more promising smORFs by linking their sequence to biochemical properties and molecular functions, greatly reducing the scope of exploratory work to be undertaken. Accordingly, we have since characterised *hemotin*, a fly smORF from the membrane-associated group, and revealed its activity in endo-phagosomes and its conservation in vertebrates[29]; and have obtained further experimental and bioinformatic data that supports a functionally-relevant smORF classification. Thus, although this preliminary classification is far from predicting smORF function with the kind of precision we enjoy with most canonical proteins, it offers heuristic value, and therefore here we refine and expand it to vertebrate smORFs. We propose the existence of at least 5 types of smORFs with distinct transcript organization, size, conservation, translation mode, amino acid usage and peptide properties. We propose that these 5 classes likely have different cellular and molecular functions, from inert DNA sequences to transcribed and translated cis-regulators of translation, and finally to expression of functional peptides with propensity to act as regulators of canonical proteins. We outline these classes next (**Table 1**), and we will then elaborate the basis for this classification, and its functional implications:

1) intergenic ORFs are the most numerous class (96% of short ORFs, some 600,000 in flies and 21.3 million in humans, see **Figure 1A**). These are stretches of DNA between an ATG and stop codon, having a median size of 22 codons (**Figure 1B**). Judging from high-throughput data, they do not appear to undergo transcription or translation and thus it is likely that most are non-functional, simple random consequences of nucleotide permutations in 'junk' or non-transcribed DNA. We suggest the term 'intergenic ORFs' to distinguish them from the transcribed, putative functional, smORFs of the other classes (Table 1).

2) uORFs (for upstream ORFs) are the second most abundant class, comprising more than 18,000 in flies and around 50.000 in humans, and thus potentially doubling the number of currently annotated coding sequences (**Figure 1A**). uORFs are smORFs found in the 5'UTRs of mRNAs encoding canonical proteins, and have a median length of 22 codons (similar to intergenic ORFs, **Figure 1B**). Close to 50% of annotated animal mRNAs contain uORFs (**Figure 2A**); see also [14, 30]), and translation of a fraction of uORFs has been repeatedly reported in all studied organisms, such as yeast, flies, zebrafish and mouse[31-34], albeit with low translational efficiency (TE) [8, 28, 35]. uORFs are regarded as regulating the translation of the downstream, canonical ORFs in their transcripts. uORFs appear lowly conserved on

average[8, 30], and their amino acid usage is clearly different from random values, yet subtly different from canonical proteins.

3) lncORFs or long non-coding RNA ORFs, are the third most abundant class (some 17,000 in flies and 172,000 in humans (**Figure 1A**). They are found in putative lncRNAs, have a median size of 24 codons (**Figure 1B**), and their translation mode is similar to uORFs: low translational efficiency, and only in a third of the lncORF assessed. Given their size, a typical lncRNA of about 3Kb could contain up to 120 lncORFs (40 per frame), and in fact, 98% of annotated lncRNAs contain at least 1 ORF in the metazoan species we have analysed, with a median of 6 smORFs per lncRNA (**Figure 2B**). Even amongst the 40 lncRNAs characterised in humans with a non-coding function[36], all contain between 1 and 15 lncORFs (not shown). lncORFs are thus typically found in polycistronic arrangements, sometimes overlapping, hindering their experimental characterization. Their amino acid usage is non-random, but different from canonical proteins. Their function is unknown at present, with considerable debate about whether lncORFs are translated, and whether such translation is productive[27, 34, 37-39]. However, several cases of RNAs initially classified as long non-coding have been shown to actually encode and translate peptides with biomedically important functions in development and physiology, and to be conserved across vast evolutionary distances[12, 13, 40].

4) short coding sequences (short CDSs) (previously called "longer smORFs"[8, 15]), have a median size of 79 codons (**Figure 1B**), and are preferentially found in functionally monocistronic transcripts with mRNA characteristics, albeit shorter and simpler in structure than canonical protein mRNAs. They appear translated as frequently and as strongly as canonical proteins[8] and appear conserved on average at the taxonomic class level. There are around 800 short CDSs in flies and 1200 in humans (**Figure 1A**) but only a fraction have been characterised functionally. The characterised examples, and the average amino-acid sequence features of the class suggest a function as regulators of canonical proteins, often involving cell membranes.

5) short isoforms are the fifth and least abundant class of smORFs (some 130 in flies and 500 in mouse, according to annotated data, **Figure 1A**), which are generated by an alternative transcript or splice form from a longer, canonical protein-coding gene. Annotated short isoforms have corroborated translation, and have a median size of 79 codons (**Figure 1B**), resembling short CDSs in size and transcript structure, although their amino acid sequences

are closer to canonical proteins, as expected (see below). Short isoforms merit separate classification and study on two bases: certainty about their origins, and potential for functions directly related to their canonical protein paralogues. Their number may be higher as their detection depends on experimental data, and very short transcript and protein isoforms can be discarded as artefactual.

Coding vs. non-coding functions of smORFs

As mentioned, due to the high number of smORFs in animal genomes, there is a pressing need to distinguish between functional and inert smORFs, a distinction that would guide the in-depth characterization of functional smORFs. Evidence of transcription and/or translation are two objective criteria for assuming function, whether coding or non-coding. We analyzed the existing data and compared the characteristics of different RNAs containing smORFs, and of the smORFs themselves.

smORF Transcription: to be or not to be

Next generation RNA sequencing (RNA-seq) obtains snapshots of entire transcriptomes, unexpectedly revealing the pervasive transcription of up to 75% of the genome in humans and flies[2, 41]. Extensive RNA-seq studies in metazoans have been carried out, especially in model organisms such as *Drosophila melanogaster* and *Mus musculus*. The repertoire of transcribed sequences may not be complete, as not all organs and cell types have been sampled, but in general most transcription must have been detected. The small population of short CDSs of around 79 codons (**Figure 1A)** are found in polyadenylated monocistronic transcripts that are often annotated as putatively coding[26, 42], even though direct experimental corroboration of their translation is lacking in most cases[8]. Their transcripts are shorter and simpler (with fewer exons) than canonical proteins (**Figure 2C-D**)**.** This could follow a trend observed in eukaryotes for fewer exons in shorter coding transcripts[43]. More surprisingly, a large proportion of transcripts detected by RNA-seq lacked a canonical "long" ORF, and have been considered long non-coding RNAs (lncRNAs)[2, 3], even though they contain lncORFs and display coding-like mRNA features, such as similar length and structure to short CDSs (**Figure 2C-D**), transcription by polymerase II, capping, polyadenylation, and accumulation in the cytoplasm[44, 45].

However, large numbers of intergenic ORFs can be found in non-transcribed regions of the genomes of flies and vertebrates (**Figure 1A**). Do they represent smORFs in uncharacterised transcripts? What is their origin and function? Their median size across species of 23 amino acids is expected by random: amongst 60 possible codons there are 3 stop codons, i.e. a 1/20 or 0.05 chance for a stop codon. Starting from an ATG codon, the length of the resulting ORF depends on its probability of encountering a STOP codon. This probability is independent at each new codon, but the accumulated probability of encountering any stop codon obviously increases with length. Thus, the following exponential decay function,

$$f(x) = \lambda e^{-\lambda x}$$

where λ (the decay rate parameter) is *0.05* and X is the length of the ORF in codons, indicates the frequency at which ORFs of each size are expected to occur by random, and generates a size distribution that fits closely that observed for intergenic ORFs (**Figure 2E**). Thus, intergenic ORFs, unlike short CDS (**Figure 2F**), appear to be randomly generated by our genomes, so it would be expected that most are not functional (just as most mutations are not advantageous).

Further data indicates that we do not have millions of genes in our genomes. Although computational gene annotation protocols are biased against smORFs[6, 46], classical estimates of gene numbers obtained by biochemical and genetic results[47] are compatible with the annotated numbers of genes obtained with said computational methods, but differ by several orders of magnitude with the high numbers of intergenic ORFs. Thus, it is likely that most of the intergenic ORF sequences are not active genes, and they should be excluded to avoid inflated estimates of functional smORFs. A starting filter to consider a smORF as 'genic', or putatively functional, must be the existence of solid transcriptional data. For example, computational and RNA-Seq evidence was used to re-examine previous computational estimates of putatively functional non-annotated smORFs in a variety of species and could only corroborate a small percentage of them[21]. This transcriptional filter does not discard that some intergenic ORFs could be transcribed or translated in tissues not yet subjected to RNA-seq, but focuses experimental and computational efforts to more likely functional targets.

Two modes of smORF translation: long non coding RNAs and short CDS mRNAs

Ribosome profiling provides quantitative and qualitative measures of translation both at the single gene and at the genomic scale. Ribosome profiling consists of next-generation

sequencing of protected mRNA fragments (or footprints) bound by translating ribosomes (stabilized with an elongation inhibitor, generally cyclohexamide), after nuclease digestion[27, 28]. Ribosome profiling offers a direct read-out of ribosome occupancy on mRNA at the single nucleotide level, and can be compared to RNA-Seq analysis of the same biological sample to provide quantitative metrics directly related to the translation rate per ORF such as translational efficiency (TE; see glossary and [48] for a review). Results in a wide variety of species show that translation occurs in a more pervasive fashion than expected, with numerous ribosome footprints detected in lncRNAs and in the untranslated regions (UTRs) of annotated transcripts. Many of these newly identified translated regions coincide with smORFs [8, 10, 27, 49]. Although non-coding functions of several lncRNAs are well established[3, 50], the functions of the vast majority are currently unknown, and it is plausible that some of these actually encode translated smORFs.

Using a variation of ribosomal profiling, smORF translation in *Drosophila melanogaster* was shown to occur in two different modes, in correlation with smORF class[8]. 220 (84%) of annotated short CDSs transcribed in a fly cell line were translated at similar frequency and efficiency (TE) as canonical proteins. Short CDS translation correlated with their mRNA abundance, and followed canonical models, with multiple ribosomes covering the ORF at regular spacing. Similarly, short mRNA isoforms are generated by an alternative RNA processing form of a longer, canonical protein-coding sequence. Although the translation of annotated short isoforms is assumed to be the default scenario, the unambiguous assignment of ribosome profiling (see glossary) reads to specific mRNA isoforms remains a difficulty. However, about 2000 uORFs and lncORFs did not follow this canonical mode of translation, but were translated at a third of the frequency and efficiency of canonical proteins[8]. These differences have also been observed in vertebrates (reviewed in [15]). Ribosomal profiling of zebrafish embryos validated the translation of 302 previously annotated smORFs, and identified 190 novel smORFs in previously uncharacterised transcripts and putative lncRNAs, as well as 311 uORFs and 93 ORFs in 3'UTRs[10]. These smORFs tended to be more than 50 amino acids long and show conservation across vertebrates, which classifies them as short CDSs. However, the authors noted the existence of a class of ORFs of less than 20 amino acids with ribosomal profiling signal in lncRNAs, which would belong to the lncORF class. Similarly, up to 50% of lncRNAs in mouse embryonic stem cells exhibited ribosome profiling signal[28], which could potentially give rise to thousands of peptides[49].

There has been a rather technical debate on whether low RiboSeq signal represents productive translation[27, 34, 37-39]. It has been suggested that some lncRNAs could be associated with the translation machinery to regulate the translation of canonical mRNAs; or that the ribosomal binding detected is incidental and non-productive; or that the footprints detected are not generated by ribosomes: in summary, that the ribosomal profiling signal in lncRNAs is noise, yielding false-positives. However, it has been shown that while there is a linear correlation between canonical mRNA levels in polysomes and their ribosomal profiling signal[8, 11, 51], such positive correlation could not be observed with lncRNAs. Many lncRNAs were present in high quantities in polysomes, yet only some produced ribosomal profiling signal[8, 11] suggesting that the ribosomal profiling signal of lncRNAs is not produced by generic background noise, but specific translation of a subset of lncRNAs, even if at a modest rate. Furthermore, lncORF translation has been corroborated by ORF tagging and proteomics[8, 10, 11, 46] (see next). Finally, it has been demonstrated that small ORFs in transcripts annotated as lncRNAs can produce bioactive peptides with important functions[12-14, 20, 25, 40, 52].

An alternative high-throughput method to detect translation is proteomics, which matches mass spectrometry signatures of digested peptides to expected protein sequences[53]. Improvements in proteomics methods have allowed the detection of SEPs, but in general proteomics lags behind ribosome profiling in smORF detection, and detection of lncORF peptides has been lacking. Even with the use of specific and new size-fractionation methods and custom libraries including non-annotated smORFs, 'peptidomic' studies only corroborated 8 new peptides in *Drosophila* brains[54], and 23 SEPs in human cell lines[46]. Peptides below 80 amino acids are preferentially targeted by proteases[55], yet have fewer amino-acids to generate two non-overlapping trypsinated peptides as currently required by standard proteomics protocols. Two ribosomal profiling studies failed to obtain parallel mass-spectrometry evidence for any peptide below 50 amino acids[8, 10], a common limitation of proteomic studies into smORF translation. These factors could explain their low detection, but it could also be possible that lncORFs only produce unstable peptides in small quantities (the latter in agreement with their ribosome profiling metrics).

Could lncORF-encoded peptides in general have a biological role, or in other words, are lncORFs functional and have a function conveyed by the peptides produced? Or are the peptides irrelevant, with the functions of these sequences conveyed by the RNA? ‘Non-

coding' lncRNAs that actually produce bioactive peptides could be a simple misclassification, such that annotated lncRNAs could actually include two sub-populations, true non-coding lncRNAs, and protein-coding short CDS mRNAs. Alternatively, some, many, or all, lncRNAs could have dual functions as coding and non-coding, as in the case of the plant pri-miR171b (which produces miRNA and peptides[56]), or the mammal Humanin and MOTS-c peptides produced by mitochondrial rRNA[57, 58], or simply produce inactive peptides that are quickly degraded. The act of lncORF translation may be the functionally relevant process, whereas the peptides themselves may convey little or no function; that is, lncORFs could have a non-coding function that involves ribosomes. A precedent for such scenario is presented by uORFs (see next).

uORFs as cis-regulators of translation in canonical genes

uORFs exploit two fail-safe features of eukaryotic translation: re-initiation and leaky scanning[31, 59]. In eukaryote translation, the small ribosomal subunit (40S) joins the mRNA at the 5' cap complex, and then scans the transcript until it encounters an AUG codon preceded by a 4nt CA-rich Kozak sequence. Upon this, the 60S ribosomal unit joins to form a complete ribosome (80S), and a Met-tRNA-Methionine complex joins to initiate translation. Once translation of this ORF is terminated at the stop codon, the ribosome dissociates, but the 40S unit can reinitiate scanning for further downstream smORFs to translate. Such re-initiation of translation can happen if the ORF just translated is no longer than 30 amino acids, and if an additional ORF is around 100-200bp downstream of its stop codon. Reciprocally, weak Kozak sequences may be not recognised occasionally, leading to 'leaky scanning' whereby ORFs may be scanned over but not translated, allowing continuing scanning and translation of downstream ORFs.

Both processes are stochastic but can facilitate the translation of polycistronic eukaryotic genes, and can act as a canonical translation regulatory mechanism[31, 32]. The classical model for uORF function is the yeast gene GCN4 (General control protein GCN4), where 4 uORFs act to repress the translation of the GCN4 protein[59, 60]. uORF translation precludes translation of the downstream GCN4 protein ORF, but under starvation conditions, the uORFs are bypassed (scanned but not translated), allowing the translation of the GCN4 protein. In this case the sequences of the peptides produced by the GCN4 uORFs are irrelevant, and their physical presence itself acts as a regulatory mechanism; in other genes, however, the nascent uORF peptides can stall the ribosomal complex upstream of the main ORFs in a sequence-

dependent manner[61]. This inhibitory cis-regulatory uORF function does not necessarily preclude that some uORF peptides could have functions in *trans*, independent of their main ORF[62]. A pure *cis*-regulatory role for uORFs fits with a) their low translation levels[8, 30, 39]; b) low sequence conservation[8, 30]; c) no propensity to form known protein domains (**Figure 4A**) and d) amino acid usage generally closest to random (**Figure 4D**). In this repressory *cis*-regulatory function, the expectation is that there should be a negative correlation between the translation of the uORF and the main ORF. Such a negative correlation at the genomic level has been found in mammals[32] and zebrafish[30] but not in *Drosophila melanogaster*[8] or yeast[35, 63].

In summary, there is evidence of translation for all types of transcribed smORFs, at different frequencies and intensities which correlate with their proposed class. i.e., their size and type of transcript-of-origin. There are also well-established smORF functions, that can be separated into *sensu stricto* 'coding' i.e. production of a bioactive peptide, and cis-regulatory 'non-coding' functions which depend on engaging ribosomes, rather than of the production of a specific peptide. In the next sections, we explore the functions of SEPs.

Molecular functions of SEPs: regulators of canonical proteins.

The small size of SEPs does not allow for the typical, multi-domain structure of canonical proteins, but rather can accommodate only one, at most two, simple domains (considering the simplest domain is a 30 amino acid-long transmembrane α-helix (TMH), and the need for an unstructured spacer region between domains)[64]. Interestingly, isolated or incomplete protein domains can display functions unrelated to those observed in their native configuration inside large multi-domain proteins[65]. For example, artificially-expressed peptides with the ANTP (Antennapedia) homeodomain or the HIV-TAT domain act as cell-penetrating peptides, a function unrelated to their native proteins[66]. It follows that even the function of smORFs containing known protein domains cannot be predicted easily; in fact, in most cases, is still unknown. A bioinformatic examination of the smORF peptide sequences, informed by examples whose functions have been characterized experimentally in detail at the molecular and cellular level, might clarify their role.

Characterized SEPs

Several smORFs producing bioactive peptides have been characterised in several metazoan species and in unicellular organisms, such that their translation has been corroborated and the peptide has been detected, their molecular and organismal function has been determined, and their conservation levels established. We have identified a group of about 60 short CDS-encoded peptides which are conserved across metazoans, from humans to flies, and in some cases even yeasts and plants[42]. The most common function of these ancient short CDSs is as positive regulators of cytoplasmic processes (**Figure 3**). These include ubiquitination (SUMO - *small ubiquitin-related modifier*, Nedd8 - *neural precursor cell expressed, developmentally down-regulated 8*)[67]; cytoskeleton dynamics (HSPC300, a.k.a. Brick1)[24]; translation (RpS21 - *40S ribosomal protein S21*, **Figure 3**A)[68]; cyclin function in mitosis (Cks85A - *Cyclin-dependent kinase subunit 85A*, **Figure 3A**)[69]. The second most common function is related to mitochondria, such as apoptosis-related Reaper and Bcl-2[70], and mitochondrial respiration (NMLDQ, UQCR10, Tim9a)[71, 72]; (**Figure 3B**; see also [16]). These peptides offer concrete examples of crucial functions conserved for hundreds of millions of years, but they constitute a minority (7-8%) of short CDSs in each species. Since most smORFs are not so widely conserved, it is unclear if these ancient smORFs offer a functional blueprint for all short CDSs, and even less so, for lncORFs and uORFs.

Other functionally and molecularly well characterised smORFs are perhaps not so ancient but still show conservation comparable to canonical proteins, and present a further repertoire of functions, this time as negative regulators. Plants contain smORFs encoding for small interfering peptides that act predominantly as transcription repressors[73, 74]. These small interfering peptides (also called 'microproteins'[75]) are shortened dominant-negative isoforms or duplications of canonical transcription factors. They are less than 100 amino acids in length and usually contain one known protein domain, and interfere with canonical transcription factors either sequestering them in unproductive dimers (i.e. if the small interfering peptide contains the dimerization domain but not the DNA-binding domain) or by competing for the DNA binding sites (i.e. if the small interfering peptide contains the DNA-binding domain but not others required for its activity)[73, 75]. There are no small interfering peptides characterized in animals, but there is no reason why they could not exist. For example, *Pgc* in *Drosophila melanogaster* (initially described as a long-noncoding RNA) represses euchromatic DNA transcription in an epigenetic manner[25]. There are, however, examples of dominant-negative isoforms in humans which are close to 100 amino acids in

length. The 119-168 amino acid-long Id (Inhibitor of DNA-binding) family of HLH-like peptides sequester basic HLH proteins into inactive complexes (**Figure 3C**), regulating various developmental processes, cell cycle and circadian rhythms from flies to humans, and have been implicated in cancer and stem cell renewal[76]. Small interfering peptides could be as prevalent in animals as in plants, but their short RNAs and peptides may have been discarded as artefactual or non-functional when experimentally detected.

Small interfering and regulatory peptides could provide a general model for smORF peptide function, since dominant-negative interference does not need to be limited to transcription factors. This model fits well with the small size of SEPs, which cannot form the large globular proteins with buried active sites characteristic of enzymes, or the large multi-domain structural proteins roles[77]. However, they could be perfect for interfering with larger proteins. Indeed, some short CDS-encoded peptides have demonstrated functions that show a negative regulatory role: in mitosis (Z600[78], **Figure 3A**); apoptosis (Humanin[79], **Figure 3B**); ubiquitination (Brd[80], Tal[52]); ER and muscle contraction **(**Scl[13]**Figure 3D**); phagocytosis (Hemotin[29], Spec2 - *CDC42 small effector 2*[81] **Figure 3E**); and as antimicrobial peptides (Defensins[82], Drosocin[83], **Figure 3E**). Interestingly, most of these functions involve cell membranes (see below).

<u>smORFs without annotated functions: protein structure and amino acid usage.</u>

In yeast and in bacteria, attempts to identify non-annotated functional smORFs at the genomic scale have identified hundreds of genes. In the baker's yeast (*S. cerevisiae*), 299 smORFs were identified with evidence of transcription, translation or sequence conservation[5]. Of these, 247 revealed a requirement for growth under different starvation and stress conditions. In *E. coli*, Hemm et al.[17] identified 217 putative smORFs by bioinformatic criteria, of which they tested experimentally 24 and confirmed 18. Of these, 10 were observed at membranes and were predicted to encode TMHs, as 65% of annotated bacterial proteins of less than 50 amino acids do[17]. We can obtain two messages from these studies in unicellular organisms: smORF requirements may not be immediately obvious; and functional SEPs that have not been characterised so far may locate to cellular membranes (which complicates their biochemical detection).

A bioinformatic analysis of short CDSs in *Drosophila* revealed a higher frequency (32%) of compatibility with TMHs than canonical proteins (25%), and unlike lncORFs or uORFs

(11%)[8]. These compatibilities are shared by vertebrate smORFs (**Figure 4A**). Trans-membrane alpha-helix compatibility could easily be related to putative functions in cell membranes and organelles. Indeed, the limited GO data available displayed an enrichment of membrane-related terms, and tagging a sample of translated longer peptides revealed a tendency for these peptides to locate to cell membranes, including mitochondria[8]. Finally, the characterisation of Hemotin[29] (**Figure 3E)**, as well as other short CDSs being characterised (E. Magny and J.I. Pueyo, pers. communication) corroborates the prediction of a membrane- and organelle-related function for TMH-carrying short CDS-encoded peptides.

Another source of information is the amino acid composition, or amino acid usage, of smORFs peptides (i.e. frequency at which each amino acid is present). We observed in *Drosophila* putatively different amino acid usage between short CDSs, canonical proteins, lncORFs, uORFs and random RNA sequences[8]. These differences could underlie different molecular functions, and be an indicator of coding potential[84]. We have studied the amino acid usage of mouse and human smORFs, intergenic ORFs, and randomised RNA sequences, and compared them to *Drosophila*. We observe a remarkable degree of similarity among fly and vertebrate smORF classes in their non-random amino acid sequence propensities (**Figure 4B-C**), in correlation with their similar size distributions (**Figure 2E-F**). Pooling the data from fly, mouse and human reveals statistically significant correlations and differences amongst smORF amino acid usage, when compared to those canonical proteins (**Figure 4C-D**). As observed for flies, metazoan canonical proteins and short CDSs resemble each other and differ significantly from randomised RNA and intergenic ORF values, yet display subtle differences amongst themselves (**Figure 4B-D**).

The amino acid usage of short CDSs is biased towards the positively-charged and against the negatively-charged amino acids (**Figure 4D**). Artificial cell-penetrating peptides have an overall positive charge and can cross plasma membranes, being of great interest to the pharmaceutical industry[85, 86]. As mentioned earlier, isolated fragments of canonical proteins can also act as cell-penetrating peptides[66], providing another function for some short isoforms. Given the prevalence of TMHs and membrane localisation of short CDS peptides, it is tempting to speculate that this charge bias similarly favours their incorporation into membranes and organelles. A favoured organelle is the mitochondria[8, 16]. Given that peptides of less than 50 amino acids can cross the mitochondria outer membrane[72], short CDS peptides could also do so.

The amino acid usage of short CDSs would also fit a role as antimicrobial peptides. Antimicrobial peptides constitute the humoral branch of the innate immune system, and production of peptides specifically tailored against the type of invading organism is regulated by a signalling mechanism conserved from flies to humans[82, 87]. Antimicrobial peptides tend to be amphipathic molecules of around 50-150 amino acids, displaying both positively charged and hydrophobic regions. These regions confer on them solubility and the ability to bind and integrate into microbial membranes, respectively; they also display a propensity to form TMHs[88, 89]. These characteristics, which are identical to those we find for short CDSs (**Figure 4A,D**), are sufficient to design artificial antimicrobial peptides that act even more efficiently than natural ones[88]. Antimicrobial peptides can form pores in the microbial membranes, leading to cell leakage and death[82], but also can behave as cell-penetrating peptides that once inside the cells, interfere with vital cellular processes[82, 83, 89]. In this regard, antimicrobial peptides could be seen as 'negative regulators' too [89]. Since the overall molecular characteristics of antimicrobial peptides are identical to those of short CDSs, some of the hundreds of short CDS-derived peptides with currently unknown function might work in this way. Indeed, several well-characterised antimicrobial peptides are encoded by short CDSs[22] (**Figure 3F**).

Another possible function of positively-charged peptides is nucleic acid binding. DNA and RNA are negatively charged, and transcription factors and other DNA-binding proteins such as histones act through positively-charged domains. It is therefore interesting that smORF isoforms in plants have been revealed as DNA-binding regulators of transcription, but animal isoforms match the amino acid usage of canonical proteins, not showing a positive charge bias. Their only significantly different amino acid is Met, whose higher frequency is attributable to their shorter length (not shown). However, SEPs with possible DNA-binding activities have been described, such as the Human MRI-2, a regulator of Ku protein in genome stability[90]; and interestingly, those produced by the dual-function (coding-non coding) RNA pri-miR171b[56].

The amino acid usage of lncORFs and uORFs are intriguing (**Figure 4D**). uORFs resemble short CDSs by favouring positively-charged amino acids while avoiding negatively-charged ones. Overall uORF propensities are third closest to canonicals after short isoforms and short CDSs (**Figure 4C**); uORFs are highly correlated with all three, and yet, their highest

correlation is with lncORFs (Pearson coefficient = 0.94; see also **Figure 4C**). If uORF-derived peptides have no function, their 'coding-like' amino acid usage might be an irrelevant consequence of their location near canonical ORFs. It is possible that uORFs are derived 'nonsense' fragments of nearby canonical ORFs. A rather interesting alternative possibility is that uORF peptide sequences reflect a specific and yet undiscovered function[62].

Regarding lncORFs, their amino acid usage resembles short CDSs, showing a higher proportion of sulfidic amino acids (Met and Cys) and lower of negatively charged Asp and Glu. It is not clear that these propensities would confer an overall positive or amphipatic nature to lncORFs-derived peptides, and thus we cannot speculate on a putative membrane function for lncORFs. However, the Scl-family peptides, previously annotated as lncRNAs, function in the ER, while Tal peptides influence adjacent cells[23, 91], implying diffusion across membranes. (**Figure 3D**) Translated lncORF peptides can locate to mitochondria and other organelles[8], and human mitochondrial rRNA produce SEPs such as Humanin (**Figure 3B**), a generic inhibitor of cell death of biomedical importance[57], and MOTS-c, which can act outside the mitochondria[58], implying the crossing of membranes. However, high degradation (as could be endured by the short lncORF peptides, see above) hinders cell-penetrating peptide function[92]. Basically, there are too few examples of lncORFs characterized, and more functional studies of lncORFs are needed. As in the case of uORFs, lncORFs might be a case of mixed identities, with some coding (i.e. being in reality short CDSs) and others not, or else represent a group of sequences poised for coding function but not yet doing so. We explore this possibility next.

Genomic function of smORFs in *de novo* gene birth

There are indications that, as a whole, smORFs have a general function at a higher, genomic level: as a source of new protein-coding genes. We have seen that there exist different classes of smORFs with different functions at the molecular and cellular levels. Despite the marked difference in average lengths, amino acid usage, translation efficiency, protein structure and conservation, these classes have a small degree of overlap, suggesting a continuum with evolutionary flow between classes (**Figure 5**). We next examine the two possibilities for the

generation of smORFs: from existing coding sequences (i.e. from canonical proteins), or from previously non-coding sequences (*de novo*).

smORFs can emerge as fragments of longer protein-coding genes through alternative RNA processing, intron-retention or premature stop codons. Short protein isoforms can be generated by alternative transcription, splicing and polyadenylation (or a combination of the three) from canonical proteins. It appears that higher eukaryotes have 'leaky' splicing mechanisms[93], and this can lead to the production of smORF isoforms with proper mRNA and translation features. We have seen that short isoforms can produce dominant-negative peptides, which could have deleterious consequences, in a similar manner to the 36-43 amino acids amyloid-beta peptides that form plaques in Alzheimer's disease[94]. However, if the deleterious consequences are small, late-onset (past reproductive age), or pleitropically-linked to positive traits[95], even such 'deleterious isoforms' could be temporarily carried by our genomes. A short isoform could acquire a positive fitness effect, and eventually become a duplicated gene in the genome following a gene duplication or retrotransposition event. After further evolution and divergence, this duplicated isoform could become a pseudogene or, alternatively a new short CDS. (**Figure 5**). The similar sizes (**Figure 2F)** and amino acid usage (**Figure 4**) of short CDSs and short isoforms could indicate that short CDSs have been generated in this way, especially those displaying protein domains or clear homology to canonical proteins. Alternatively, such 'paralogue' short CDSs could also arise directly from canonical ORFs. Splice junction mutations or intron retention can introduce intronic sequences into a protein, but introns contain stop codons either by chance, or by selection presumably to stop the production of abnormal long proteins[93]. This would lead to the production of proteins with shorter and new C-termini, and long 3' trailers. In this scenario, evolution would be expected to optimise smORF-producing transcripts, if the peptide produced would be advantageous. Altogether, 'paralogue' short CDSs emerging from canonical proteins could be part of canonical protein evolution, an opposing mechanism to processes that usually inactivate proteins and result in the formation of pseudogenes (**Figure 5D**).

Other short CDSs such as *hemotin*[29, 96] or *toddler* (a.k.a. *elabela*)[97, 98] seem to have no paralogues in the genome, so we cannot assign their origin from a pre-existing canonical protein-coding gene. Where do these 'singletons' come from? Short CDSs could evolve from shorter lncORFs and uORFs, by mechanisms favouring 'ORF extension'. In principle,

extension of any ORF only necessitates changing the stop codon to another amino acid, the most likely outcome (95%) in the event of stop codon mutation; and an outcome observed in the *Tal* genes[12] (Figure 5B). Stop codon read-through, an event readily detected *in vivo* by ribosome profiling[99], offers an alternative, or an intermediate step, to stop codon mutation. In either case, such ORF elongations will be again subjected to the exponential function $\lambda e^{-\lambda x}$ (see above), moderating elongation to 25 amino acid-long steps. Alternatively, N-terminal elongation (as observed in the Pln peptides of the Scl family, **Figure 5C**) could also occur, although this necessitates a new in-frame ATG in an appropriate Kozak context. Either way, if such elongated peptides preserve their original function, they could be positively selected due to the higher stability that comes with increased length[55]. In time, the elongated ends of the peptide would provide new material for selection to improve the peptide function, or add new functions.

However, how do lncORFs and uORFs arise? The size distribution of lncORFs and uORFs fits closely the exponential random distribution of intergenic ORFs (**Figure 2F**), suggesting that lncORFs and uORFs also appear randomly in the genome (**Figure 5D**). They may appear as intergenic ORFs that then become transcribed (ORF first), or in non-coding RNA regions with 'space' available (RNA first). The possible *de novo* generation of proteins from previously non-coding sequences is increasingly debated[100, 101]. Up to 1% of protein-coding genes could be species-specific (without homologues) and hence of recent origin[102-105], but this notion is controversial[106] and it critically depends on computational ability to detect homologues. The mechanism for the emergence and spread of such *de novo* genes has not been clarified, although a role for lncORFs has been proposed[9, 107]. Most lncORFs and *de novo* genes seem devoid of distant homologues, appearing and disappearing in the genome of species in the same Order[9, 104, 107, 108] suggesting both recent origins and dynamic evolutionary behaviour. However, the *scl* and *tal* families have been conserved through hundreds of millions of years[12, 13] (**Figure 5B-C)**, showing that lncORFs can become short CDSs fixed in the genome, perhaps in correlation with increased peptide coding potential[9, 21]. Further observations are compatible with the evolution of lncORF and uORFs towards short CDSs and full canonical coding content. First, short CDSs have an intermediate evolutionary age, being more conserved than uORF and lncORFs, but less than canonical ORFs (**Figure 5A**). Second, smORF amino acid usage, reveals a progression from intergenic ORFs to lncORF and uORFs, and from these to short CDSs, and then to short isoforms and canonical proteins (**Figure 4C-D**). Finally, although the size of lncORFs and uORF suggest an origin at random

in intergenic or non-coding RNA sequences (**Figure 3D**), their amino acid sequences do not reflect such origin but resemble coding ORFs (**Figure 4B, D**). The conclusion is that lncORFs and uORFs may appear at random, but that once appeared, their nucleotide sequences are subjected to selection. Whether this selection is initially due to a coding or non-coding function needs to be ascertained; however, it can end up producing amino acid sequences with full peptide function[12, 13, 20, 40, 52].

In summary, it is possible that, while some smORFs emerged from canonical proteins, others could emerge from non-coding sequences. Either way, the evolutionary processes that act on canonical protein evolution (duplication, neo-functionalization[109]) should also act on smORFs and give rise to smORF 'families' offering yet more raw materials for smORF evolution. Indeed, short CDSs and short isoforms can duplicate both inside their transcript and in gene families [12, 13, 75, 76], (**Figure 5B-C**) and can be generally assorted according to their sequence similarity (unp. obs.). Finally, smORFs could not only appear from or grow into new canonical proteins, but could also be attached to existing canonical proteins by exon-shuffling thus providing a source of new protein domains[110]. In our view, smORFs could represent a *genomic protein factory*, using both new (lncORFs, intergenic) and recycled (canonical protein) materials, constantly bubbling out putative new peptides and protein domains.

Conclusions and future perspectives

The study of smORFs and the identification of functional smORFs has been hampered by bioinformatic and experimental limitations. Overcoming these limitations and increasing the pool of experimentally characterised smORFs is the foremost challenge in the field. CRISPR gene editing should herald a new phase of faster progress, by allowing targeted manipulation of individual smORFs. This is especially important in the case of lncORFs, uORFs and small isoforms, in which specific ORFs within the transcript or gene (in the latter example) have to be mutated separately. Nonetheless, our analysis of the accumulated data suggests some emerging principles of smORF classification and function.

smORFs across animals can be classified according to sequence length and transcript structure. These features correlate with other characteristics, such as their evolutionary conservation and amino acid usage (**Table 1**). Further, these smORF classes seem to display preferred cellular and molecular functions, facilitating more detailed studies and an

understanding of the genomic role of smORFs as a whole. Non-transcribed intergenic ORFs likely have no function; uORFs act as cis-regulators of translation of downstream canonical proteins; lncORFs can give rise to novel bioactive peptides; short CDSs produce peptide regulators of canonical proteins in the cytoplasm and membranes (**Figure 3**); short isoforms can produce peptides interfering with homologous transcription factors (**Figure 3C**). Altogether, smORFs may generate new protein sequences during evolution, with the different smORF classes representing steps in the evolution of proteins from inert intergenic sequences (**Figure 5**).

Understanding the origin, evolution and function of smORFs would be needed to clarify this crucial, and so far, underappreciated, function of the genome; a far more dynamic and living genome than we currently contemplate. Additionally, considering the current interest in artificial peptides as new pharmacological agents as new drugs, delivery vectors, and antimicrobial peptides, smORFs could provide an unexplored reservoir of peptides either with such functions, (such as short CDS) or currently inactive but naturally primed for such functions by virtue of their amino acid composition (as uORF and lncORF peptides). The conservation of individual smORFs and classes across animals that we show here allows a model-system-based experimental approach to these fascinating research topics.

Figure Legends

Table 1 – Properties of smORFs in Fruit Flies and Mammals. Animal genomes contain millions of open reading frames (ORFs) between an ATG (start) and a stop codon. Most of these are found in untranscribed regions (intergenic ORFs, top row shaded pale yellow) and are deemed non-functional. Genomes also contain canonical ORFs of 101 codons or more (bottom row, green) which are translated and produce annotated proteins with well-known functions. In between these two extremes, our genomes also contain transcribed and putatively functional short ORFs of 10 to 100 codons (smORFs) which can be divided into different classes, according to their transcript type: lncORFs - ORFs present in long noncoding RNAs (grey); upstream ORFs, or uORFs, which occur in the 5'UTRs of canonical mRNAs (pink); short CDSs - annotated ORFs of 100 codons or less present in short mRNAs (red); small isoform ORFs of 100 codons or less generated by alternative splicing of canonical mRNAs (shaded blue). We extracted all AUG-STOP ORFs from both the annotated transcriptomes and the non-transcribed regions of *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*, and divided them into these classes. We find that the distinction between smORF classes correlates with differential biochemical markers: size (indicated as the median number of amino acids per ORF), average rate of translation, average taxonomic level of ORF conservation, and features of the encoded amino acid sequence (TMHs indicates prevalence of trans-membrane alpha-helices). We propose that these characteristics correlate with a favoured function for each smORF class and are conserved in Flies, Mice and Humans. Data from this work unless indicated. Short isoform translation and conservation includes their long isoforms. See text for details and references.

Figure 1 – Conservation of smORF numbers and lengths across animals. A) smORF classes are colour-coded as in Table 1. Circles are proportional to number of smORFs. The number of annotated, short CDSs (red) is similar in all animals analysed, and low when compared to both annotated canonical ORFs (green), and transcribed, non-annotated smORFs (uORFs -pink- and lncORFs, -grey-). The higher number of uORFs and lncORFs in mammals is related to their higher number of mRNAs and lncRNAs, whereas the number of intergenic, non-transcribed ORFs is related to the genome size in each species. **B)** The lengths of different ORF classes are conserved across Metazoans. ORF classes can be subdivided into three classes according to their median size: intergenic ORFs, uORFs and lncORFs all have median lengths of 22-23 codons; short CDSs and short isoforms have a

median length of 79 codons; finally, canonical ORFs show a median length of around 450 codons ("F" – Fruit Fly; "M" – Mouse; "H" – Human).

Figure 2 – Transcribed smORFs: RNA characteristics. Pooled fly, mouse and human data.

A) Number of predicted uORFs per annotated mRNA. Mean and standard error to the mean (SEM) plotted in purple and pink, respectively. About 50% of the animal mRNAs analysed contain no uORFs in their 5'UTRs, while half of mRNAs contain 1 or more uORFs. **B)** Number of predicted lncORFs per annotated lncRNA. Mean and standard error to the mean (SEM) plotted in dark grey and light grey, respectively. Most annotated lncRNAs contain one or more smORF (median = 6). **C)** Short CDS mRNAs display low transcript complexity, as measured by the number of different exons in annotated transcripts of each class. lncRNAs display a similar pattern, being less complex than shorter isoforms, while canonical mRNAs are on average more complex than all other classes. **D)** Metazoan RNA transcript length according to ORF classes. Coloured areas indicate the mean frequency of each RNA size, plus its standard error. mRNAs containing short CDSs and lncRNAs are similar in size being, on average, 400 bp long, although the lncRNAs class includes few larger transcripts. Short isoform transcripts are on average 600 bp long, while canonical ORF mRNAs are on average larger than all transcript classes above. **E)** uORFs and lncORF size distributions in animal genomes. Mean relative frequency of each ORF size (coloured lines) and standard error (coloured area). uORFs and lncORFs display a similar size-distribution to that of intergenic ORFs. Interestingly, their distributions fit a exponential decay distribution $f(x) = \lambda e^{-\lambda x}$ where λ (the decay rate parameter) is 0.05 (dotted curve; r^2 fit above 0.9 in all cases), suggesting that intergenic ORFs, uORFs and lncORFs are randomly and unavoidably generated by animal genomes (see text). Median=23 of the exponential decay distribution for sizes 10 to 100 codons is indicated. **F)** short CDS and short isoform size distribution in three animal genomes. Median=79 (short CDSs). Short CDS and short isoform smORFs have similar size distributions that are different from the exponential distribution of other smORFs (dotted line and panel B).

Figure 3 – Function of SEPs. Ancestral smORF peptides conserved across eukaryotes (fungi, insects and vertebrates) and acting as positive regulators of canonical proteins are

represented by blue elipses. More evolutionarily recent smORF peptides act as negative regulators and are represented by red elipses. **A)** Ancestral cytoplasmic smORF peptides promote the activity of canonical protein complexes. Small ribosomal proteins such as RpS21 - *40S ribosomal protein S21* (83 amino acids), a structural constituent of the small ribosomal subunit (40S, green), aid in the translation of RNAs by ribosomes, while Cks85A (Cyclin-dependent kinase subunit 85A, 96 amino acids) interacts with Cdk1 and Cdk2 (green), and promotes cell cycle progression. Its activity is opposed by the negative regulator Z600, which represses Cdk1. **B)** Ancestral smORF peptides produced in the cytoplasm can promote mitochondrial processes. Bcl peptides (left) allocate to the mitochondrial outer membrane to promote mitochondrial outer membrane permeabilization -MOMP-, and the subsequent release of apoptotic signals. Other smORF peptides (such as UQCR10, or Ubiquinol-cytochrome-c reductase complex III subunit 10, bottom right), are involved in electron transport at the inner mitochondrial membrane. The evolutionarily recent humanin peptide is produced by a mitochondrial rRNA and represses Bcl function. **C)** Small nuclear interfering peptides act as dominant negative repressors of transcription factors. The Id peptides contain an HLH domain, but lack a basic DNA-binding domain. Ids bind HLH transcription-factor proteins, sequestering them in inactive complexes that cannot bind DNA. **D**) SclA (Sarcolamban A) is a 28 amino acid peptide encoded by a smORF in *Drosophila melanogaster*[13]. Similarly to its human orthologue (PLN), SclA regulates calcium uptake in the sarcoplasmic reticulum (SR) of muscle cells through repression of SERCA (green), the Ca^{2+} pump which transfers Ca^{2+} from the cytosol to the ER to terminate muscle contraction[13]. In the absence of Scl or Pln, both fly and human hearts develop arrythmias. **E)** Negative smORF regulators repress the activity of canonical proteins in cell membranes and organelles. Hemotin[29] represses the Pi3K68D (PI3K, green) activator 14-3-3z (yellow), slowing down endosomal maturation to allow phagocytic digestion, whereas Spec2 represses Cdc42 to moderate the formation of phagocytic pockets. **F)** Antimicrobial peptides penetrate plasma membranes to attack invading microorganisms, either creating membrane pores, as Defensin (left), leading to leakage of cytoplasmic contents, or binding and interfering with cytoplasmic proteins, as Drosocin, which represses DnaK to slow bacterial metabolic rate.

Figure 4 – Coding features of smORFs. A) Number of putative trans-membrane alpha-helix domains (predicted by TMHMM 2.0) encoded per 100 amino acids in each ORF class. Short CDSs are significantly enriched in TMHs with respect to canonical ORFs (*p* values for *t*-tests indicated), whereas uORFs are significantly depleted. **B)** amino acid frequencies, or

propensities, in smORFs and canonical ORFs normalized to expected frequencies in a randomised transcriptome[8, 84]. smORFs of all classes, like canonical ORFs, exhibit non-random amino acid composition. S: sulphidic amino acids; +: positively charged amino acids; – : negatively-charged; P, H: polar and hydrophobic. **C)** Similarity of smORF amino acid propensities with canonical proteins (left axis, green graph) and randomised RNA (right axis, black graph) quantified as correlation coefficients. Short CDSs and short isoforms are most similar to canonical proteins and most different from random sequences; uORFs and lncORFs display an intermediate amino acid composition between canonical ORFs and intergenic and random values. **D)** smORF amino acid propensities, normalized to canonical propensities. Short CDS amino acid frequencies differ significantly from canonical ORFs (multiple t-tests with Bonferroni corrections, $p<0.05$) by encoding more sulphidic amino acids, as well an enrichment in more positively-charged and fewer negatively charged amino acids. uORFs and lncORFs resemble each other and display a pattern related to short CDSs, but with more extreme variations. Amino acids with frequencies significantly different from canonical are indicated.

Figure 5 – Evolution of smORFs: a dynamic continuum?

A) Stepwise model for smORF evolution within a dynamic continuum. It has been proposed that coding genes can emerge from non-functional sequences ('protogenes'[100]). We propose that the different classes of smORFs (Table 1) represent different steps in this process. Intergenic ORFs (red boxes) appear and disappear at random (**Figure 2E**) in non-transcribed DNA but can become part of a transcription unit under the control of Pol II over evolutionary time, giving rise to lncORFs. lncORFs (and uORFs) can also appear randomly in transcribed sequences; either way, they have been shown to be translated at low frequency and efficiency. The main functional outcome of this low translational profile may not be to produce bioactive peptides, but it provides the cell with a reservoir of lowly-translated peptides that could integrate, fuse, grow and be selected for function (as the Scl and Tal families, see B-C). These coding lncORFs could increase in TE, giving rise to short CDSs. In turn, short CDSs could integrate, fuse and grow into canonical proteins. This mechanism would increase the number of genes through evolution, and although such an increase can be appreciated (**Figure 1**) creation of new canonical ORFs is counter-balanced by conversion of canonical ORFs into shorter isoforms, pseudogenes, and perhaps lncRNAs[111], as well as the random disappearance of intergenic ORFs by transcription and ATG codon loss. **B)** The Scl family of lncORFs show loose amino acids sequence conservation, but their structure and

molecular function is conserved[13, 20]. The family includes two fly peptides (only SclA shown) and four mammal peptides. They all repress SERCA activity (see **Figure 3D**), except DWORF, which on the contrary promotes it by acting as a competitive inhibitor of the other Scl-family SEPs (double negative)[40]. Both Myoregulin (MRLN) and Phospholamban (PLN) show N-terminal extension of their sequences. **C)** Evolution of *tal* smORFs. *tal* (*tarsal-less*) is a *Drosophila melanogaster* polycistronic protein-coding gene, which had been initially annotated as a lncRNA but contains three 11 and 12 amino acid-long short CDSs (Tal-1A, Tal-2A, and Tal-3A). These short CDSs include a conserved functional heptapeptide LDPTGXY, indicating an origin through tandem duplications, corroborated by their phylogenetic tree[12]. Tal-AA (32 amino acids) contains two heptapeptides separated by a 17 amino acid-long sequence containing degenerate STOP and START codons (red and green, respectively), consistent with elongation and fusion of two smORFs within the same transcriptional unit by STOP codon loss. **D)** Evolutionary conservation of canonical ORFs (green) and short CDSs (red), indicating the proportion of ORFs conserved at each taxonomic level (data extracted from Ensembl and Flybase). Solid graph lines indicate the average, and coloured areas the SEM, for flies, mouse and human data. 50% of Canonical ORFs are conserved across the Animal Kingdom (see intersection of black and green dotted lines), while short CDSs display a 50% conservation at the Class level (intersection of red and black dotted lines). Note, however, that some short CDSs (±60, see text and **Figure 3**) are conserved across the Eukaryotic Domain. Conservation in benchmarking species was used to indicate conservation at each taxonomic. level. ***Drosophila melanogaster ORFs:*** kingdom if conserved in Mouse or Human; phylum if in *Daphnia pulex or Ixodes scapularis;* class if in *Tribolium castaneum;* order if in *Anopheles gambiae.* ***Mus musculus ORFs:*** kingdom conservation if in *Drosophila melanogaster*; phylum, *Danio rerio;* class, *Homo sapiens;* order, *Rattus norvegicus.* ***Homo sapiens ORFs:*** kingdom, *Drosophila*; phylum, *Danio rerio; class, Mus musculus*; order, *Pan troglodytes.* For ***all ORFs,*** conservation in *Saccaromycies cerevisiae* indicated conservation across the eukaryotic domain.

References

1. Gerstein, M.B. et al. What is a gene, post-ENCODE? History and updated definition. *Genome Res* **17**, 669-81 (2007).

2. Derrien, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* **22**, 1775-1789 (2012).
3. Guttman, M. & Rinn, J.L. Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339-46 (2012).
4. Basrai, M.A., Hieter, P. & Boeke, J.D. Small Open Reading Frames: Beautiful Needles in the Haystack. *Genome Research* **7**, 768-771 (1997).
5. Kastenmayer, J.P. et al. Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Res.* **16**, 365-373 (2006).
6. Ladoukakis, E., Pereira, V., Magny, E.G., Eyre-Walker, A. & Couso, J.P. Hundreds of putatively functional small open reading frames in Drosophila. *Genome Biol* **12**, R118 (2011).
7. Frith, M.C. et al. The abundance of short proteins in the mammalian proteome. *Plos Genetics* **2**, 515-528 (2006).
8. Aspden, J.L. et al. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq *eLife* **3**, e03528 (2014).
9. Ruiz-Orera, J., Messeguer, X., Subirana, J.A. & Alba, M.M. Long non-coding RNAs as a source of new peptides. *Elife* **3**, e03523 (2014).
10. Bazzini, A.A. et al. in EMBO J 981-993 (2014).
11. Smith, J.E. et al. Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell Rep* **7**, 1858-66 (2014).
12. Galindo, M.I., Pueyo, J.I., Fouix, S., Bishop, S.A. & Couso, J.P. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biology* **5**, 1052-1062 (2007).
13. Magny, E.G. et al. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* **341**, 1116-20 (2013).
14. Andrews, S.J. & Rothnagel, J.A. Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* **15**, 193–204 (2014).
15. Pueyo, J.I., Magny, E.G. & Couso, J.P. New peptides under the s(ORF)ace of the genome. *Trends in Biochemical Sciences* **41**, 665-678 (2016).
16. Saghatelian, A. & Couso, J.P. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat Chem Biol* **11**, 909-16 (2015).
17. Hemm, M.R., Paul, B.J., Schneider, T.D., Storz, G. & Rudd, K.E. Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol* **70**, 1487-501 (2008).
18. Hanada, K. et al. Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc Natl Acad Sci U S A* **110**, 2395-400 (2013).
19. Ma, J. et al. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res* **13**, 1757-65 (2014).
20. Anderson, Douglas M. et al. A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* **160**, 595-606 (2015).
21. Mackowiak, S.D. et al. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol* **16**, 179 (2015).
22. Lemaitre, B., Reichhart, J.M. & Hoffmann, J.A. Drosophila host defense: differential induction of antimicrobial peptide genes after infection by various classes of microorganisms. *Proc Natl Acad Sci U S A* **94**, 14614-9 (1997).

23. Pueyo, J.I. & Couso, J.P. The 11-aminoacid long Tarsal-less peptides trigger a cell signal in Drosophila leg development. *Developmental Biology* **324**, 192-201 (2008).
24. Djakovic, S., Dyachok, J., Burke, M., Frank, M.J. & Smith, L.G. BRICK1/HSPC300 functions with SCAR and the ARP2/3 complex to regulate epidermal cell shape in Arabidopsis. *Development* **133**, 1091-1100 (2006).
25. Hanyu-Nakamura, K., Sonobe-Nojima, H., Tanigawa, A., Lasko, P. & Nakamura, A. Drosophila Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. *Nature* **451**, 730-3 (2008).
26. Consortium., T.F. The FlyBase Database of the Drosophila Genome Projects and community literature [http://flybase.net/]. *Nucleic Acids Research* **27**, 85-88 (1999).
27. Ingolia, N.T. et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**, 1365-79 (2014).
28. Ingolia, N.T., Lareau, L.F. & Weissman, J.S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of Mammalian proteomes. *Cell* **147**, 789-802 (2011).
29. Pueyo, J.I. et al. Hemotin, a regulator of phagocytosis encoded by a small ORF and conserved across metazoans. *PloS Biology* **14**, e1002395 (2016).
30. Johnstone, T.G., Bazzini, A.A. & Giraldez, A.J. Upstream ORFs are prevalent translational repressors in vertebrates. *Embo J* **35**, 706-23 (2016).
31. Wang, X.Q. & Rothnagel, J.A. 5'-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. *Nucleic Acids Res* **32**, 1382-91 (2004).
32. Calvo, S.E., Pagliarini, D.J. & Mootha, V.K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A* **106**, 7507-12 (2009).
33. Fritsch, C. et al. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res* **22**, 2208-18 (2012).
34. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* **4** (2015).
35. Duncan, C.D. & Mata, J. The translational landscape of fission-yeast meiosis and sporulation. *Nat Struct Mol Biol* **21**, 641-7 (2014).
36. Pegueroles, C. & Gabaldon, T. Secondary structure impacts patterns of selection in human lncRNAs. *BMC biology* **14**, 60 (2016).
37. Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S. & Lander, E.S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**, 240-51 (2013).
38. Banfai, B. et al. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res* **22**, 1646-57 (2012).
39. Chew, G.L. et al. Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* **140**, 2828-34 (2013).
40. Nelson, B.R. et al. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **351**, 271-5 (2016).
41. Li, Z. et al. Detection of intergenic non-coding RNAs expressed in the main developmental stages in Drosophila melanogaster. *Nucleic Acids Res* **37**, 4308-14 (2009).
42. Yates, A. et al. Ensembl 2016. *Nucleic Acids Research* **44**, D710-D716 (2016).

43. Rubin, G.M. et al. Comparative genomics of the eukaryotes. *Science* **287**, 2204-2215 (2000).
44. van Heesch, S. et al. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol* **15**, R6 (2014).
45. Wang, H., Wang, Y., Xie, S., Liu, Y. & Xie, Z. Global and cell-type specific properties of lincRNAs with ribosome occupancy. *Nucleic Acids Research* (2016).
46. Slavoff, S.A. et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **9**, 59-64 (2013).
47. Miklos, G.L.G. & Rubin, G.M. The Role of the Genome Project in Determining Gene Function: Insights from Model Organisms. *Cell* **86**, 521-529 (1996).
48. Mumtaz, M.A. & Couso, J.P. Ribosomal profiling adds new coding sequences to the proteome. *Biochem Soc Trans* **43**, 1271-6 (2015).
49. Crappe, J. et al. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics* **14**, 648 (2013).
50. Fatica, A. & Bozzoni, I. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* **15**, 7-21 (2013).
51. Kronja, I. et al. Widespread changes in the posttranscriptional landscape at the Drosophila oocyte-to-embryo transition. *Cell Rep* **7**, 1495-508 (2014).
52. Zanet, J. et al. Pri sORF peptides induce selective proteasome-mediated protein processing. *Science* **349**, 1356-8 (2015).
53. Kessler, M.M. et al. Systematic Discovery of New Genes in the Saccharomyces cerevisiae Genome. *Genome Res.* **13**, 264-271 (2003).
54. Baggerman, G., Cerstiaens, A., De Loof, A. & Schoofs, L. Peptidomics of the larval Drosophila melanogaster central nervous system. *J Biol Chem* **277**, 40368-74 (2002).
55. Loose, C.R., Langer, R.S. & Stephanopoulos, G.N. Optimization of protein fusion partner length for maximizing in vitro translation of peptides. *Biotechnol Prog* **23**, 444-51 (2007).
56. Lauressergues, D. et al. Primary transcripts of microRNAs encode regulatory peptides. *Nature* **520**, 90-3 (2015).
57. Paharkova, V., Alvarez, G., Nakamura, H., Cohen, P. & Lee, K.W. 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 (2015).
58. Lee, C. et al. The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. *Cell Metab* **21**, 443-54 (2015).
59. Kozak, M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361**, 13-37 (2005).
60. Szamecz, B. et al. eIF3a cooperates with sequences 5' of uORF1 to promote resumption of scanning by post-termination ribosomes for reinitiation on GCN4 mRNA. *Genes Dev* **22**, 2414-25 (2008).
61. Ebina, I. et al. 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 (2015).
62. Combier, J.P., de Billy, F., Gamas, P., Niebel, A. & Rivas, S. Trans-regulation of the expression of the transcription factor MtHAP2-1 by a uORF controls root nodule development. *Genes & Development* **22**, 1549-1559 (2008).

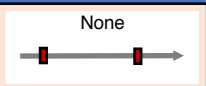
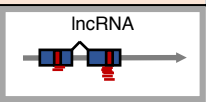
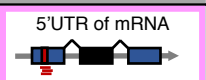
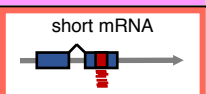
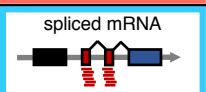
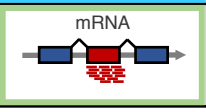
63. Zhang, Z. & Dietrich, F. Identification and characterization of upstream open reading frames (uORF) in the 5' untranslated regions (UTR) of genes in Saccharomyces cerevisiae. *Current Genetics* **48**, 77-87 (2005).
64. Abrusan, G. Integration of new genes into cellular networks, and their structural maturation. *Genetics* **195**, 1407-17 (2013).
65. Kelley, L.A. & Sternberg, M.J. Partial protein domains: evolutionary insights and bioinformatics challenges. *Genome Biol* **16**, 100 (2015).
66. Joliot, A. & Prochiantz, A. Transduction peptides: from technology to physiology. *Nature Cell Biology* **6**, 189-196 (2004).
67. Schulman, B.A. & Harper, J.W. Ubiquitin-like protein activation by E1 enzymes: the apex for downstream signalling pathways. *Nature reviews. Molecular cell biology* **10**, 319-331 (2009).
68. Alonso, J. & Santaren, J.F. Characterization of the Drosophila melanogaster ribosomal proteome. *J Proteome Res* **5**, 2025-32 (2006).
69. Ghiglione, C., Perrimon, N. & Perkins, L.A. Quantitative variations in the level of MAPK activity control patterning of the embryonic termini in Drosophila. *J Dev Biol* **205**, 181-193 (1999).
70. Vaux, D.L. & Korsmeyer, S.J. Cell death in development. *Cell* **96**, 245-254 (1999).
71. Itoh, K., Nakamura, K., Iijima, M. & Sesaki, H. Mitochondrial dynamics in neurodegeneration. *Trends Cell Biol* **23**, 64-71 (2012).
72. Michelakis, E.D. Mitochondrial Medicine. *Circulation* **117**, 2431-2434 (2008).
73. Seo, P.J., Hong, S.Y., Kim, S.G. & Park, C.M. Competitive inhibition of transcription factors by small interfering peptides. *Trends Plant Sci* **16**, 541-9 (2011).
74. Graeff, M. et al. MicroProtein-Mediated Recruitment of CONSTANS into a TOPLESS Trimeric Complex Represses Flowering in Arabidopsis. *PLoS Genet* **12**, e1005959 (2016).
75. Staudt, A.C. & Wenkel, S. Regulation of protein function by 'microProteins'. *EMBO Rep* **12**, 35-42 (2011).
76. Ling, F., Kang, B. & Sun, X.-H. Id Proteins: Small Molecules, Mighty Regulators. *Current Topics in Developmental Biology* **Volume 110**, 189-216 (2014).
77. Au, Y. The muscle ultrastructure: a structural perspective of the sarcomere. *Cell Mol Life Sci* **61**, 3016-33 (2004).
78. Gawliński, P. et al. The *Drosophila* mitotic inhibitor Frühstart specifically binds to the hydrophobic patch of cyclins. *EMBO reports* **8**, 490-496 (2007).
79. Guo, B. et al. Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature* **423**, 456-461 (2003).
80. Weinmaster, G. & Fischer, Janice A. Notch Ligand Ubiquitylation: What Is It Good For? *Developmental Cell* **21**, 134-144 (2011).
81. Ching, K.H., Kisailus, A.E. & Burbelo, P.D. Biochemical characterization of distinct regions of SPEC molecules and their role in phagocytosis. *Experimental Cell Research* **313**, 10-21 (2007).
82. Zasloff, M. Antimicrobial peptides of multicellular organisms. *Nature* **415**, 389-395 (2002).
83. Brogden, K.A. Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nat Rev Micro* **3**, 238-250 (2005).

84. D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C. & Bernardi, G. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J Mol Evol* **32**, 504-10 (1991).
85. Hansen, M., Kilk, K. & Langel, U. Predicting cell-penetrating peptides. *Advanced Drug Delivery Reviews*, 1-8 (2007).
86. Jones, S.W. et al. Characterisation of cell-penetrating peptide-mediated peptide delivery. *British Journal of Pharmacology* **145**, 1093-1102 (2005).
87. Hoffmann, J.A., Kafatos, F.C., Janeway, C.A. & Ezekowitz, R.A. Phylogenetic perspectives in innate immunity. *Science* **284**, 1313-8 (1999).
88. Fan, L. et al. DRAMP: a comprehensive data repository of antimicrobial peptides. *Sci Rep* **6**, 24482 (2016).
89. Wenzel, M. et al. Small cationic antimicrobial peptides delocalize peripheral membrane proteins. *Proc Natl Acad Sci U S A* **111**, E1409-18 (2014).
90. Slavoff, S.A., Heo, J., Budnik, B.A., Hanakahi, L.A. & Saghatelian, A. A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J Biol Chem* **289**, 10950-7 (2014).
91. Pueyo, J.I. & Couso, J.P. Tarsal-less peptides control Notch signalling through the Shavenbaby transcription factor. *Dev Biol* **355**, 183-93 (2011).
92. Palm, C., Jayamanne, M., Kjellander, M. & Hallbrink, M. Peptide degradation is a critical determinant for cell-penetrating peptide uptake. *Biochimica Et Biophysica Acta-Biomembranes* **1768**, 1769-1776 (2007).
93. Jaillon, O. et al. Translational control of intron splicing in eukaryotes. *Nature* **451**, 359-362 (2008).
94. Pimplikar, S.W. Reassessing the amyloid cascade hypothesis of Alzheimer's disease. *Int J Biochem Cell Biol* **41**, 1261-8 (2009).
95. Waxman, D. & Peck, J.R. Pleiotropy and the preservation of perfection. *Science* **279**, 1210-3 (1998).
96. Billingsley, M.L. et al. Functional and structural properties of stannin: roles in cellular growth, selective toxicity, and mitochondrial responses to injury. *J Cell Biochem* **98**, 243-50 (2006).
97. Pauli, A. et al. Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* **343**, 1248636 (2014).
98. Chng, S.C., Ho, L., Tian, J. & Reversade, B. ELABELA: a hormone essential for heart development signals via the apelin receptor. *Dev Cell* **27**, 672-80 (2013).
99. Dunn, J.G., Foo, C.K., Belletier, N.G., Gavis, E.R. & Weissman, J.S. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife* **2**, e01179 (2013).
100. Carvunis, A.R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370-4 (2012).
101. McLysaght, A. & Guerzoni, D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci* **370**, 20140332 (2015).
102. Zhao, L., Saelao, P., Jones, C.D. & Begun, D.J. Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. *Science* (2014).
103. Zhou, Q. et al. On the origin of new genes in Drosophila. *Genome Research* **18**, 1446-55 (2008).
104. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14**, 117 (2013).

105. Reinhardt, J.A. et al. De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. *PLoS Genet* **9**, e1003860 (2013).
106. Moyers, B.A. & Zhang, J. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. *Mol Biol Evol* **33**, 1245-56 (2016).
107. Xie, C. et al. Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding RNAs. *PLoS Genet* **8**, e1002942 (2012).
108. Schlotterer, C. Genes from scratch--the evolutionary fate of de novo genes. *Trends Genet* **31**, 215-9 (2015).
109. Sommer, R.J. The future of evo-devo: model systems and evolutionary theory. *Nat Rev Genet* **10**, 416-22 (2009).
110. Yang, S. & Bourne, P.E. The Evolutionary History of Protein Domains Viewed by Species Phylogeny. *PLoS One* **4**, e8378 (2009).
111. Milligan, M.J. et al. Global Intersection of Long Non-Coding RNAs with Processed and Unprocessed Pseudogenes in the Human Genome. *Front Genet* **7**, 26 (2016).

Table 1 - Properties of smORFs in Fruit Flies and Mammals.

Transcribed smORFs

ORF Class	RNA type	AA Size	Translation[13]	Conservation	Coding Features	Function
intergenic ORFs		22	None	None[6]	Random AA	None
lncORFs		24	Low	None[3]	Non-Random AA No Domains	Non-coding / Coding
uORFs		22	Low	Order[22]	Non-Random AA No Domains	Translational regulation
short CDSs		79	High	Class	Positive charge TMHs	Coding Regulators of canonical proteins
short isoforms		79	High	Kingdom	Canonical AA Protein Domain loss	Coding small interfering peptides
Canonical		491	High	Kingdom	Canonical AA Multiple protein Domains	Coding[Yates 2016] Structural, enzymatic, regulatory

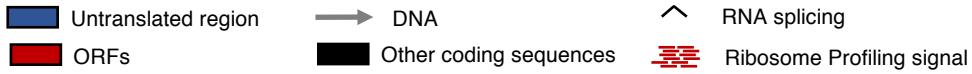
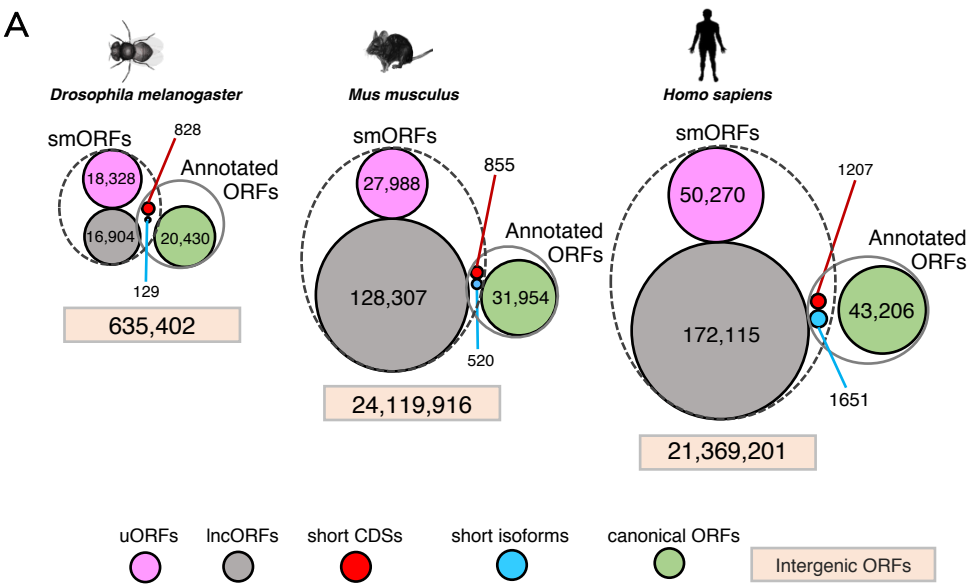


Figure 1 – Conservation of smORF numbers and lengths across animals.



B

Transcribed

ORF Class	intergenic ORFs			lncORFs			uORFs			short CDSs			short isoforms			Canonical		
	F	M	H	F	M	H	F	M	H	F	M	H	F	M	H	F	M	H
Median Length (codons)	23	22	22	25	23	24	20	22	23	79	81	78	82	79	75	490	424	421

Predicted

Annotated

Figure 2 – Transcribed smORFs: RNA characteristics

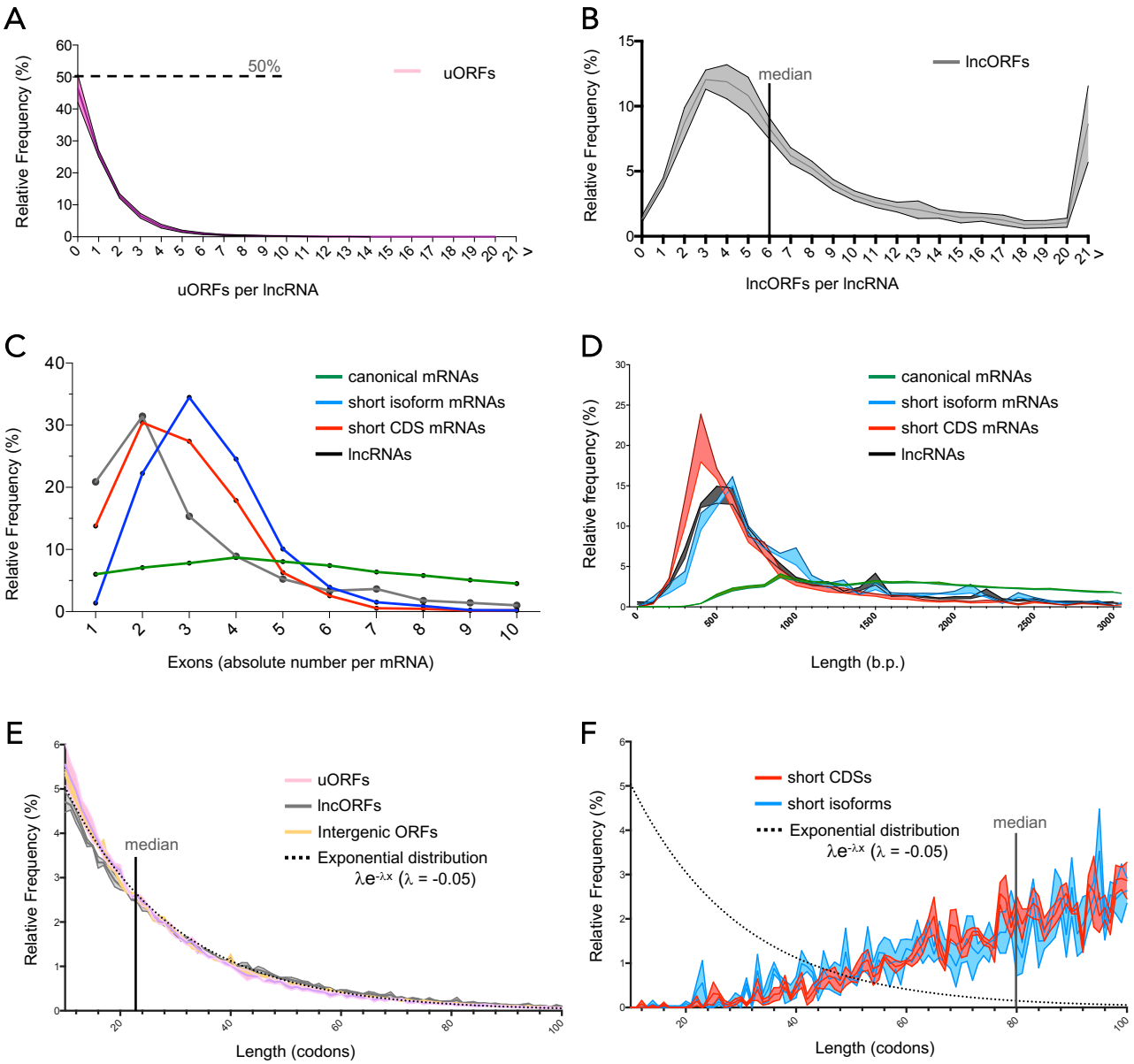


Figure 3 – Function of SEPs.

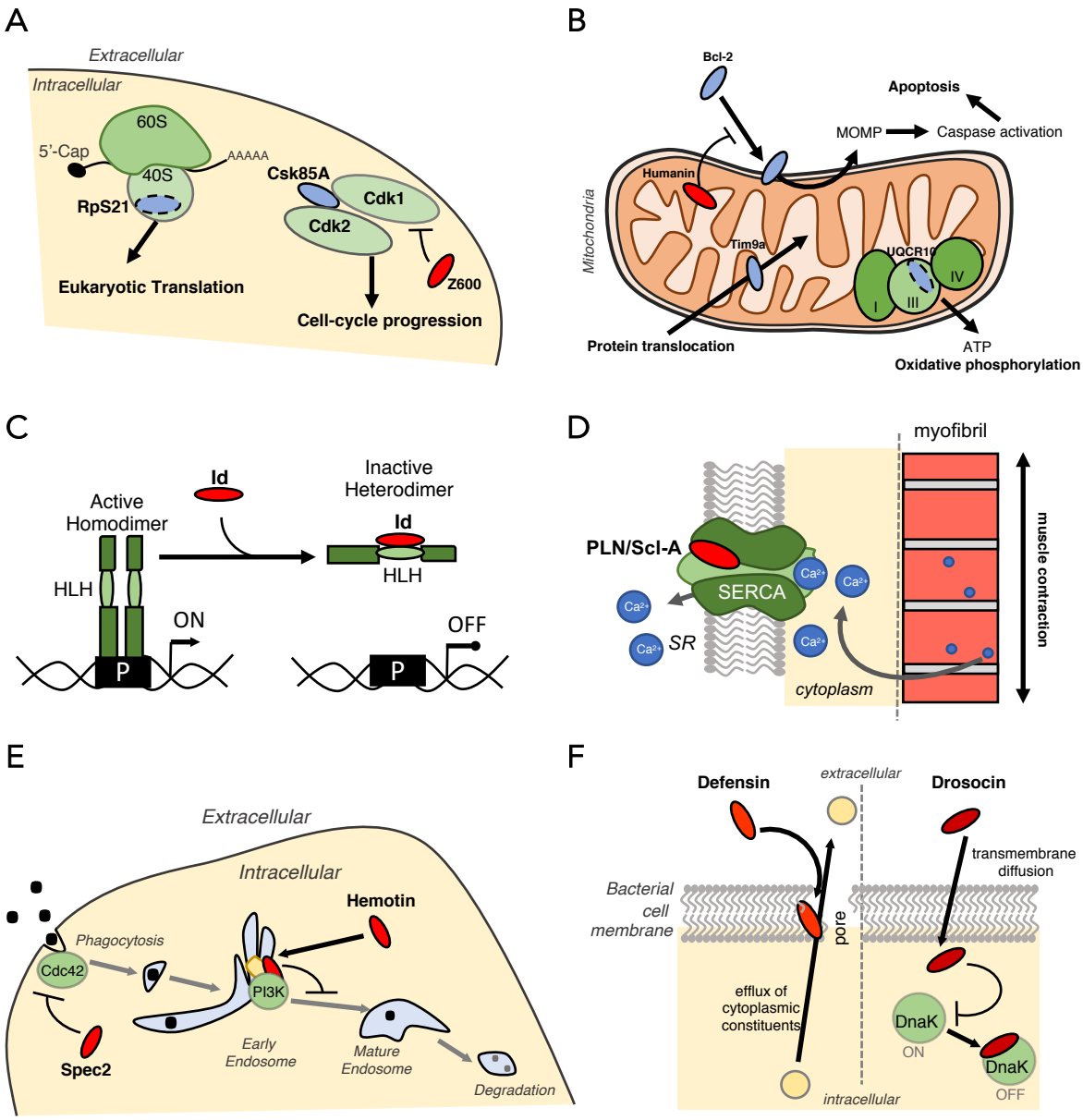


Figure 4 – Coding features of smORFs

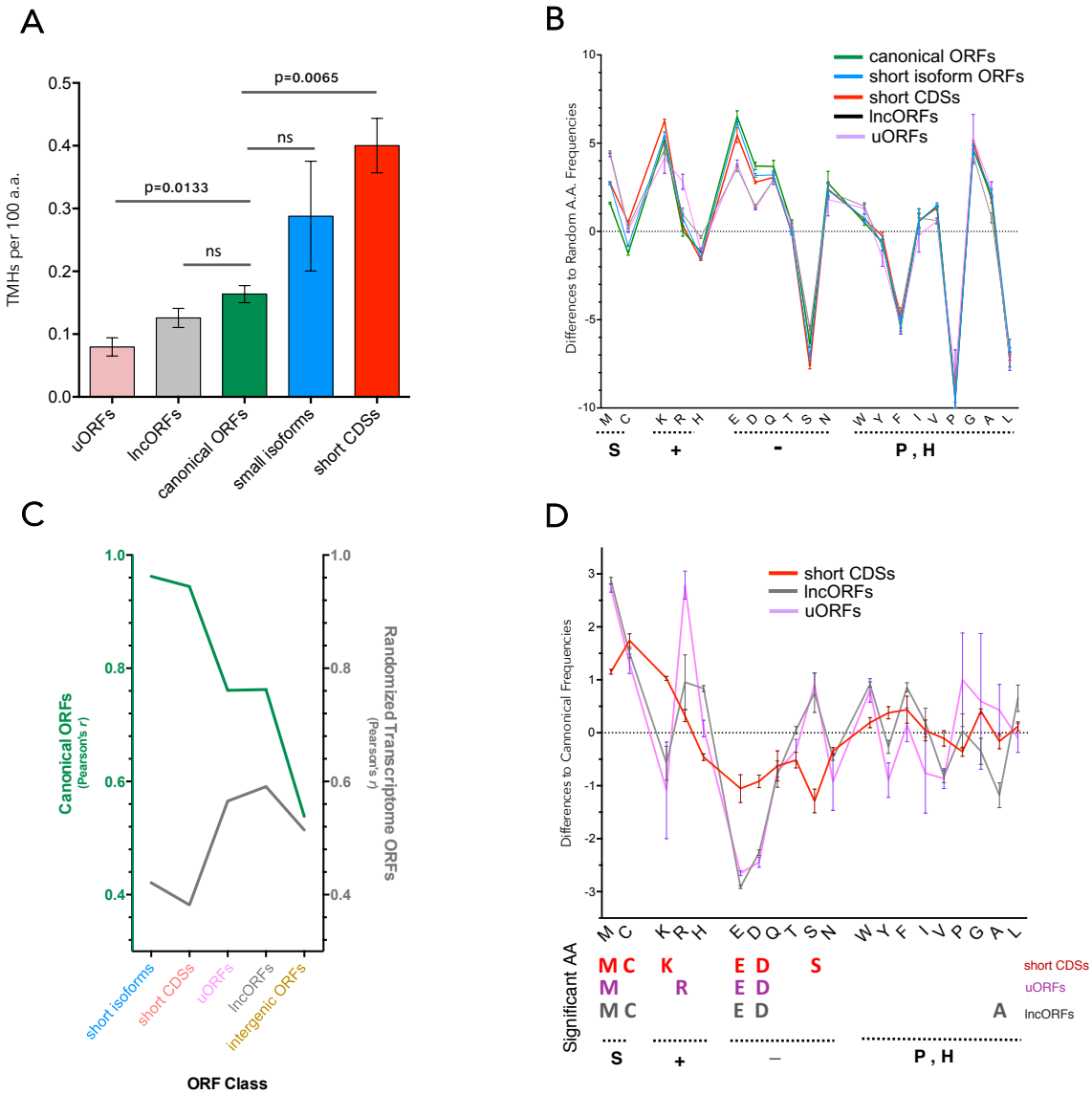


Figure 5 – Evolution of smORFs: a dynamic continuum?