Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://www.elsevier.com/locate/euprot

Little things make big things happen: A summary of micropeptide encoding genes

Jeroen Crappé, Wim Van Criekinge, Gerben Menschaert**

Lab of Bioinformatics and Computational Genomics (BioBix), Department of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, 9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 31 October 2013
Received in revised form
6 December 2013
Accepted 14 February 2014

Keywords:
Micropeptide
sORF
Ribosome profiling
Peptidomics
(l)ncRNA
Bioactive peptide

ABSTRACT

Classical bioactive peptides are cleaved from larger precursor proteins and are targeted toward the secretory pathway by means of an N-terminal signaling sequence. In contrast, micropeptides encoded from small open reading frames, lack such signaling sequence and are immediately released in the cytoplasm after translation. Over the past few years many such non-canonical genes (including open reading frames, ORFs smaller than 100 AAs) have been discovered and functionally characterized in different eukaryotic organisms. Furthermore, *in silico* approaches enabled the prediction of the existence of many more putatively coding small ORFs in the genomes of *Sacharomyces cerevisiae*, *Arabidopsis thaliana*, *Drosophila melanogaster* and *Mus musculus*. However, questions remain as to what the functional role of this new class of eukaryotic genes might be, and how widespread they are. In the future, approaches integrating *in silico*, conservation-based prediction and a combination of genomic, proteomic and functional validation methods will prove to be indispensable to answer these open questions.

1. Introduction

It is a well-known fact that small peptides play important roles in all kinds of biological processes [1]. The largest and most extensively studied class of small peptides comprises classical bioactive peptides. These are enzymatically cleaved from longer protein precursors containing an N-terminal signal sequence, hence directing the translation product toward the secretory pathway (see Fig. 1). Once released from the secretory vesicles, most of these peptides act as ligands of membrane receptors (mostly G protein-coupled receptors) and exert their extra-cellular biological signaling function in a autocrine, paracrine or endocrine way [2]. Examples are neuropeptides, peptide hormones and growth factors [3–6]. Other secreted peptides exercise their function in host defense systems having antimicrobial or toxic properties or show anti-hypertensive, antithrombotic or antioxidative activity [7,8]. Recently, other (non-classical) peptides – encoded by small open reading frames (sORFs) – have been discovered, presumably defining a new eukaryotic gene family [9–16]. These so-called micropeptides are translated from sORFs shorter

* *Corresponding authors at:* Lab of Bioinformatics and Computational Genomics (BioBix), Department of Mathematical Modelling, Statistics and Bioinformatics, FBE, Ghent University, Coupure Links 653, 9000 Ghent, Belgium. Tel.: +32 9 264 99 22.
E-mail addresses: Jeroen.Crappe@Ugent.be (J. Crappé), Gerben.Menschaert@Ugent.be, Gerben.Menschaert@gmail.com (G. Menschaert).

http://dx.doi.org/10.1016/j.euprot.2014.02.006

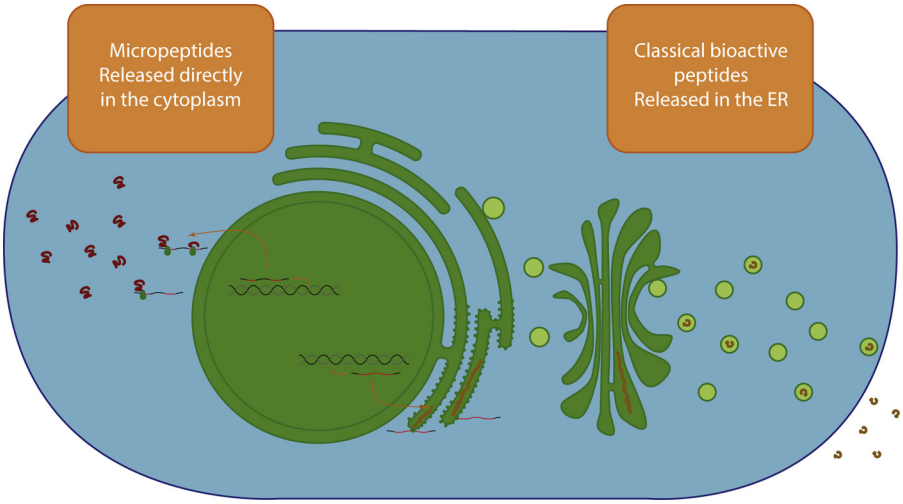

Fig. 1 – Localization of classical bioactive peptides and micropeptides. Classical bioactive peptides contain an N-terminal signal sequence directing the translation product toward the secretory pathway. As a consequence, these biologically active peptides exert an extra-cellular function. In contrast, micropeptides lack an N-terminal signaling sequence, and are consequently released in the cytoplasm immediately after translation.

than 100 AAs [14,17]. Sometimes these are also referred to as polycistronic peptides in the case were they are translated from a polycistronic mRNA [13] or as short open reading frame (sORF)-encoded polypeptides (SEPs) [18]. In contrast to other bioactive peptides, micropeptides are not cleaved from a larger precursor protein and lack an N-terminal signaling sequence. As such they are in principle released in the cytoplasm immediately after translation. This review focuses on this new class of peptides (see Fig. 1).

In the past, many molecules have been overlooked because of various biases and/or simplifications introduced in the performed discovery strategy. For example, it was only in 1993 that the first microRNA (lin-4) was discovered in *Caenorhabditis elegans* [19]. In the 2 subsequent decades more than 2500 microRNAs were identified in human alone [20]. Since micropeptides came into the limelight, ever more research is conducted to this new type of biomolecules, providing increasing evidence that this type of biomolecules is possibly also long overlooked [17]. It was assumed, especially for comprehensive cDNA annotation studies, that protein-coding genes do not code for translation products shorter than 100 AAs [21]. This arbitrarily chosen minimum length reduces the likelihood of false-positive detection by gene-prediction software and genome annotation algorithms, but at the same time vastly underestimates the true number of (atypical) small proteins [22,23]. This generalization is also noticeable in (manually curated) protein databases such as SwissProt-KB, where at the time of writing only 680 (3.4%) out of a total of 20,271 reviewed human proteins have a length shorter than 100 AAs. Although micropeptide research is not yet widespread and much remains to be learned about their abundance, functional activity and localization, a handful of these peptides have recently been functionally annotated in different eukaryotic organisms (see next paragraph for an extensive overview or Table 1 for a brief summary). Though important to an argument of the general conservation and function across all kingdoms of life, sORFs in bacteria and viruses [24–28], will not be covered in this review. The above-mentioned references can serve as a brief overview of putatively coding sORF (pcsORF) detection in lower organisms.

2. Overview of functionally annotated micropeptides

The first eukaryotic micropeptide was only described in 1996. While investigating the function of *early nodulin 40* (*Enod40*), formerly annotated as a ncRNA gene in legumes, van de Sande et al. transformed tobacco plants with a soybean GmENOD40-2 construct [29], proving that this construct was active in the non-legume tobacco, modulating the action of auxin. Sequence comparison of the tobacco and legume *Enod40* clones revealed a highly conserved sORF coding for a 10 (tobacco) or 12 (soybean) AAs long peptide [29]. Later on, a second overlapping coding sORF of 24 AA was identified in soybean, categorizing *Enod40* as a polycistronic mRNA. *Enod40* is a well-known factor that functions in root nodule organogenesis in legumes and also displays a high sequence conservation among other plant species including monocots, suggesting a more general biological function [9]. In addition, *Enod40* shows a highly conserved secondary topology, giving it the characteristics of a structural RNA [30]. The presence of peptide encoding sORFs and of structured RNA, both playing a role in developmental processes, indicates that *Enod40* acts as a bi-functional or dual RNA [31,32]. Since the discovery of this first micropeptide in plants, others have been functionally annotated. In Arabidopsis, the *POLARIS* (*PLS*) gene, identified as a promoter trap transgenic line predominantly showing expression in the embryonic basal region, affects root growth and vascular development [10]. Mutation analysis has shown that the 36 AAs peptide encoded by *PLS* interacts with PIN proteins, forming a network that plays an important role in the hormonal crosstalk between auxin, ethylene and cytokinin [33,34]. In maize, the recessive mutation of *Brick1* (*Brk1*) leads

Table 1 – Functionally annotated micropeptides in Eukaryotes.

Gene name	sORF length[a]	Species[b]	Proposed function	Conservation	References
Early nodulin 40(Enod40)	12, 24	*G. max*	root nodule organogenesis	Legumes & Monocots	[9,29,30]
POLARIS (PLS)	36	*A. thaliana*	hormonal crosstalk during embryogenesis		[10,33,34]
Brick1 (Brk1)	76	*Z. mays*	morphogenesis of leaf epithelia	Plants and Animals (*HSPC300* homolog)	[11,35,36,46]
Rotundifolia (ROT4)	53	*A. thaliana*	leaf shape morphogenesis	Plants	[12,37]
Zm401p10	89	*Z. mays*	tapetum development		[38,39]
Zm908p11	97	Z. mays	pollen tube growth	Poaceae	[40]
tarsal-less (tal)	3 × 11, 32	*D. melanogaster*	leg and actin-based cell morphogenesis during embryogenesis	Arthropods and *Daphnia* (>440 MM years)	[13–15,43–45]
sarcolamban (scl)	28, 29	*D. melanogaster*	SER Ca^{2+} trafficking	Vertebrates and Arthropods (>550 MM years)	[16]

[a] Number of amino acids in the different translated and functional open reading frames.
[b] The organism in which the function has been best studied experimentally.

to several morphological defects of leaf epithelia [11]. The gene is highly conserved in plants as well as in animals and encodes a 76 AA peptide that lacks any targeting sequence. Research in *Arabidopsis* has shown that *Brk1* is a critical WAVE-complex subunit functioning in a pathway with the ARP2/3 complex [35,36]. *ROTUNDIFOLIA* (*ROT4*) was identified as an overexpressed novel single exon gene encoding a small 53 AAs peptide in an *Arabidopsis* mutant with short leafs and floral organs [37]. Phylogenetic analysis in *Arabidopsis* indicates that *ROT4* defines a novel seed-plant specific small peptide gene family, comprising 22 *ROT FOUR LIKE* (*RTFL*) genes, sharing a conserved 29 AAs region [12]. More recently, two novel maize sORF genes, *Zm401p10* and *Zm908p11* respectively encoding 89 and 97 AA peptides were identified, playing a required role in pollen development [38–40].

Micropeptide research is not limited to plants; some of the best-studied sORF genes have been identified in the animal kingdom. *In silico* prediction analysis of cDNAs in *Drosophila melanogaster* identified several mRNA-like ncRNA candidates putatively encoding ORFs [41,42]. Extensive study on one of these candidates by several groups revealed that the evolutionary conserved *tarsal-less* (*tal*) or *polished rice* (*pri*) in *Drosphila* and the orthologous *mille-pattes* (*mlpt*) in *Tribolium* is in fact a polycistronic gene encoding small peptides [13–15]. The *tal* gene contains a total of 4 sORFs encoding functionally redundant peptides of length 11–32 AAs, playing a role in *Drosophila* embryogenesis. The absence of a functional *tal* gene, either by knock-down or *via* mutation, leads to the absence of trichomes on the body surface, a missing tarsal region, ectopic leg morphogenesis and abnormal dentical belt and tracheal formation [13,14,43]. On the other hand, the overexpression of *tal* peptides negatively modulates the Notch signaling pathway [44]. Extensive molecular analysis showed that *tal* controls epidermal differentiation by modifying the transcription factor *Shavenbaby* (*Svb*) from a transcriptional activator to a repressor, thus dominantly inhibiting its downstream function in trichome formation [45]. Similarly, RNAi depleted *mlpt* embryos alter gap gene expression, generally leading to shortened embryos with additional pairs of legs, also missing some posterior abdominal segments [15].

Recently, another member of this set of mRNA like ncRNAs (also comprising the *tal* gene) led to the identification of two new coding sORFs (28 and 29 AAs long) in the *putative noncoding RNA 003* gene (*pncr003:2L*) [16,41]. Phylogenetic conservation analysis indicated that the *pncr003:2L* gene shares a common sORF encoding ancestor gene with the human *sarcolipin* (*sln*) and its longer paralogue *phospholamban* (*pln*). In order to reflect their similarity the researchers suggested to rename the *pncr003:3L* gene and its arthropod homologs to *sarcolamban* (*scl*). Visualizing intracellular Ca^{2+} levels in *scl* mutants and wild-type controls identified a primary role for *scl* encoded peptides during the Ca^{2+} trafficking at the sarco-endoplasmatic reticulum (SER) which is required for heart muscle contraction [16].

3. Micropeptide-specific characteristics

In contrast to classical bioactive peptides, one of the properties micropeptides share is the absence of an N-terminal signaling sequence; as such they are not destined toward the secretory pathway, but are immediately released in the cytoplasm (see Fig. 1) [17]. While one would expect an intracellular, cytoplasmic function for peptides lacking such signal sequences, research indicates that this is not always the case and that some of the identified micropeptides act non-cell-autonomously. Examining clonal sectors of *brk1* mutant cells in otherwise wild-type leaves, showed that *brk1* mutant cells in direct contact with wild-type cells appeared to be wild-type with normal stomata, indicating non-cell-autonomous functioning [46]. *Tal* is another example of a non-celll-autonomous functioning micropeptide. After introducing frame-shift mutations in the 4 ORFs of a full-length *tal* transcript and expressing *tal* in a subset of epithelial cells, denticle formation was completely rescued in *tal*-expressing as well as neighboring cells [13]. Different mechanisms of such micropeptide-regulated morphogenesis of neighboring cells can be postulated. Their action can be either directly as an extra-cellular signaling molecule, or indirectly *via* a downstream target, or by intracellular regulation of

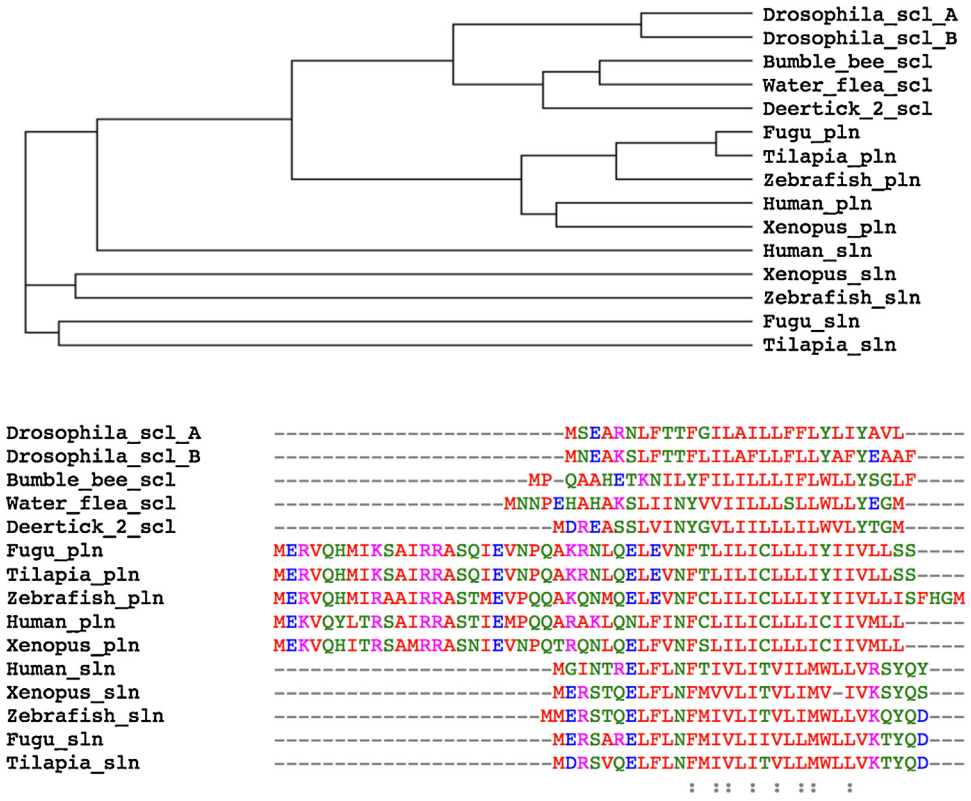

Fig. 2 – Analysis of *Drosophila Sarcolamban* homology. Phylogenetic tree and multiple protein sequence alignment of vertebrate and arthropod *Sarcolamban* (Scl), *Sarcolipin* (Sln) and *Phospholamban* (Pln) peptides, constructed with Clustal Omega and ClustalW2-Phylogeny [102–104].

one or more intercellular signaling pathways [13,17,47]. More speculative, their small molecular size makes them ideal candidates for cellular translocation as gap peptides, to function as membrane permeable peptides, or to leave the cell *via* cell-derived (micro) vesicles such as exosomes [48–52].

Conservation of the coding sequence across very large evolutionary distances is another peculiar feature of micropeptides. The recently identified *sarcolamban* shows conservation of more than 550 million years (see Fig. 2). Analysis of the related peptides indicates a conserved peptide sequence and molecular structure from flies to vertebrates all involved in the regulation of Ca^{2+} traffic [16]. *Tarsal-less* is another example of a highly conserved micropeptide. *Tal* homologs could be identified in other insects and even in *Daphnia pulex*, a crustacean species. This gene family is at least 440 million years old and shows a varying number of sORFs with an evolutionary trend toward accumulation of more ORFs [14]. A systematic search for new genes in *Saccharomyces cerevisiae* [53] (see also next paragraph) led to the identification of, among others, *smORF2*. Functional homologs for this gene with a temperature-sensitive phenotype could be identified in many organisms from yeast to human [54]. To our knowledge, *Brk1* is however by far the most conserved sORF encoding gene. Next to being highly conserved throughout the plant kingdom, homologs for the maize *brk1* gene have been identified in almost all studied animal eukaryotic genomes, including human, where *HSPC300* (mammalian homolog of *brk1*) also functions in the Scar/WAVE complex [11,36]. It remains to be confirmed that conservation is a typical characteristic of sORF-encoding genes. Most discovered micropeptide genes result from large genetic or computational screenings, focusing on phylogenetic conservation as a proxy to functionality, in this way choosing interesting targets for further (elaborate) downstream *in vivo* research. This might introduce a bias toward discovery of highly conserved sORFs. On the other hand, a highly conserved state can also point to the role most of these genes play in (embryonic) development, morphogenesis and other very basic and important biological processes. As such, the high sequence and functional conservation of this new type of eukaryotic gene products, might explain many basic but very important functions shared by a plethora of species over different kingdoms.

4. Systematic searches for putatively coding sORFs and micropeptides

sORF encoding genes have, in our opinion, long been overlooked. However, the past decade has seen some important advances in the (genome-wide) identification of pcsORFs. To identify those new and interesting candidates in the vast amount of random sORFs scattered al over the genome, *in silico* strategies (often making use of expression data) have been devised. *S. cerevisiae* was the first eukaryotic species to be the subject of such a systematic and elaborate scan. Many yeast sORFs were identified based on comprehensive sequence

database searching. Homology, comparative genomics and expression features from serial analysis of gene expression (SAGE), Northern blotting, RT-PCR and ORF tagging experiments were taken into account [54–61]. Later, Kastenmayer et al. concluded that, based on the above-mentioned independent experimental approaches and computational analyses, at least 299 pcsORFs are present in the *S. cerevisiae* genome, many of which have potential orthologs in other eukaryotic species. This represents a significant percentage (~5%) of the amount of annotated genes in *S. cerevisiae*. Of these identified pcsORFs, four were confirmed to produce a translation product and 22 seem to regulate growth [53].

Arabidopsis was the first plant species undergoing a thorough *in silico* analysis in the search of new sORF encoding genes. Because common gene-finding algorithms have a hard time identifying small protein products and are prone to a high number of false negatives (as already mentioned in the introduction) Hanada et al., developed the Coding Index (CI) measure for pcsORF prediction based on the hexamer composition bias, a general measure to distinguish CDS from non-CDS [62,63]. This CI measure would later form the basis for a specific program package to identify sORFs, named sORFfinder [64]. After performing a six reading frame translation of intergenic sequences of *Arabidopsis thaliana*, pcsORFs were only assigned as being coding when they demonstrated qualifying CI values, above background tiling array hybridization intensities, evidence of purifying selection based on Ka/Ks values and overlapping ESTs. Using these criteria, 7159 pcsORFs with high coding potential, of which 2376 are subject to purifying selection, were identified in *A. thaliana* [62,64]. In a recently published follow-up study, elaborating on the function of these new pcsORFs, an array was designed to generate an expression atlas at several developmental stages and under multiple environmental conditions for the 7901 identified pcsORFs [65]. 473 pcsORFs showed a high number of homologs in other plant species and were overexpressed. 49 of those expressed and significantly conserved pcsORFs induced various morphological changes and visible phenotypic effects [65].

Arabidopsis is not the only plant species subject to an integrative procedure to identify pcsORFs at the genome level. After obtaining ~2.6 million expressed sequence tag (EST) reads from a *Populus deltoides* leaf transcriptome, full-length transcripts from the EST sequences could be reconstructed. Using a computational approach based on coding potential, evolutionary conservation and gene family clustering, and by showing evidence of protein domains, ncRNA motifs, sequence length distribution or mass-spectrometry data, at least 56 pcsORF encoding genes (<200 AAs) new to the *Populus* genome annotation could be identified [66]. Very recently, work was published exploiting publicly available genome sequences of *Phaseolus vulgaris, Medicago tuncatula, Gycine max* and *Lotus japonicus* in a search for pcsORFs (30–150 AAs) [67]. A bioinformatics analysis was performed based on evidence of expression (transcription level), presence of known protein regions or domains (translation level) and identification of orthologues genes in the explored genomes. Respectively 6170, 10461, 30521 and 23599 pcsORFs were uncovered within the *P. vulgaris*, *G. max*, *M. truncatula* and *L. japonicus* genomes. Based on specific EST expression analysis in *P. vulgaris*, 2336 of the identified pcsORFs showed evidence of gene expression.

In the animal kingdom, the first *in silico* and systematic search for new pcsORFs was carried out for the model organism *D. melanogaster* [68]. Starting from putatively non-coding euchromatic DNA, an initial set of 593 586 open reading frames between 30 and 300 basepairs (bps) long could be identified. Using tBlastn, all pcsORFs showing significant similarity with annotated coding sequences or transposons were removed, at the same time only retaining pcsORFs showing significant amino acid sequence similarity with *Drosophila pseudoobscura*. After realigning extended versions of the conserved pcsORFs with ClustalW, an upper estimate of 4561 pcsORFs were identified in *Drosophila*. 72% of the in *D. pseudoobscura* conserved pcsORFs appeared to be true homologs as they were conserved in syntenic regions with regard to the original *D. melanogaster* pcsORF. Only taking into account syntenic pcsORFs with favorable Ka/Ks values (having a ratio below 0.1), and with transcriptional evidence (based on combining both publicly available RNA-seq and tiling array data), the authors postulate that at least 401 functional sORFs exist in the *D. melanogaster* genome [68].

Very recently, Crappé et al. combined an *in silico* approach and experimental evidence by means of ribosome profiling data (see also next paragraph) for a genome-wide search to detect novel pcsORFs in the *Mus musculus* genome [69]. First, the genome was scanned for sORFs with high coding potential using the sORFfinder package. Secondly, a comprehensive feature matrix with peptide conservation measures, based on UCSC multiple species alignments, was constructed. In a third step, the coding capabilities of these pcsORFs were assessed by means of a machine-learning algorithm. Afterwards, the sORFs with a high coding score were verified for the presence of experimental ribosome profiling signals obtained from mouse Embryonic Stem Cells (mESCs), hinting to sORF translation. Using this combined genome-wide approach dozens of both highly conserved and ribosome-targeted pcsORFs (possibly encoding micropeptides) were identified [69].

5. sORF identification using ribosome profiling

Ribosome profiling is a recently described new strategy to monitor protein synthesis based on deep sequencing of ribosome protected mRNA fragments [70–74]. By exploiting the properties of drugs as harringtonine, puromycin or lactimidomycin, stalling ribosomes at Translation Initiation Sites (TIS), the study of (alternative) (a)TIS with subcodon to single-nucleotide resolution is now possible [28,73,75–79]. In an attempt to provide a genome-wide map of protein synthesis, Ingolia et al. exploited a machine learning algorithm on top of their ribosome profiling data to systematically delineate protein products in mESCs [75]. Special attention was paid to a recently identified and apparently abundant class of RNAs, referred to as long non-coding RNAs (lncRNAs). These lncRNAs were scanned for translated regions, by defining the most highly ribosome-occupied 90 nucleotide window

[80,81]. In this way, Ingolia et al. were able to identify many lncRNA regions, displaying high ribosomal occupancy and containing small open reading frames, classifying them as short polycistronic ribosome-associated coding RNAs (sprcRNAs) [75]. Motivated by these findings and while performing a global translation initiation analysis using ribosome profiling (GTI-seq) in HEK293 cells, Lee et al. also specifically characterized the translation of ncRNAs. They were able to identify 228 ncRNAs associated with GTI-seq sequencing reads, often overspanning evolutionary conserved sORFs (median length of 54 nt) and frequently showing alternative initiation at non-AUG start codons [78]. Our own research on mESC ribosome profiling data strengthens the idea that some lncRNAs actually contain putatively coding sORFs. While investigating sORFs within annotated lncRNA regions, we were also able to detect very well-conserved and ribosome targeted pcsORFs [69].

The question if and more specifically to what extent lncRNAs act through their translational sORF products remains up for debate and is one that will not find an answer based on ribosome profiling data alone. For example, the mouse H19 lncRNA transcript functions as a true ncRNA [82,83], even though demonstrating ribosome occupancy [75,84]. This proves that simply ribosome profiling does not suffice as evidence of protein synthesis nor can be proposed as a fool-proof method to distinguish between coding and non-coding transcripts [85,86]. In addition, one has to keep in mind that spurious association of ribosomes could lead to translational noise [87]. The fact that most of the predicted lncRNA transcripts that encode sORFs lack any significant conservation and that lncRNAs are rarely translated in human cell lines seems consistent with these observations [80,81,88,89]. In a follow up study Guttman et al. developed a metric, the ribosome release score (RRS), enabling sensitive identification of functional protein-coding transcripts based on the termination of translation at the end of the ORF [90,91]. With this metric it is possible to discriminate between protein-coding transcripts and other classes of non-coding transcripts, including lncRNAs. Because the class of lncRNAs closely resembled the ribosome occupancy of other classes of non-coding transcripts with respect to this metric, it was deemed unlikely that lncRNAs, as a class, produce functional products [91]. Future measures will certainly be devised to assess the true coding potential of ribosome profiling occupied mRNA on a case per case basis.

Although most research at the moment points to the true non-coding state of lncRNAs, a subclass could still comprise pcsORFs, considering that some of them have proven to be highly conserved [69]. The absence of detectable peptide products does furthermore not rule out their existence, as the expression of lncRNA encoded sORFs could be very specific in time as well as in space [45,92]. An attractive hypothesis could be that lncRNAs are generally non-coding, but under specific circumstances, enclosed sORFs can be translated (presumably at very low levels), thus rendering these lncRNAs as bifunctional or dual RNAs [31,32,93]. In the end, only scrutinized and functional *in vivo* analysis will be able to proof if and what lncRNA transcripts give rise to functional small protein products [16,45,94].

6. Detection of micropeptides using mass spectrometry

Few studies exist where mass spectrometry is used for the direct detection of micropeptides, although this still is the gold standard when looking for protein or peptide products. Using a newly developed strategy, combining peptidomics and massive parallel RNA-seq, Slavoff et al. claim the discovery of many previously uncharacterized human sORF-encoded polypeptides (SEPs) in K562 (human leukemia) cells [18]. First, custom databases were constructed containing all possible polypeptides based on the annotated human transcriptome (RefSeq) and an experimental RNA-seq derived K562 transcriptome. Identifying the peptidomics mass spectrometry fragmentation spectra (MS/MS) using these custom polypeptide databases and four previously reported SEPs as positive controls, an extra 86 still uncharacterized SEPs were discovered, bringing the total of unannotated human SEPs to 90 [18,95].

Although this study is one of the early attempts to systematically identify micropeptides by means of mass spectrometry and subsequent peptide-to-spectrum matching strategy, they largely failed to prove the mature forms of micropeptides. Alternative strategies, for which the above approach could serve as a guideline, should lead to the true identification of mature and native forms of endogenous micropeptides as this is still one of the most important aspects of peptidomics.

7. Outlook

Although there are many sophisticated gene prediction programs available, the majority is optimized to predict genes with 100 or more codons, rendering them inappropriate for sORF detection [96–98]. Development of *ab initio* single-sequence methods (based on codon patterns) and discriminative metrics (pairwise and multi-species alignment-based comparative metrics), suited for the detection of small ORFs, is still in its infancy. On shorter exons, comparative metrics clearly outperform single-sequence based methods, adding discriminatory power as additional species are used. Using hybrid metrics, exploiting the relative independence of their input metrics, further increases performance [94,98,99]. Based on these findings, Crappé et al. combined several metrics, computed from a multi-species alignment and subsequently built a classifier model (using a Support Vector Machine) to classify coding *versus* non-coding. Although the combination of different metrics generally leads to better performance, it is still dependent on the correctness of the multiple sequence alignment, does not incorporate near-cognate start sites and possibly misses a lot of highly divergent and/or quickly diverging pcsORFs [69]. New and promising metrics with regard to this growing field of sORF detection will certainly emerge. In this respect, PhyloCSF is certainly noteworthy. It is a comparative genomics method that analyzes a multiple sequence alignment using phylogenetic codon models to correctly distinguish between protein-coding and non-coding regions. PhyloCSF clearly outperforms other methods for the analysis of short exons [100].

An integrated approach combining computational and experimental validation stands a better chance to result in meaningful findings than merely performing an *in silico* prediction [94]. Slavoff et al. compiled a custom mRNA-seq derived polypeptide database to identify mass spectrometry fragmentation spectra and were able to identify 86 uncharacterized SEPs [18]. However, for reasons already mentioned in this paper, the ribosome profiling technique is more suitable than mRNA-seq to delineate the exact ORFs and thus derive putative micropeptide sequences. Menschaert et al. prove that a ribosome profiling (RIBO-seq) derived custom database yields a highly informative search space of translation products for MS/MS-based peptide identification [79]. An automated pipeline converting RIBO-seq information into a custom pcsORF sequence database, by delineating open reading frames from calling the translation start sites and detecting SNP mutations, will prove to be very beneficial in future MS-based studies.

Known micropeptides have a very narrow expression in time as well as in space [45]. These characteristics are probably part of the reason why *tarsal-less*, one of the best-studied micropeptides to date, has never been identified using mass spectrometry. New and alternative extraction methods should prove to be more effective at extracting cytoplasm bound micropeptides [3]. For example, Schwaid et al. recently reported on an affinity-based approach to enrich and identify cysteine-containing human sORF encoded polypeptides (ccSEPs) from cells. Using this approach they were able to identify 16 novel ccSEPs, derived from uncharacterized sORFs [101]. The development of new mass spectrometry based techniques, such as the reported chemoproteomic approach, will prove indispensable in order to identify and characterize the biological function of micropeptides.

The mere existence of a peptide does not imply that it has a function. Evolutionary conservation is definitely suggestive for functionality, but to pinpoint the actual function, experimental demonstration of a biological effect is required [94]. Approaches used to functionally describe micropeptides such as *tarsal-less* or *sarcolamban* can be seen as a general guide for further *in vivo* analyses [16,45].

8. Conclusion

Research on short peptides, encoded by small open reading frames, is still in its infancy. Nevertheless, growing evidence points to the existence of these so-called micropeptides, but to what extent and how important this class of new biomolecules is, still needs to be seen. Approaches integrating *in silico*, conservation-based prediction and a combination of genomic, proteomic and functional validation methods will prove to be indispensible to further explore this micropeptide research field.

Acknowledgments

The financial support of the Institute for the Promotion of Innovation in Flanders (IWT) and the Belgian National Fund for Scientific Research (FWO-Flanders) is gratefully acknowledged. Dr. G. Menschaert is supported by a postdoctoral fellowship of FWO-Flanders, J. Crappé is supported by a fellowship of IWT-Flanders.

REFERENCES

[1] Fricker LD. Neuropeptide-processing enzymes: applications for drug discovery. AAPS J 2005;7:E449–55.
[2] Boonen K, Creemers JW, Schoofs L. Bioactive peptides, networks and systems biology. BioEssays 2009;31:300–14.
[3] Fricker LD. Analysis of mouse brain peptides using mass spectrometry-based peptidomics: implications for novel functions ranging from non-classical neuropeptides to microproteins. Mol BioSyst 2010;6:1355–65.
[4] Nassel DR, Winther AM. Drosophila neuropeptides in regulation of physiology and behavior. Prog Neurobiol 2010;92:42–104.
[5] Hummon AB, Amare A, Sweedler JV. Discovering new invertebrate neuropeptides using mass spectrometry. Mass Spectrom Rev 2006;25:77–98.
[6] Baggerman G, Boonen K, Verleyen P, De Loof A, Schoofs L. Peptidomic analysis of the larval *Drosophila melanogaster* central nervous system by two-dimensional capillary liquid chromatography quadrupole time-of-flight mass spectrometry. J Mass Spectrom 2005;40:250–60.
[7] Kim SK, Wijesekara I. Development and biological activities of marine-derived bioactive perptides: a review. J Funct Foods 2010;2:1–9.
[8] Sarmadi BH, Ismail A. Antioxidative peptides from food proteins: a review. Peptides 2010;31:1949–56.
[9] Rohrig H, Schmidt J, Miklashevichs E, Schell J, John M. Soybean ENOD40 encodes two peptides that bind to sucrose synthase. Proc Natl Acad Sci U S A 2002;99:1915–20.
[10] Casson SA, Chilley PM, Topping JF, Evans IM, Souter MA, Lindsey K. The *POLARIS* gene of Arabidopsis encodes a predicted peptide required for correct root growth and leaf vascular patterning. Plant Cell 2002;14:1705–21.
[11] Frank MJ, Smith LG. A small, novel protein highly conserved in plants and animals promotes the polarized growth and division of maize leaf epidermal cells. Curr Biol 2002;12:849–53.
[12] Narita NN, Moore S, Horiguchi G, Kubo M, Demura T, Fukuda H, et al. Overexpression of a novel small peptide ROTUNDIFOLIA4 decreases cell proliferation and alters leaf shape in *Arabidopsis thaliana*. Plant J 2004;38:699–713.
[13] Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, Kageyama Y. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat Cell Biol 2007;9:660–5.
[14] 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.
[15] Savard J, Marques-Souza H, Aranda M, Tautz D. A segmentation gene in tribolium produces a polycistronic mRNA that codes for multiple conserved peptides. Cell 2006;126:559–69.
[16] Magny EG, Pueyo JI, Pearl FM, Cespedes MA, Niven JE, Bishop SA, et al. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science 2013;341:1116–20.
[17] Hashimoto Y, Kondo T, Kageyama Y. Lilliputians get into the limelight: novel class of small peptide genes in morphogenesis. Dev Growth Differ 2008;50(Suppl. 1):S269–76.
[18] Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 2013;9:59–64.

[19] Lee RC, Feinbaum RL, Ambros V, The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993;75: 843–54.
[20] Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 2011;39:D152–7.
[21] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The transcriptional landscape of the mammalian genome. Science 2005;309:1559–63.
[22] Dinger ME, Pang KC, Mercer TR, Mattick JS. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol 2008;4:1–5.
[23] Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, et al. The abundance of short proteins in the mammalian proteome. PLoS Genet 2006;2:e52.
[24] Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. Small membrane proteins found by comparative genomics and ribosome binding site models. Mol Microbiol 2008;70:1487–501.
[25] Hemm MR, Paul BJ, Miranda-Ríos J, Zhang A, Soltanzad N, Storz G. Small stress response proteins in *Escherichia coli*: proteins missed by classical proteomic studies. J Bacteriol 2010;192:46–58.
[26] Boekhorst J, Wilson G, Siezen RJ. Searching in microbial genomes for encoded small proteins. Microb Biotechnol 2011;4:308–13.
[27] Hobbs EC, Fontaine F, Yin X, Storz G. An expanding universe of small proteins. Curr Opin Microbiol 2011;14:167–73.
[28] Stern-Ginossar N, Weisburd B, Michalski A, Le VT, Hein MY, Huang SX, et al. Decoding human cytomegalovirus. Science 2012;338:1088–93.
[29] van de Sande K, Pawlowski K, Czaja I, Wieneke U, Schell J, Schmidt J, et al. Modification of phytohormone response by a peptide encoded by ENOD40 of legumes and a nonlegume. Science 1996;273:370–3.
[30] Gultyaev AP, Roussis A. Identification of conserved secondary structures and expansion segments in enod40 RNAs reveals new enod40 homologues in plants. Nucleic Acids Res 2007;35:3144–52.
[31] Ulveling D, Francastel C, Hube F. When one is better than two: RNA with dual functions. Biochimie 2011;93:633–44.
[32] Bardou F, Merchan F, Ariel F, Crespi M. Dual RNAs in plants. Biochimie 2011;93:1950–4.
[33] Chilley PM, Casson SA, Tarkowski P, Hawkins N, Wang KL, Hussey PJ, et al. The POLARIS peptide of Arabidopsis regulates auxin transport and root growth via effects on ethylene signaling. Plant Cell 2006;18:3058–72.
[34] Liu J, Mehdi S, Topping J, Friml J, Lindsey K. Interaction of PLS and PIN and hormonal crosstalk in Arabidopsis root development. Front Plant Sci 2013;4:75.
[35] Le J, Mallery EL, Zhang C, Brankle S, Szymanski DB. Arabidopsis BRICK1/HSPC300 is an essential WAVE-complex subunit that selectively stabilizes the Arp2/3 activator SCAR2. Curr Biol 2006;16:895–901.
[36] Djakovic S, Dyachok J, Burke M, Frank MJ, Smith LG. BRICK1/HSPC300 functions with SCAR and the ARP2/3 complex to regulate epidermal cell shape in Arabidopsis. Development 2006;133:1091–100.
[37] Ikeuchi M, Yamaguchi T, Kazama T, Ito T, Horiguchi G, Tsukaya H. ROTUNDIFOLIA4 regulates cell proliferation along the body axis in Arabidopsis shoot. Plant Cell Physiol 2011;52:59–69.
[38] Ma J, Yan B, Qu Y, Qin F, Yang Y, Hao X, et al. Zm401, a short-open reading-frame mRNA or noncoding RNA, is essential for tapetum and microspore development and can regulate the floret formation in maize. J Cell Biochem 2008;105:136–46.
[39] Wanga DX, Lia CX, Zhao Q, Zhao LN, Wang MZ, Zhu DY, et al. Zm401p10, encoded by an anther-specific gene with short open reading frames, is essential for tapetum degeneration and anther development in maize. Funct Plant Biol 2009;36:73–85.
[40] Dong X, Wang D, Liu P, Li C, Zhao Q, Zhu D, et al. Zm908p11, encoded by a short open reading frame (sORF) gene, functions in pollen tube growth as a profilin ligand in maize. J Exp Bot 2013;64:2359–72.
[41] Tupy JL, Bailey AM, Dailey G, Evans-Holm M, Siebel CW, Misra S, et al. Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. Proc Natl Acad Sci U S A 2005;102:5495–500.
[42] Inagaki S, Numata K, Kondo T, Tomita M, Yasuda K, Kanai A, et al. Identification and expression analysis of putative mRNA-like non-coding RNA in *Drosophila*. Genes Cells 2005;10:1163–73.
[43] 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.
[44] Pi H, Huang YC, Chen IC, Lin CD, Yeh HF, Pai LM. Identification of 11-amino acid peptides that disrupt Notch-mediated processes in *Drosophila*. J Biomed Sci 2011;18:42.
[45] Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, et al. Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila embryogenesis*. Science 2010;329:336–9.
[46] Frank MJ, Cartwright HN, Smith LG. Three Brick genes have distinct functions in a common pathway promoting polarized cell division and cell morphogenesis in the maize leaf epidermis. Development 2003;130:753–62.
[47] Ghabrial A. Coding RNAs: separating the grain from the chaff. Nat Cell Biol 2007;9:617–9.
[48] Tour E, McGinnis W. Gap peptides: a new way to control embryonic patterning? Cell 2006;126(3):448–9.
[49] Joliot A, Prochiantz A. Transduction peptides: from technology to physiology. Nat Cell Biol 2004;6:189–96.
[50] Neijssen J, Herberts C, Drijfhout JW, Reits E, Janssen L, Neefjes J. Cross-presentation by intercellular peptide transfer through gap junctions. Nature 2005;434:83–8.
[51] Ludwig A-K, Giebel B, Exosomes:. small vesicles participating in intercellular communication. Int J Biochem Cell Biol 2012;44:11–5.
[52] Chua CEL, Lim YS, Lee MG, Tang BL. Non-classical membrane trafficking processes galore. J Cell Physiol 2012;227:3722–30.
[53] Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au WC, Yang H, et al. Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. Genome Res 2006;16:365–73.
[54] Kessler MM, Zeng Q, Hogan S, Cook R, Morales AJ, Cottarel G. Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome. Genome Res 2003;13:264–71.
[55] Velculescu VE, Zhang L, Vogelstein B. Serial analysis of gene expression. Science-AAAS-Weekly 1995.
[56] Basrai MA, Hieter P, Boeke JD. Small open reading frames: beautiful needles in the haystack. Genome Res 1997;7:768–71.
[57] Olivas WM, Muhlrad D, Parker R. Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. Nucleic Acids Res 1997;25:4619–25.
[58] Blandin G, Durrens P, Tekaia F, Aigle M, Bolotin-Fukuhara M, Bon E, et al. Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. FEBS Lett 2000;487:31–6.
[59] Brachat S, Dietrich FS, Voegeli S, Zhang Z, Stuart L, Lerch A, et al. Reinvestigation of the *Saccharomyces cerevisiae*

genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. Genome Biol 2003;4:R45.
[60] Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, et al. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science 2003;301:71–6.
[61] Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 2003;423:241–54.
[62] Hanada K, Zhang X, Borevitz JO, Li WH, Shiu SH. A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. Genome Res 2007;17:632–40.
[63] Fickett JW, Tung C-S. Assessment of protein coding measures. Nucleic Acids Res 1992;20:6441–50.
[64] Hanada K, Akiyama K, Sakurai T, Toyoda T, Shinozaki K, Shiu SH. sORF finder: a program package to identify small open reading frames with high coding potential. Bioinformatics 2010;26:399–400.
[65] Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, et al. Small open reading frames associated with morphogenesis are hidden in plant genomes. Proc Natl Acad Sci U S A 2013;110:2395–400.
[66] Yang XH, Tschaplinski TJ, Hurst GB, Jawdy S, Abraham PE, Lankford PK, et al. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome Res 2011;21:634–41.
[67] Guillen G, Diaz-Camino C, Loyola-Torres CA, Aparicio-Fabre R, Hernandez-Lopez A, Diaz-Sanchez M, et al. Detailed analysis of putative genes encoding small proteins in legume genomes. Front Plant Sci 2013;4:208.
[68] Ladoukakis E, Pereira V, Magny E, Eyre-Walker A, Couso JP. Hundreds of putatively functional small open reading frames in *Drosophila*. Genome Biol 2011;12:R118.
[69] Crappe J, Van Criekinge W, Trooskens G, Hayakawa E, Luyten W, Baggerman G, et al. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC Genomics 2013;14:648.
[70] Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 2009;324:218–23.
[71] Ingolia NT. Genome-wide translational profiling by ribosome footprinting. Methods Enzymol 2010;470:119–42.
[72] Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 2010;466:835–40.
[73] Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc 2012;7:1534–50.
[74] Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. Genome-wide annotation and quantitation of translation by ribosome profiling. Curr Protoc Mol Biol 2013;4.18:1–19. Chapter 4: Unit418.
[75] 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.
[76] Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science 2012;335:552–7.
[77] Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, Schumann F, et al. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res 2012.
[78] Lee S, Liu B, Huang SX, Shen B, Qian SB. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A 2012.
[79] Menschaert G, Van Criekinge W, Notelaers T, Koch A, Crappe J, Gevaert K, et al. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. Mol Cell Proteomics 2013;12:1780–90.
[80] Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009;458:223–7.
[81] Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs (vol. 28, pp. 503, 2010). Nat Biotechnol 2010;28:756–66.
[82] Brannan CI, Dees EC, Ingram RS. The product of the H19 gene may function as an RNA. Mol Cell Biol 1990;10(1):28–36.
[83] Cai XZ, Cullen BR. The imprinted H19 noncoding RNA is a primary microRNA precursor. RNA 2007;13:313–6.
[84] Li YM, Franklin G, Cui HM, Svensson K, He XB, Adam G, et al. The H19 transcript is associated with polysomes and may regulate IGF2 expression in trans. J Biol Chem 1998;273:28247–52.
[85] Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature 2012;482:339–46.
[86] Volders P-J, Helsens K, Wang X, Menten B, Martens L, Gevaert K, et al. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res 2013;41:D246–51.
[87] Struhl K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 2007;14:103–5.
[88] Banfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WEJ, et al. Long noncoding RNAs are rarely translated in two human cell lines. Genome Res 2012;22:1646–57.
[89] Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 2012;22:1775–89.
[90] Jackson RJ, Hellen CUT, Pestova TV. The mechanism of eukaryotic translation initiation and principles of its regulation. Nat Rev Mol Cell Biol 2010;11:113–27.
[91] Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell 2013;154:240–51.
[92] Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[93] Dinger ME, Gascoigne DK, Mattick JS. The evolution of RNAs with multiple functions. Biochimie 2011;93:2013–8.
[94] Kageyama Y, Kondo T, Hashimoto Y. Coding vs. non-coding: translatability of short ORFs found in putative non-coding transcripts. Biochimie 2011;93:1981–6.
[95] Oyama M, Kozuka-Hata H, Suzuki Y, Semba K, Yamamoto T, Sugano S. Diversity of translation start sites may define increased complexity of the human short ORFeome. Mol Cell Proteomics 2007;6:1000–6.
[96] Do JH, Choi D. Computational approaches to gene prediction. J Microbiol 2006;44(2):137–44.
[97] Sleator RD. An overview of the current status of eukaryote gene prediction strategies. Gene 2010;461:1–4.
[98] Cheng H, Chan WS, Li Z, Wang D, Liu S, Zhou Y. Small open reading frames: current prediction techniques and future prospect. Curr Protein Pept Sci 2011;12:503–7.

[99] Lin MF, Deoras AN, Rasmussen MD, Kellis M. Performance and scalability of discriminative metrics for comparative gene identification in 12 *Drosophila* genomes. PLoS Comput Biol 2008;4:e1000067.
[100] Lin MF, Jungreis I, Kellis M, PhyloCSF:. a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 2011;27:i275–82.
[101] Schwaid AG, Shannon DA, Ma J, Slavoff SA, Levin JZ, Weerapana E, et al. Chemoproteomic discovery of cysteine-containing human sORFs. J Am Chem Soc 2013;135(45):16750–3.
[102] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947–8.
[103] Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, et al. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res 2010;38: W695–9.
[104] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539.