Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation

Ariel A Bazzini1,*,†, Timothy G Johnstone1,‡, Romain Christiano2, Sebastian D Mackowiak3, Benedikt Obermayer3, Elizabeth S Fleming3, Charles E Vejnar1, Miler T Lee3, Nikolaus Rajewsky3,***, Tobias C Walther2 & Antonio J Giraldez1,4,***

Abstract

Identification of the coding elements in the genome is a fundamental step to understanding the building blocks of living systems. Short peptides (< 100 aa) have emerged as important regulators of development and physiology, but their identification has been limited by their size. We have leveraged the periodicity of ribosome movement on the mRNA to define actively translated ORFs by ribosome footprinting. This approach identifies several hundred translated small ORFs in zebrafish and human. Computational prediction of small ORFs from codon conservation patterns corroborates and extends these findings and identifies conserved sequences in zebrafish and human, suggesting functional peptide products (micropeptides). These results identify micropeptide-encoding genes in vertebrates, providing an entry point to define their function in vivo.

Keywords micropeptides; non-coding RNA; ribosome profiling; small ORFs; translation

Subject Categories Protein Biosynthesis & Quality Control; RNA Biology; Methods & Resources

DOI 10.1002/embj.201488411 | Received 6 March 2014 | Revised 13 March 2014 | Accepted 14 March 2014 | Published online 4 April 2014 | The EMBO Journal (2014) 33: 981–993

See also: SM Cohen (May 2014)

Introduction

Analysis of the genome has identified many putative transcripts that lack the classical hallmark of eukaryotic protein-coding genes: a single, long, conserved coding sequence (CDS) encoding a protein of more than 100 amino acids (Carninci et al, 2005; Birney et al, 2007; Tautz, 2009; Ulitsky et al, 2011; Derrien et al, 2012; Pauli et al, 2012). However, many of these transcripts (including lincRNAs) (Khalil et al, 2009; Guttman et al, 2010; Ingolia et al, 2011) contain multiple putative small open reading frames (smORFs, ≤ 100 aa) that can potentially be translated and thus might have a coding function (Ingolia et al, 2011; Chew et al, 2013; Slavoff et al, 2013). Recent examples have revealed functional, protein-coding smORFs across various genomes in RNAs previously thought to be non-coding (Savard et al, 2006; Magny et al, 2013). mille-pattes and tarsal-less/polished-rice were found to encode several micropeptides required during development in Tribolium and Drosophila, respectively (Savard et al, 2006; Kondo et al, 2007, 2010; Pueyo & Couso, 2008; Magny et al, 2013). Similarly, the predicted non-coding pncr003:2L gene encodes two micropeptides, each smaller than 30 aa, that regulate cardiac contraction in Drosophila (Magny et al, 2013).

Comprehensive identification of smORFs has been challenging and has mainly relied on evolutionary conservation (Stark et al, 2007; Lin et al, 2011), known patterns of codon occurrence and mass spectrometry (Schwaid et al, 2013; Slavoff et al, 2013). However, these approaches can be limited by the size, abundance and amino acid composition of the polypeptide. Ribosome footprinting measures translation by direct quantification of mRNA fragments protected by the 80S ribosome (ribosome-protected fragments, RPFs) after nuclease digestion (Fig 1A) (Wolin & Walter, 1988; Ingolia et al, 2009). Recent studies have used ribosome footprinting (Ingolia et al, 2009; Bazzini et al, 2012) to characterize the coding potential of different transcripts (Chew et al, 2013; Guttman et al, 2013) and identify translated protein-coding...
sequences (Ingolia et al., 2011; Brar et al., 2012; Michel et al., 2012; Stern-Ginossar et al., 2012; Chew et al., 2013; Crappe et al., 2013; Menschaert et al., 2013; Pauli et al., 2014). However, it has been questioned whether fragments recovered from ribosome profiling libraries always reflect a translating ribosome as opposed to regions protected by other RNA-binding proteins, or spurious binding to the ribosome (Guttman et al., 2013), and therefore, it is unclear how well these methods perform to identify individual smORFs. Unlike other interactions between mRNA, proteins, or scanning ribosomes, actively translating ribosomes have a unique property: the discrete movement along the message in three-nucleotide steps (phasing) (Ingolia et al., 2009; Guo et al., 2010; Michel et al., 2012), a feature that has been used to identify frame shifts and dually decoded regions in the genome (Michel et al., 2012). Since phased ribosome binding is a direct consequence of active translation, we reasoned that using phased binding as a criterion would reduce the noise in conventional ribosome profiling analyses and would allow us to identify smORFs undergoing translation in vivo.

Results

Ribosome footprinting in zebrafish with subcodon resolution

To define the coding potential of the transcriptome and identify translated smORFs, we analyzed the positional distribution of active ribosomes during zebrafish development. To this end, we generated high-depth ribosome footprinting with subcodon resolution across embryogenesis (at 2, 5, 12, 24 and 48 h post-fertilization, hpf) (Fig 1B). This resulted in approximately 200 million mapped reads after filtering for ribosomal RNAs, tRNAs and snoRNAs (Supplementary Table S1). 95% of reads within RefSeq protein-coding genes overlapped the CDS. Because phasing of the ribosome footprints can vary with fragment size, we first analyzed the distribution of RPFs within a composite RefSeq transcript (Supplementary Fig S1). Metagene analysis of the reads mapping to the annotated CDS revealed that 84.6% of the 28 and 29 nt RPFs were in-frame relative to their 5' ends (position 1 in the codon of the P-site, offset +12 nt), whereas
the RNA input fragments did not present any bias in their distribution (Fig 1C). The periodic distribution of RPFs observed along the CDS within each codon (Fig 1D, Supplementary Fig S1) reflects the stepwise translocation of active ribosomes (Ingolia et al., 2009). We reasoned that this pattern should derive from a biased in-frame distribution of RPFs within each individual CDS (Fig 1E). Thus, we hypothesized that this pattern could be used to define actively translated regions and distinguish them from background signal.

**ORFscore, a method to identify actively translated smORFs**

Guided by this hypothesis, we developed a method (ORFscore) that quantifies the biased distribution of RPFs toward the first frame of a given CDS (Fig 2). Given a putative ORF in the transcriptome (AUG to stop), we quantified the number of RPFs in each frame and determined whether RPFs were uniformly distributed or preferentially accumulated in one frame. We assigned a negative value to RPF distributions inconsistent with the frame of the ORF (Fig 2A). To filter ORFs with single or few codons covered by reads, we calculated the proportion of codons with in-frame reads (coverage) (Fig 2B). Next, we tested the predictive value of our methodology. Several lines of evidence suggest that our method identifies individual ORFs with coding potential. First, we analyzed all possible ORFs in annotated coding RefSeq transcripts. ORFscore was generally high across RefSeq CDS regions, with 85% of the expressed genes (> 1RPKM) having ORFs score ≥ 6.044 and coverage ≥ 10% (Fig 3A, C). In contrast, scores for most ORFs in the 5’UTR, the 3’UTR or overlapping the annotated CDS out of frame fell below these levels, reflecting their lack of coding potential (Fig 3A). Selecting the ORF with the highest ORFs score per transcript correctly identified the annotated CDS in 99% of the expressed coding transcripts, clearly distinguishing them from other possible ORFs in each transcript (P < 2.2e-16, Chi-squared test) (Fig 3B).

Annotated CDSs are usually longer than ORFs in the 5’- and 3’-UTRs (Supplementary Fig S2). To ensure that the correct identification of annotated CDSs is not simply due to this bias in size, we restricted the analysis to transcripts containing known coding regions ≤ 100 aa, using the same parameters (Fig 3D). This analysis identified 86% (208 out of 241) of short annotated CDSs, distinguishing them from 74,669 other putative short ORFs in expressed RefSeq coding transcripts (20-to-100 aa, P < 2.2e-16, Chi-squared test) (Fig 3E). Thus, combining frame bias and coverage provides a measure of coding potential that can be used to confidently identify small translated ORFs.

**Identification of novel smORFs by ORFscore**

To identify novel translated ORFs, we applied the ORFscore method to transcripts without defined coding sequences, including previously annotated long non-coding RNAs (Ulitsky et al., 2011; Pauli et al., 2012; Howe et al., 2013) and uncharacterized processed transcripts from Ensembl (Howe et al., 2013) (Fig 4A). In this analysis, Ensembl-annotated smORFs were used as a positive control (Fig 4A). Out of 2450 genes without previously defined coding sequences, many of which are thought to be non-coding, our analysis found experimentally supported coding ORFs in 303 genes. Of these, 214 (71%) encode smORFs between 20 and 100 aa long corresponding to 190 non-redundant smORF loci (Fig 4C and E, Supplementary Fig S3, Supplementary Table S2 and Supplementary File S1) and 89 (29%) encode for proteins longer than 100 aa. The majority of defined smORFs do not share significant amino acid sequence homology with known proteins in zebrafish (Fig 4D). An additional set of 53 non-redundant smORFs (Fig 4E) was identified after relaxing the coverage requirements while maintaining requirement for phasing of the ribosomes through the ORFs score. In contrast, 959 expression-matched transcripts lacked evidence for coding ORFs, including the known non-coding RNAs cyrano and megamin (Ulitsky et al., 2011; Chew et al., 2013) (Fig 4B, Supplementary Fig S4). Our analysis also provided experimental support for translation of 302 (52%) of the smORFs that were previously predicted by Ensembl and RefSeq (Howe et al., 2013) (Fig 4E) and distinguished them from size-matched ORFs in the 3’UTR that were used as control for non-coding regions (Fig 4A). Gene expression analysis revealed developmental regulation of mRNA levels for both smORF-containing and non-coding RNAs during embryogenesis (Fig 4K, Supplementary Fig S4). As an independent analysis, we
determined whether the polypeptide products from translated smORFs are detected by mass spectrometry (MS) (Supplementary Fig S5). We identified peptides for 98 annotated smORFs (~32% out of 302) and 6 novel smORFs (~3% out of 190) (Fig 4F), including those encoded by ENSDART00000145781 and linc-brsk1 (Fig 4I and J, Supplementary Fig S3). Identification of proteins by shotgun proteomics depends, among other factors, on protein and peptide lengths and abundances (Slavoff et al, 2013), which may explain why novel smORFs appear to be underrepresented in our recovered set, since they are shorter than previously annotated peptides (P = 1.6e-43, Wilcoxon test) (Fig 4G). Translated smORFs are also present in canonical protein-coding transcripts. Using ORFscore, we identify 311 (5'UTR) and 93 (3'UTR) translated ORFs, of which 17 and 10, respectively, were also identified by mass spectrometry (Fig 4H, Supplementary Fig S3). Future studies will be needed to further characterize the function of this large set of upstream and downstream ORFs, as they may regulate mRNA stability or translation of the main CDS (Barbosa et al, 2013). Taken together, these results reveal expression of several hundred smORFs present in transcripts with previously undefined coding sequences.

Figure 3. ORFscore discriminates translated from non-translated regions.
A–D Scatterplot of the ORFscore and coverage for all ORFs (A), the subset of ORFs with the highest ORFscore per transcript (B) and short (20–100 aa) annotated CDS (D). Relative density plots (scaled to the maximum value for each group) of the ORFscore and coverage are shown for each ORF type. Note the separation between annotated ORFs from the rest of the ORFs, even for short (20–100 aa) annotated CDS. (C) Color code used to label different ORF types found in Refseq protein-coding transcripts: annotated CDS (green), 5'UTR ORFs (purple), 3'UTR ORFs (red) and ORFs overlapping the annotated CDS (orange).
E Bar plots representing the number of ORFs identified on the basis of their ORFscore and coverage and defined as translated for each ORF type as in (C). Among all putative ORFs, the distribution of annotated ORFs was significantly different from the overall set (P = 2.2e-16, chi-squared test) with long and short CDS showing the highest fold-change enrichment in translated ORFs compared to other ORF types.
Computational prediction of smORFs from codon conservation patterns

Translation by ribosomes per se does not imply that a smORF encodes a functional peptide. For example, the peptide could be unstable or translation could function to regulate transcript stability. Moreover, a fraction of smORFs may have been missed using ribosome footprinting if they were not sufficiently expressed in the stages analyzed. Alternatively, conservation analysis can reveal evolutionary pressure to maintain the amino acid sequence of functional peptides. Thus, as an independent and complementary approach, we developed a computational pipeline (micropetide detection pipeline, micPDP) to search for smORFs and evaluate the evidence for negative selection on the encoded amino acid sequence from codon substitutions observed in whole-genome alignments. We filtered candidate alignments by coverage and reading frame conservation and then used phyloCSF (Lin et al., 2011) to score the coding potential from codon substitutions observed in whole-genome multiple alignments. We also used a simple parameter-free method (Ka/Ks) as control, which yields comparable results (Supplementary Fig S6). From published catalogs of zebrafish transcripts without annotated coding sequences, including lincRNAs (Ulitsky et al., 2011; Pauli et al., 2012; Howe et al., 2013), we evaluated 15,743 ORFs and from these predicted 63 conserved smORFs on 60 different transcripts (Fig 5A). 23 of these were also found by ORFscore ($P < 2e-22$, Fisher’s exact test) (Fig 5B). Note that only 45 of the experimentally detected smORFs in zebrafish had sufficient sequence alignment and could be scored by phyloCSF. Experimental and computational scores are correlated in zebrafish (Fig 5B and C), and smORFs score better by one method if they were detected by the other than if they were not ($P < 3e-15$, Mann–Whitney U-test, for both cases). Analysis of 33,961 human lincRNAs from Ensembl and RefSeq (Cabilly et al., 2011; Derrien et al., 2012) yields comparable results, predicting 173 smORFs on 160 different transcripts (Fig 5A and C). Using ORFscore to analyze previously published ribosome footprinting data in HeLa cells (Guo et al., 2010) (Supplementary Fig S6) defines 135 translated smORFs (118 unique loci) in human lincRNAs (Fig 5D, Supplementary Table S2 and Supplementary File S2) and a small overlap of seven candidates out of 95,780 smORFs with the computational results ($P < 6.3e-9$, Fisher’s exact test) (Fig 5C and E). Taken together, we identify hundreds of translated smORFs in human and fish and define an overlapping set of smORFs encoding evolutionarily conserved peptides.

Discussion

Our analysis of the zebrafish transcriptome using ribosome profiling provides two key insights into the genome-wide expression of smORFs in vertebrates. First, smORFs are widely distributed and are translated from a large body of transcripts, many of which were thought to lack coding potential. We experimentally identified hundreds of translated smORF regions that encode small proteins (micropeptides), defining 190 smORFs, 311 ORFs in the 5’UTR and 93 in the 3’UTR, and validated a portion of these by mass spectrometry. Previous studies have used ribosome footprinting to define classes of transcripts within non-coding RNAs based on the pattern of ribosome footprints when compared to known coding genes (Chew et al., 2013; Gutttman et al., 2013), but in most cases this classification does not define the translated ORF. In contrast to existing methods (Supplementary Fig S7) (Ingolia et al., 2011; Michel et al., 2012; Chew et al., 2013; Gutttman et al., 2013), the ORFscore leverages the periodicity of high-quality ribosome-protected fragments to define small translated ORFs (Supplementary Fig S7) independent of the surrounding sequence context in zebrafish and humans (Supplementary Fig S7) (Gutttman et al., 2013). While our method provides strong support for translation of individual ORFs using the parameters defined, we observe that relaxing the coverage cutoff recovers an additional set of ORFs that are defined with lower confidence but maintain strong phasing of the ribosome, suggesting active translation (Fig 4E; Supplementary Fig S6). Indeed, the Translated ORF Classifier designed by Chew et al. (2013) identifies a fraction of these transcripts as coding, providing a complementary method to define the coding potential of RNA (Supplementary Fig S6). The ORFscore strongly depends on the phasing of the ribosomes, therefore overlapping ORFs that are translated on different reading frames (Michel et al., 2012) can be missed by this method depending on the region of overlap for these ORFs; thus, future refinement will be necessary to define overlapping translated ORFs in zebrafish (Supplementary Fig S3). Applying the ORFscore analysis to previously published ribosome footprinting data in human cell lines (Guo et al., 2010) provides evidence for translation of smORFs present in human RNAs previously classified as non-coding (Supplementary Table S2). The presence of these small translated regions does not rule out a direct function of the mRNA transcript independent of the encoded peptide. Translation of small regions in these transcripts may be necessary for RNA function through localization, folding or triggering non-sense-mediated decay (Medenbach et al., 2011; Chew et al., 2013; Gutttman et al., 2013; Somers et al., 2013). Indeed, we have observed in lincRNAs a class of tiny ORFs (< 20 aa) supported by in-frame RPFs: For example, cyrano contains a 2aa ORF in zebrafish that displays high in-frame translation (Fig 4B). Due to the small size of these ORFs, further work will be needed to characterize their functions in vivo. Our method also defines a subset of transcripts with no evidence of translation, supporting a non-coding function for these transcripts.

Second, independent of the ribosomal footprint analyses, we developed a computational pipeline (micPDP) that identified a set of micropeptides which are likely under natural selection by computationally analyzing codon conservation patterns in multiple species alignments of annotated human lincRNAs and fish transcripts without previously defined coding sequences, including lincRNAs. Half of the translated micropeptides analyzed by ribosome footprinting with sequence alignment across species (23 out of 45) present strong patterns of evolutionary conservation (Fig 5B and C), supporting a functional role of these coding sequences. A small group of the identified smORFs (25%) have predicted homologs (e.g. RPL41) (Fig 4D), complementing current genome annotations (Howe et al., 2013). We note that in human, the number of micropeptides with good conservation scores, but low ORFscore was much higher than in zebrafish. This could be explained by lower depth and phasing of the human ribosomal profiling data used or by micropeptides expressed specifically in developmental stages and not in the human cell line. Further, the larger number of species in the human genome alignment allowed us to score smORF coding potential more comprehensively than in zebrafish. Still, smORFs might have different coding usage and conservation patterns.
Identification of small ORFs in vertebrates

smORFs encoding for lineage-specific or fast-evolving peptides, or smORFs with purely regulatory function, will be missed since our comparative method is tailored toward the identification of smORFs with conservation of their encoded amino acid sequence over larger
evolutionary distances. The conceptual differences between experimental and computational approaches lend strong support to jointly identified smORFs but translate into an inability to use one method to estimate false-positive or false-negative rate of the other. In sum, the computational results reported here reveal numerous smORFs in zebrafish and human and suggest that they have evolutionarily conserved sequence and function.

Central roles of small peptides have been known for decades based on the importance of neuropeptides, peptide hormones (such as secretin or insulin) and secreted signaling molecules (such as FGF1). However, the majority of these small peptides are encoded as secretin or insulin) and secreted signaling molecules (such as FGF1). However, the majority of these small peptides are encoded by our in silico or in vivo identified smORFs are biologically relevant. However, ELABELA/toddler (Chng et al., 2013; Pauli et al., 2014) (also identified in our set of smORFs; Supplementary Fig S3) are directly translated from small ORFs. Indeed, Pauli et al. (2014) recently reported the independent identification of more than 300 previously unannotated zebrafish proteins. Small proteins are also found in viral genomes, where transmembrane proteins (often shorter than 50 aa) have been shown to play vital roles in virus replication and virulence (Dimaio, 2014). It remains unclear how many of the peptides encoded by our in silico or in vivo identified smORFs are biologically relevant. However, ELABELA/toddler (Chng et al., 2013; Pauli et al, 2014), together with small peptides from other organisms and viruses (Savard et al, 2006; Galindo et al, 2007; Kondo et al, 2007, 2010; Pueyo & Couso, 2008; Magny et al, 2013), show that smORFs can play important roles across development and physiology. In sum, our identification of hundreds of translated smORFs significantly expands the set of micropeptide-encoding vertebrate genes providing an entry point to investigate their function in vivo.

Materials and Methods

RPF Library/Prep/SEQ

Ribosome-protected fragments: 50 embryos were collected for each time point and immediately frozen in liquid nitrogen. Each sample was lysed in 800 μl of Lysis buffer (1× Polysonse Buffer, 1% Triton X-100, 1 mM DTT, 25 U/ml DNasel and 100 μg/ml cycloheximide) according to manufacturer specifications (Epicentre, Artseq Ribosome Profiling Kit Mammalian, HRP1MR12126). The mix was clarified for 10 minutes (min.) at 10,000 g at 4°C. Three microliters of ARTseq Nuclease (Epicentre, HRP1MR12126) was added to 400 μl of the lysed supernatant and incubated at room temperature for 45 min with gentle mixing. Nuclease digestion was stopped with 15 μl of SUPERase-In™ RNase Inhibitor (Life Technologies, cat. no. AM2696) and chilled on ice. Ribosomes were purified by Sepahcryl S400 spin column chromatography (GE Healthcare, cat. no. 27-5140-01) or surcoose cushion ultracentrifugation according to manufacturer specifications (Epicentre, Artseq Ribosome Profiling Kit). The 400 μl of treated lysate was divided into four S400 spin columns (1 min/735 g). One hundred microliters of 1× Polysonse Buffer was added to the same columns and spin again (1 min/735 g). RNA from the 400 μl samples was extracted with Trizol, and RNA was subtracted using Rib0Zero (Rib0-Zero™ Magnetic Kit (Human/Mouse/Rat)) (Epicentre; cat no. MRZ11124), following the manufacturer’s protocol, omitting the 50°C incubation step. Ribosome-protected fragments (RPFs) were separated in a 15% Urea gel, and the region from 28- to 30- nucleotides was excised. RNA was eluted overnight in 300 mM NaOAc pH 5.5; 1 mM EDTA; 0.1 U/μl SUPERase In (Ambion #AM2694), followed by Ethanol precipitation.

RNA input

Total RNA was isolated from 400 μl of the 800 μl of clarified extract before ARTseq Nuclease treatment. Poly-A selection was done according to the manufacturer guidelines (Dynabeads mRNA Purification Kit, Cat no.610.06), and RNA was fragmented using the ARTseq Ribosome Profiling Kit Mammalian protocol.

Library preparation

Both RPF and RNA input fragments were cloned according to the ARTseq Ribosome Profiling Kit, Mammalian. The final PCR was carried out with an initial 15-s denaturation at 98°C, followed by 9-12 cycles of 15 s at 98°C, 5 s at 55°C and extension at 72°C for 10 s. Reactions were separated on a non-denaturing 8% polyacrylamide TBE gel, and DNA fragments of the correct size (113 ± 28~29 nt) were extracted and sequenced using an Illumina HiSeq 2000 sequencer.
Filtering and alignment of ribosome profiling reads

Base calling was performed using CASAVA-1.8.2. The Illumina TruSeq index adaptor sequence was then trimmed from raw reads by aligning its sequence, requiring 100% match of the first five base pairs and a minimum global alignment score of 60 (Matches: 5, Mismatches: 4, Gap opening: 7, Gap extension: 7, Cost-free ends gaps). Trimmed reads were then depleted of rRNA, tRNA, snRNA, snoRNA and misc_RNA from Ensembl and RepeatMasker annotations using strand-specific alignment performed with Bowtie2 v2.1.0 (Flicek et al, 2013) with default parameters. Filtered reads were finally aligned strand-specifically to the Zebrafish Zv9 (danRer7) genome assembly using Tophat v2.0.8 (Kim et al, 2013) with default parameters and the exon-junction coordinates from Ensembl r70. The same procedure was applied to the “input” reads.

The GEO GSE22004 filtered sequence set (Guo et al, 2010) was analyzed with the same procedure and the following modifications. The trimming required only five base pairs of the adaptor matching more than 80%. The filtering step used the tRNA from the Genomic tRNA database (Chan & Lowe, 2009) and Mt_rRNA, Mt_tRNA, rRNA, snoRNA, snRNA from Ensembl r73 non-coding RNAs. Sequencing reads were combined for the 12 and 32 h time points.

Defining transcript sets

To define annotated zebrafish transcripts, Ensembl transcripts (release 73) were downloaded from the Ensembl FTP repository and imported into R using the GenomicFeatures package. RefSeq transcripts were retrieved from UCSC and imported using the R GenomicFeatures package (Gentleman et al, 2004). To assemble a set of published lincRNA transcripts, supplemental data were downloaded from (Ulitsky et al, 2011) and (Pauli et al, 2012) and loaded into R using the RTrackLayer package. To define annotated human transcripts, the GenomicFeatures package was used to retrieve Ensembl (r73) transcript sets, and scaffolds were filtered out leaving...
Defining ORFs

Using the spliced version of each transcript, all possible stop codons in each of the three reading frames were defined. Each stop codon was paired with the most distal in-frame AUG codon without an intervening stop. Each region from most distal AUG to stop was thus defined as an ORF. ORFs were stored in both genomic and transcript-relative coordinates for further analyses. For position-related calculation of ORFscore and coverage of position 1, only reads within the ORF (defined as the region excluding the reads aligning to the start and stop (−1) codons) were counted.

Processing alignments/reads

Aligned reads were imported into R using the GenomicFeatures package. For downstream analysis, all sequencing experiment replicates were combined per time point. For “global” values, all time points were combined. RPF reads were filtered by size (28 & 29 nt). RNA expression was calculated in RPKM using corresponding input samples, using reads mapped to each transcript set as a total. For each transcript, all reads aligning to that transcript were mapped to transcript coordinates. For a given open reading frame, the position of each RPF read was designated as the +12 offset position of the read, corresponding to the P-site of the ribosome (Lawrence et al., 2009).

Calculating ORFscore

To calculate the ORFscore, reads were counted at each position within the ORF, excluding the first and last coding codons. To filter out putative artificial peaks, the most abundant read position was masked if reads aligning to that position comprised more than 70% of the total reads in the ORF. This filter was determined empirically by applying a variable filter and minimizing 3′UTR ORFs that were misclassified as coding based on such peaks (Supplementary Fig S1). The ORFscore was then calculated as:

\[
\text{ORFscore} = \log_2 \left( \frac{\sum_{i=1}^{3} (F_i - \bar{F})^2}{\bar{F}} \right) + 1
\]

where \(F_n\) is the number of reads in reading frame \(n\), \(\bar{F}\) is the total number of reads across all three frames divided by 3.

Calculating coverage

To calculate coverage in position 1, each position was considered covered if the P-site of at least one RPF read aligned to that position. For a given ORF, the coverage in position 1 is the resulting ratio of first frame positions covered versus all possible first frame positions in the ORF.

Calculating RRS

RRS (Ribosome Release Score) values for the 5 h time point were calculated to compare with ORFscore values. The RRS calculation and FindORF programs were provided by Gutmann et al (2013). To calculate the RRS score of annotated CDSs, a raw bed file was provided to the calculation program; to calculate the scores of other ORFs, the FindORF program was used to generate an input file. For coding genes, the program was run using various 3′UTR options (using either the whole 3′UTR or the minimal 3′UTR until next ORF start), resulting in a slight improvement in RRS score of CDS ORFs when using the whole 3′UTR, but similar clustering of results with both methods.

Calculating translation efficiency

Translation efficiency (TE) was calculated as the base 2 logarithmic ratio of normalized RPFs over the normalized input for each ORF (Ingolia et al, 2011; Bazzini et al, 2012).

Metagene analysis

Metagene plots were generated using R, taking read sets and CDS boundaries in transcript coordinates as input. For each RefSeq transcript with an annotated CDS, reads were counted at each position within two windows (surrounding the start and stop codons). These counts were normalized per transcript by dividing each position’s count by the sum of all reads in that window. These normalized counts were summed across all genes.

Defining coding ORFs

Translated ORFs were defined using the following parameters, in order to capture 85% of non-genome duplicated RefSeq annotated coding genes \(N = 9,559\): minimum size of 20 aa, ORFscore > 6.044, coverage of position 1 > 10% and expression greater than 1 RPKM at any time point. The less stringent criteria to define the relaxed set of translated ORFs used only ORFscore > 6.044.

Defining alternative (altORFs)

To define altORFs in 5′ and 3′ UTR, ORFs were excluded for which the genomic position of the start or stop codon overlapped any coding regions annotated by RefSeq or Ensembl. To define translated ORFs, see parameters above.

Defining Non-coding RNA

To define all the genes with no coding region (non-coding RNA), several filters were imposed to the transcript set described above. First, transcripts overlapping a coding exon were excluded. Second, all the transcripts from a genes with at least one ORF defined as coding (strict or relaxed) were excluded. Third, transcripts with coding potential predicted by Chew et al, 2013 were removed. Fourth, genes with RNA level lower than the first quartile of the genes encoding for defined smORFs (20–100 aa) were discarded. Finally, genes with any ORF that could encode for a protein larger than 100 aa were also excluded.
Defining translated small ORFs (smORFs)

Annotated smORFs were defined as all protein-coding ORFs annotated in RefSeq and Ensembl sized 20–100 aa that passed the same ORF score and position 1 coverage thresholds (6.044, 10%) mentioned previously. To define smORFs, these thresholds were applied to “processed transcripts” and lincRNAs from Ensembl in addition to the lincRNAs previously described in (Ulitsky et al., 2011) and (Pauli et al., 2012). Any ORFs that overlapped an annotated coding exon were excluded. smORFs residing in the same gene as an ORF > 100 aa defined as coding were also excluded. To avoid multiply counting isoforms/variants that encode for the same peptide, only one of these transcripts was selected when more than one ORF shared the same genomic start and stop coordinates and coding length. For final smORF counts, only one transcript per gene was counted.

Embryo preparation for MSMS detection

About 2,500 embryos were collected at 5 h and deyolked. Groups of 100 embryos were transferred from an agar coated dish to a 1.5-ml tube filled with 1 ml of deyolking buffer (55 mM NaCl, 1.8 mM KCl, 1.25 mM NaHCO3), the yolk sac was disrupted by pipetting with a 200 μl narrow tip, and the embryos were shaken for 5 min at 1,100 rpm to dissolve the yolk (Thermomixer, Eppendorf) at 4°C. Cells were pelleted at 300 g for 30 s and the supernatant was discarded. Subsequently, two additional washes were performed with 1 ml of wash buffer (110 mM NaCl, 3.5 mM KCl, 2.7 CaCl2, 10 mM Tris/Cl pH 8.5) shaking for 2 min at 1100 rpm and centrifugation. Cell pellets were frozen in liquid nitrogen.

MS sample preparation

To optimize detection of small proteins, unfractionated samples and small-protein-enriched fractions were prepared from embryo extracts (Supplementary Fig S5). Fractions were either digested or not using different enzymes (LysC, trypsin or GluC) to increase sequence coverage. Unfractionated samples were prepared following two different protocols: filter-aided sample preparation (FASP) (Winstead et al., 2009) and in-solution protein digestion. For FASP, approximately 300 frozen cells were boiled at 95°C in 5% (w/v) SDS, 100 mM DTT in 100 mM Tris/Cl pH 7.6 for 15 min and then submitted to FASP protocol as previously described (Frohlich et al., 2013). For this fraction only, trypsin (Promega) digestion and 4 h chromatographic runs were used.

For in-solution digestion, approximately 700 frozen cells were boiled at 95°C in lysis buffer 100 mM Tris/Cl (300 μl, pH 7.6). After cooling to room temperature, urea powder (160 mg) was directly added to the boiled embryos solution and cells were lysed using a Dounce homogenizer on ice. Samples were prepared for standard in-solution digestion protocol using LysC (Wako), trypsin (Promega) and GluC (Promega) according to the manufacturers’ recommendations.

The small-protein-enriched fractions were prepared from the rest of the embryos (~1,500) by acid extraction (Oyama et al., 2004). Briefly, frozen cells were boiled in water (600 μl) at 95°C for 15 min. After cooling to room temperature, cells were lysed in 1 M acetic acid using a Dounce homogenizer on ice. Precipitates and cell debris were pelleted by centrifugation at 17,000 g for 10 min. The extracted peptides were adjusted to pH 8 with NH4OH and then subjected to standard reduction/alkylation steps. Peptides were incubated for 30 min with DTT (1 mM) at room temperature and alkylated with iodoacetamide (55 mM) in the dark for 20 min. Peptides were submitted to size exclusion centrifugation using a 10-kDa filter unit (Millipore) at 14,000 g for 10 min. Resulting fractions “<~10 kDa” and “>~10 kDa” were dried in a Speedvac and resolubilized in NH4HCO3 (100 μl). One-fourth of each fraction was kept for direct MS analysis without further treatment and the remainder of the fractions was submitted to three proteolytic digests using LysC (Wako), trypsin (Promega) and GluC (Promega) according to the manufacturers’ recommendations.

Each peptide fraction was acidified by trifluoroacetic acid (TFA) and cleared of precipitates by centrifugation at 17,000 g for 5 min. Two micrograms of peptides was desalted following the protocol for StageTip purification (47). Prior to MS injection, samples were eluted with 70 μl buffer B (80% ACN, 0.1% formic acid in H2O) and reduced to a final volume of 5 μl in a Speedvac.

Chromatography and mass spectrometry

Each peptide fraction was separated by reversed-phase chromatography on a Thermo Easy nLC 1000 system connected to a Q Exactive mass spectrometer (Thermo) through a nano-electrospray ion source. Peptides were separated on 50-cm columns (New Objective) with an inner diameter of 75 μm packed in house with 1.9 μm C18 resin (Dr. Maisch GmbH). For FASP samples, 265-min chromatographic runs were used and peptides were eluted with a linear gradient of acetonitrile from 5 to 30% in 0.1% formic acid for 240 min at a constant flow rate of 250 nl/min. For the rest of the samples, 120-min chromatographic runs were used and peptides were eluted with a linear gradient of acetonitrile from 5 to 30% in 0.1% formic acid for 95 min at a constant flow rate of 250 nl/min. The column temperature was kept at 45°C. Eluted peptides were directly electrosprayed into the mass spectrometer. Mass spectra were acquired on the Q Exactive in a data-dependent mode to automatically switch between full scan MS and up to 5 (10 for FASP samples) data-dependent MS/MS scans. The maximum injection time for full scans was 10 ms (20 ms for FASP samples) with a target value of 3,000,000 at a resolution of 70,000 at m/z = 200. The five or ten most intense multiple charged ions (z ≥ 2) from the survey scan were selected with an isolation width of 2Th (3Th for FASP samples) and fragmented with higher energy collision dissociation (HCD) with normalized collision energies of 25. Target values for MS/MS were set to 10,000 with a maximum injection time of 120 ms at a resolution of 17,500 at m/z = 200. To avoid repetitive sequencing, the dynamic exclusion of sequenced peptides was set to 45 s for 265 min runs and to 35 s for 120 min runs.

MS data analysis

MS and MS/MS spectra were analyzed using MaxQuant (version 1.4.0.5), utilizing its integrated ANDROMEDA search algorithms (Cox et al., 2011). Scoring of peptides for identification was carried out with an initial allowed mass deviation of the precursor ion of up to 4.5 ppm for the search for peptides with a minimum length of
five amino acids. The allowed fragment mass deviation was 20 ppm. The false discovery rate (FDR) was set to 0.01 for proteins and peptides. Peak lists were searched against manually curated databases corresponding to all annotated RefSeq proteins as well as all smORF and altORF putative peptides encoded from all isoforms predicted by ribosome profiling combined with 262 common contaminants. Maximum missed cleavages were set to 4. The search included carbamidomethylation of cysteine as fixed modification, methionine oxidation, N-terminal acetylation and phosphorylation as variable modifications.

**ORF similarity analysis**

To analyze the similarity of newly detected smORFs to existing proteins, the amino acid sequences of all smORFs defined as coding by ORFscore and coverage were collapsed and input to NCBI BLASTp with start (Met) and stop codon not included. Default parameters were used, with species restricted to a number of model and studied organisms: *Danio rerio*, *Homo sapiens*, *Mus musculus*, *Xenopus tropicalis*, *Drosophila melanogaster*, *Rattus norvegicus*, *Strongylocentrotus purpuratus*, *Caenorhabditis elegans* and *Arabidopsis thaliana*. Results were manually curated and classified into one of the following categories:

- High similarity: 90% or more of the smORF aligns with score > 50 to known protein.
- High similarity; predicted: 90% or more of the smORF aligns with score > 50 to predicted protein.
- Partial similarity: < 90% of the smORF aligns with score > 50 to known protein.
- Partial similarity: predicted: < 90% of the smORF aligns with score > 50 to predicted protein.
- Non-significant: no part aligns with an alignment score > 50, but some hits are returned.
- None: no hits.

**Analysis of smORF coding signatures across species**

Computational prediction of conserved smORFs was based on published catalogs of lincRNAs together with whole-genome alignments from the UCSC browser (human: alignment of 45 vertebrates to hg19; zebrafish: alignment of seven vertebrates to danRer7). lincRNAs for human were taken from Gencode v18 (Derrien et al., 2012) and from Cabili and colleagues (Cabili et al., 2011). For zebrafish, lincRNAs from Ensembl (v73), as well as catalogs from Ullitsky et al., (2011) and Pauli et al., (2012), were used. The spliced sequence of lincRNAs was then scanned for the longest putative open reading frame longer than 63 nt (≥ 20 aa), with a canonical ATG start codon closest to the next upstream stop. ORFs in human and zebrafish which overlapped (sense or antisense) with coding exons from annotated genes from RefSeq (downloaded Nov. 14, 2013) or Ensembl (v73) were filtered out. For each smORF, the corresponding multiple alignment block (“stitched”) was then extracted from the whole-genome alignment. Sequences from species where less than 50% could be aligned, or with frameshift inducing insertions or deletions, were discarded, using an index for nucleotide insertions prepared from the original unstitched alignment blocks. Only smORFs where at least 50% of species with enough alignable sequence have no frameshift inducing indels were considered. These species were then used to calculate the phyloCSF score using the omega test (with –strategy = omega) (Pauli et al., 2012). A cutoff of 50 on the phyloCSF score was used (Pauli et al., 2012; Gutman et al., 2013). For ORF counts used in calculations, ORFs were filtered by unique genomic location.

As a control, the coding potential of smORFs was scored with a method that had not been previously used to annotate lincRNAs. A simple parameter-free empirical Ka/Ks calculation was implemented: Ka was defined as the number of non-synonymous mutations per non-synonymous site (adding pseudocounts equal to the fraction of non-synonymous sites in the sequence to numerator and denominator), and Ks similarly as the number of synonymous mutations per synonymous site (where pseudocounts equal to the fraction of synonymous sites are added). Ka/Ks ratios for pairs of species were summarized by taking the median. For significance estimation, the columns of the multiple species alignment of each putative smORF were permuted 1000 times. For each shuffled smORF, median Ka/Ks values were calculated as before. The empirical P-value was then defined as the percentage of shuffles with same or better median Ka/Ks value than the original smORF.

**Data deposition**

Sequencing data are deposited in GEO with accession number GSE53693.

**Supplementary information** for this article is available online: http://embj.embopress.org

**Acknowledgements**

We thank J. Weissman, N. Ingolia, M. Gutman and S. Kuersten for reagents, scripts and discussion, D. Cifuentes, C. Takacs and all the members of the Giraldez laboratory for intellectual and technical support and M. Hammarlund and A.F. Schier for comments on the manuscript. NR and SDM thank F. Payne for initial discussions and support as well as all members of the Rajewsky laboratory for stimulating feedback. This work was supported by T32GM007499 (TG), Pew Fellows Program in Biomedical Sciences (AAB), the Helmholtz Alliance on Systems Biology (MDC Systems Biology Network) and MDC funding (SDM), the Dobléck Fellows program of the MDC (BO), the Swiss National Science Foundation (CEV), R01GM081602-06 (AJC), R01GM105789-01 (AJC), R01HD074078-02 (AJC), F32HD071697-02 (MTL), R01GM095982 (TCW), R01GM097194 (TCW), the Pew Scholars Program in the Biomedical Sciences (AJC), the March of Dimes (AJC) and the Yale Scholars Program (AJC, TCW).

**Author contributions**

AAB and AJC conceived the project and together with TGJ and ESF designed and performed the experiments and data analysis. AAB, TGJ and AJC with input from MTL developed the ORFscore to define micropeptides using ribosome footprinting. TGJ and CEV analyzed the human data. AAB, TGJ and AJC wrote the paper with input from the other authors. AAB, RC and TCW designed and performed the MS experiments. SDM, BO and NR independently conceived, designed and implemented the computational micropeptide detection pipeline (micPDP).
Conflict of interest

The authors declare that they have no conflict of interest.

References


near-cognate translation initiation events. Mol Cell Proteomics 12: 1780–1790
Tautz D (2009) Polycistronic peptide coding genes in eukaryotes–how widespread are they? Brief Funct Genomic Proteomic 8: 68–74