Nucleic Acids Research, 2018 **1**
doi: 10.1093/nar/gky978

RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling

Hongwei Wang[1,*,†], Ludong Yang[1,†], Yan Wang[1,†], Leshi Chen[2,†], Huihui Li[1] and Zhi Xie[1,*]

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China and [2]Centre for Advanced Computational Solutions (C-fACS), Lincoln University, New Zealand

Received September 15, 2018; Revised October 04, 2018; Editorial Decision October 05, 2018; Accepted October 06, 2018

ABSTRACT

RPFdb (http://www.rpfdb.org or http://sysbio.sysu.edu.cn/rpfdb) is a public database for hosting, analyzing and visualizing ribosome profiling (ribo-seq) data. Since its initial release in 2015, the amount of new ribo-seq data has been considerably enlarged with the increasing popularity of ribo-seq technique. Here, we describe an updated version, RPFdb v2.0, which brings significant data expansion, feature improvements, and functionality optimization: (i) RPFdb v2.0 currently hosts 2884 ribo-seq datasets from 293 studies, covering 29 different species, in comparison with 777 datasets from 82 studies and 8 species in the previous version; (ii) A refined analysis pipeline with multi-step quality controls has been applied to improve the pre-processing and alignment of ribo-seq data; (iii) New functional modules have been added to provide actively translated open reading frames (ORFs) information for each ribo-seq data; (iv) More features have been made available to increase database usability. With these additions and enhancements, RPFdb v2.0 will represent a more valuable and comprehensive database for the gene regulation community.

INTRODUCTION

Ribosome profiling is emerging as a powerful technique that enables genome-wide investigation of *in vivo* translation at sub-codon resolution (1,2). By precisely pinpointing ribosomes during translation, this technique can provide deeper insights into the composition, regulation and mechanism of translation (3–6). The discriminatory power of ribo-seq technique leads to its widespread applications in various organisms that are enabling data generation at an unprecedented scale. However, vast amounts of ribo-seq data are mostly stored in raw data format so that they cannot be easily inspected and analyzed. Furthermore, comparisons across different datasets cannot be reliably made without unified data processing and analytics. As a result, efficient storage, retrieval and management of these large amounts of publicly available ribo-seq data are urgently needed to the research community. To this end, several dedicated database of ribo-seq data have been built, including GWIPS-viz (7), a visualization tool for ribo-seq data; RiboSeqDB (8), a repository of selected human, mouse and rat ribo-seq data; TranslatomeDB (9), a collection of ribo-seq, RNC-seq and mRNA-seq data with emphasis on differential gene expression analysis, and our RPFdb (10). In addition, several databases are designed specifically for storing genomic elements detected from ribo-seq, such as sORFs.org (11) and uORFdb (12), hosting small ORFs (sORFs) and upstream ORFs respectively.

RPFdb is dedicated to host, analyze and visualize ribo-seq data, processed by a unified pipeline (10). Since the initial release of RPFdb in 2015, the volume of raw ribo-seq data in repositories such as the Sequence Read Archive (SRA) (13) has grown rapidly because of the increasing popularity of ribo-seq technique and reduction of sequencing costs. At the same time, substantial progress has been made in our understanding the what, when, where and how of protein synthesis (5,14). Moreover, a growing number of studies on the coding potential of genomes have showed that a diverse set of non-canonical translation products such as sORF-encoded micro-peptides do exist (15–18). Some of micro-peptides are even known to have important physiological functions (19–23). However, despite receiving more attentions, the full repertoire of non-canonical translation products is still unknown and waits for further exploration. The rapidly accumulating ribo-seq data provide a basis for discovering the 'dark matter' in the genome and understanding the components of the translation apparatus. Therefore, all these premises lead us to make timely update on RPFdb.

*To whom correspondence should be addressed. Tel: +86 20 6667 7086; Email: xiezhi@gmail.com
Correspondence may also be addressed to Hongwei Wang. Email: bioccwhw@126.com.
†The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

Herein, we present an updated version of PRFdb (v2.0) that currently hosts 2884 ribo-seq datasets from 293 studies, covering 29 different species. In line with the significant expansion of ribo-seq data available in the database, we also have made several major improvements in this release, including a refined analysis pipeline with multi-step quality control applied for improving the pre-processing and alignment of ribo-seq data, new functional modules providing actively translated ORFs information, and more web features for better database usability. Further details on these additions and enhancements made to PRFdb are described below.

SUMMARY OF FEATURES IN THE INITIAL RPFdb

The initial release of RPFdb (v1.0) provided ribo-seq data for 777 samples from 82 studies covering 8 species, including *Arabidopsis*, *Caenorhabditis elegans*, *Drosophila*, *Escherichia coli*, *Saccharomyces cerevisiae*, zebrafish, mouse and human (Table 1). The web interface for RPFdb was constructed with three main functional modules: 'Browse', 'Search', and 'Download'. The 'Browse' module provides a brief description for each ribo-seq dataset such as accession identifier, source name, reference genome and the associated paper; a graphical overview of mapping statistics, distribution of reads mapped to genomic regions and RPKM (Reads Per Kilobase per Million) values of each genomic region; and a tabular display for translation levels of the top-ranked genes. The 'Search' module provides two types of queries: 'Gene query' by HGNC symbol or Ensembl gene ID and 'Study query' by keywords. In response to the 'Gene query', the webpage returns mRNA translation represented by normalized RPKM values from different samples of different studies and an interactive JBrowse genome browser for visualizing read-alignment data stored in bam files. In response to the 'Study query', the webpage returns a meta-information on the resulting study and an extended statistical summary of the study. The 'Download' module provides tabulated summary table of gene translation in different genomic regions for each study.

DATABASE ADDITIONS AND ENHANCEMENTS

Significant expansion of ribo-seq datasets

To obtain a more comprehensive set of ribo-seq data, we have included more data sources. The SRA (13), the European Nucleotide Archive (ENA) (24) and DDBJ Sequence Read Archive (DRA) (25) were queried with keywords, including 'ribosome profiling', 'ribo-seq', 'riboseq', 'ribosome protected fragments', 'ribosome footprints', 'ribosome footprinting' and 'RPF'. After manual screening, we retrieved additional 2107 ribo-seq datasets from 211 studies, bringing the current total number of available datasets to 2884 and the number of studies to 293 (Figure 1). Besides the eight species included in the previous version, 21 other species have been added, including *Bacillus subtilis*, *Chlamydomonas reinhardtii*, *Candida albicans*, Chinese hamster, *Caulobacter crescentus*, Chicken, *Halobacterium salinarum*, *Mycobacterium abscessus*, *Mycobacterium smegmatis*, *Neurospora crassa*, *Plasmodium falciparum*, *Pseudomonas aeruginosa*, *Pseudomonas fluorescens*, Rat, *Salmonella enterica*, *Schizosaccharomyces pombe*, *Streptomyces coelicolor*, *Staphylococcus aureus*, *Toxoplasma gondii*, *Trypanosoma brucei*, and *Xenopus laevis*. Notably among these species, human has shown the most dramatic increase not only in number of studies (from 27 to 101) but also in number of datasets (from 202 to 894) (Figure 1B and C). The other major species are mouse, *Saccharomyces cerevisiae* and *E. coli*. The rest of the species, particularly for the new included species, have only a limited number of datasets available. Nevertheless, the inclusion of these new species provides increased flexibility and enhanced options for a wide range of evolutionary and comparative studies.

Pre-processing optimization of ribo-seq data

The complicated experimental procedures of ribo-seq may introduce some inherent noise in the output data, which requires careful filtering that will not only minimize errors introduced by contamination but also enhance data efficiency. In the initial version of RPFdb, we directly selected the first 26 nucleotides of each sequencing read for alignment. In this update, we made more considerations on data pre-processing. To avoid adapter interference with downstream analysis, the 3′ adapter sequences were manually extracted for each dataset either from the publications or from the corresponding MultiQC outputs (26). If present in the ends of sequencing reads, the linked adapter sequences were removed using Cutadapt (version 1.16) (27). To eliminate rRNA and tRNA contamination, rRNA and tRNA sequences were fetched for each species from ENSEMBL (28) and UCSC (29), and then removed after mapping by Bowtie2 (version 2.3.4.1) (30). Considering that high-quality footprints are expected to have a characteristic distribution of read-lengths reflecting the size of a translating ribosome on the RNA, footprints with non-typical size were further excluded, and only those with 25–34 nucleotides in length were kept after contaminant removal and alignment. It is indubitable that these improvements will achieve more precise quantification of gene translation levels and more accurate visualization of gene translation. In addition, a number of quality control assessment metrices are provided to all the datasets for each study, including detailed FastQC reports, plots of read size and reading frame distributions, as well as P-site offset and enrichment of RPF reads in different genomic regions. These features will give users a better idea about the datasets before utilizing the data.

New contents presented by ribo-seq analysis

Recent advances in global translatome analyses have demonstrated that a much larger proportion of genomes have protein-coding potential than originally thought, revealing the existence of numerous alternative ORFs (altORFs) in addition to annotated protein coding sequences (31,32). These altORFs include upstream ORFs (uORFs) in the 5′ untranslated regions, downstream ORFs (dORFs)

Table 1. Comparison between RPFdb v2.0 and v1.0

	RPFdb v1.0	RPFdb v2.0
Summary of data		
Data source	SRA	SRA, ENA, DDBJ
No. of datasets	777	2884
No. of studies	82	293
No. of species	8	29
Data processing		
Pre-processing steps	• Quality control • Keep the first 26 nucleotides of each sequencing read	• Quality control • Adapter removal • Low-quality sequence trimming • rRNA and tRNA filtration • Read-length selection (25–34 nt)
Functionalities		
Browse	• Study browser(overview of dataset-meta description and summary statistics)	• Study browser (overview of dataset-meta description, summary statistics, and quality control assessment) • ORF browser (overview and detailed annotation information on actively translated ORFs)
Search and visualization	• RPKM values • Footprint coverage at different genomic regions	• RPKM values • ORF entry • Footprint coverage at different genomic regions
Download	• RPKM table	• Raw read count table • RPKM table • ORF annotation table
Website compatibility		
Applicability	• Desktop computers	• Desktop computers • Mobile devices

in the 3′ untranslated regions, and long non-coding derived ORFs (lncORFs). Some of them have been shown to actively undergo translation (15,18,33–35), which may play critical roles in fine-tuning translation program by serving as either translational repressors or dampeners (36,37). Hence, in this update we performed a systematic detection of actively translated regions by RibORF (34). To increase the power of footprint signal, we combined ribo-seq data with duplicate samples. To reduce the false positive rate in ORF finding, we only used those footprints with clear sub-codon phasing or triplet periodicity to predict actively translated ORFs. In total, we predicted canonical and non-canonical translated ORFs for 1405 aggregated ribo-seq datasets. More detailed information on each ORF entry is provided, such as genomic position, strand, annotated category and encoded amino acid length. The inclusion of these information in our database will facilitate systematic characterization, functional exploration and downstream analysis of ORFs.

Enhanced user interface features

The addition of actively translated ORFs was required for the development of a new interface for browsing and searching them in our database. To this end, we developed a dual-interface webpage to achieve both browse and search functions, which was then integrated into the RPFdb. This new webpage allows users to easily browse ORF content and to quickly search for a specific entry. Users can specify species, study and experimental conditions of interest. In addition, the ORF browser webpage provides a built-in filter that can help users further narrow down the results. The returned results of ORFs are presented in tabular form, which are also available for download by clicking the download button on the page.

The significant changes in both new ribo-seq data and new contents have demanded an enhanced web-interface. To facilitate count-based analyses such as differential translation detection, the download webpage was modified to support retrieving matrix of raw read counts. To facilitate inspection of translation of non-coding genes, the gene search webpage was modified to provide RPKM values of non-coding genes. Moreover, to achieve a better compatibility view, all the pages were enhanced to support facile viewing of page contents on different terminals, particularly for mobile devices.

CONCLUSIONS AND FUTURE PLAN

We presented RPFdb v2.0, an updated database for genome-wide information of translated mRNA generated from ribosome profiling, which is dedicated to providing more accurate, comprehensive and convenient way for translatome research. In this update, one of the most significant characteristics is that the scale of its collected ribo-seq data has been greatly expanded. Furthermore, a refined analysis pipeline has been applied for the pro-processing and alignment of ribo-seq data to make more precise quantification and visualization. In addition to providing quantitative information and intuitive visualization-oriented information about gene translation, RPFdb v2.0 now provides annotation information about actively translated ORFs. Moreover, more web features have been made available, including RPKM value of non-coding genes and

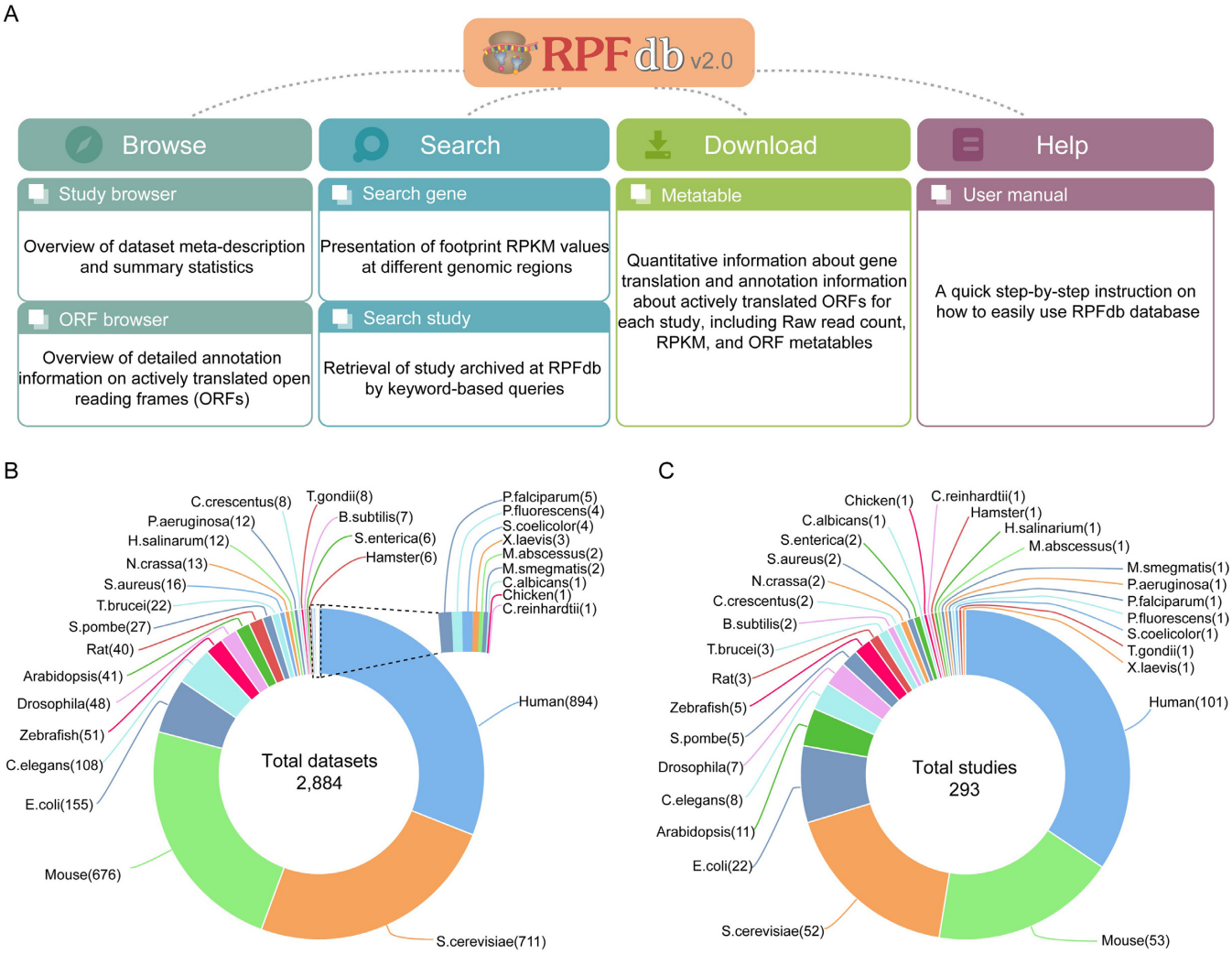

Figure 1. Overview of RPFdb v2.0. (**A**) Functional modules in RPFdb v2.0. (**B**) The number of ribo-seq datasets in species. (**C**) The number of studies in species.

matrix of raw read counts. The web interface has been also optimized for delivering an improved user experience.

RPFdb v2.0 will be periodically updated with the latest release of ribo-seq data. Also, we plan to include paired RNA-seq datasets with ribo-seq datasets in our database to facilitate downstream analysis such as translational efficiency analysis, considering that it is usually a common practice in ribosome profiling that the RPF and fragmented RNA would be profiled in parallel. Moreover, we will continue to expand the database contents based on newly developed computational frameworks and to fine-tune the database features and functionalities. In conclusion, with these additions and enhancements, RPFdb will continue serving as a valuable resource and make important contributions to the gene regulation community.

DATA AVAILABILITY

RPFdb is publicly available at http://www.rpfdb.org or http://sysbio.sysu.edu.cn/rpfdb.

ACKNOWLEDGEMENTS

We would like to thank all the members of Zhi Xie's lab for their testing and assistance as well as many users of RPFdb for their feedback and suggestions.

FUNDING

National Natural Science Foundation of China [31871302, 31829002 to Z.X.]. Funding for open access charge: National Natural Science Foundation of China [31871302, 31829002].

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Ingolia,N.T., Brar,G.A., Rouskin,S., McGeachy,A.M. and Weissman,J.S. (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.*, **7**, 1534–1550.

3. Ingolia,N.T. (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.*, **15**, 205–213.
4. Ingolia,N.T. (2016) Ribosome footprint profiling of translation throughout the genome. *Cell*, **165**, 22–33.
5. Brar,G.A. and Weissman,J.S. (2015) Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.*, **16**, 651–664.
6. Gobet,C. and Naef,F. (2017) Ribosome profiling and dynamic regulation of translation in mammals. *Curr. Opin. Genet. Dev.*, **43**, 120–127.
7. Michel,A.M., Kiniry,S.J., O'Connor,P.B.F., Mullan,J.P. and Baranov,P.V. (2018) GWIPS-viz: 2018 update. *Nucleic Acids Res.*, **46**, D823–D830.
8. R.N.,S., I.S.,Y., Y.V.,K. and O.A.,V. (2014) RiboSeqDB-a repository of selected human and mouse ribosome footprint and rna-seq data. *Virtual Biol.*, **1**, 37–46.
9. Liu,W., Xiang,L., Zheng,T., Jin,J. and Zhang,G. (2018) TranslatomeDB: a comprehensive database and cloud-based analysis platform for translatome sequencing data. *Nucleic Acids Res.*, **46**, D206–D212.
10. Xie,S.Q., Nie,P., Wang,Y., Wang,H., Li,H., Yang,Z., Liu,Y., Ren,J. and Xie,Z. (2016) RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res.*, **44**, D254–258.
11. Olexiouk,V., Van Criekinge,W. and Menschaert,G. (2018) An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.*, **46**, D497–D502.
12. Wethmar,K., Barbosa-Silva,A., Andrade-Navarro,M.A. and Leutz,A. (2014) uORFdb–a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res.*, **42**, D60–D67.
13. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database, C. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
14. Wang,H., Wang,Y. and Xie,Z. (2017) Computational resources for ribosome profiling: from database to Web server and software. *Brief. Bioinform.*, doi:10.1093/bib/bbx093.
15. 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.
16. Makarewich,C.A. and Olson,E.N. (2017) Mining for Micropeptides. *Trends Cell Biol.*, **27**, 685–696.
17. Couso,J.P. and Patraquim,P. (2017) Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.*, **18**, 575–589.
18. Wang,H., Wang,Y., Xie,S., Liu,Y. and Xie,Z. (2017) Global and cell-type specific properties of lincRNAs with ribosome occupancy. *Nucleic Acids Res.*, **45**, 2786–2796.
19. Anderson,D.M., Anderson,K.M., Chang,C.L., Makarewich,C.A., Nelson,B.R., McAnally,J.R., Kasaragod,P., Shelton,J.M., Liou,J., Bassel-Duby,R. *et al.* (2015) A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*, **160**, 595–606.
20. Nelson,B.R., Makarewich,C.A., Anderson,D.M., Winders,B.R., Troupes,C.D., Wu,F., Reese,A.L., McAnally,J.R., Chen,X., Kavalali,E.T. *et al.* (2016) A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science*, **351**, 271–275.
21. Matsumoto,A., Pasut,A., Matsumoto,M., Yamashita,R., Fung,J., Monteleone,E., Saghatelian,A., Nakayama,K.I., Clohessy,J.G. and Pandolfi,P.P. (2017) mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature*, **541**, 228–232.
22. Cai,B., Li,Z., Ma,M., Wang,Z., Han,P., Abdalla,B.A., Nie,Q. and Zhang,X. (2017) LncRNA-Six1 encodes a micropeptide to activate Six1 in Cis and is involved in cell proliferation and muscle growth. *Front. Physiol.*, **8**, 230.
23. Huang,J.Z., Chen,M., Chen, Gao,X.C., Zhu,S., Huang,H., Hu,M., Zhu,H. and Yan,G.R. (2017) A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth. *Mol. Cell*, **68**, 171–184.
24. Silvester,N., Alako,B., Amid,C., Cerdeno-Tarraga,A., Clarke,L., Cleland,I., Harrison,P.W., Jayathilaka,S., Kay,S., Keane,T. *et al.* (2018) The European nucleotide archive in 2017. *Nucleic Acids Res.*, **46**, D36–D40.
25. Kodama,Y., Mashima,J., Kosuge,T., Kaminuma,E., Ogasawara,O., Okubo,K., Nakamura,Y. and Takagi,T. (2018) DNA Data Bank of Japan: 30th anniversary. *Nucleic Acids Res.*, **46**, D30–D35.
26. Ewels,P., Magnusson,M., Lundin,S. and Kaller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
27. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, **17**, 10–12.
28. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Giron,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
29. Casper,J., Zweig,A.S., Villarreal,C., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Karolchik,D. *et al.* (2018) The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.*, **46**, D762–D769.
30. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
31. 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.
32. 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.
33. Ruiz-Orera,J., Messeguer,X., Subirana,J.A. and Alba,M.M. (2014) Long non-coding RNAs as a source of new peptides. *eLife*, **3**, e03523.
34. Ji,Z., Song,R., Regev,A. and Struhl,K. (2015) Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife*, **4**, e08890.
35. Raj,A., Wang,S.H., Shim,H., Harpak,A., Li,Y.I., Engelmann,B., Stephens,M., Gilad,Y. and Pritchard,J.K. (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife*, **5**, e13328.
36. Johnstone,T.G., Bazzini,A.A. and Giraldez,A.J. (2016) Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.*, **35**, 706–723.
37. Young,S.K. and Wek,R.C. (2016) Upstream open reading frames differentially regulate Gene-specific translation in the integrated stress response. *J. Biol. Chem.*, **291**, 16927–16935.