Published in final edited form as:

Nat Chem Biol. 2015 December ; 11(12): 909–916. doi:10.1038/nchembio.1964.

Discovery and Characterization of smORF-Encoded Bioactive Polypeptides

Alan Saghatelian[1] and **Juan Pablo Couso**[2]

Alan Saghatelian: asaghatelian@salk.edu; Juan Pablo Couso: j.p.couso@sussex.ac.uk

[1]Clayton Foundation Laboratories for Peptide Biology, Helmsley Center for Genomic Medicine, Salk Institute for Biological Studies, San Diego, CA 92037

[2]School of Life Sciences, University of Sussex, Falmer, Brighton, BN1 6PU, UK

Abstract

Analysis of genomes, transcriptomes, and proteomes reveals the existence of hundreds to thousands of translated, yet non-annotated short open reading frames (small ORFs or smORFs). The discovery of smORFs, and their protein products, smORF-encoded polypeptides (SEPs), reveals a fundamental gap in our knowledge of protein-coding genes. Different studies have identified central roles for smORFs in metabolism, apoptosis, and development. The discovery of these bioactive SEPs emphasizes the functional potential of this unexplored class of biomolecules. Here, we provide an overview of this emerging field and highlight the opportunities for chemical biology to answer fundamental questions about these novel genes. Such studies will provide new insights into the protein-coding potential of genomes and identify functional genes with roles in biology and disease.

Introduction

Peptides and small proteins are an important class of molecules with essential roles in biology[1–3]. For instance, the discovery and use of the peptide hormone insulin to treat diabetes is one of the great accomplishments of 20th century research[4,5]. The body contains a myriad of other endogenous peptides and small proteins that regulate sleep (orexins)[6,7], stress (CRF)[8], metabolism (leptin)[9], and more[2]. Furthermore, molecules that activate or inhibit receptors for these hormones[10–13] or control the levels of endogenous hormones[14,15] have successfully been translated into novel therapeutics.

Peptides are typically defined as greater than two but fewer than 50 amino acids (aa), while any peptide larger than 50 aa is considered a protein, and Eukaryotes have a median protein length of 361 aa. Until recently, most known peptides and small proteins were known to arise from the processing of longer precursors (see below). However, in genomes there exist hundreds of thousands to millions of short Open Reading Frames of less than 100 codons, potentially able to be translated into peptides and small proteins. The name smORF (for small ORF) was introduced to identify those short ORFs of less than 100 codons that are

Competing financial interests. We have no conflicts of interest.

actually translated[16], and here we use the term smORF-encoded polypeptide (SEP) to mean a protein product of less than 100 aa arising from a smORF. We will focus on SEPs identified as bioactive using the same criteria that were used for peptide hormones: activity in biochemical, cellular, or physiological experiments. In cells or in vivo, we are primarily interested in loss of function experiments, which indicate biological relevance.

The search for new bioactive peptides and small proteins has led to the discovery of hundreds to thousands of previously non-annotated smORFs in genomes from various kingdoms (animals, plants, bacteria)[17–27]. The remarkable finding of so many translated smORFs indicate that functional smORF-encoding genes comprise at least 5–10% of genomes. And some of these smORFs have already been shown to have fundamental biological activities mediated by the encoded peptides[28–32]. Undoubtedly, many more smORFs producing bioactive SEPs are bound to be identified. Classical bioactive peptides, neuropeptides and peptide hormones, and SEPs differ in specific ways (Fig. 1a). Classical bioactive peptides are produced from proteolysis of longer polypeptides called prepropeptides (Fig. 1a). For example, the 29-amino acid glucagon peptide is generated by proteolysis of preproglucagon, which is 180aa long[33]. The additional sequence in the prepropeptide contains a signal sequence that directs these peptides through the secretory pathway, where they undergo proteolysis, before eventual release from the cell.

Bioactive SEPs, on the other hand, are produced directly from ribosomal translation of smORFs (Fig. 1a), not from proteolysis of a precursor longer than 100aa. This does not exclude that some SEPs might be post-translationally modified and act upon neighbouring cells[30,34], but their initial translation as short products poses significant challenges for the detection of SEPs and the identification of their encoding smORFs, as we will see below. These difficulties have precluded the systematic characterisation of smORFs and SEPs and stimulated the ongoing development of a field focused on their study.

At a deeper level, smORFs challenge our current understanding of the coding and information content of genomes. Genes were conceptually defined by genetics as units of function and inheritance[35]. Next, molecular genetics established that the genetic information is encoded in DNA, then expressed into peptides and proteins *via* RNA. Genome sequencing allowed the physical characterization of genomes and completed the re-definition of protein-coding genes as DNA sequences containing Open Reading Frames (ORFs) potentially translatable into proteins. And today, we understand that other genes produce functional non-coding RNAs, such as microRNAs and long-non-coding-RNAs.

Thus, genome annotations have indicated gene numbers, that although initially surprising, are not out of kilter with estimates from genetics: tens of thousands of genes in animal genomes, potentially encoding up to 100,000 protein variants in humans and other mammals (Ensembl, August 2015). However, these annotations have excluded millions of short ORFs found in the genomic DNA[18,22,26]. Do genomes contain millions, or more, genes? If not, how do we identify which smORFs are functional genes, actually producing bioactive peptides?

With such a large set of putative smORFs and SEPs, chemical biology is bound to have a significant role in such identification, and in ascertaining the molecular biology of the newly identified SEPs.

Preliminary Evidence for the existence of smORFs

During the assignment of protein-coding status to ORFs, several parameters are included to reduce false positives[18,22]. These parameters include the requirement for an ATG start codon, a minimum length of a 100 codons for the ORF, and the prediction of a single ORF per transcript[22,36,37]. The choice of a 100 codons cut-off was made to distinguish *bona fide* protein-coding ORFs from the numerous[38] random in-frame arrangements of start and stop codons in genomes[18,21,22]. Due to this criteria, a histogram analysis of the number of ORFs versus ORF length has a predictable cliff at 100 codons[18]. Computational and experimental analysis of genomes that remove of this length criteria indicate that there are many more potential protein-coding ORFs, many of which are smORFs[17,18,22,25,26,38–40].

Empirical data in support of smORFs emerged from early studies into protein translation. Some mRNAs contain multiple open reading frames, with a short ORF present in the 5′-UTR of a much longer downstream ORF. These short ORFs were named upstream ORFs or uORFs[41–44]. Intially, uORFs were not considered to be protein-coding, but were thought to be cis-acting elements that mediate ribosomal scanning to regulate the translation of longer downstream ORFs[45,46]. The deletion of uORFs resulted in increased translation of longer downstream ORFs[46] to support this hypothesis (Fig. 1c). More recent studies have revealed that at least some uORFs are translated (i.e. they are smORFs) and translation is necessary to regulate downstream ORF expression.

Several mechanisms of uORF regulation of downstream translation are suspected[45]. Thousands of uORFs are translated in mouse stem cells[47] and in flies[17]. The translation of the uORFs causes the ribosomes to slow down[48] (Fig. 1c). The net effect is decreased translation of the downstream ORF. Extensive work has demonstrated a cis-regulatory function for the uORFs in the yeast gene GCN4. Translation of the uORFs facilitates the translation of the GNC4 gene, but this role does not require the uORF peptides[49]. In this case the process of making the polypeptide (i.e. translation) is important and the peptide products do not appear to participate in the regulation.

There is some evidence that in some cases the uORF peptide sequence is important. The mRNA for the mammalian gene *Chop*, for example, contains a 31-codon uORF that reduces CHOP protein translation under basal conditions[44]. Mutations to uORF that change its amino acid sequence no longer inhibit CHOP translation and this dependence on the uORF amino acid sequence demonstrate not only that this uORF is also being translated[44], but also that the uORF peptide itself is involved in the regulation of CHOP. The hypothesis is that the nascent 31-amino acid uORF peptide interacts with the peptide exit tunnel on the ribosome to pause or disassociate the ribosome from the mRNA, and perturbations to this sequence inhibit this function. Other mechanisms for uORF regulation of downstream genes has also been reported, including uORF peptide inhibition of translation[50].

uORFs are prevalent. A recent analysis of mouse and human genomes revealed that nearly 50% of all genes contain uORFs in their 5′-UTRs[46], and deletion of these uORFs amplified downstream ORF translation. This data supports a general function for uORFs as cis-acting translational regulators of downstream ORFs. While uORFs provided early evidence of smORF translation, real interest in discovering how many smORFs are in the genome emerged after the discovery of smORFs with functions outside of translation regulation. The discovery of a 36-bp smORF that encodes an 11-amino acid peptide that controls fly development [refs. 28, 29] indicated that smORFs influence fundamental biology, and catalyzed the development of new strategies to discover smORFs in various genomes.

Systematic smORF and SEP discovery

Knowing how many smORFs and SEPs are present in the genome and proteome, respectively, is of fundamental interest. There have been several approaches taken to systematically annotate smORFs and SEPs in the genome. These methods have all led to the identification of additional smORFs and SEPs.

Computational methods

The computational annotation of smORFs has been challenging because it is difficult to distinguish smORFs from chance in-frame start and stop codons. Moreover, some smORFs[24], and ORFs in general[47], have been reported to use non-ATG start codons, which makes these assignments even more difficult. Nevertheless, several reports have attempted to computationally annotate smORFs[17,18,20–22,25,26,38–40]. In mammals, new algorithms that removed the length dependence of ORFs and identified approximately 3000 candidate smORFs transcribed in mammalian genomes[18]. This study and others indicate that genomes may contain as many several thousand non-annotated smORFs.

In flies, a combination of short ORF prediction and conservation was used to identify novel smORFs[22]. Comparison of non-coding regions for conservation between *Drosophila melanogaster* and *Drosophila pseudoobscura* identified many new smORFs of less than a 100 codons which had conserved sequences and in-frame start and stop codons in both species. These regions were then analyzed to ensure that the smORF RNAs are transcribed and that the nucleotide substitutions were indicative of translated proteins[51]. This led to the identification of at least 401 conserved smORFs, a 3% increase the coding potential of the fly genome. A less conservative estimate suggests that the upper limit might be closer to ~4,500 smORF-coding genes, which highlights the potential biology mediated by smORFs. This library of potential smORFs was used to subsequently identify the conserved sarcolamban peptides (see below). More generally, this approach provides a reliable outline for approaching smORF discovery. A similar approach has been recently applied in other animals, leading to the identification of 800 conserved putative smORFs in humans[40].

Proteomics methods

Proteomics enables the detection of the translation product to reveal protein-coding smORFs. In a typical proteomics experiment, mass spectrometry data is searched against a database of annotated genes to identify proteins[52]. Many translated smORFs are not

annotated and, therefore, a different strategy was required. An early example of this utilized RefSeq mRNAs that were translated in all six possible reading frames (3 forward and 3 reverse to account for antisense RNAs) to generate a list of all possible proteins encoded by the RefSeq database[23]. Analysis of proteomics data using this database identified several new ORFs, including four ORFs under 150 codons. One of these four ORFs contained less than a 100 codons and was a smORF[23].

Modern sequencing methods offer a way to improve this approach by creating proteomics databases from RNA-Seq data, which presumably includes all of possible protein-producing RNAs. This field is commonly referred to as proteogenomics to indicate the integration of two different types of –omics datasets[53,54]. For example, many additional SEPs were identified in K562 cells by creating a custom proteomics database from RNA-Seq[24].

In this approach, the proteome is enriched to isolate lower molecular weight peptides and small proteins[24] prior to proteomics analysis. The proteomics data is then searched against the RNA-Seq-derived custom proteomics database (Fig. 2a). Annotated proteins from this search are removed, and the remaining non-annotated proteins are manually curated to validate that the proteins are indeed SEPs. This analysis revealed 86 novel human smORFs[24], the largest number reported at the time.

The potential for the number of human coding sORFs was expanded through an approach that predicted alternate open reading frames (AltORFs) in the human genome[26], and then used these predicted AltORFs to generate a protein database for subsequent analysis[25]. The average putative AltORF protein is 57 amino acids long versus 344 for the reference database, indicating that most missed ORFs are smORFs. Analysis of human cell lines, lung, ovary, CSF, urine, plasma, and serum revealed many new smORFs[25].

Ribosomal profiling and other genomics methods

Application of ribosome profiling has provided an overview of the protein-coding potential of entire transcriptomes[17,39]. These studies had several key findings regarding global protein translation. This includes the prodigious use of non-ATG initiation codons, as well as the identification of polycistronic genes, uORFs, and overlapping ORFs. Moreover, the mouse studies observed changes in translation as cells undergo differentiation[47], suggesting that uORFs serve a broad regulatory role in gene expression.

In flies, a modified ribosome profiling method called Poly-Ribo-Seq was used to experimentally identify smORFs in the fly genome[17] (Fig. 2b). Poly-Ribo-Seq enriches polysomes, which are more likely to be actively translating RNA into protein. These experiments began by validating the method using the small polysome fraction enriched translated smORFs in the S2 fly cell line. This analysis identified a total of 228 smORFs, a four-fold increase from the validated proteome, and they used proteomics to identify 60 smORF products, 40 of which are novel[17]. Additional analysis of these data identified hundreds of additional smORFs within putative long non-coding RNAs and in 5′-UTRs (uORFs). In total, this approach led to the confident assignment of ~700 smORFs in Drosophila.

The overall lesson of these works is that polycistronic arrangements in animals are common, such as translation can occur at multiple uORFs and initiation codons[17,39,47], and that there are likely many putative long-non-coding RNAs which are actually protein-coding[17,39]. In synthesis, that the protein coding landscape is complex and dynamic.

Ribosome profiling can be combined with proteomics to identify smORFs and validate their expression[17,39,55]. Ribosome profiling of cytomegalovirus (CMV) infected cells, for example, identified hundreds of new smORFs in the CMV genome[55] (Fig. 2c). Proteomics was then used to detect SEPs generated from these smORFs to validate their translation. These studies revealed an highly dynamic virus proteome that utilizes temporal regulation of genes to enable the expression of hundreds of genes in a compact genome. Moreover, these studies led to the discovery of many new smORFs that can be studied for their role in viral replication.

Functional Validation Approaches

A variety of different experimental approaches for smORF and SEP validation at the genomic scale have also been performed. In E. coli, epitope tags were added to annotated as well as predicted smORFs to validate the production of SEPs[20]. These efforts identified 18 novel smORFs, and demonstrated that many of the SEPs are membrane associated. Subsequent work revealed that many of these SEPs are regulated by cell stress, such as heat shock, identifying these particular genes as a group of stress-response genes[56]. These studies highlight the existence bacterial smORFs controlled by changes in physiological conditions (i.e. glucose[57] and stress[56]). Given that several SEPs with critical functions in bacteria have been identified these findings indicate that there is substantially more molecular and cell biology to be learned from these smORFs. More limited use of epitope tags has also been useful in animals to validate smORF translation and detect smORF peptides[17,24,41], and although truly whole-genome tagging studies have not been carried out, an association with membranes has also been reported[17].

Characterization of Functional smORFs

With methods in place to discover smORFs and SEPs in genomes, the next question is how to identify and characterize functional smORFs, i.e. those producing bioactive SEPs. Several functional smORFs were discovered serendipitously, but improved methods are leading to the identification and characterization of functional smORFs at a much higher rate.

smORFs that regulate growth and metabolism of unicellular organisms

smORFs with biological activity have been discovered in bacteria[57], yeast[21], and human cells[58,59], and more. In *E. coli*, a smORF plays a central role in cell survival under conditions of glucose toxicity[57]. An RNA called SgrS, a 227-nucleotide RNA, is rapidly increased during glucose toxicity. SgrS RNA has two activities (Fig. 3). First, the SgrS mRNA sequence enables it to hybridize to the *ptsG* mRNA, which encodes the primary *E. coli* glucose transporter, to inhibit translation of PtsG[57]. The reduction in PtsG protein results in the lower glucose flux. Second, SgrS mRNA conatins a smORF that produces an SEP called SgrT, a 43-amino acid polypeptide. While SgrS inhibits *ptsG* translation, SgrT is

an inhibitor of PtsG glucose transport activity, providing a two-pronged mechanism to efficiently inhibit glucose influx during times of glucose toxicity.

Yeast is the organism with the largest number of functionally characterized smORFs. A collection of 247 yeast deletion strains were used to study smORF function[21]. Some of these smORFs were known before this work while others were identified for the first time. By growing these deletion strains under different temperatures, carbon sources, and chemical agents the cellular functions of these smORFs were identified. The loss of some smORFs led to lethality or slow growth in haploid strains, and others were now temperature sensitive.

In addition, smORFs are important in other unicellular organisms as well. Indeed, 53% of all smORFs in Mycoplasma pneumonia are essential, while another 11% effect the fitness of the organism[60]. This experiment provides a genome-wide window into the role of smORFs. The characterization of these unicellular smORFs indicated that functional smORFs are not rare, which has helped drive continued interest in these genes.

smORFs that build and regulate animal bodies

A seminal example of the physiological function of smORFs comes from the discovery of a fly gene called *tarsal-less*[29] or *polished rice*[30] (tal/pri) (Fig. 4a). Mutation of this gene resulted in flies having truncated limbs with a missing tarsus[29]. Analysis of this gene revealed the existence of a polycistronic gene with three smORFs that encode 11-amino acid SEPs and a fourth smORF that produces a 32-amino acid polypeptide. Heterologous expression of the 11-amino acid peptide can reverse the phenotype in tal/pri null flies to validate the polypeptide as the bioactive molecule. An investigation into how tal/pri regulates development revealed that this gene is a regulator of the transcription factor shavenbaby (SvB) degradation[61,62].

The tal/pri gene has homologs in other arthropods, suggesting that conservation could be a powerful tool in the discovery of new smORFs. Indeed, subsequent bioinformatic study of conserved short ORFs in flies[22] identified the sarcolamban gene, encoding two new functional smORFs. These newly discovered smORFs produce 28- and 29-amino acid peptides respectively, which are highly similar and adopt an alpha-helix stucture[32]. Null flies lacking both peptides had no overt morphological defects in their structures of their muscles, but did have a heart arrhythmia[32] (Fig. 4b). The search for peptides with similar structure in other organisms revealed homology to the known 30-amino acid mammalian peptide sarcolipin, a peptide with roles in thermogenesis and muscle contraction in mice[63], and phospholamban, a 52-amino acid paralog of sarcolipin[64,65]. Because sarcolipin, phospholamban, and the smORF-encoded 28- and 29-amino acid fly polypeptides were shown to likely derive from a common precursor, the fly peptides were named sarcolambans. And like sarcolipin and phospholamban, sarcolambans bind and inhibit the sacro-endoplasmic reticulum calcium ATPase (SERCA)[66], which regulates calcium signaling and muscle contraction.

Furthermore, myoregulin is a newly discovered mouse homolog of sarcolipin and phospholamban[28]. Sarcolipin, phospholamban, and myoregulin have unique expression patterns, with myoregulin specifically expressed in skeletal muscle (Fig. 5a). Modeling and

functional assays demonstrated that myoregulin also interacts with SERCA. Loss of function studies revealed the importance of myoregulin *in vivo*, as mice lacking this SEP had increased endurance when compared to their WT counterparts[28]. These studies reveal new research avenues into the specific regulation of contraction in different muscles and muscle types, and may prove important in the future development of therapeutic approaches to muscle diseases or aging.

smORFs and the mitochondria

A screen in human cells for genes that protected against beta amyloid (Aβ)-mediated cell death led to the discovery of a smORF on mitochondrial 16S RNA that produces a SEP called humanin[59]. Humanin is a 24-amino acid peptide that is encoded by a 75-bp smORFs encoded in the mitochondrial 16s RNA. Expression of humanin protects cells from Abeta-mediated cell death and apoptosis in general. Subsequent work revealed that humanin operates through the inhibition of the pro-apoptotic BCL-2 protein BAX[58], providing a mechanistic explanation for humanin activity.

A search for additional mitochondrial encoded smORFs led to the discovery of a novel 16-amino acid SEP with anti-diabetic activity. The peptide is named mitochondrial open reading frame of the 12S rRNA-c or MOTS-C[31] (Fig. 5b). Bioinformatics analysis of a human 12s rRNA revealed a 51-bp smORF, MOTS-C, is conserved between 14 different mammalian species. MOTS-c RNA is transported out of the mitochondria where it is translated. Characterization of MOTS-C revealed that this SEP regulates cellular metabolism through changes in the methionine-folate cycle and an increase in AMPK[31] activity. AMPK activation is imperative in whole body metabolism[67], which prompted *in vivo* metabolic studies with MOTS-c. Acute administration of MOTS-c (i.p.) reduced glucose levels, improved muscle insulin sensitivity, and prevented weight gain on a high-fat diet (Fig. 5b). This biology supports continued studies to determine the therapeutic potential of MOTS-c[31].

Another example is the identification of the smORF-encoding gene *Boymaw*, which is linked to an inherited form of schizophrenia[68]. *Boymaw* activity affects rRNA expression and protein translation, and is found at high levels in the post-mortem brains of people with neuropsychiatric diseases. Interestingly, the *Boymaw* peptide also localizes to mitochondria[68], and in flies, both mitochondrial localisation and putative electron transport functions appeared as favoured amongst translated smORF peptides[17]. These highly interesting findings reveal the potential for SEPs to regulate mitochondrial-based physiology, and highlight smORFs as potential biologic therapeutic agents and targets.

Potential opportunities for chemical biology

Many important questions about smORFs and SEPs remain and chemical biology is poised to make significant contributions to our understanding of these atypical genes. smORFs and SEPs that are not homologous to known peptides and proteins must be characterized from scratch. smORF conservation can be an important sign that a gene is functional but if none of the homologs is characterized this only serves as an initial filter. General strategies for deciphering the molecular functions of SEPs are necessary for their characterization, especially in cases where it is not straightforward to screen. Chemical biology offers a

plethora of methods for the molecular characterization of protein function, and these methods will be of tremendous value in characterizing smORFs and SEPs. Also, once bioactive SEPs are identified, chemical biology will enable the production of chemical matter that can be used to investigate these molecules in cells and tissues. For some smORFs and SEPs, these agents may eventually be of therapeutic value.

SEP screens

Screening has been used to identify activities for shorter isoforms of longer proteins, such as catalytic nulls of tRNA synthetases[69]. Along these lines, a potential use of synthetic SEPs is in a functional screen using cell lines or tissue cultures. SEP peptides could be added to the media, and their effect observed. A similar 'gain of function' screen was used in plants to validate smORF function, albeit using transgenes to induce peptide overexpression in vivo[19]. In this study, 800 smORFs were selected computationally, and some 200 overexpressed in the plant laboratory model Arabidopsis. Of these, near 50 produced morphological abnormalities in the resulting plants. However, knockdown experiments would be needed to verify the endogenous functions of SEPs and smORFs identified as potentially functional by these approaches.

SEP-Protein Interactions

To date, bioactive smORFs and SEPs that have been well characterized operate through protein-protein interactions (PPIs)[28,32,57,61] (Table 1). The importance of PPIs in known SEP function means that the identification of SEP-protein interactions will be an expedient route to characterize the molecular functions of uncharacterized SEPs. Of course, there is no reason to believe that SEPs are limited to PPIs, they may interact with DNA, RNA or small molecules as well[70], but current evidence points to a role in protein complexes.

For instance, a potential function for modulator of retroviral infection 2 (MRI-2), a 69-amino acid SEP was revealed through PPI studies[71]. MRI-2 was identified by proteomics and is a shorter isoform of the 156 amino acid MRI-1 gene[72]. MRI-1 was identified as a regulator of retroviral but its molecular mechanism was unknown. Therefore, the characterization of the MRI-2-SEP required the use of an unbiased strategy. Immunoprecipitation mass spectrometry experiments with MRI-2 revealed that this SEP interacts with Ku70 and Ku80, or the Ku heterodimer[73] (Table 1).

Ku70 and Ku80 are the key proteins involved in a DNA repair process called non-homologous end join repair (NHEJ), the predominant form of double strand-break repair in mammalian cells[74]. The interaction between MRI and Ku70 and Ku80 was validated in cells and the addition of MRI-2 to cellular extracts promoted NHEJ to indicate that the peptide interacts with the Ku heterodimer. The function of MRI-2 in cells remains to be determined, but the identification the MRI-2 binding partner provides an operative starting point for developing and testing hypothesis about SEP functions.

Chemical biologists have developed some effective methods for detection of protein-protein interactions in living cells, including transient interactions, which should be amenable to study SEP-protein interactions. One method that has successfully been developed for intracellular interactions is the proximity-labeling approach using the enzyme ascorbate

peroxidase (APEX) and biotin phenol[75,76]. In this approach, the gene of interest is tagged with APEX and the cells are then treated with hydrogen peroxide and biotin phenol. APEX oxidizes the biotin phenol to create a radical species that can covalently label nearby proteins, which results in these proteins being biotin labeled. These proteins can then be enriched and analyzed by mass spectrometry. This method has successfully been used to study protein complexes in the mitochondria, and using SEP-APEX fusions will help identify SEP-protein interactions (Fig. 6a).

Furthermore, a number of suitable methods exist to validate interactions in cells. One of the newest techniques developed is called ReBIL, which is an inducible system that uses luciferase complementation to observe protein-protein interactions in living cells with superb signal-to-noise[77]. ReBIL has been successfully applied to several important biological questions, and such a system can be used to validate SEP-protein interactions in cells. Moreover, once an interaction has been determined, mutagenesis of the SEP sequence would enable the binding site to be rapidly mapped within the context of a living cell using ReBil. In aggregate, the use of chemical biology approaches to discover and validate SEP-protein interactions will greatly accelerate the functional characterization of these molecules.

SEP analogs and small-molecule SEP modulators

As more smORFs and SEPs are characterized, it will be interesting to see if the information gleaned from these studies will lead to the development of synthetic compounds that regulate biology. One avenue that has been taken with natural peptide hormones has been the development of non-natural peptides to develop clinical candidates. For example, Symlin is an analog of the peptide hormone of amylin, which inhibits glucose flux from the stomach to the bloodstream. When given before a meal Symlin reduces postprandial blood glucose levels[78].

Based on the terrific physiological results obtained in mice with MOTS-c, this SEP might be used therapeutically. The therapeutic development of MOTS-c, or eventually other SEPs, will benefit from modifications of the structure to improve stability or pharmacokinetic properties. The use of unnatural amino acid analogs, for instance, has proven useful in obtaining peptide analogs that are active but proteolytically stable[79–81]. A terrific example uses peptide stapling to stabilize the HIV drug enfuvirtide and improve its pharmacokinetic properties[81]. Likewise, analogs of the insulinotropic peptide GLP-1 have been acylated, and this allows them to bind to albumin to have a longer half-life in vivo[82]. Similar strategies can be applied to SEPs such as MOTS-c, which can be readily synthesized (Fig. 6b).

In some cases, the SEPs will need to enter cells to be functional. Chemical biologists have also developed different cell-penetrating strategies to carry protein cargo into cells[34,83]. For example, supercharged versions of green fluorescent proteins, as well as naturally supercharged proteins, are able to transport a variety of protein cargo into cells and tissues[84,85]. For SEPs that operate within mammalian cells conjugation of these molecules to supercharged proteins will enable them to be used as chemical agents for transport into cells (Fig. 6c). SEPs might even be delivery agents that are similar to positively-charged cell-

penetrating peptides[86]. This possibility has not been rigorously tested, although the tal/pri peptides have been reported to affect neighbouring cells[87].

Also, SEPs might provide an ideal type of protein to develop protein-protein interaction inhibitors. A primary issue in developing small-molecule inhibitors of protein-protein interactions is the challenge of using a small molecule of limited size to inhibit a large and energetically favorable protein interaction. By contrast, SEPs, which are much smaller than average proteins, might be more susceptible to inhibition by small molecules. Indeed, a recent review by Arkin, Tang, and Wells divides protein-protein interactions into three categories: primary, secondary, and teritiary[88].

Primary interactions use the primary sequence of a protein to bind to its target. These are the easiest to block with inhibitors since they involve the least surface area. Secondary and tertiary interactions utilize increasingly complex structures, with larger surface areas, at the protein interface and, therefore, are harder to block. Because of the short length of SEPs they are more likely to partake in primary and secondary interactions. Thus, SEP-protein interactions might reveal a group of protein interactions that are particularly rich in targets that can be inhibited by small-molecules. For SEPs that facilitate processes that are involved in disease, blocking these interactions with small molecules will provide novel therapeutic targets.

Conclusions

smORFs represent an under-explored group of genes, but the few examples that have been well characterized indicate that these molecules have important functions. Given the expertise of chemists as synthesizing and working with proteins and peptides, this field is ripe for chemical biology to make a lasting impact. In particular, methods in chemical biology are especially useful for the functional elucidation of SEPs. Furthermore, many SEPs can be synthesized to improve their physiological or cellular uptake properties to enable their use in cell culture and in vivo. Lastly, SEP-protein interactions might prove useful for targeting by small-molecules, which could be used as an alternative method to target these pathways. These studies will identify additional functional molecules and begin to answer broader questions such as the complexity of protein-coding genes in genomes.

Acknowledgments

The authors acknowledge support from the US National Institutes of Health (GM102491 to A.S.), The Leona M. and Harry B. Helmsley Charitable Trust (grant #2012-PG-MED002 to A.S.), and Wellcome Trust Senior Fellowship (08756 to J.P.C.).

References

1. Gardner, D.; Shoback, D. Greenspan’s Basic and Clinical Endocrinology. 9. McGraw-Hill Education; 2011.
2. Kastin, A. Handbook of Biologically Active Peptides. Elsevier Science; 2013.
3. Wilkinson, M.; Brown, RE. An Introduction to Neuroendocrinology. Cambridge University Press; 2015.
4. Bliss, M. The Discovery of Insulin. University of Chicago Press; 2013.
5. Bliss, M.; Purkis, R. The discovery of insulin. University of Chicago Press; Chicago: 1982.

6. De Lecea L, et al. The hypocretins: hypothalamus-specific peptides with neuroexcitatory activity. Proceedings of the National Academy of Sciences. 1998; 95:322–327.
7. Sakurai T, et al. Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. Cell. 1998; 92:573–585. [PubMed: 9491897]
8. Vale W, Spiess J, Rivier C, Rivier J. Characterization of a 41-residue ovine hypothalamic peptide that stimulates secretion of corticotropin and beta-endorphin. Science. 1981; 213:1394–1397. [PubMed: 6267699]
9. Zhang Y, et al. Positional cloning of the mouse obese gene and its human homologue. nature. 1994; 372:425–432. [PubMed: 7984236]
10. Eng J, Kleinman W, Singh L, Singh G, Raufman J. Isolation and characterization of exendin-4, an exendin-3 analogue, from Heloderma suspectum venom. Further evidence for an exendin receptor on dispersed acini from guinea pig pancreas. Journal of Biological Chemistry. 1992; 267:7402–7405. [PubMed: 1313797]
11. Finan B, et al. A rationally designed monomeric peptide triagonist corrects obesity and diabetes in rodents. Nature medicine. 2014
12. Hruby VJ. Designing peptide receptor agonists and antagonists. Nature Reviews Drug Discovery. 2002; 1:847–858. [PubMed: 12415245]
13. Sammons MF, Lee EC. Recent progress in the development of small-molecule glucagon receptor antagonists. Bioorganic & medicinal chemistry letters. 2015
14. Brown NJ, Vaughan DE. Angiotensin-converting enzyme inhibitors. Circulation. 1998; 97:1411–1420. [PubMed: 9577953]
15. Thornberry NA, Weber AE. Discovery of JANUVIA™ (Sitagliptin), a Selective Dipeptidyl Peptidase IV Inhibitor for the Treatment of Type2 Diabetes. Current topics in medicinal chemistry. 2007; 7:557–568. [PubMed: 17352677]
16. Basrai MA, Hieter P, Boeke JD. Small Open Reading Frames: Beautiful Needles in the Haystack. Genome Research. 1997; 7:768–771. DOI: 10.1101/gr.7.8.768 [PubMed: 9267801]
17. Aspden JL, et al. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. Elife. 2014; 3:e03528. [PubMed: 25144939]
18. Frith MC, et al. The abundance of short proteins in the mammalian proteome. PLoS Genet. 2006; 2:e52. [PubMed: 16683031]
19. Hanada K, et al. Small open reading frames associated with morphogenesis are hidden in plant genomes. Proceedings of the National Academy of Sciences. 2013; 110:2395–2400.
20. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. Small membrane proteins found by comparative genomics and ribosome binding site models. Molecular microbiology. 2008; 70:1487–1501. [PubMed: 19121005]
21. Kastenmayer JP, et al. Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. Genome Res. 2006; 16:365–373. DOI: 10.1101/gr.4355406 [PubMed: 16510898]
22. Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. Hundreds of putatively functional small open reading frames in Drosophila. Genome Biol. 2011; 12:R118. [PubMed: 22118156]
23. Oyama M, et al. Diversity of translation start sites may define increased complexity of the human short ORFeome. Molecular & Cellular Proteomics. 2007; 6:1000–1006. [PubMed: 17317662]
24. Slavoff SA, et al. Peptidomic discovery of short open reading frame–encoded peptides in human cells. Nature chemical biology. 2013; 9:59–64. [PubMed: 23160002]
25. Vanderperre B, et al. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PLoS One. 2013; 8:e70698. [PubMed: 23950983]
26. Vanderperre B, Lucier JF, Roucou X. HAltORF: a database of predicted out-of-frame alternative open reading frames in human. Database (Oxford). 2012; :bas025.doi: 10.1093/database/bas025 [PubMed: 22613085]
27. Yang X, et al. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome research. 2011; 21:634–641. [PubMed: 21367939]
28. Anderson DM, et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell. 2015; 160:595–606. DOI: 10.1016/j.cell.2015.01.009 [PubMed: 25640239]

29. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol. 2007; 5:e106. [PubMed: 17439302]
30. Kondo T, et al. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nature Cell Biology. 2007; 9:660–665. [PubMed: 17486114]
31. Lee C, et al. The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. Cell Metab. 2015; 21:443–454. DOI: 10.1016/j.cmet.2015.02.009 [PubMed: 25738459]
32. Magny EG, et al. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science. 2013; 341:1116–1120. [PubMed: 23970561]
33. White JW, Saunders GF. Structure of the human glucagon gene. Nucleic acids research. 1986; 14:4719–4730. [PubMed: 3725587]
34. Richard JP, et al. Cell-penetrating peptides A reevaluation of the mechanism of cellular uptake. Journal of Biological Chemistry. 2003; 278:585–590. [PubMed: 12411431]
35. Lodish, H. Molecular Cell Biology. W. H. Freeman; 2008.
36. Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet. 2014; 15:193–204. DOI: 10.1038/nrg3520 [PubMed: 24514441]
37. Cheng H, et al. Small open reading frames: current prediction techniques and future prospect. Curr Protein Pept Sci. 2011; 12:503–507. [PubMed: 21787300]
38. Crappé J, et al. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC genomics. 2013; 14:648. [PubMed: 24059539]
39. Bazzini AA, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. 2014; 33:981–993. DOI: 10.1002/embj.201488411 [PubMed: 24705786]
40. Mackowiak SD, et al. Extensive identification and analysis of conserved small ORFs in animals. Genome Biol. 2015; 16:179. [PubMed: 26364619]
41. Fritsch C, et al. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome research. 2012; 22:2208–2218. [PubMed: 22879431]
42. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. science. 2009; 324:218–223. [PubMed: 19213877]
43. Jackson RJ, Hellen CU, Pestova TV. The mechanism of eukaryotic translation initiation and principles of its regulation. Nature reviews Molecular cell biology. 2010; 11:113–127. [PubMed: 20094052]
44. Jousse C, et al. Inhibition of CHOP translation by a peptide encoded by an open reading frame localized in the chop 5′ UTR. Nucleic acids research. 2001; 29:4341–4351. [PubMed: 11691921]
45. Iacono M, Mignone F, Pesole G. uAUG and uORFs in human and rodent 5′ untranslated mRNAs. Gene. 2005; 349:97–105. [PubMed: 15777708]
46. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proceedings of the National Academy of Sciences. 2009; 106:7507–7512.
47. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011; 147:789–802. [PubMed: 22056041]
48. Morris DR, Geballe AP. Upstream open reading frames as regulators of mRNA translation. Molecular and Cellular Biology. 2000; 20:8635–8642. [PubMed: 11073965]
49. 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. 2008; 22:2414–2425. [PubMed: 18765792]
50. Parola AL, Kobilka BK. The peptide product of a 5′leader cistron in the beta 2 adrenergic receptor mRNA inhibits receptor synthesis. Journal of Biological Chemistry. 1994; 269:4497–4505. [PubMed: 8308019]

51. Nekrutenko A, Makova KD, Li WH. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. Genome Res. 2002; 12:198–202. DOI: 10.1101/gr.200901 [PubMed: 11779845]
52. Washburn MP, Wolters D, Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature biotechnology. 2001; 19:242–247.
53. Castellana NE, et al. Discovery and revision of Arabidopsis genes by proteogenomics. Proceedings of the National Academy of Sciences. 2008; 105:21034–21038.
54. Branca RM, et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nature methods. 2014; 11:59–62. [PubMed: 24240322]
55. Stern-Ginossar N, et al. Decoding human cytomegalovirus. Science. 2012; 338:1088–1093. DOI: 10.1126/science.1227919 [PubMed: 23180859]
56. Hemm MR, et al. Small stress response proteins in Escherichia coli: proteins missed by classical proteomic studies. Journal of bacteriology. 2010; 192:46–58. [PubMed: 19734316]
57. Wadler CS, Vanderpool CK. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. Proceedings of the National Academy of Sciences. 2007; 104:20454–20459. DOI: 10.1073/pnas.0708102104
58. Guo B, et al. Humanin peptide suppresses apoptosis by interfering with Bax activation. Nature. 2003; 423:456–461. DOI: 10.1038/nature01627 [PubMed: 12732850]
59. Hashimoto Y, et al. A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta. Proc Natl Acad Sci U S A. 2001; 98:6336–6341. DOI: 10.1073/pnas.101133498 [PubMed: 11371646]
60. Lluch-Senar M, et al. Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. Molecular systems biology. 2015; 11:780. [PubMed: 25609650]
61. Kondo T, et al. Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. Science. 2010; 329:336–339. [PubMed: 20647469]
62. Zanet J, et al. Pri sORF peptides induce selective proteasome-mediated protein processing. Science. 2015; 349:1356–1358. [PubMed: 26383956]
63. Bal NC, et al. Sarcolipin is a newly identified regulator of muscle-based thermogenesis in mammals. Nat Med. 2012; 18:1575–1579. DOI: 10.1038/nm.2897 [PubMed: 22961106]
64. MacLennan DH, Kranias EG. Phospholamban: a crucial regulator of cardiac contractility. Nature reviews Molecular cell biology. 2003; 4:566–577. [PubMed: 12838339]
65. Schmitt JP, et al. Dilated cardiomyopathy and heart failure caused by a mutation in phospholamban. Science. 2003; 299:1410–1413. [PubMed: 12610310]
66. Odermatt A, et al. Characterization of the gene encoding human sarcolipin (SLN), a proteolipid associated with SERCA1: absence of structural mutations in five patients with Brody disease. Genomics. 1997; 45:541–553. [PubMed: 9367679]
67. Shackelford DB, Shaw RJ. The LKB1–AMPK pathway: metabolism and growth control in tumour suppression. Nature Reviews Cancer. 2009; 9:563–575.
68. Ji B, Kim M, Higa KK, Zhou X. Boymaw, overexpressed in brains with major psychiatric disorders, may encode a small protein to inhibit mitochondrial function and protein translation. Am J Med Genet B Neuropsychiatr Genet. 2015; 168B:284–295. DOI: 10.1002/ajmg.b.32311 [PubMed: 25943690]
69. Lo WS, et al. Human tRNA synthetase catalytic nulls with diverse functions. Science. 2014; 345:328–332. DOI: 10.1126/science.1252943 [PubMed: 25035493]
70. Lauressergues D, et al. Primary transcripts of microRNAs encode regulatory peptides. Nature. 2015
71. Agarwal S, et al. Isolation, characterization, and genetic complementation of a cellular mutant resistant to retroviral infection. Proceedings of the National Academy of Sciences. 2006; 103:15933–15938.
72. Slavoff SA, Heo J, Budnik BA, Hanakahi LA, Saghatelian A. A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. J Biol Chem. 2014; 289:10950–10957. C113.533968 [pii]. DOI: 10.1074/jbc.C113.533968 [PubMed: 24610814]

73. Walker JR, Corpina RA, Goldberg J. Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature. 2001; 412:607–614. [PubMed: 11493912]
74. Pierce AJ, Hu P, Han M, Ellis N, Jasin M. Ku DNA end-binding protein modulates homologous repair of double-strand breaks in mammalian cells. Genes & development. 2001; 15:3237–3242. [PubMed: 11751629]
75. Lam SS, et al. Directed evolution of APEX2 for electron microscopy and proximity labeling. Nature methods. 2015; 12:51–54. [PubMed: 25419960]
76. Rhee HW, et al. Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. Science. 2013; 339:1328–1331. [PubMed: 23371551]
77. Li YC, et al. A Versatile Platform to Analyze Low-Affinity and Transient Protein-Protein Interactions in Living Cells in Real Time. Cell reports. 2014; 9:1946–1958. [PubMed: 25464845]
78. Hollander PA, et al. Pramlintide as an Adjunct to Insulin Therapy Improves Long-Term Glycemic and Weight Control in Patients With Type 2 Diabetes A 1-year randomized controlled trial. Diabetes care. 2003; 26:784–790. [PubMed: 12610038]
79. Johnson LM, et al. A potent *α*/*β*-peptide analogue of GLP-1 with prolonged action in vivo. Journal of the American Chemical Society. 2014; 136:12848–12851. [PubMed: 25191938]
80. Denton EV, et al. A *β*-peptide agonist of the GLP-1 receptor, a class B GPCR. Organic letters. 2013; 15:5318–5321. [PubMed: 24087900]
81. Bird GH, et al. Hydrocarbon double-stapling remedies the proteolytic instability of a lengthy peptide therapeutic. Proc Natl Acad Sci U S A. 2010; 107:14093–14098. DOI: 10.1073/pnas.1002713107 [PubMed: 20660316]
82. Buse JB, et al. Liraglutide once a day versus exenatide twice a day for type 2 diabetes: a 26-week randomised, parallel-group, multinational, open-label trial (LEAD-6). The Lancet. 2009; 374:39–47.
83. Lindgren, M.; Langel, Ü. Cell-Penetrating Peptides. Springer; 2011. p. 3-19.
84. Cronican JJ, et al. Potent delivery of functional proteins into Mammalian cells in vitro and in vivo using a supercharged protein. ACS chemical biology. 2010; 5:747–752. [PubMed: 20545362]
85. Cronican JJ, et al. A class of human proteins that deliver functional proteins into mammalian cells in vitro and in vivo. Chemistry & biology. 2011; 18:833–838. [PubMed: 21802004]
86. Joliot A, Prochiantz A. Transduction peptides: from technology to physiology. Nat Cell Biol. 2004; 6:189–196. DOI: 10.1038/ncb0304-189 [PubMed: 15039791]
87. Pueyo JI, Couso JP. The 11-aminoacid long Tarsal-less peptides trigger a cell signal in Drosophila leg development. Dev Biol. 2008; 324:192–201. DOI: 10.1016/j.ydbio.2008.08.025 [PubMed: 18801356]
88. Arkin MR, Tang Y, Wells JA. Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. Chemistry & biology. 2014; 21:1102–1114. [PubMed: 25237857]

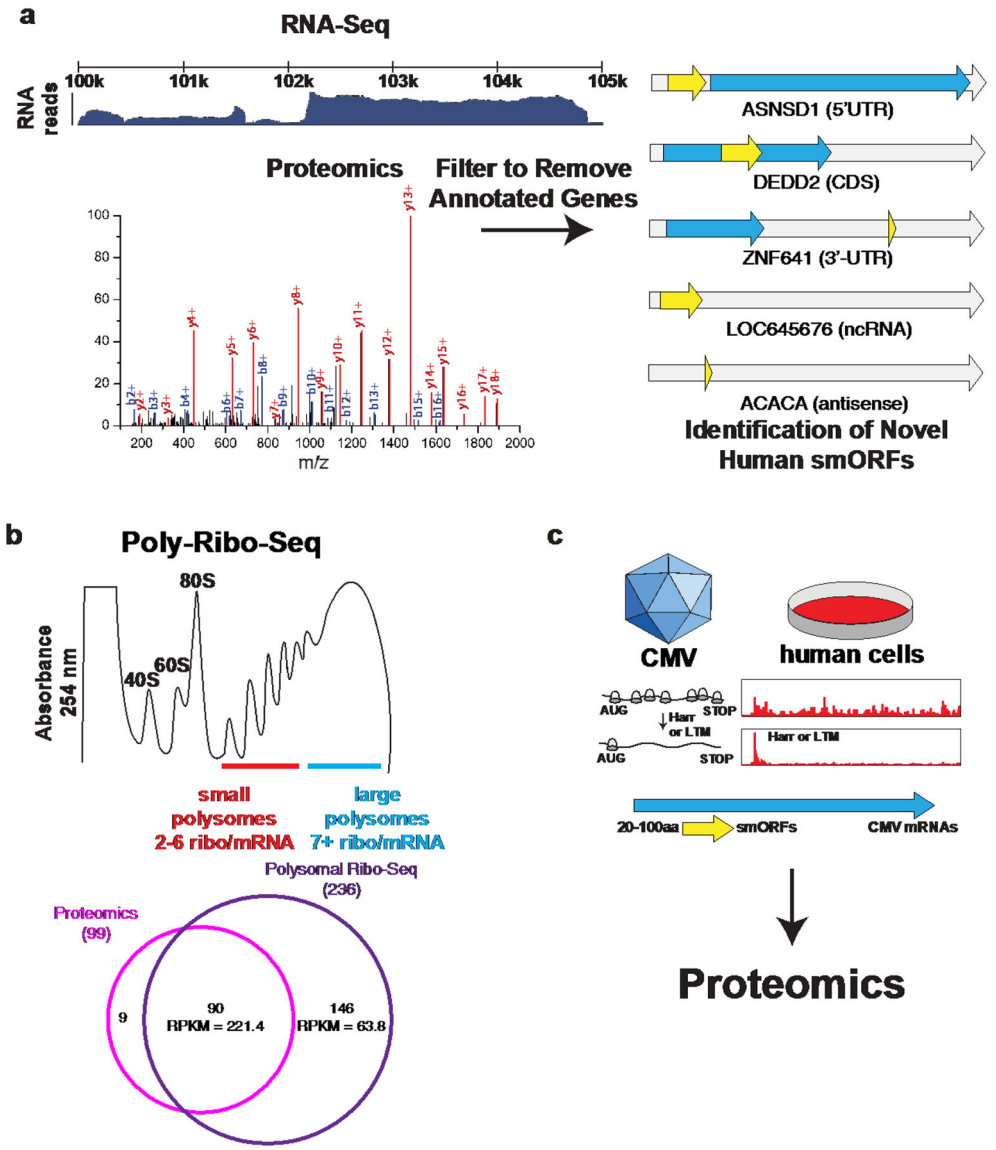

Figure 1. Overview of smORFs and SEPs

a) Ribosomal translation of smORFs produces bioactive SEPs (left) while classical peptide and small protein hormones are produced from limited proteolysis of a much longer prohormone gene (right). b) SEPs partake in protein-protein interactions with a variety of different proteins, such as ion channels, transporters, and other complexes (left). Classical polypeptide hormones, on the other hand, are secreted and primarily interact with two receptor classes g protein-coupled receptors (GPCRs) and receptor tyrosine kinases (RTKs). c) uORFs regulate downstream ORF translation by different mechanisms including stalling or enhancing dissociation of the ribosome. Removal of the uORF leads increased translation of downstream ORFs. Not all uORFs are smORFs (i.e. protein-coding) but in some cases the sequence of protein-coding uORFs are important in regulating downstream ORF translation.

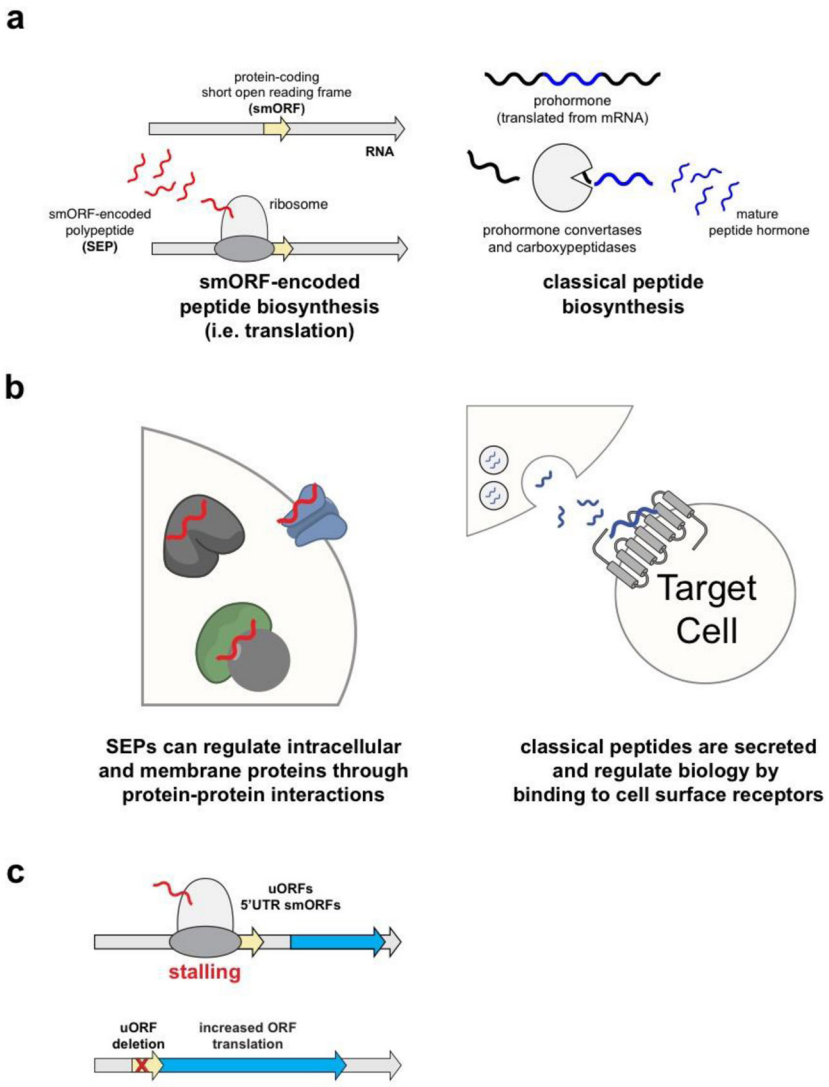

Figure 2. Integrated genomic and proteomic discovery and validation of smORFs

a) Combining RNA-Seq data with proteomics identifies novel human smORFs. The RNA-Seq data is assembled in to transcripts using Cufflinks and then 3-frame translated in silico to generate a searchable proteomics database that contains non-annotated transcripts. Because all possible RNA-produced proteins are included, non-annotated proteins including SEPs can be found. This approach identified 86 novel human smORFs in the 5′-UTR, 3′-UTR, the coding sequence (CDS), non-coding RNAs, and antisense RNAs. b) Polysomes contain strings of ribosomes attached to RNAs. While longer ORFs can have many ribosomes attached simultaneously, smORFs should only have a handful (2–6) ribosomes per RNA. The ribosomal profiling of these short polysomes (referred to as Poly-Ribo-Seq) successfully enriched smORF-containing RNAs and identified 236 smORFs, including 146 whose SEPs have not been identified by proteomics, due to either their lower level of translation (expressed as RPKM, ribosomal-protected reads per million per kilobase) or lower peptide stability. c) Ribosome sequencing (Ribo-Seq) of cytomegalovirus (CMV) infected human cells revealed many novel viral smORFs. Ribo-Seq is a way to measure

ribosome occupancy on mRNA and is used as a surrogate to indicate protein translation. The addition of harringtonine and lactimidomycin (LTM) stall the ribosome on the translation initiation codon, which allows the start of an ORF to be defined. Using this approach led to the identification of many novel CMV ORFs, most of which were sORFs. Several of these were validated via proteomics.

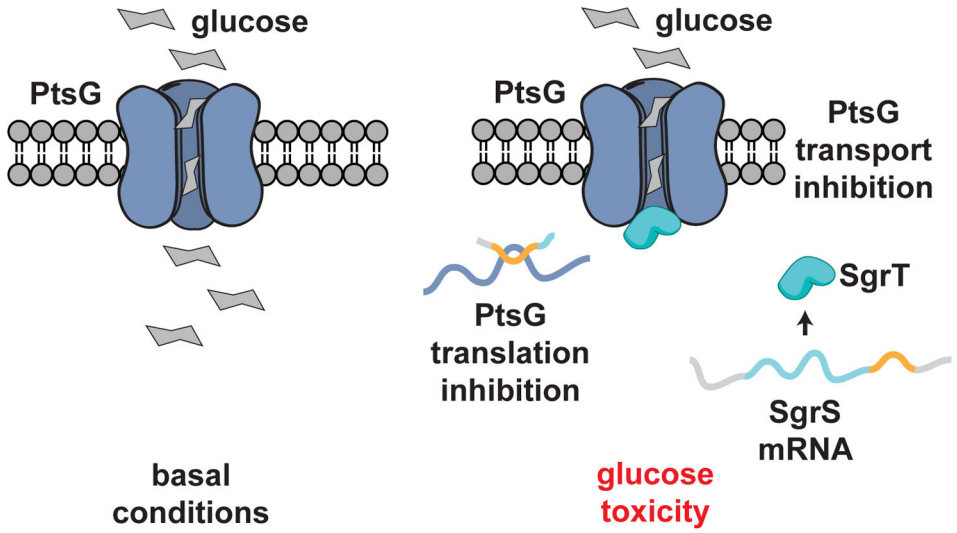

Figure 3. A bacterial smORFs with metabolic function

In bacteria, the cellular response to glucotoxicity includes a unique RNA that includes a smORF[57]. SgrS is an RNA that inhibits the translation of the glucose transporter *ptsG* mRNA through an antisense interaction. Also, a smORF present on the *SgrS* RNA produces an SEP called SgrT that inhibits glucose transport activity of the PtsG protein thereby inhibiting glucose uptake via two distinct mechanisms.

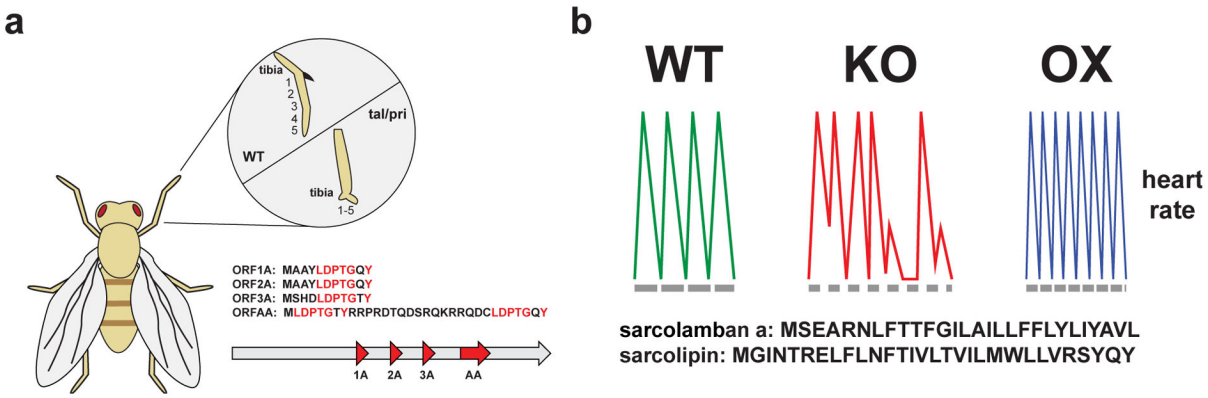

Figure 4. smORFs have varied functions flies

a) smORF discovery was stimulated by the identification of the tarsal-less (*tal*) gene[29], which is also known as polished rice (*pri*)[30]. The loss of *tal/pri* which leads to a truncated leg and a loss of the tarsus. Analysis of the tal gene revealed four homologous smORFs (ORF1A, 2A, 3A, and AA). The SEPs generated from these smORFs are as short as 11 amino acids and are conserved throughout evolution. b) Gene conservation led to the discovery of the smORF encoding sarcolamban[32]. The Sarcolamban regulates calcium flux by binding to the P60A SERCA ion channel. In the flies, loss of sarcolamban (KO) results in cardiac arrhythmia and overexpression (OX) leads to an increased heart rate to show a direct effect for these smORF in fly muscle function.

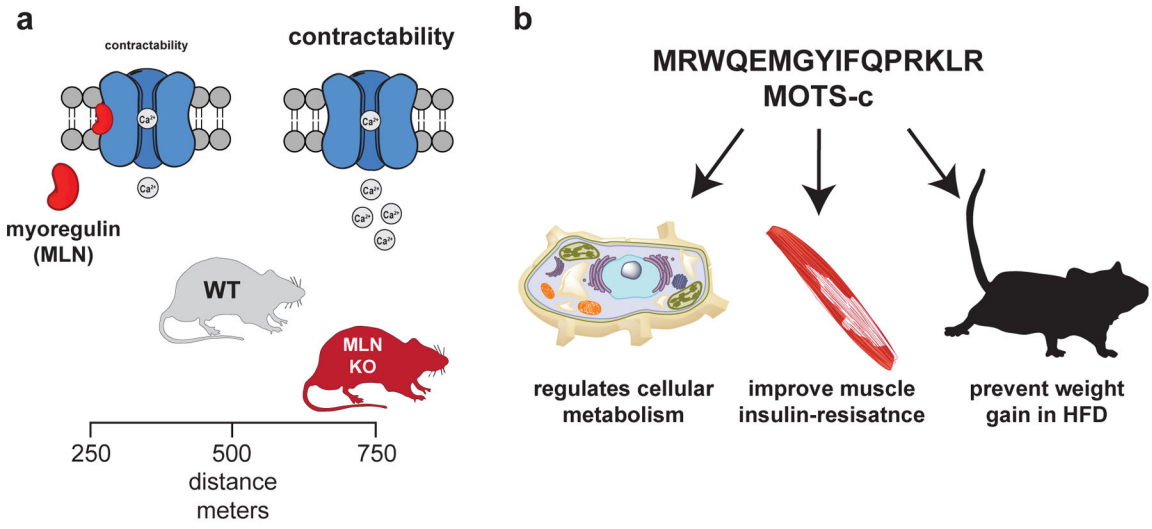

Figure 5. smORFs with functions in mice

a) Myoregulin (MLN) is a sarcolamban homolog that is restricted to the skeletal muscle of mammals. This SEP binds and inhibits the SERCA calcium channel to affect calcium flux and, therefore, calcium contraction in the muscle. Loss of myoregulin increases contractibility and this improves endurance as myoregulin knockout mice (MLN KO) run longer and significantly further than their wild type (WT) counterparts. b) MOTS-c is a newly discovered SEP that is produced from a smORF found on mitochondrial RNA. Cellular and physiological studies identified some biological functions for MOTS-c, such as regulation of cell metabolism. When administered to mice MOTS-c improves muscle insulin resistance and can prevent weight gain and the onset of metabolic disease in mice fed a high-fat diet (HFD).

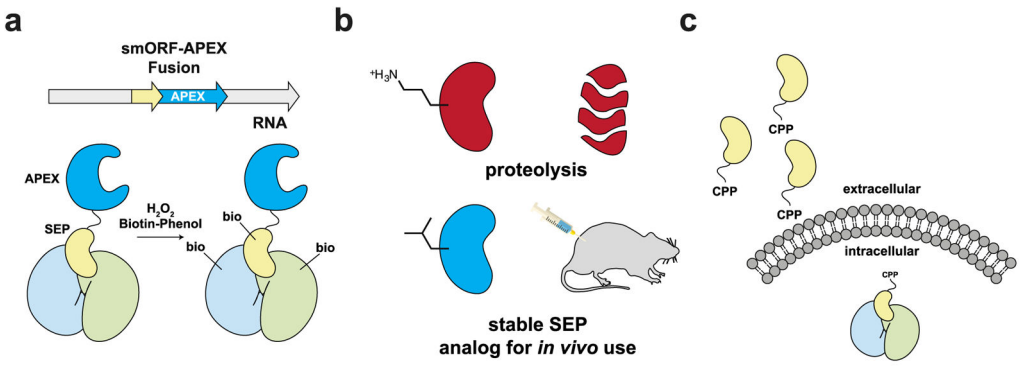

Figure 6. The possible impact of chemical biology in smORF/SEP research

a) Functional SEPs appear to operate through protein-protein interactions (PPIs). The characterization of new SEPs will benefit from proximity labeling methods such as APEX oxidation of a biotin phenol to generate a reactive intermediate that labels nearby proteins to identify PPIs in cells. SEP-APEX fusions will help identify SEP-protein complexes, providing key information in characterizing the SEP. b) SEPs, and peptides in general, are going to be limited as therapeutics because peptides have poor stability and pharmacokinetic properties. The chemical synthesis of SEPs will enable the development of analogs with superior properties that will enable the therapeutic potential of these peptides to be realized. c) Most SEPs operate intracellularly, but most peptides cannot cross biological membranes. The addition of a cell penetrating peptide (CPP) to a SEP will enable these proteins to affect biology inside the cells.

Table 1

smORFs and SEPs involved in interactions with canonical proteins.

smORF/SEP	Length	Protein Interaction Partner	Biology
Bacteria			
SgrT	43 aa	glucose transporter (PtsG)	Glucose metabolism
Flies			
Tal/Pri	11–32 aa	Ubr3	Development
Scl	28–29 aa	calcium transporter (Ca-P60A SERCA)	Muscle (heart) contraction
Mice			
Mln	46 aa	calcium transporter (SERCA1)	Muscle (skeletal) contraction/endurance
Human			
Humanin	24 aa	BAX, IGFBP3	Apoptosis
MRI-2	69 aa	Ku70/Ku80	DNA repair