Ribosomal profiling adds new coding sequences to the proteome

Muhammad Ali S. Mumtaz* **and Juan Pablo Couso***[1]

*School of Life Sciences, JMS Building, University of Sussex, Brighton BN1 9QG, U.K.

Abstract

Next generation sequencing (NGS) has enabled an in-depth look into genes, transcripts and their translation at the genomic scale. The application of NGS sequencing of ribosome footprints (Ribo-Seq) reveals translation with single nucleotide (nt) resolution, through the deep sequencing of ribosome-bound fragments (RBFs). Some results of Ribo-Seq challenge our understanding of the protein-coding potential of the genome. Earlier bioinformatic approaches had shown the presence of hundreds of thousands of putative small ORFs (smORFs) in eukaryotic genomes, but they had been largely ignored due to their large numbers and difficulty in determining their translation and function. Ribo-Seq has revealed that hundreds of putative smORFs within previously assumed long non-coding RNAs (lncRNAs) and UTRs of canonical mRNAs are associated with ribosomes, appearing to be translated. Here we review some of the approaches used to define translation within Ribo-Seq experiments and the challenges in defining translation of these novel smORFs in lncRNAs and UTRs. We also look at some of the bioinformatic and biochemical approaches used to independently corroborate these exciting new findings and elucidate real translation events.

Introduction

The advent of mass sequencing has provided us with the full sequence of genomes and has given rise to the field of genomics. Initially, gene sequences were identified as containing ORFs, but when the first eukaryotic genome (yeast) was sequenced, an arbitrary cut-off of 100 amino acids (aa) was introduced for the annotation of ORFs as protein-coding genes [1]. Thus, small ORFs (smORFs) of less than 100 codons, without experimental evidence of function or homology with other protein-coding genes, were simply discarded. This arbitrary measure made sense, since smORFs outweigh long ORFs by several orders of magnitude in every genome. Classic, pre-sequencing genetics data could be accommodated with the thousands of annotated genes in sequenced animals, but it seemed unconceivable that actual protein-coding genes were in the hundreds of thousands in yeast and fruit flies and in the millions in mammals [2–4] (Table 1). Since then, further techniques have allowed the detection of transcription at a genomic scale, starting with microarray technology and next generation sequencing (NGS) of cDNAs or NGS sequencing of total mRNA (RNA-Seq) [5]. Surprisingly, RNA-Seq reveals that up to 85 % of a mammal genome can be transcribed [6], far more than the 4 % that accounts for annotated protein-coding genes. Therefore, most of these novel transcripts have been annotated as non-coding RNAs, because they do not contain ORFs longer than 100 codons. However, thousands of non-coding RNAs longer than 200 nts, which do not produce short regulatory RNAs (such as siRNAs or miRNAs [7] are expressed in animals. These long non-coding RNAs (lncRNAs) are capped, spliced and polyadenylated-like mRNAs [8] and are currently estimated to number approximately 10000 in mammals and 2000 in *Drosophila* [9]. Most research on lncRNAs has focused on their role in transcriptional regulation [8], based on the assumption that they localize to the nucleus. In fact most lncRNAs are also present in the cytoplasm and even enriched there [10] and most, if not all, lncRNAs contain smORFs. Currently there is great interest in ascertaining the role(s) lncRNAs may play in the cell and whether they are truly non-coding.

The new technique of NGS sequencing of ribosome footprints (Ribo-Seq), which detects translation at a genome-wide level by deep sequencing of mRNA sequences bound by ribosomes [11] has the potential to resolve this issue. In a very short time, Ribo-Seq has provided evidence for:

1. Alternative START and STOP codons for canonical proteins [11,12]
2. Use of non-AUG START codons [11,13]
3. Polycistronic arrangements in eukaryote transcripts [11,14], including overlapping reading frames [15]
4. Translation of non-annotated ORFs [11,13–16]
5. Translational mechanics, i.e. elongation, pausing and frame-shifting [17]

A detailed discussion of each of these topics warrants a review on its own, but here we will focus on two related

Key words: long non-coding ribonucleic acid (ncRNA), NGS sequencing of footprints from polysome fractions (Poly-Ribo-Seq), proteome, ribosome profiling, small open reading frames (smORFs), translation.

Abbreviations: aa, amino acids; lncRNA, long non-coding RNA; NGS, next generation sequencing; nt, nucleotide; PhyloCSF, phylogenetic codon substitution frequencies; Poly-Ribo-Seq, NGS sequencing of footprints from polysome fractions; RBF, ribosome-bound fragment; Ribo-Seq, NGS sequencing of ribosome footprints; RNA-Seq, NGS sequencing of total mRNA; RPKM, reads per kilobase per million; smORF, small ORF; TE, translation efficiency; UTR, untranslated region.

[1] **To whom correspondence should be addressed (email j.p.couso@sussex.ac.uk).**

Biochem. Soc. Trans. (2015) 43, 1271–1276; doi:10.1042/BST20150170

Table 1 | **Approximated number of smORFs and annotated protein-coding genes in three model organisms**
Data obtained from Basrai et al. [2] and the saccharomyces genome database (yeast); Flybase and Ladoukakis et al. [3] (fruit fly); Ensembl and Crappe et al. 2015 [4] (mouse).

Species	Protein-coding genes	smORFs in genome
Saccharomyces cerevisiae (Brewer's yeast)	6600	265000
Drosophila melanogaster (Fruitfly)	18000	556000
Mus musculus (Mouse)	22000	40000000

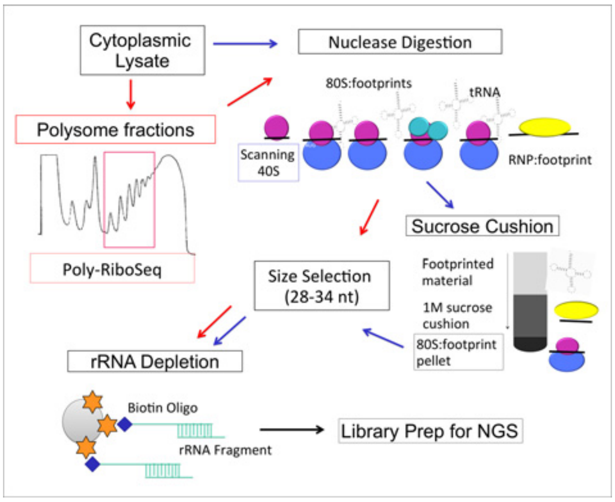
issues that are researched in our laboratory: translation of smORFs and non-coding RNAs.

Ribosomal profiling methods

Ribosome profiling entails the sequencing of ribosome-bound fragments (RBFs) from mRNAs, in order to map translated sequences within the transcriptome [11,13]. Elongating mRNA-associated ribosomes are trapped by a translation inhibitor, followed by nuclease treatment to generate RBFs by digestion of unprotected mRNA. The RBFs (28–34 nt long) are then purified from the ribosomes for cDNA library preparation, deep sequencing and mapping to the transcriptome or the genome (Figure 1). Ribo-Seq offers a direct read-out of ribosome occupancy on mRNA at the single nucleotide (nt) level and provides quantitative metrics directly related to the translation rate (see below; Figure 2A). This is far more precise than previous techniques such as polysomal profiling, which simply detected mRNAs associated with polysomes. Due to translational regulation, RNAs may not be translated in direct proportion to their presence in polysomes, as seen with regulatory non-coding RNAs that associate with polysomes without being translated [14,18].

Detection of a single RBF sequence cannot be equated with translation. A degree of background signal has to be accounted for, whether technical background (detection of signal when there is none) or biological background (generation of signal in the absence of the detected phenomena, in our case detection of mRNA binding not produced by translating ribosomes). Regarding technical background, the procedure for generation of NGS libraries follows the progress in RNA-Seq and thus, although a distortion in the number of reads sequenced is conceivable (due for example to PCR amplification bias [19]), *de novo* generation of false reads in detectable quantities is unlikely. Two sources of biological background can be identified: protection of RNA digestion by binding by proteins not constituting a ribosome (RNA-binding proteins, ribosomal subunits) or by ribosomes not engaged in translation, (either scanning, assembling, involved in nonsense-mediated decay or simply stalled by translational regulation). However, by far the most prevalent source of background is provided by rRNA and tRNAs, which are isolated with the mRNA–ribosome complexes and can constitute up to 85 % of the reads, depending on the library preparation method used (unpub. obs).

Figure 1 | **Ribosomal profiling methods**
An overview of the original Ribo-Seq method (blue arrows), comparing it with the Poly-Ribo-Seq technique (red arrows) for generating RBF libraries, for the original Ribo-Seq, a cytoplasmic cell lysate is prepared in the presence of translation inhibitor to trap mRNA–ribosome complexes and then treated with nuclease to generate RBFs. The ribosome–RBF complexes are then purified on a sucrose cushion before RNA extraction, size selection and NGS Library prep including ribosomal RNA depletion using subtractive hybridization. Poly-Ribo-Seq introduces a further purification/fractionation step before nuclease digestion in order to isolate RBFs from actively translating mRNAs (bound by 2+ Ribosomes) thus excluding signal from non-productive 80S binding and mRNP complexes.



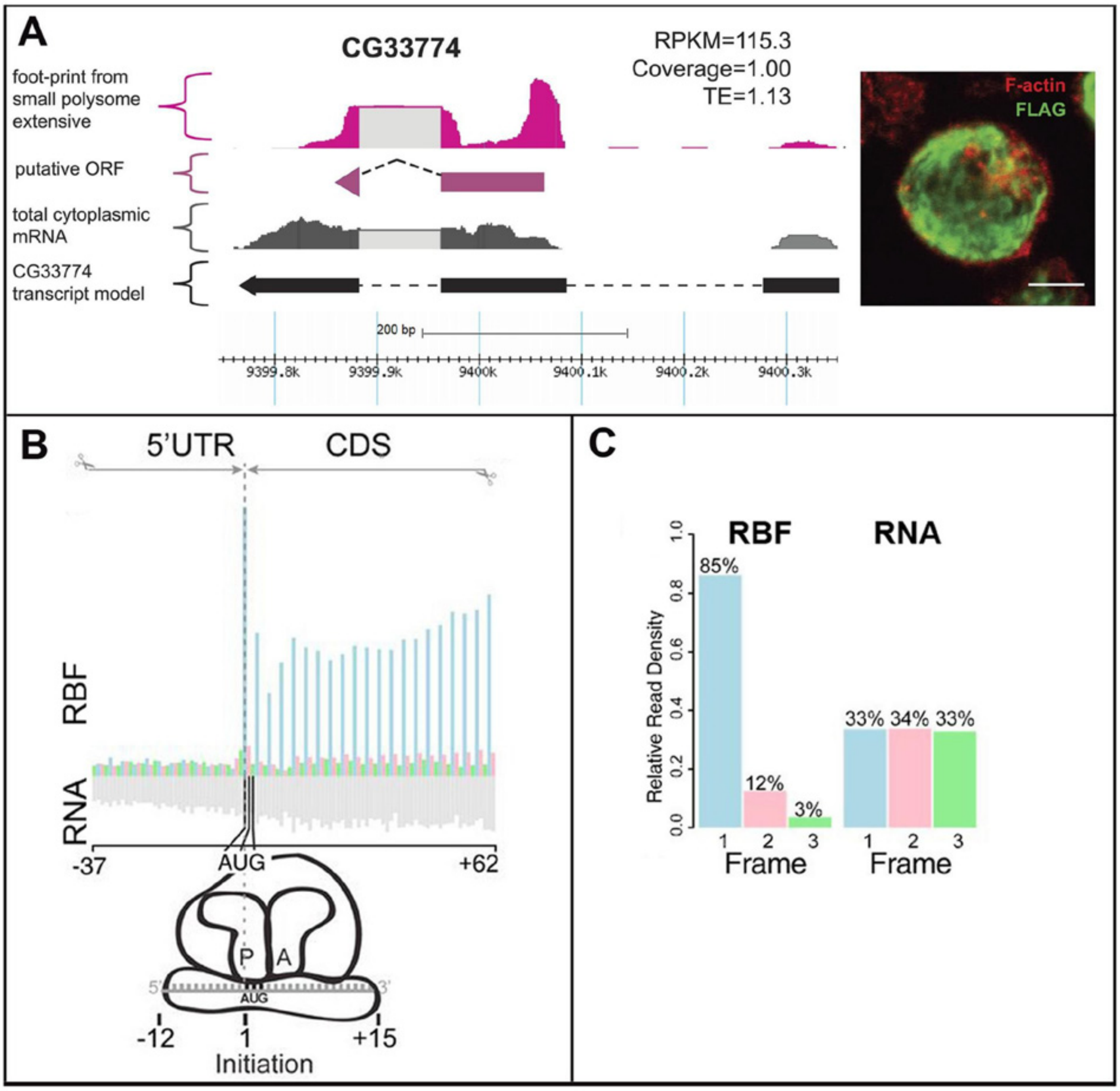
A combination of experimental and bioinformatic approaches have been devised to deal with Ribo-Seq noise. Experimentally, after RNA digestion, ribosome–mRNA units can be separated from other RNA-binding protein complexes using either a sucrose cushion [11] or a column [16]. However, this purification still retains 80S monosomes not engaged in productive translation (i.e. scanning, assembling or stalled at AUG codons). rRNA and tRNA can be depleted by subtractive hybridization using biotinylated oligonucleotides targeted to their sequences (Figure 1).

Ribosomal profiling metrics

Bioinformatically, a number of filters and metrics are used to ascertain productive translation from raw RBF sequenced data (Figure 2A). Firstly, reads matching rRNA and tRNA sequences are discarded and the remaining reads are then mapped to an annotated transcriptome or genome. Next, reads mapping to multiple sites in the genome and

Figure 2 | Visualization of Ribo-Seq data

(**A**) Visualization of Ribo-Seq and RNA-Seq data for the fruitfly smORF CG33774. The RBFs (pink) map almost exclusively to the ORF whereas the RNA-Seq data is evenly distributed across the transcript. Also apparent is the pile-up of reads at the start codon due to initiating ribosomes and to a lesser extent at the stop codon. A tagging assay (left) corroborates the translation of the peptide (FLAG antibody: green colour, F-actin stained with phalloidin: red, scale bars = 5 μm) together with Poly-Ribo-Seq metrics (RPKM, coverage and TE). (**B**) Schematic of P-site assignment of the RBFs calculated from reads aligning at the start codon. The distance from the 5′-end of the read to the start codon is used to calculate the offset (12 nt) for each read length. (**C**) Ribo-Seq data shows trinucleotide periodicity due to ribosomes moving from one codon to the next, with the majority of reads mapping to the AUG frame, whereas RNA-Seq show a uniform distribution across frames. [(**A**) Modified from [14]: Aspden, J.L., Eyre-Walker, Y.C., Phillips, R.J., Amin, U., Mumtaz, M.A., Brocard, M. and Couso, J.P. (2014) Extensive translation of small open reading frames revealed by poly-ribo-seq. Elife **3**, e03528; (**B** and **C**) modified from [16]: Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C. and Giraldez, A.J. (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. **33**, 981–993].



reads with mismatches above a certain threshold (usually 0–2 nts) with the reference genomic sequence are also discarded. Typically only ∼20 % of the initial reads pass these filters.

As with most NGS data, relative signal abundance is estimated in a metric called RPKM (reads per kilobase per million) that allows for normalization to total mapped library size (millions of reads), as well as for transcript length (per kb), since longer transcripts can accrue larger numbers of reads [20]. However, high Ribo-Seq RPKMs can result from either intensive translation of a moderately abundant transcript or from low translation of a very abundant transcript. Therefore, to ascertain the translation level for a particular protein, its Ribo-Seq RPKM must be normalized to its mRNA levels. For this, RNA-Seq must be performed in parallel to Ribo-Seq to calculate translation efficiency (TE), which is the ratio $\text{RPKM}^{\text{Ribo-Seq}}/\text{RPKM}^{\text{RNA-Seq}}$ across the ORF.

The coverage metric describes the proportion of an ORF that is covered by reads. The greater the proportion, the greater the likelihood that it is translated, thus excluding for example, cases resulting from a pile-up of stalled, non-translating ribosomes at the start codon. Coverage can be measured by measuring the fraction of ORF length covered by RBFs or by counting the fraction of codons (or nts), harbouring putative ribosome P-sites (see next).

P-site assignment

Given the length constraint of RBFs, it is possible to estimate the ribosomal P-sites, as the most prevalent offset (around 12 nts) from the 5′ end of reads to well-defined start codons (Figure 2B). This estimation allows the mapping of the P-site of any other read of similar length with single nt resolution and hence determine (1) codon coverage (see above); (2) the number of reads in frame for a given ORF (framing; see below) and (3) new or poorly-defined start and stop codons. However, P-site assignments are an estimation and suffer from various problems; for example, there is no consensus as to what should be the exact length of the RBFs used and various studies have used different values within the 25–36 nt range. RBF length and P-site offset must reflect the degree of packing and folding of the mRNA that gets ‘tucked away’ inside/around the ribosome and protected from digestion and this could be species-specific and sensitive to the experimental setup. Usually, P-site assignment can only be reliably ascertained in a subset of reads within a narrow length range and framing becomes less clear in the middle of long ORFs, away from the start and stop codons where reads pile-up and ‘synchronize’.

Framing score

Reads over protein-producing regions tend show a trinucleotide periodicity based on the frame of translation for a given ORF, reflecting the trinucleotide codon-by-codon movement of the ribosome [11] (Figure 2A). This behaviour of translating ribosomes can be used to define a framing score for ambiguous or novel ORFs. It is calculated as the proportion of reads across the ORF that have their P-site (or 5′-end) in frame with the ORF [16] (Figure 2C) and although it shares the limitations of P-site assignment, it is still the most unambiguous signature of translation.

These metrics allow for the robust identification of canonical translated sequences and many novel ones, as mentioned before. However, many non-coding RNAs and the smORFs they contain remain controversial.

Translation of putative non-coding RNAs and smORFs

Known non-coding functions of cytoplasmic lncRNAs include roles in mRNA translation elongation, as miRNA molecular sponges [21] or altering mRNA stability and affecting cap-independent translation (reviewed in [22]). However, ∼50 % of lncRNAs in mouse embryonic stem cells exhibit ribosome profiling signal [13], ∼34 % in flies [14] and (∼14 %) in zebrafish [16]. Some of these lncRNAs are probably regulating translation, but others contain smORFs that can be shown to be translated by other methods [14,16,23–24] and produce peptides with biological functions (reviewed in [25]). The high number of putative translated lncRNA detected by Ribo-Seq has generated discussion as to the level of background signal from ribosome profiling and whether the detected RBFs correspond to meaningful protein production [26–29]. The smORFs in lncRNAs obtain few profiling reads (partially arising from their very short length), perhaps indicating noise or a biologically irrelevant level of translation. Simply increasing the number of total Ribo-Seq reads does not help, since ‘bona-fide noise’ (i.e. in UTRs or small non-coding RNAs) also increases. Further bioinformatic metrics have been introduced beyond the standard ones discussed above, in order to further distinguish noise from ‘real’ lncRNA translation (i.e. Ribosomal Release Score [26]; Fragment Length Organization Similarity Score [29]). However, these metrics suffer from their own problems and do not seem superior to a framing score.

Two main experimental refinements have been introduced to deal with controversial findings: Harringtonine and other drugs trap Ribosomes at start codons during the initiation to elongation transition. This enables the identification of alternative Translation Initiation Sites, such as N-terminally extended or downstream internal AUG of many ORFs and non-AUG start codons. Additionally, new putatively translated ORFs are highlighted, particularly upstream ORFs (uORFs), short ORFs located in the 5’ leaders of up to 60 % of canonical coding genes and smORFs in lncRNAs [13].

Poly-Ribo-Seq

We introduced NGS sequencing of footprints from polysome fractions (Poly-Ribo-Seq) to target ribosomal profiling specifically to the detection of translated smORFs, by profiling only those RNAs bound by a precise number of ribosomes. Polysomal fractions were isolated following a classic sucrose gradient and fractions containing 2–6 ribosomes (the maximum number of ribosomes packed into a 300 nt ORF [14]) were profiled (Figure 1). In principle, this method ensures that the RBFs detected arise from genuine translation events rather than from sporadic binding by non-productive single ribosomes. However, it discriminates against very smORFs of less than 90 nt (30 codons) only able to accommodate one ribosome at a time (allowing for lax ribosomal packing and/or low translation levels [30]). Twenty-nine aa would appear to be a very smORF size; however, we have identified 11-aa peptides controlling leg and embryonic development arising from 36-nt smORFs [23] and, more recently, 28–32 aa ORFs producing peptides controlling heart rhythm [24]. Reciprocally, many such 'dwarf' smORFs are found in putative polycistronic transcripts containing several such smORFs (like, typically, lncRNAs) and thus their presence in RNAs harbouring multiple ribosomes is still compatible with an accumulation of single ribosomes binding, non-productively, to different non-translated ORFs. Despite these limitations, we were able to prove the translation of 'longer' smORFs with a median size of 80 aa contained in short monocistronic transcripts. These 'longer' smORFs exist in the hundreds in animal genomes and were translated with similar frequency (~80 %) and intensity (TE) as canonical proteins [14].

Independent corroboration

Two main options exist for the genome-wide corroboration of surprising or ambiguous Ribo-Seq data: bioinformatics and proteomics.

Sequence conservation

Canonical long ORFs display conservation of their aa sequence despite having nt changes across evolution, in other words, comparison of functional ORFs across related species are expected to display a prevalence of synonymous compared with non-synonymous codon substitutions (a ratio $K_a/K_s < 1$). However, it is difficult to score statistically significant values for very short sequences because their number of possible changes is low [31]. PhyloCSF (phylogenetic codon substitution frequencies) is a new method that takes into account other types of substitutions, such as conservative (replacing an aa for a similar one) and non-sense (introducing a stop codon) [32]. PhyloCSF then assigns a score to each codon substitution in an alignment, based on the relative frequency of that substitution in known coding and non-coding regions. This method has been shown to score better for sequences 30–180-nt long. However, the act of translation is different from the evolutionary conservation of the produced peptide. For example, uORFs may function as translational regulators without a further function for their peptides [25] and novel genes may be producing functional peptides in a given species, but not in related ones [33].

Proteomics obtains mass spec signatures of trypsinized peptides and these are then matched to a database of theoretical signatures resulting from the *in silico* translation and digestion of a previously defined library of coding sequences (either annotated or putative). However, *de novo* discovery of new smORF peptides is difficult, because this entails generating a putative library for all possible genomic smORFs (Table 1) and this is both computationally expensive and prone to false positives [25]. Further, standard procedures require the detection of several non-overlapping signatures to validate detection of a given protein and short peptides have less room for doing so. Finally, the detection of a peptide is influenced not only by its translation rate, but also its stability and detectability (mostly dictated by size [16,34] and, thus accordingly, proteomics only reliably detects abundantly translated peptides [14]). In summary, proteomics can provide a useful corroboration of Ribo-Seq data, but absence of (proteomic) evidence is not evidence of absence (of translation).

Future avenues

A scenario that could explain apparent controversies about lncRNA translation is that of highly regulated translation of a fraction of putative non-coding RNAs. Alternatively, lncRNAs could be translated stochastically/randomly. The first scenario is likely to produce functional peptides, the second less so. For example, since lncRNAs can be highly tissue-specific [8], the current ribosomal profiling techniques may underestimate their translation; a RNA transcribed and translated at a low rate across a multicellular organism gives the same low RPKM and TE signal as an RNA both highly transcribed and translated, but only in a few specific cells. Developments in ribosomal profiling techniques are likely to clarify this issue. Ribosomal immunopurification already allows detection of translation in different sub-cellular localizations [35,36] and could be adapted to study specific organs and tissues.

The availability of an *in vivo* high-throughput experimental test of translation and function would be very useful; approaches such as tagging have been used extensively in yeast and bacteria, but are more limited in eukaryotes [14]. An integrated approach with ribosomal profiling and independent corroboration methods has the best potential to clarify the issues discussed and transform our understanding of the proteome. The differences between coding and non-coding RNA are likely to be blurred. Already, mRNAs contain sequences with non-coding functions (for stability, splicing etc.) and similarly even short regulatory RNAs have been shown to have coding functions [37], perhaps in a more regulated and sophisticated way than the 'fully-on' model we imagine mRNAs to have. The phenomenon of pervasive translation of putative non-coding sequences is likely to expand and lead to a change in our understanding of the nature of translation and of its functions.

Acknowledgements

We thank Julie Aspden for suggestions and Sarah Bishop for comments to the manuscript.

Funding

This work was funded by the Wellcome Trust [grant number 087516] to J.P.C.

References

1 Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. et al. (1996) Life with 6000 genes. Science **274**, 563–567 CrossRef

2 Basrai, M.A., Hieter, P. and Boeke, J.D. (1997) Small open reading frames: beautiful needles in the haystack. Genome Res. **7**, 768–771 PubMed

3 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 CrossRef PubMed

4 Crappé, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De Keulenaer, S., De Meester, E., De Meyer, T., Van Criekinge, W. et al. (2015) Proteoformer: deep proteome coverage through ribosome profiling and MS integration. Nucleic Acids Res. **43**, e29 CrossRef PubMed

5 Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. **10**, 57–63 CrossRef PubMed

6 Hangauer, M.J., Vaughn, I.W. and McManus, M.T. (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. **9**, e1003569 CrossRef PubMed

7 Ghildiyal, M. and Zamore, P.D. (2009) Small silencing RNAs: an expanding universe. Nat. Rev. Genet. **10**, 94–108 CrossRef PubMed

8 Guttman, M. and Rinn, J.L. (2012) Modular regulatory principles of large non-coding RNAs. Nature **482**, 339–346 CrossRef PubMed

9 Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.-L. and Ponting, C.P. (2012) Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. Genome Biol. Evol. **4**, 427–442 CrossRef PubMed

10 Ulitsky, I. and Bartel, D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. Cell **154**, 26–46 CrossRef PubMed

11 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**, 218–223 CrossRef PubMed

12 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 CrossRef PubMed

13 Ingolia, N.T., Lareau, L.F. and Weissman, J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell **147**, 789–802 CrossRef PubMed

14 Aspden, J.L., Eyre-Walker, Y.C., Phillips, R.J., Amin, U., Mumtaz, M.A., Brocard, M. and Couso, J.P. (2014) Extensive translation of small open reading frames revealed by poly-ribo-seq. Elife **3**, e03528 CrossRef PubMed

15 Duncan, C.D. and Mata, J. (2014) The translational landscape of fission-yeast meiosis and sporulation. Nat Struct. Mol. Biol. **21**, 641–647 CrossRef PubMed

16 Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C. and Giraldez, A.J. (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. **33**, 981–993 CrossRef PubMed

17 Ingolia, N.T. (2014) Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Rev. Genet. **15**, 205–213 CrossRef PubMed

18 van Heesch, S., van Iterson, M., Jacobi, J., Boymans, S., Essers, P.B., de Bruijn, E., Hao, W., MacInnes, A.W., Cuppen, E. and Simonis, M. (2014) Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. Genome Biol. **15**, R6 CrossRef PubMed

19 Zheng, W., Chung, L.M. and Zhao, H. (2011) Bias detection and correction in RNA-sequencing data. BMC Bioinformatics **12**, 290 CrossRef PubMed

20 Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Methods **5**, 621–628 CrossRef PubMed

21 Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H. and Bartel, D.P. (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell **147**, 1537–1550 CrossRef PubMed

22 Fatica, A. and Bozzoni, I. (2013) Long non-coding RNAs: new players in cell differentiation and development. Nat. Rev. Genet. **15**, 7–21 CrossRef PubMed

23 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**, 1052–1062 CrossRef

24 Magny, E.G., Pueyo, J.I., Pearl, F.M., Cespedes, M.A., Niven, J.E., Bishop, S.A. and Couso, J.P. (2013) Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science **341**, 1116–1120 CrossRef PubMed

25 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 CrossRef PubMed

26 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**, 240–251 CrossRef PubMed

27 Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W.E., Kundaje, A., Gunawardena, H.P., Yu, Y., Xie, L. et al. (2012) Long noncoding RNAs are rarely translated in two human cell lines. Genome Res. **22**, 1646–1657 CrossRef PubMed

28 Chew, G.L., Pauli, A., Rinn, J.L., Regev, A., Schier, A.F. and Valen, E. (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. Development **140**, 2828–2834 CrossRef PubMed

29 Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J., Jackson, S.E., Wills, M.R. and Weissman, J.S. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. **8**, 1365–1379 CrossRef PubMed

30 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. U.S.A. **100**, 3889–3894 CrossRef PubMed

31 Housman, G. and Ulitsky, I. (2015) Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. Biochim. Biophys. Acta, in the press

32 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–i282 CrossRef PubMed

33 Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B. et al. (2012) Proto-genes and *de novo* gene birth. Nature **487**, 370–374 CrossRef PubMed

34 Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L. and Saghatelian, A. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat. Chem. Biol. **9**, 59–64 CrossRef PubMed

35 Williams, C.C., Jan, C.H. and Weissman, J.S. (2014) Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. Science **346**, 748–751 CrossRef PubMed

36 Jan, C.H., Williams, C.C. and Weissman, J.S. (2014) Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. Science **346**, 1257521 CrossRef PubMed

37 Lauressergues, D., Couzigou, J.M., Clemente, H.S., Martinez, Y., Dunand, C., Bécard, G. and Combier, J.P. (2015) Primary transcripts of microRNAs encode regulatory peptides. Nature **520**, 90–93 CrossRef PubMed

Received 7 August 2015
doi:10.1042/BST20150170