Dna Research, 2016, 23(6), 571–580
doi: 10.1093/dnares/dsw040
Advance Access Publication Date: 24 August 2016
Full Paper

OXFORD

Full Paper

Ultra-deep sequencing of ribosome-associated poly-adenylated RNA in early *Drosophila* embryos reveals hundreds of conserved translated sORFs

Hongmei Li[1,†], Chuansheng Hu[2,†], Ling Bai[2,†], Hua Li[2], Mingfa Li[3], Xiaodong Zhao[2], Daniel M. Czajkowsky[2,*], and Zhifeng Shao[2,*]

[1]Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai 200240, China, [2]Bio-ID Center, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and [3]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

*To whom correspondence should be addressed. Tel: (86) 21-3420-5188; Fax: (86) 21-3420-6059; Email: zfshao@sjtu.edu.cn (Z.S.); Email: dczaj@sjtu.edu.cn (D.M.C.)

[†]These authors contributed equally to the article.

Edited by Dr Osamu Ohara

Received 9 March 2016; Accepted 11 July 2016

Abstract

There is growing recognition that small open reading frames (sORFs) encoding peptides shorter than 100 amino acids are an important class of functional elements in the eukaryotic genome, with several already identified to play critical roles in growth, development, and disease. However, our understanding of their biological importance has been hindered owing to the significant technical challenges limiting their annotation. Here we combined ultra-deep sequencing of ribosome-associated poly-adenylated RNAs with rigorous conservation analysis to identify a comprehensive population of translated sORFs during early *Drosophila* embryogenesis. In total, we identify 399 sORFs, including those previously annotated but without evidence of translational capacity, those found within transcripts previously classified as non-coding, and those not previously known to be transcribed. Further, we find, for the first time, evidence for translation of many sORFs with different isoforms, suggesting their regulation is as complex as longer ORFs. Furthermore, many sORFs are found not associated with ribosomes in late-stage *Drosophila* S2 cells, suggesting that many of the translated sORFs may have stage-specific functions during embryogenesis. These results thus provide the first comprehensive annotation of the sORFs present during early *Drosophila* embryogenesis, a necessary basis for a detailed delineation of their function in embryogenesis and other biological processes.

Key words: small open reading frames, sORFs, PhyloCSF, translatome, early Drosophila embryo

1. Introduction

Now that the genomes of many organisms have been sequenced, their comprehensive annotation is required to fully understand the functional elements that are encoded therein.[1–3] Although the focus of much of this annotation is of longer open reading frames (ORFs), there is growing appreciation that the much less studied small ORFs (sORFs), historically defined as <100 amino acids (aa) in length, may prove to be as widely significant in many biological processes as the larger class.[4–16] For example, the *tarsal-less* (*tal*) gene in *Drosophila*, which was previously thought to be a long non-coding RNA (lncRNA), has been found to contain four conserved sORFs encoding three 11-aa and one 32-aa peptides, that are required for embryonic tracheal development and leg morphogenesis.[5,13,17] Similarly, a conserved 56-aa peptide encoded by the *Toddler* gene in zebrafish has been found to function as a mitogen to promote cell migration during gastrulation.[18]

However, what has primarily hindered our appreciation of the extent to which the sORFs are biologically significant is, in fact, their reliable identification, the necessary first step before any detailed functional characterization.[19,20] In general, strictly bioinformatic approaches require a large number of experimentally validated ORFs to serve as a training set to enable subsequent *de novo* prediction, but there are presently not a sufficient number of identified sORFs for this purpose. Alternatively, ORFs can be identified based on conservation alone, but for that, reliable prediction generally requires sequences that encode peptides that are longer than 100-aa.[21–26] Direct identification of translated peptides in the cell by proteomic methods is also often highly effective, however these methods are well known to be inefficient at detecting proteins of small size.[10,27,28] Thus, it is highly likely that many sORFs have been misclassified as lncRNAs or missed entirely from present annotation.

Several attempts have been made to identify sORFs on a genome-wide scale in model organisms such as *Drosophila* using more targeted bioinformatics approaches.[29–32] Yet, while this work suggests that there may be thousands of sORFs translated in these organisms, there is presently a lack of experimental translational verification for the majority of the predictions. Perhaps the most successful experimental method to identify sORFs so far has been deep sequencing of ribosome-associated RNA.[10,31–33] However, recent work has suggested that ribosome occupancy alone is insufficient to unequivocally conclude that a potential sORF is indeed translated.[34,35] Instead, combining deep sequencing of ribosome-associated RNAs with bioinformatics analysis has emerged as a powerful approach to identify translated sORFs, genome-wide.[34–37]

Of the few well characterized sORFs, a surprisingly large fraction, like *tal* and *Toddler* mentioned above, has been found to play vital roles during development.[14,16–18,38,39] We thus speculated that there might be a large number of presently unannotated sORFs that perform critical functions during development, and that identifying the complete repertoire of translated sORFs during embryogenesis would prove both a useful strategy to identify a large set of sORFs to aid *de novo* sORF prediction as well as a necessary resource to understand this fundamental biological process.

Thus, to this end, we have performed ultra-deep sequencing of ribosome-associated poly-adenylated RNAs together with conservation analyses to identify conserved translated sORFs during the first 4 hours of embryogenesis in *Drosophila*, the period during which control shifts from maternal- to zygotic-encoded transcripts.[40] Overall, we have identified 399 sORFs that are translated during early embryogenesis, which substantially increases the number of verified sORFs in *Drosophila*. Of these, 128 were previously predicted sORFs but lacked experimental support, 22 are located in transcripts previously classified as lncRNAs, and 45 are novel sORFs found on transcripts not previously known to be transcribed. Further, among the sORFs that were previously annotated, we provide the first evidence of translation for sORFs with multiple isoforms. We tested the translational capability of randomly selected sORFs identified here in *Drosophila* S2R+ cell lines using an eGFP-tagged transfection assay, and found that most (22 out of 23) were highly translated, attesting to the validity of our combined experimental and bioinformatics approach. Thus, with this work, we provide the first comprehensive annotation of sORFs during early *Drosophila* embryogenesis, which we anticipate will aid our understanding of early development as well as the functions of sORFs in biological processes in general.

2. Materials and methods

2.1. *Drosophila* embryo collections

Early (0–4 h) Canton S embryos were collected from egg laying dishes. The embryos were then dechorionated by treatment with 50% bleach for 3–5 min and washed thoroughly with phosphate buffer solution (PBS buffer, pH7.4).[41] The embryos were then transferred into Eppendorf tubes.

2.2. Ribosome material preparation

To obtain the ribosome material, the dechorionated embryos were immediately incubated with 100 μg/ml cycloheximide in PBS for 5 min on ice, then the embryos were homogenized with a plastic pellet pestle in 100 μl of a mild ribosome lysis buffer (20 mM Tris-HCl, pH 7.4, 140 mM KCl, 5 mM $MgCl_2$, 0.5 mM DTT, 1% Triton X-100, 100 μg/ml cycloheximide, 0.5 U/ml RNasin) and incubated for 10 min on ice. The nuclei and whole cells were removed by centrifugation at 16,000 × g for 10 min at 4°C, and the lipid and other membranes were filtered out from the supernatant using a 100 μm sieve mesh. Finally, the cytosolic supernatant was loaded onto 20–50% continuous sucrose density gradients with the lysis buffer (20 mM Tris-HCl, pH 7.4, 140 mM KCl, 5 mM $MgCl_2$, 0.5 mM DTT), followed by ultracentrifugation at 35,000 rpm for 2.5 h in a SW41 rotor at 4°C (OptimaL-100 XP 2100 Ultracentrifuge, Beckman). Absorbance in each layer of the sucrose density gradient was measured at an optical density of 254 nm using the Piston Gradient Fractionater (BioComp). The RNA components and the amount of ribosomes were determined based on distinct peaks in the polysome profiling. Ribosome material was collected by pooling both the 80S (monosome) and the polysome peaks identified in the profile. The polysome profiling of the material treated with EDTA was obtained as above, except that the cytosolic supernatant was treated with 50 mM EDTA for 5 min on ice before loading onto the sucrose gradient.

2.3. Strand-specific RNA-seq library construction

The ribosome-associated RNA was extracted by adding an equal volume of Trizol reagent (Invitrogen) to the ribosome material, followed by chloroform extraction and ethanol precipitation. The RNA concentration was quantified by Nanodrop2000 (Thermo Scientific) and the RNA quality was detected by Agilent Bioanalyzer 2100 (Shanghai Biotechnology Corporation). The method to purify poly-adenylated RNA of ribosome-associated RNA was optimized using

the RiboMinus Eukaryote Kit for RNA-Seq (no. A10837-08, Ambion) to delete ribosomal RNAs (rRNAs) and Dynabeads oligo $(dT)_{25}$ (no. 61002, Life Technologies) purification to select RNAs with poly-adenylated tails (Supplementary Fig. S1). The strand-specific RNA-seq library of the ribosome-associated poly-adenylated RNA was prepared using the Illumina TruSeq Stranded mRNA Sample Preparation Kit (A10837-08, Ambion). The library was sequenced on the Illumina HiSeq 2000 (Shanghai Biotechnology Corporation) to a depth of about 70 M reads per library.

2.4. Preparation of total RNA and cytosolic RNA

Total RNA was isolated by homogenization of dechorionated embryos, followed by the RNA extraction protocol with Trizol. The total RNA was then treated with DNase I to eliminate genomic DNA contamination. Cytosolic RNA was isolated from the supernatant of the embryo extract before loading onto sucrose gradient. The enrichment of both total RNA and cytosolic RNA for poly-adenylated RNAs was also performed using Dynabeads oligo $(dT)_{25}$ purification. Both samples were sequenced as the ribosome-associated RNA above.

2.5. Transcript assembly

The ribosome-associated RNA, total RNA, and cytosolic RNA deep sequencing reads were each separately aligned using the TopHat v2.0.9 package.[42] We used a built-in strategy of TopHat for a higher mapping rate. Raw data were first mapped to the *Drosophila melanogaster* transcriptome (FlyBase r5.57 release) using the '-G' parameter with other parameters set as default. The read pairs that were not completely mapped were then aligned to the *D. melanogaster* genome (dm3) for a second run. Only the uniquely aligned and concordant read pairs were used for further analysis. We used Cufflinks v2.2.1 with parameters set as default except using '-G' parameter to assemble transcripts in the ribosome RNA, total RNA, and cytosolic RNA separately,[43] and merged the three assembled gtf files and Flybase r5.57 annotation into a comprehensive transcriptome annotation, from which all further analysis was based. The raw read counts were normalized by the FPKM (fragments per kilobase of exon model per million mapped fragments) for each transcript based on the Flybase r5.57 gene models. The FPKM values were calculated using the Cufflinks v2.2.1 package with parameters set as default.

2.6. Estimation of detectable level of transcription

A machine learning algorithm described by Ramsköld *et al.*[44] was first used to determine an optimal FPKM cutoff by comparing the expression levels of all annotated genes with that of randomly selected intergenic regions. The intergenic regions were at least 5 kb away from any annotated genes of FlyBase and the length distribution of the selected intergenic regions was the same as the distribution of the annotated exons to avoid a FPKM calculation bias. In this way, we identified a threshold FPKM value of 0.15 (Supplementary Fig. S2). To increase the confidence of the expression, we further required that the read counts be >20 reads. This value was determined based on an analysis of the annotated genes in our data of length between 0.2 and 0.6 kb, assuming that background reads for a gene would follow a geometric distribution.[45] Small RNAs (shorter than 200 nt) were excluded because these were not efficiently captured and would be more likely resulting in assembly artifacts. For translation detection, we also required that the transcripts exhibit an expression level of $\geq$ 0.15 FPKM and $\geq$ 20 reads in our ribosome-associated RNA data.

2.7. Reverse transcription polymerase chain reaction

Total RNA or ribosome RNA was reverse transcribed into cDNAs with SuperScript III Reverse Transcriptase (Invitrogen) and oligo $(dT)_{20}VN$ primer. cDNAs were used to amplify the RNA targets by polymerase chain reaction (PCR) using the internal gene-specific primers and DNA Taq polymerase (no. DR100A, Takara).

2.8. Evaluation of translational evidence for annotated sORFs

We considered as evidence of translation for a given annotated sORF as either: (i) the peptide fragment corresponding to the sORF was present in the most-recent proteomic screen (FlyBase r5.57 release); or (ii) the sORF was identified as a translated sORF in a previous ribosome profiling study in *Drosophila* S2 cells.[46]

2.9. Analysis of translated potential by PhyloCSF

To identify conserved translated sORFs, we utilized two annotated datasets to estimate the PhyloCSF threshold, comparing with 11 other *Drosophila* species: (i) the positive dataset was the annotated sORFs in FlyBase; and (ii) the negative dataset was all the sORFs contained in annotated lincRNAs based on the assumption that all of these sORFs are non-translated. From the histogram of the PhyloCSF score distribution in each set, we found an optimal PhyloCSF score of 50 which could discriminate annotated sORFs from untranslated sORFs (Supplementary Fig. S3). We then used the PhyloCSF package to evaluate the coding potential of ORFs contained in ribosome-associated lncRNAs and novel transcripts with parameters '12flies -orf = ATGStop –frames = 3 –minCodons = 10', which was intended to find all ORFs longer than 10-aa in frame. To avoid any influence of annotated ORFs, we excluded the ORFs which overlapped with annotated ORFs in the same or opposite strand. The multi-alignment file of 12 flies species were downloaded from the Galaxy cloud tool.[47–49]

2.10. Identification of embryo specific sORFs

For the annotated sORFs, we compared them directly with the translated sORFs identified in the S2 cell line. Due to the lack information of sORFs in lncRNA and novel transcripts, we manually examined for the presence of at least one read in the S2 ribosome data mapped to the identified sORF regions.

2.11. Calculation of arginine frequency of the identified sORFs

We calculated the arginine frequency by counting the number of arginines within all sORFs in each set. For the random control, we determined the expected frequencies of arginine based on that encoded from a random distribution of nucleotides.[50] For this calculation we used, as observed frequencies of the four DNA bases in nature, as 0.22 of uracil, 0.303 of adenine, 0.217 of cytosine, and 0.261 of guanine.[50]

2.12. eGFP-tagged sORF construction

The eGFP-tagged sORF vectors were based on the full-length cDNA of the corresponding sORFs. The full-length cDNAs were amplified by gene-specific full-length primers that introduced two different restriction enzyme digestion sites at the two ends. The cDNAs were then cloned into the pGEM T-easy vector, inserting an AvrII enzyme digestion site before the stop codon. The sequence of the eGFP

coding regions (CDS) which did not contain start or stop codons of the CDS was amplified-tagged with AvrII digestion sites at the two ends. CDS sequences of eGFP were digested by AvrII and cloned into the AvrII linearized sORF vector in-frame. These eGFP-sORF sequences were excised by double restriction enzyme digestion and directionally cloned into pUAST.

2.13. Transfections and immunoblotting

S2R+ cells were grown in Schneider's medium (no. 21720-024, Invitrogen) with 10% heat-inactivated fetal bovine serum (no. 16140-071, Gibco). S2R+ cells were transfected with reconstructed pUAST plasmid using X-tremeGENE HP DNA Transfection Reagent (no. 06366244001, Roche). After 48 h, proteins were extracted with RIPA Buffer (no. R0278, Sigma) containing protease inhibitor (no. 04693159001, Roche). The cell extract was then run in 12% Bis-Tris gels. Immunoblots were incubated with anti-GFP (1:1,000; no. M048-3, MBL) and then the secondary antibody Alexa Fluor 680 donkey anti-mouse lgG (1:1,000; no. A10038, Life Technologies). Controls were incubated with anti-α-tubulin (1:1,000; no. PM054, MBL) and then secondary antibody Alexa Fluor 680 goat anti-rabbit lgG (1:1000; no. A21076, Life Technologies).

2.14. Localization of eGFP-fusion peptide

After 48 hr post transfection, the cells were fixed for 10 min with 4% paraformaldehyde and then mounted in antifade mountant with DAPI (no. P36962, Life Technologies). Imaging was acquired using a Nikon A1Si confocal microscope with a CFI Plan Fluor 40 × objective.

3. Results and discussion

3.1. Ultra-deep sequencing identifies an exhaustive set of ribosome-associated RNAs

To globally identify translated sORFs in the 0–4 h *Drosophila* embryos, we performed ultra-deep sequencing of ribosome-associated poly-adenylated RNAs isolated using density gradient velocity sedimentation (see Materials and methods) (Fig. 1A).[41] Our ribosome profile revealed the presence of ribosomal subunits (40S and 60S) as well as monosomes (80S) and polysomes (Fig. 1B). To verify the identity of this ribosome material, prior to the sucrose gradient separation, we treated the embryo extract with 50 mM EDTA, which is known to dissociate intact ribosomes into their constituent subunits.[51,52] As expected, this treatment completely eliminated the peaks of the intact ribosomes in the profile, while the peaks associated with the 40S and 60S subunits increased significantly (Fig. 1C). For our analysis, we collected not only polysomes as is typical with this approach but also monosomes as well, with which some short transcripts are only associated.[53,54] Further, we purified only the poly-adenylated RNAs from this ribosome material, since non-poly-adenylated RNAs, a large proportion of the transcriptome, do not likely contain translated ORFs.[55] The ribosome-associated poly-adenylated RNAs were converted into cDNA libraries for strand-specific, paired-end 100 base-pair (bp) sequencing with HiSeq 2000.

Overall, we obtained a total of 71.2 million (M) aligned reads, of which 68.9 M (96.8%) were uniquely mapped to the *Drosophila* genome (Dm3), representing a ∼460-fold coverage of the *Drosophila* transcriptome. To enable the calculation of ribosome association efficiency (see below), we also deep sequenced the poly-adenylated RNAs from the total RNA population from the 0-4 hr *Drosophila* embryos in a similar way as described for the ribosome-associated RNAs. We also deep sequenced the cytosolic RNA from this same embryonic sample to maximize the annotation of the translated transcripts (see 'Materials and methods' section). In total, 213.9 M paired-end aligned reads were obtained in the combined dataset, of which 91.7% were uniquely aligned to Dm3, nearly 10-fold higher in depth than previously obtained transcriptomic datasets of this stage.[46,56,57] Following a procedure detailed in the Materials and methods, we finally identified 20,614 unique transcripts from 9,582 loci with high confidence in the 0–4 h *Drosophila* embryo (Supplementary Table S1). Comparing this with the latest release of the *Drosophila* transcriptome (FlyBase r5.57), we found that 18,613 transcripts (90.3%) are identical to the annotated transcripts, attesting to the high quality of our assembly. We further validated this assembly by randomly selecting a set of 18 of the novel transcripts using reverse transcription-PCR (RT-PCR) and found that 17 transcripts are indeed detected in the 0–4 h embryo (Supplementary Fig. S4 and Supplementary Table S3).

From among the 20,614 unique transcripts identified in the combined dataset, we found that 17,166 from 8,803 loci are ribosome-associated (Supplementary Table S1). We directly compared the FPKM values for each transcript in the ribosome RNA data with the corresponding average intensities measured in a previous microarray study of 0–2 h embryos,[41] and found a high degree of correlation (Spearman correlation coefficient 0.76). These ribosome-associated transcripts thus represent a large fraction of the *Drosophila* transcriptome (83.3%). We classified these transcripts into three categories based on the annotation in FlyBase: protein-coding transcripts including novel variant isoforms (16,576), lncRNAs that also include novel variant isoforms (349), and assembled novel transcripts (241). We validated the ribosome association of 35 randomly selected transcripts in the latter two categories with RT-PCR, and found that 34 transcripts could indeed be detected in the 0–4 h embryo (Supplementary Fig. S5 and Supplementary Table S3).

3.2. Ribosome association provides translational evidence for annotated sORFs

The 16,576 ribosome-associated protein-coding transcripts correspond to 11,092 different ORFs, including 332 sORFs (Supplementary Table S2). Of the latter, inspection of FlyBase revealed that there was prior evidence of translation for only 204 annotated sORFs (Fig. 2A).[46] Thus, our data provides the first necessary translational evidence for 128 annotated sORFs, substantially increasing the number of sORFs in *Drosophila* with evidence of translation. We note that the well-studied functional sORFs in *Drosophila*, *tal* and *scl*, were highly enriched in our ribosome-associated fraction (Fig. 2B), lending further confidence in the quality of our data.

Much of this previous translational evidence for the annotated sORFs was obtained from *Drosophila* S2 cells, which are derived from late stage embryos (20–24 h).[46,58] Of the 332 sORFs in our data, 148 are not translated in the S2 cell line. Though cultured cells can exhibit phenotypes different from the original cells from which they are derived, these results indicate that a substantial fraction of the sORFs might be translated specifically during defined developmental stages. This comparison also suggests that a detailed characterization of the ribosome-associated RNAs at the other stages of development might uncover evidence for the translation of other sORFs.

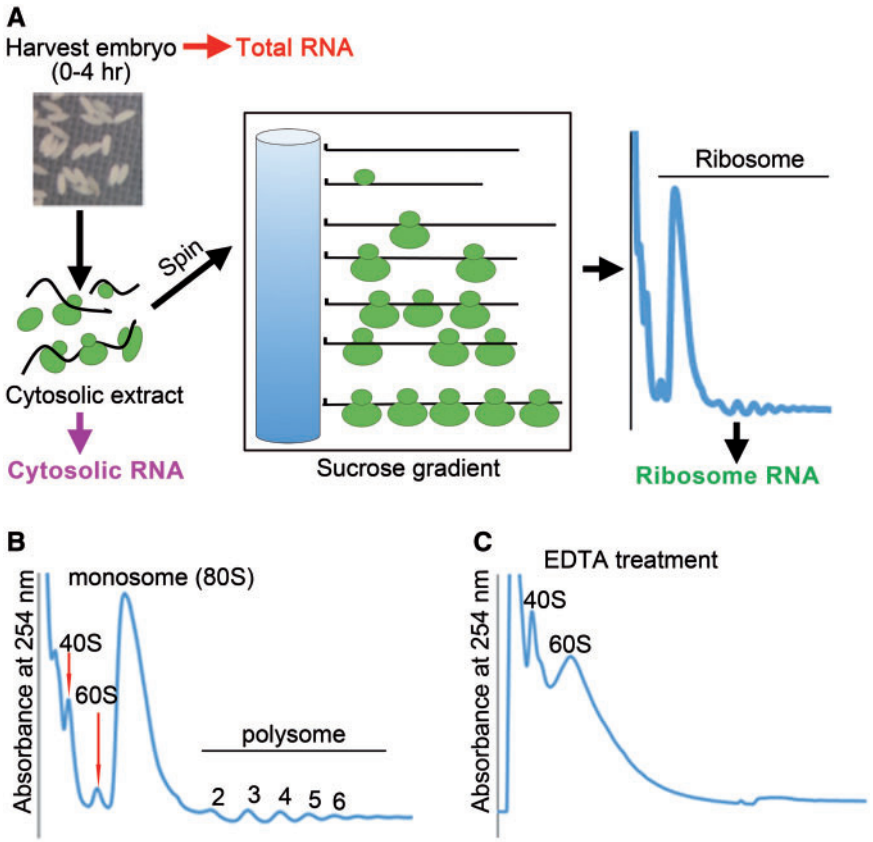

Figure 1. Isolation procedure of ribosome-associated RNA. **(A)** Schematic procedure for the preparation of ribosome material from 0-4 hr *Drosophila* embryos. **(B)** Polysome profiling of this sample enables clear resolution of the monosome and polysome fractions that were isolated for deep sequencing of the ribosome-associated RNA. **(C)** Validation of the identification of the monosome and polysome peaks. As expected, treating the sample with 50 mM EDTA before loading the sample onto the sucrose gradient caused the intact ribosomal peaks to disappear and those of the 40S and 60S subunits to increase.

One advantage of our experimental approach is that we obtain the identity of the full-length of the ribosome-associated transcripts, enabling the detection of specific gene isoforms. Of the 332 different sORFs (corresponding to 313 genes), there were 17 genes with two different sORFs and 1 gene with three different sORFs. In addition, there were 177 genes with more than one annotated isoforms and 86 genes with variant isoforms not previously described. This is, in fact, the first detection of variant isoforms of translated sORFs. For example, with the gene *CG40228*, we found five isoforms with (in total) 3 different sORFs, including a variant isoform not previously characterized with a unique 5′ untranslated region (UTR) that is 19 bp upstream of all other isoforms. Variant isoforms in longer ORFs are usually associated with their highly regulated, differential translation,[59,60] and so this observation of variant sORF isoforms suggests that the translational processing of the sORFs may be as complex as that governing their longer counterpart. Ribosome profiling might prove useful in this regard, since a recent study employing this method thoroughly characterized stop codon readthrough in many genes in early *Drosophila* embryos.[61]

3.3. Ribosome-associated RNA sequencing identifies translated sORFs among the lncRNAs

Transcripts that do not encode peptides and lack a 100-aa ORF are usually classified as lncRNAs.[62,63] Although a number of these transcripts have been found to indeed function in a non-coding capacity,[64–67] it remains to be determined whether at least some of these transcripts are misclassified and actually encode sORFs.[20,68] In this regard, it is interesting that we have found 349 lncRNA transcripts associated with 264 genes that are associated to ribosomes, which corresponds to 76.9% of the expressed lncRNAs in these early embryos (Fig. 2C), an amount that is consistent with previous findings in different species.[69]

We reasoned that if these lncRNAs actually encode for peptides, then they may be associated to ribosomes to a similar extent as established protein-coding genes. We thus evaluated the degree to which these lncRNAs and the protein-coding genes are associated with ribosomes compared with their total level of expression (their 'ribosome association efficiency'). We found that, although these ribosome-associated lncRNAs are expressed at a much lower level than the protein-coding genes (Wilcoxon test, P–value $< 2.2e\text{-}16$) (Fig. 3A), their ribosome association efficiency is not significantly different from that of the protein-coding genes (Wilcoxon test, P–value > 0.05) (Fig. 3B). Thus, these apparently lncRNA transcripts are indeed associated to ribosomes to a similar extent as bona fide protein-coding transcripts.

Although it is likely that many of these transcripts are indeed non-coding,[70–72] we reasoned that at least some of these transcripts may encode sORFs and thus examined these transcripts with a bioinformatics approach for potential sORFs. In particular, we first identified ORFs with an ATG start codon and an in-frame stop codon, using the longest ORF for each stop codon. We then discarded those ORFs

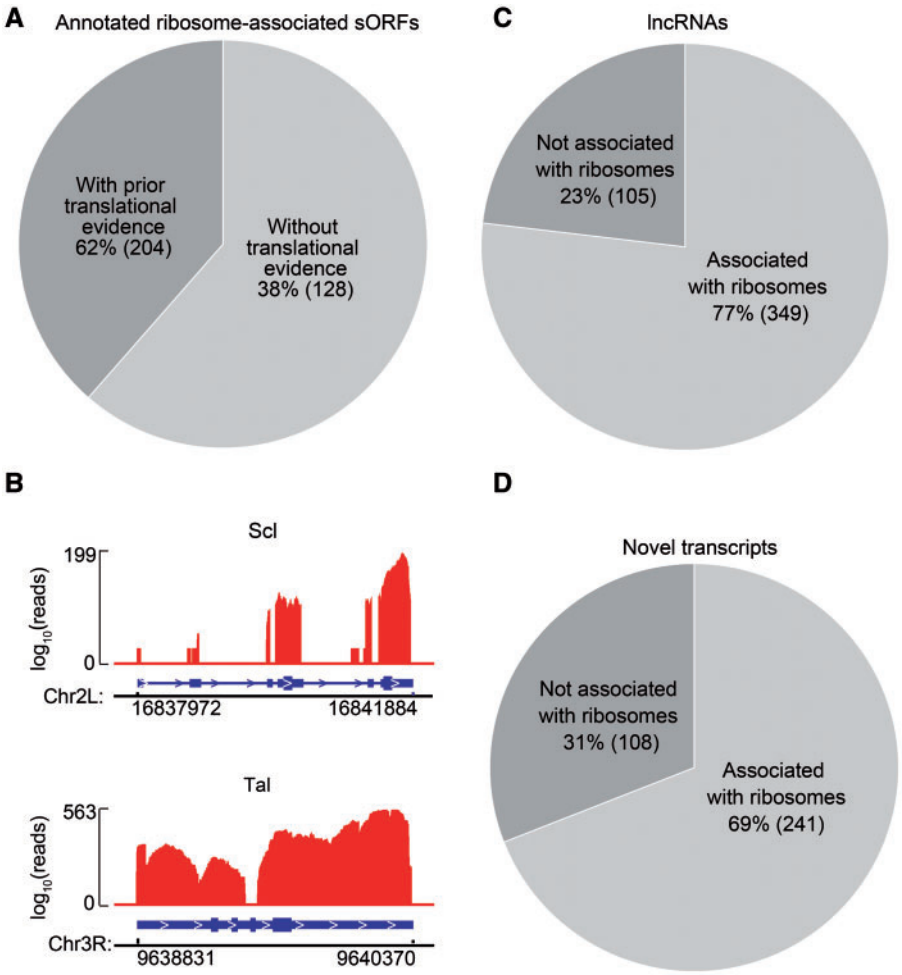

Figure 2. Ribosome-associated sORFs, lncRNAs and novel transcripts. **(A)** Among the annotated sORFs associated with ribosomes, we provide evidence of translation for 128 previously predicted *Drosophila* sORFs. **(B)** The well-studied *Drosophila* genes, *tal* and *scl* which were previously thought to be lincRNAs but then shown to encode functional sORFs, are well-resolved in our ribosome-associated RNA data. **(C)** Proportion of expressed lncRNAs that are ribosome-associated. **(D)** Proportion of expressed novel transcripts that are ribosome-associated. For both the lncRNAs and the novel transcripts, there was a large fraction of the total number of expressed transcripts that were found to be associated to the ribosomes.

that encode for peptides smaller than 10-aa or that overlapped annotated ORFs in the same or opposite strand. In this way, we identified 1,784 putative sORFs (median length of 20-aa) and 9 potential long-ORFs in 347 out of the 349 ribosome-associated lncRNAs (Fig. 3C). As a stricter criterion, we examined the conservation at the amino-acid level of these putative sORFs using PhyloCSF, which has been demonstrated to be a highly effective method to identify sORFs.[69,70] Using annotated sORFs and lincRNAs to establish rigorous threshold values in this approach (see 'Materials and methods' section), we identified 28 ORFs located in 21 ribosome-associated lncRNA genes as conserved translated ORFs. Of these, 22 are sORFs (Supplementary Table S2), including 3 that are poly-cistronic like *tal* and *scl*.[5,16,17] We note that among these 22 novel sORFs, 64% are not found in the late-stage *Drosophila* S2 cell line, and thus may be specific for early embryos. The finding of a fair number of ribosome-associated lncRNAs that lack coding potential is intriguing and remains to be resolved. Based on the current understanding, one might speculate that they may play roles in RNA localization, RNA nonsense mediated decay, translational regulation, or they may produce non-canonical proteins that are quickly degraded.[71–73] Alternatively, these lncRNAs may encode proteins with non-AUG start codons.[74,75]. However, one should also not exclude the possibility that a further analysis would yield functions that are not yet known.

3.4. Ribosome-associated RNA sequencing identifies translated sORFs in novel transcripts

Of the 350 assembled novel transcripts identified in this study, 241 are associated with the ribosomes (Fig. 2D). Similar to the ribosome-associated lncRNAs, we found that these ribosome-associated novel transcripts are as tightly associated to the ribosomes as the protein-coding genes (Wilcoxon test, P-value > 0.05) (Fig. 3B), although their expression level is much lower than those known to encode for peptides (Wilcoxon test, P-value $= 3.638\text{e-}12$) (Fig. 3A). Examining these transcripts for potential ORFs similarly to the lncRNAs described above, we identified 2,521 putative ORFs, most of which (98.5%) were sORFs (median length of 24-aa) (Fig. 3C). Analysing these putative ORFs in terms of their conservation using PhyloCSF, we identified 66 different conserved translated ORFs contained within 32 of these novel transcripts (Supplementary Table S2). Of these, 45 are sORFs, most (87%) of which were present only in the early embryos and not in the S2 cell line.

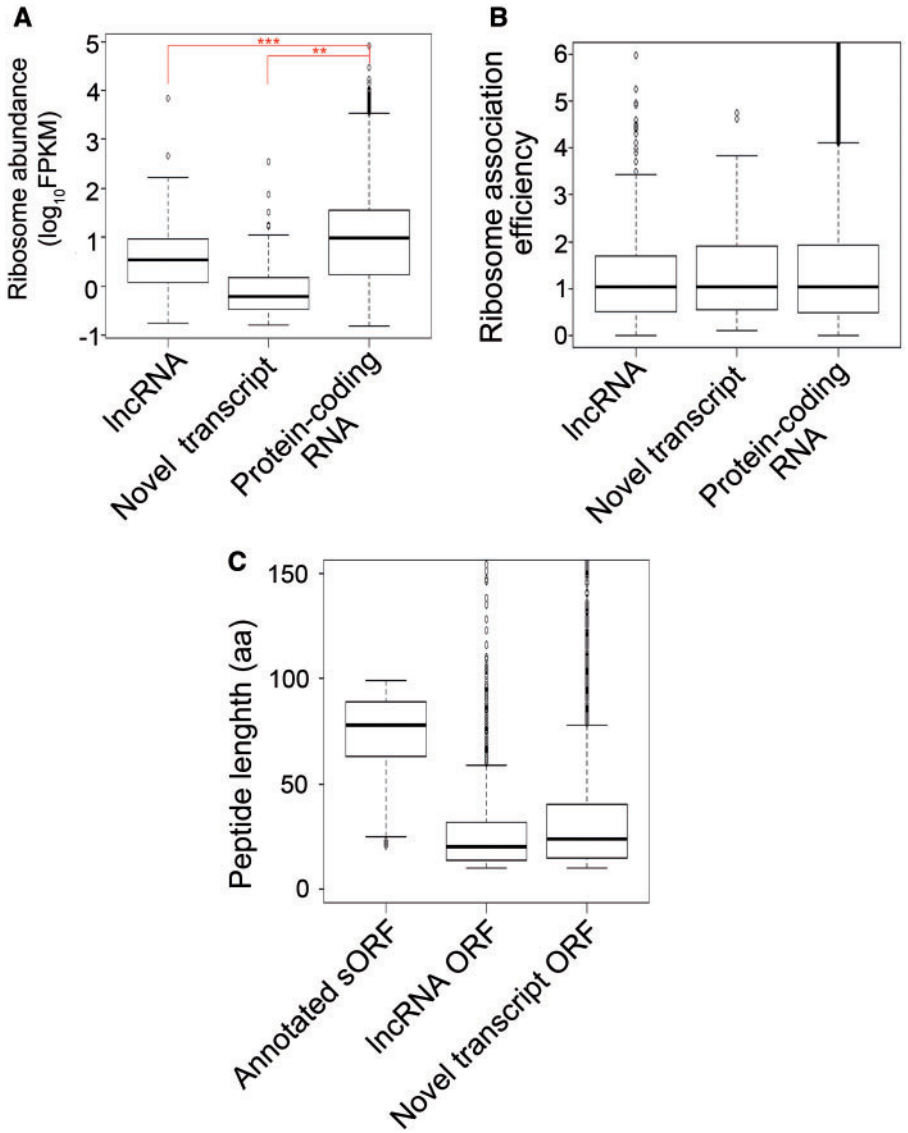

Figure 3. Characterization of the ribosome-associated lncRNAs and novel transcripts. **(A)** There was a lower abundance of the lncRNAs and novel transcripts on the ribosomes than the protein-coding RNAs. (*P*-value < 2.2e-16 (***); *P*-value = 3.638e-12 (**); Wilcoxon rank-sum test). **(B)** Despite their lower ribosome occupation, the ribosome association efficiency of the lncRNAs and novel transcripts was not significantly different from that of the protein-coding RNAs (all *P*-values > 0.05 relative to the protein coding RNA). Here, the ribosome association efficiency is defined as the ratio of the abundance of the ribosome-associated RNA to the total RNA. **(C)** Overall, the length of ORFs contained within these lncRNAs and novel transcripts are, in general, shorter than the annotated sORFs.

3.5. Translational validation of identified translated sORFs

As a low frequency of arginine occurrence is a common feature of proteins, and has been used as an indicator of the translational capacity of potential ORFs,[46] we compared the arginine usage of our identified translated ORFs with that expected from the aa frequencies associated with randomly distributed nucleotides.[50] We indeed found that, like the annotated sORFs, the novel translated ORFs in both the previously identified lncRNAs and in the novel transcripts exhibit a much lower usage of arginine than this random distribution, consistent with the notion that these are indeed translatable sORFs (Fig. 4A).

To provide additional support for this translational capacity, we examined the ability of 23 randomly selected sORFs, including 15 annotated sORFs without evidence for translation, 4 sORFs in lncRNAs, and 4 sORFs in novel transcripts, to be translated in *Drosophila* S2R+ cells (Supplementary Table S3). We generated eGFP-fusion vectors that contained all of the translation-related elements of the sORF, including the 5′UTR and 3′UTR, together with the enhanced green fluorescent protein (eGFP) coding sequence (CDS) in-frame following the sORF (Fig. 4B). Thus, translation of this eGFP-tagged sORF would produce an eGFP-fusion protein, for which we examined using Western blotting and fluorescence microscopy. Overall, we found that 22 of the 23 candidates were well translated (Fig. 4C–E and Supplementary Fig. S6). Of these, 3 were clearly localized in both the nucleus and cytoplasm as observed with eGFP control (Supplementary Fig. S6), while the other 19 were mainly localized in cytoplasm (Fig. 4C–E and Supplementary Fig. S6). Of note though, the cytoplasmic localization did not appear to be the same for all of the fusions, with some enriched in a single, large subsection

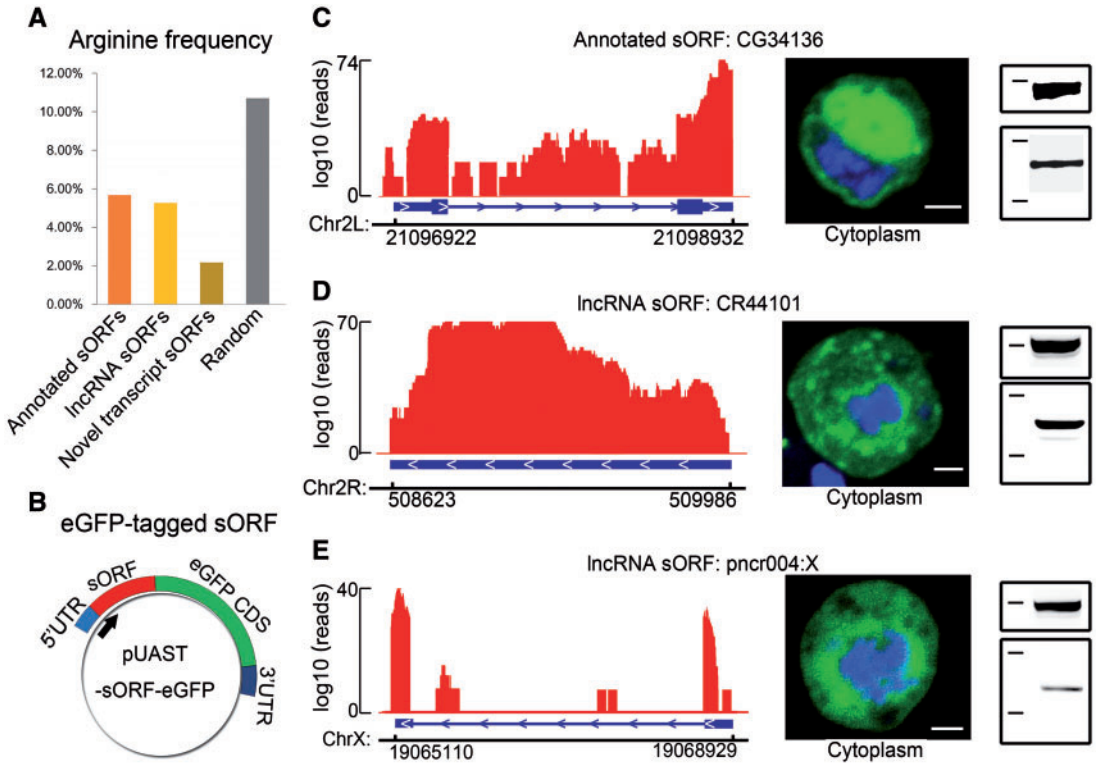

Figure 4. Translation validation of identified sORFs. **(A)** The novel identified translated sORFs contained in lncRNAs and novel transcripts have a much lower arginine frequency compared with that expected from random sequences of nucleotides, similar to annotated sORFs. **(B)** Schematic of the transfection construct of the eGFP-tagged sORFs. **(C–E)** Translational validation of CG34136, CR44101 and pncr004:X. For each of the fusions, the number of reads in the ribosome-associated RNA data of the corresponding sORF is shown in the left panel, whereas a typical image from fluorescence microscopy (DAPI: DNA; eGFP: translated sORF) and the results from the Western blot are presented in the middle and right panels, respectively. In each Western DNA, the upper panel is the α-tubulin control detected with an anti-α-tubulin antibody and the molecular weight standard corresponds to 49 kDa. The lower panel in each Western blot is the fused-sORF detected with anti-eGFP and the molecular weight standards refer to 38 kDa (top) and 28 kDa (bottom). The expected molecular weights of the sORFs are 36.6, 30.8, and 30.3 kDa, respectively. Scale bar: 2.5 μm.

of the cytoplasm (Fig. 4C and Supplementary Fig. S6d, e, h, i, s), others with a more punctate distribution scattered throughout the cytoplasm (Fig. 4D and Supplementary Fig. S6c, f, g, o), and others with a more uniform distribution within the cytoplasm except for punctate locations (Fig. 4E and Supplementary Fig. S6f, l, m, n, p, q, t, u). Such a wide range of localizations is thus likely owing to the sORF-encoded peptide and not the eGFP, since the latter is present in all of the fusions (Supplementary Fig. S6b). Taken together, these results indicate that most of the identified translated sORFs could indeed be translated into peptides *in vivo*.

4. Conclusion

In this study, we provide the first genome-wide annotation of the translated sORFs population that is present during the very early stages of *Drosophila* embryogenesis, thus setting the stage for detailed characterizations of their functions during this fundamental biological process. The 399 sORFs identified here significantly expands the population of known sORFs in this model organism, which we anticipate will aid in future bioinformatics approaches for *de novo* predictions of sORFs both in *Drosophila*, as well as in much less well studied organisms, such as humans.[10,76] Determining if their translation is indeed as complex as the longer ORFs, or if they form and evolve by mechanisms distinct from their longer counterparts, or indeed if their spectrums of biological functions are as diverse as the longer ORFs will be fascinating to now resolve.

5. Availability

RNA-seq data have been submitted to the EMBL with the accession numbers E-MTAB-4571.

Acknowledgements

The authors thank Professor Nan Liu from the Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, for providing the S2R+ cell line and vectors. We also thank Professor Andrew C. B. Cato from the Karlsruhe Institute of Technology, Institute of Toxicology and Genetics, for providing the pUAST vector.

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by the Ministry of Science and Technology of China (2014YQ090709 and 2013CB967402); the National Natural Science Foundation of China (91129000, 11374207, 91229108, 31370750, 21273148, 31501054, 91529302, and 21303104); Science and Technology Commission of

Shanghai Municipality (15142201200), Shanghai Jiao Tong University (YG2012MS58), and the K.C. Wong Education Foundation (H.K.).

References

1. Harrow, J., Frankish, A., Gonzalez, J. M., et al. 2012, GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, **22**, 1760–74.
2. Roy, S., Ernst, J., Kharchenko, P. V., et al. 2010, Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, **330**, 1787–97.
3. Landry, C. R., Zhong, X., Nielly-Thibault, L. and Roucou, X. 2015, Found in translation: functions and evolution of a recently discovered alternative proteome. *Curr. Opin. Struct. Biol.*, **32**, 74–80.
4. Zanet, J., Chanut-Delalande H., Plaza, S. and Payre, F. 2016, Small peptides as newcomers in. –The control of *Drosophila* development. *Curr. Top. Dev. Biol.*, **117**, 199–219.
5. Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S., Hayashi, S. and Kageyama, Y. 2007, Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat. Cell Biol.*, **9**, 660–5.
6. Slavoff, S. A., Heo J., Budnik, B. A., Hanakahi, L. A. and Saghatelian, A. 2014, A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J. Biol. Chem.*, **289**, 10950–7.
7. Guo, B., Zhai, D., Cabezas, E., et al. 2003, Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature*, **423**, 456–61.
8. Crappé, J., Van Criekinge, W. and Menschaert, G. 2014, Little things make big things happen: a summary of micropeptide encoding genes. *EuPA Open Proteom.*, **3**, 128–37.
9. Rohrig, H., Schmidt J., Miklashevichs, E., Miklashevichs J., Schell J. and John, M. 2002, Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proc. Natl. Acad. Sci. USA*, **99**, 1915–20.
10. Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., et al. 2013, Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.*, **9**, 59–64.
11. Okamoto, M., Higuchi-Takeuchi, M., Shimizu, M., Shinozaki, K. and Hanada, K. 2014, Substantial expression of novel small open reading frames in *Oryza sativa*. *Plant Signal. Behav.*, **9**, 2395–2400.
12. Saghatelian, A. and Couso, J. P. 2015, Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.*, **11**, 909–16.
13. Kondo, T., Plaza S., Zanet, J., Benrabah, E., et al. 2010, Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science*, **329**, 336–9.
14. Hanyu-Nakamura, K., Sonobe-Nojima, H., Tanigawa, A., Lasko, P. and Nakamura, A. 2008, *Drosophila* Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. *Nature*, **451**, 730–3.
15. Hanada, K., Higuchi-Takeuchi, M., Okamoto, M., et al. 2013, Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc. Natl. Acad. Sci. USA*, **110**, 2395–400.
16. Magny, E. G., Pueyo, J. I., Pearl, F. M., et al. 2013, Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science*, **341**, 1116–20.
17. Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A. and Couso, J. P. 2007, Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.*, **5**, e106.
18. Pauli, A., Norris, M. L., Valen, E., et al. 2014, Toddler: an embryonic signal that promotes cell movement via apelin receptors. *Science*, **343**, 1248636.
19. Pauli, A., Valen, E. and Schier, A. F. 2015, Identifying (non-)coding RNAs and small peptides: challenges and opportunities. *BioEssays*, **37**, 103–12.
20. Kageyama, Y., Kondo, T. and Hashimoto, Y. 2011, Coding vs noncoding: Translatability of short ORFs found in putative non-coding transcripts. *Biochimie*, **93**, 1981–6.
21. Chu, Q., Ma, J. and Saghatelian, A. 2015, Identification and characterization of sORF-encoded polypeptides. *Crit Rev Biochem Mol*, **50**, 134–141.
22. Sleator, R. D. 2010, An overview of the current status of eukaryote gene prediction strategies. *Gene*, **461**, 1–4.
23. Brent, M. R. and Guigo, R. 2004, Recent advances in gene structure prediction. *Curr. Opin. Struct. Biol.*, **14**, 264–72.
24. Wang, J., Li, S., Zhang, Y., et al. 2003, Vertebrate gene predictions and the problem of large genes. *Nat. Rev. Genet.*, **4**, 741–9.
25. Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P. and Li, W. 2013, CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74.
26. Kong, L., Zhang, Y., Ye, Z. Q., et al. 2007, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–9.
27. Andrews, S. J. and Rothnagel, J. A. 2014, Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.*, **15**, 193–204.
28. Fälth, M., Sköld, K., Norrman, M., Svensson, M., Fenyö, D. and Andren, P. E. 2006, SwePep, a database designed for endogenous peptides and mass spectrometry. *Mol. Cell Proteomics*, **5**, 998–1005.
29. Lin, M. F., Jungreis, I. and Kellis, M. 2011, PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–82.
30. Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K. and Shiu, S. H. 2010, sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics*, **26**, 399–400.
31. Ladoukakis, E., Pereira, V., Magny, E. G., Eyre-Walker, A. and Couso, J. P. 2011, Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol.*, **12**, R118.
32. Hanada, K., Zhang X., Borevitz, J. O., Li, W. H. and Shiu, S. H. 2007, A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. *Genome Res.*, **17**, 632–40.
33. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. and Weissman, J. S. 2009, Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–23.
34. Chew, G. L., Pauli, A., Rinn, J. L., Regev, A., Schier, A. F. and Valen, E. 2013, Ribosome profiling reveals resemblance between long noncoding RNAs and 5' leaders of coding RNAs. *Development*, **140**, 2828–34.
35. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. and Lander, E. S. 2013, Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*, **154**, 240–51.
36. Crappé, J., Van Criekinge, W., Trooskens, G., et al. 2013, Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics*, **14**, 648.
37. Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., et al. 2014, Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.*, **8**, 1365–79.
38. Nelson, B. R., Makarewich, C. A., Anderson, D. M., et al. 2016, A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science*, **351**, 271–5.
39. Galewsky, S., Xie, X. and Schulz, R. A. 1990, The *Drosophila melanogaster* z600 gene encodes a chromatin-associated protein synthesized in the syncytial blastoderm. *Gene*, **96**, 227–32.
40. Lécuyer, E., Yoshida, H., Parthasarathy, N., et al. 2007, Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, **131**, 174–87.
41. Qin, X., Ahn, S., Speed, T. P. and Rubin, G. M. 2007, Global analyses of mRNA translational control during early *Drosophila* embryogenesis. *Genome Biol.*, **8**, R63.
42. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. 2013, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
43. Trapnell, C., Roberts, A., Goff, L., et al. 2012, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–78.
44. Ramsköld, D., Wang, E. T., Burge, C. B. and Sandberg, R. 2009, An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, **5**, e1000598.

45. Peano C, Pietrelli A, Consolandi C, et al. 2013, An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb. Inform. Exp.*, **3**, 1.
46. Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., et al. 2014, Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *Elife*, **3**, e03528.
47. Goecks, J., Nekrutenko, A. and Taylor, J. 2010, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
48. Blankenberg, D., Kuster, G. V., Coraor, N., et al. 2010, Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.*, **19**, 10.
49. Giardine, B., Riemer, C., Hardison, R. C., et al. 2005, Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–5.
50. King, J. L. and Jukes, T. H. 1969, Non-darwinian evolution. *Science*, **164**, 788–98.
51. Bonander, N., Darby, R. A., Grgic, L., et al. 2009, Altering the ribosomal subunit ratio in yeast maximizes recombinant protein yield. *Microb. Cell. Fact*, **8**, 10.
52. Yang, F., Peng, Y., Murray, E. L., Otsuka, Y., Kedersha, N. and Schoenberg, D. R. 2006, Polysome-bound endonuclease PMR1 is targeted to stress granules via stress-specific binding to TIA-1. *Mol. Cell. Biol.*, **26**, 8803–13.
53. Arava, Y., Wang, Y., Storey, J. D., Liu, C. L., Brown, P. O. and Herschlag, D. 2003, Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA*, **100**, 3889–94.
54. Heyer, E. E. and Moore, M. J. 2016, Redefining the Translational Status of 80S Monosomes. *Cell*, **164**, 757–69.
55. Cheng, J., Kapranov, P., Drenkow, J., et al. 2005, Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–54.
56. Brown, J. B., Boley, N., Eisman, R., et al. 2014, Diversity and dynamics of the *Drosophila* transcriptome. *Nature*, **512**, 393–9.
57. Graveley, B. R., Brooks, A. N., Carlson, J. W., et al. 2011, The developmental transcriptome of *Drosophila melanogaster*. *Nature*, **471**, 473–9.
58. Schneider, I. 1972, Cell lines derived from late embryonic stages of *Drosophila melanogaster*. *Development*, **27**, 353–65.
59. Sterne-Weiler, T., Martinez-Nunez, R. T., Howard, J. M., et al. 2013, Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res.*, **23**, 1615–1623.
60. Gawron, D., Gevaert, K. and Van Damme, P. 2014, The proteome under translational control. *Proteomics*, **14**, 2647–62.
61. Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R. and Weissman, J. S. 2013, Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila* melanogaster. *elife*, **2**, e01179.
62. Inagaki, S., Numata, K., Kondo, T., et al. 2005, Identification and expression analysis of putative mRNA-like non-coding RNA in *Drosophila*. *Genes Cells*, **10**, 1163–73.
63. Tupy, J. L., Bailey, A. M., Dailey, G., et al. 2005, Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA*, **102**, 5495–500.
64. Mercer, T. R., Dinger, M. E. and Mattick, J. S. 2009, Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–9.
65. Mercer, T. R. and Mattick, J. S. 2013, Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.*, **20**, 300–7.
66. Rinn, J. L. and Chang, H. Y. 2012, Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, **81**, 145–66.
67. Guttman, M. and Rinn, J. L. 2012, Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–46.
68. Bassett, A. R., Akhtar, A., Barlow, D. P., et al. 2014, Considerations when investigating lncRNA function in vivo. *Elife*, **3**, e03058.
69. Mackowiak, S. D., Zauber, H., Bielow, C., et al. 2015, Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.*, **16**, 1–21.
70. Bazzini, A. A., Johnstone, T. G., Christiano, R., et al. 2014, Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.*, **33**, 981–93.
71. Carrieri, C., Cimatti L., Biagioli, M., Beugnet, A., et al. 2012, Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*, **491**, 454–7.
72. Tani, H., Torimura M. and Akimitsu, N. 2013, The RNA degradation pathway regulates the function of GAS5 a non-coding RNA in mammalian cells. *PLoS One*, **8**, e55684.
73. van Heesch, S., van Iterson, M., Jacobi, J., et al. 2014, Extensive localization of long noncoding RNAs to the cytosol and mono-and polyribosomal complexes. *Genome Biol.*, **15**, 1–12.
74. Asano K. 2014, Why is start codon selection so precise in eukaryotes? *Translation*, **2**, e28387.
75. Hinnebusch A G, Ivanov I P, Sonenberg N. 2016, Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science*, **352**, 1413–6.
76. Ma, J., Ward, C. C., Jungreis, I., et al. 2014, Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J. Proteome Res.*, **13**, 1757–65.