D254–D258 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 3 October 2015*
doi: 10.1093/nar/gkv972

RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling

Shang-Qian Xie[1,2,†], Peng Nie[3,†], Yan Wang[1,†], Hongwei Wang[1], Hongyu Li[3], Zhilong Yang[4], Yizhi Liu[1,*], Jian Ren[3,*] and Zhi Xie[1,2,*]

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China, [2]Scientific Center for Precision Medicine, Sun Yat-sen University, Guangzhou 510000, China, [3]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, China and [4]Division of Biology, Kansas State University, Manhattan, KS 66506, USA

Received July 29, 2015; Revised September 10, 2015; Accepted September 15, 2015

ABSTRACT

Translational control is crucial in the regulation of gene expression and deregulation of translation is associated with a wide range of cancers and human diseases. Ribosome profiling is a technique that provides genome wide information of mRNA in translation based on deep sequencing of ribosome protected mRNA fragments (RPF). RPFdb is a comprehensive resource for hosting, analyzing and visualizing RPF data, available at www.rpfdb.org or http://sysbio.sysu.edu.cn/rpfdb/index.html. The current version of database contains 777 samples from 82 studies in 8 species, processed and reanalyzed by a unified pipeline. There are two ways to query the database: by keywords of studies or by genes. The outputs are presented in three levels. (i) Study level: including meta information of studies and reprocessed data for gene expression of translated mRNAs; (ii) Sample level: including global perspective of translated mRNA and a list of the most translated mRNA of each sample from a study; (iii) Gene level: including normalized sequence counts of translated mRNA on different genomic location of a gene from multiple samples and studies. To explore rich information provided by RPF, RPFdb also provides a genome browser to query and visualize context-specific translated mRNA. Overall our database provides a simple way to search, analyze, compare, visualize and download RPF data sets.

INTRODUCTION

Translational control is crucial in the regulation of gene expression. Deregulation of translation is associated with a wide range of cancers and human diseases (1). For example, hereditary hyperferritinaemia-cataract syndrome (HHCS) is an autosomal dominant disorder caused by mutations in the iron-response elements of ferritin. The mutation causes increased translation of ferritin mRNA and, hence, elevated serum levels of ferritin. This leads to nuclear cataract, an eye disease that eventually progresses to total blindness (2). Therefore, accurate measurement of translated mRNA is invaluable to better understand cellular functions and human diseases.

Nuclease footprinting is a conventional way to determine ribosome positions on mRNA, where the 28–30 nucleotides of mRNAs protected by a ribosome indicate translated mRNA (3,4). Ribosome profiling is a recently developed high-throughput strategy based on deep sequencing of ribosome-protected mRNA fragments (RPF) (5,6), that provides genome-wide information of mRNA in translation. Since its inception in 2009, RPF technique has been utilized in a range of studies in both prokaryotic and eukaryotic organisms and the number of studies increases rapidly every year (4,6–8).

To date, GWIPS-viz is the only database specifically designed for RPF data sets (9). GWIPS-viz provides a very valuable online genome browser to view the coverage and distribution of RPF reads. Currently it hosts RPF and mRNA data sets from 45 studies. Another relevant database is TISdb that is based on the recently developed Global Translation Initiation sequencing (GTI-seq) technology which provides global mapping of translation initiation codons (10). TISdb provides tools to search for translation initiation sites and the associated open reading frames (ORFs) based on multiple GTI-seq datasets. Since the num-

*To whom correspondence should be addressed. Tel: +86 20 87335131. Email: xiezhi@gmail.com
Correspondence may also be addressed to Jian Ren. Tel: +86 20 87343088. Email: renjian.sysu@gmail.com
Correspondence may also be addressed to Yizhi Liu. Tel: +86 20 87335131. Email: yizhi_liu@aliyun.com
†These authors contributed equally to the paper as first authors.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://nar.oxfordjournals.org/ at Sun Yat-Sen University on January 8, 2017

bers of studies using RPF technique have been growing significantly in the recent years, there is a strong need for an integrated database that facilitates the exploration of data from these studies. Furthermore, in addition to visualization of RPF reads, there is also an emerging database demand both for hosting the meta information of the studies but also for in-depth analysis of studies in a consistent way.

Herein, we present RPFdb, a comprehensive resource for hosting, analyzing and visualizing RPF data sets, available at www.rpfdb.org or http://sysbio.sysu.edu.cn/rpfdb/index.html. The current version of database contains 777 samples from 82 studies in 8 species, reprocessed by a unified pipeline. The main functions of the database include Browse, Search and Download, summarized in Figure 1. There are two ways to query the database: by keywords of studies or by genes. The outputs are presented in three levels. (i) Study level: including meta information of studies and reprocessed data for gene expression of translated mRNAs; (ii) Sample level: including global perspective of translated mRNA and a list of the most translated mRNA of each sample from a study; (iii) Gene level: including normalized sequence counts of translated mRNA on different genomic location of a gene from multiple samples and studies. To explore the rich information provided by RPF, RPFdb also provides a genome browser to query and visualize context-specific translated mRNA. Overall our database provides a simple way to search, analyze, compare, visualize and download RPF data sets.

MATERIALS AND METHODS

Data sources

The RPF sequencing data were collected from Gene Expression Omnibus (GEO) and Short Read Archive (SRA) databases. The current version contains 777 samples from 82 studies in 8 species: Arabidopsis, C. elegans, Drosophila, E. coli, Human, Mouse, Yeast and Zebrafish. Figure 2 shows the distribution of studies and samples by species. Human and yeast are the two species that were most studied.

Data processing

The pipeline for data processing is summarized in Figure 3. Specifically, SRAToolkit v2.4.3 (http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software) was used to convert sra files to fastq format. And FastQC v0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) was used to access the base quality of raw data. Since ribosome encloses a ∼30 nucleotides of mRNA and the 3′ sequences are usually linkers (5), we selected the first 26 nucleotides of the sequences for the subsequent alignment against reference genome as described in the literature (11). STAR v2.4.0i was used for alignment where one mismatch was allowed and multiple alignments were accepted (12). Supplemental Table S1 shows reference genome and related annotation files of eight species. We removed any sample if the number of its uniquely mapped reads is less than 1 million.

Data analysis

RNA–SeQC was used to generate two types of statistics of mapped reads: (i) the numbers of mapped and unmapped reads of each sample and (ii) the mapped ratio in exonic, intronic and intergenic regions of each sample (13). Reads Per Kilobase per Million mapped reads (RPKM) was used to measure RPF abundance, as defined below (14).

$$RPKM = r_f \times 10^9/(R \times fl_f),$$

where R is the total mapped reads in all genes, fl_f is the feature length and r_f is the raw mRNA counts.

The raw mRNA counts was calculated by HTseq-count v0.6.1p1 (15). Exonic loci of CDS and exon, were retrieved directly from annotation file against reference genome (Supplemental Table S1), and exonic loci of 3′ UTR and 5′ UTR were extracted by R package GenomicFeatures (16). If an exonic locus was classified as UTR for one transcript and as CDS for another transcript of the same gene, UTR or CDS region for each transcript was annotated separately, and the same exon was counted as both UTR and CDS. The feature length from multiple transcripts of the gene were merged.

Database implementation

The database was implemented by PHP, MySQL and JavaScript. The study and gene information were stored and queried by MySQL and PHP. The JavaScript JQuery and D3.js library were used for producing dynamic and interactive data visualization in the web browser. In addition, we integrated JBrowse in our database for visualizing context-specific translated mRNA intuitively. And the aligned RPF sequences in each species against their reference and annotation information are hosted in the genome browser.

RESULTS

Usage and access

The RPFdb includes Home, Browse, Search, Download and Help pages.

Search. This page provides two ways to query the database. (i) Search gene: by selecting a species and entering a gene symbol or Ensembl ID in the search box of the search page (also appears in the home page), the output shows the genome information of this gene from all the samples for the selected species, including RPKM of the gene and RPKM of the 5′ UTR, CDS and 3′ UTR regions. The users can sort the table by clicking the column names. The users may also use a search box on the output page to filter the results. A snapshot of output of 'Search gene' is shown in Supplemental Figure S1. The JBrowse icon provides hyperlink to a genome browser, which is described in the next section. (ii) Search study: users can search studies by keywords. This feature is useful to retrieve data set from the curated studies in the database. A snapshot of output of 'Search study' is shown in Supplemental Figure S2.

Genome browser. To explore the distribution of RPF reads for a given gene, RPFdb provides a genome browser to query and visualize context-specific translated mRNA. A

Downloaded from http://nar.oxfordjournals.org/ at Sun Yat-Sen University on January 8, 2017

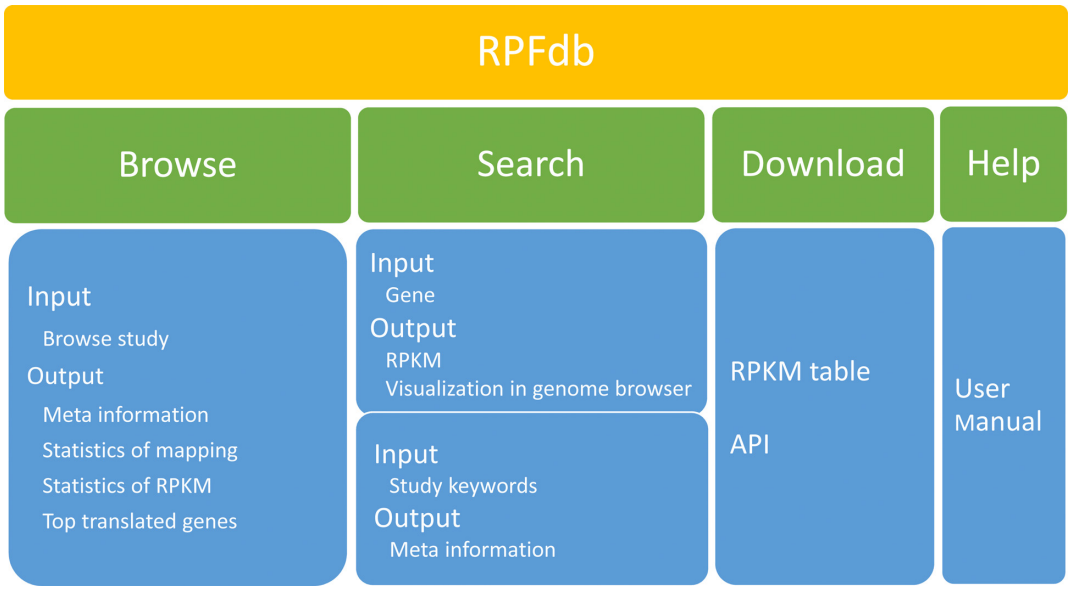

Figure 1. Structure and contents of RPFdb.

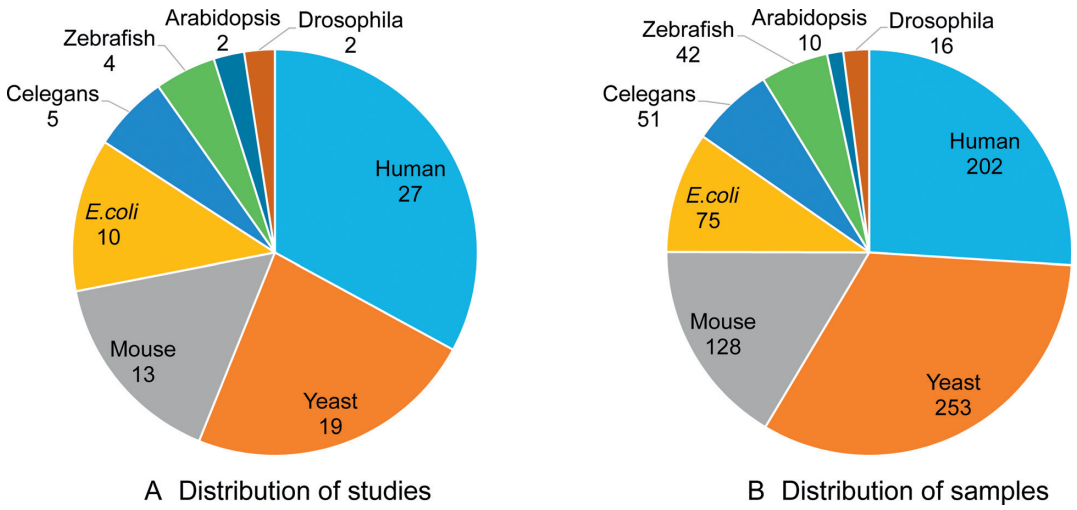

Figure 2. Distribution of studies and samples in species

Table 1. Comparison of RPFdb and GWIPS-viz

	RPFdb	GWIPS-viz
No. of studies	82	45
Type of data sets	RPF only	RPF and mRNA-seq
Data processing	• The first 26 nucleotides kept; • One mismatch allowed	• The adaptor linker sequence or poly-(A) tails trimmed from the 3′ ends of reads; • Three mismatches with alignment allowed
Aligner	STAR	Bowtie
Genome browser	Jbrowse	UCSC genome browser
Meta information	Searchable	Not searchable
Main features	• Statistics of studies and samples; • RPKM of RPF on different genomic location; • Visualization of RPF	• Visualization of RPF, inferred A-sites and mRNA; • Comparison of RPF and mRNA from the same sample

Downloaded from http://nar.oxfordjournals.org/ at Sun Yat-Sen University on January 8, 2017

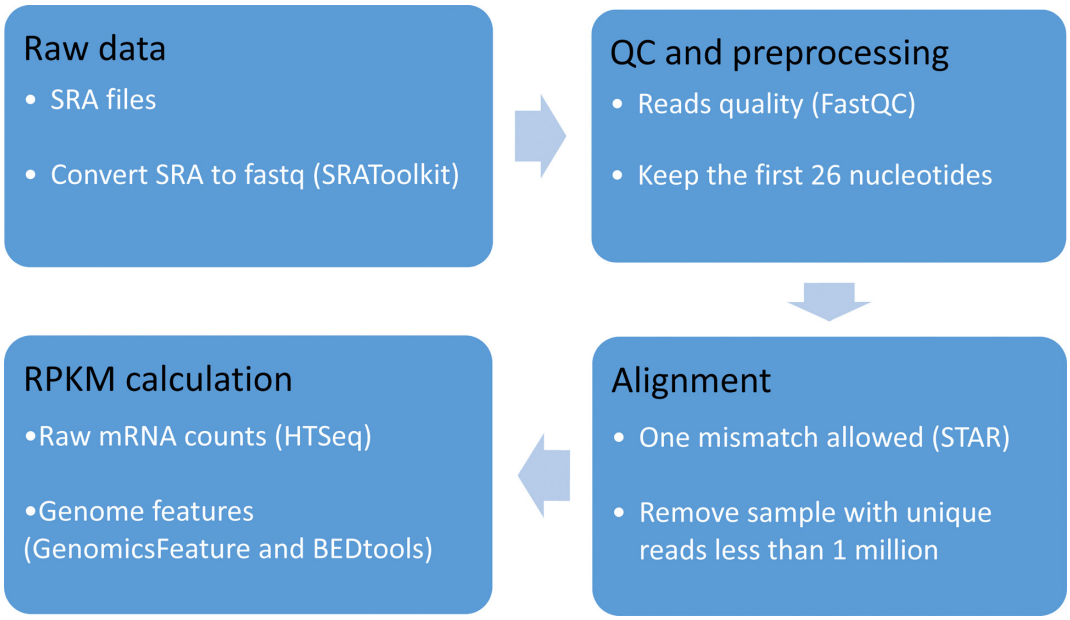

Figure 3. Data processing and RPKM calculation

snapshot of an example of 'Genome browser' is shown in Supplemental Figure S1. The annotated gene track and reference are displayed on the top of browser. RPF reads of the selected gene are shown at the bottom. If users select multiple samples, they can easily compare reads distribution on the genome.

Browse. For each study, this page displays: (i) Meta information of the study, including abstract of study, tissue or cell source, treatment for RPF experiment and reference genome for alignment (Supplemental Figure S3); (ii) The top 200 most translated genes. In addition, the whole RPKM table of the study is also downloadable (Supplemental Figure S4); (iii) Plots showing overall statistics of each sample, including the numbers and fraction of mapped and unmapped reads in each sample, and statistics of RPKM on different genomic regions of each sample, including 5′ UTR, CDS, 3′ UTR and gene (Supplemental Figures S3 and S4).

Download. Download page also has search function so that users can quickly find out their interested data set for downloading. In addition, we also support the Application Programming Interface (API), which allows developers to obtain the analysis result from RPFdb by using a HTTP client. The server-side programs in RPFdb accept a fixed URL syntax for retrieval operations. For example, the search result of gene THI1 in Arabidopsis can be returned by using the URL http://sysbio.sysu.edu.cn/rpfdb/fetchExpression.php?gene=THI1&species=Arabidopsis. Besides, both gene symbol and Ensembl ID can be used as query keys.

A case study

Here we show how to explore genes or studies of interest. For example, we want to know how mouse gene *Swi5* is translated under different conditions. On either the home page or the search page, we can select the species, mouse, and enter *Swi5* or Ensembl gene ID *ENSMUSG00000044627*. The output page displays RPKM of *Swi5* from 128 samples. We next want to know how *Swi5* is translated in mouse embryonic stem cells, 'E14'. Entering 'E14' into the search box of the output page, it displays the results from the studies using 'E14' (11,17) (Supplemental Figure S5A). The output shows that RPKM in 3′ UTR is lower than that in 5′ UTR and CDS for all samples. And RPKM in 5′ UTR is much higher than that in CDS for the harringtonine-treated samples (Supplemental Figure S5A). This can be explained that ribosomes dropped after stop codon in 3′ UTR whereas ribosomes accumulate at initiation sites in 5′ UTR for the harringtonine treated cells (11). If we want to explore reads distribution of cycloheximide and harringtonine treated cells in the genome browser. We select samples of interest and click 'Genome Browse'. The output page shows the reads coverage on genome (Supplemental Figure S5B). It shows that the harringonine-treat cells have one clear peak in the 5′ UTR region, in contrast, cycloheximide treated cells have reads in multiple exon regions and 5′ UTR region (Figure Supplemental Figure S5C and SD). These features are consistent with the results presented in the original publication (11).

Another useful function of RPFdb is to query studies of interest. For example, we enter 'cell cycle' in the search study page. The output page shows the relevant study of Stumpf CR *et al.* (18). Click 'Details', it leads to the page displaying the summary of the study and global perspective of the genome-wide distribution of RPF reads (Supplemental Figure S2).

DISCUSSION

The numbers of studies using RPF technique have been growing significantly in the recent years. There is a strong

Downloaded from http://nar.oxfordjournals.org/ at Sun Yat-Sen University on January 8, 2017

need for an integrated database that facilitates the exploration of data from these studies. Here we present RPFdb, a comprehensive resource for hosting and analyzing the publicly available RPF data sets. The main functions of RPFdb include to search studies of interest, to explore basic statistics of reads of the studies, to compare reads of translated mRNA of given genes from multiple studies and to visualize RPF data in a genome browser.

GWIPS-viz (9) and RPFdb are both useful resources for users who are interested in translated mRNA but with little computational knowledge. In order to help users to choose an appropriate database, we compared RPFdb and GWIPS-viz regarding the number of studies, the type of collected data sets, data processing method, aligner, genome browser and main features (Table 1).

Owing to the increasing interest in ribosome profiling and in translational regulation in general, we envision that ribosome profiling technology will be applied to a broader set of species and conditions and more publications will be released in future. RPFdb will be updated in a timely manner with new released data from public studies. We will also make efforts to improve the database. We hope that RPFdb will be a valuable resource for both experimental and computational biologists who are interested in understanding translational regulation and gene regulation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Dr. Nicholas Ingolia, Dr. Pasha Baranov and all the members of Zhi Xie's lab to provide suggestions to help improve the database.

FUNDING

Recruitment Program of Global Experts and National Natural Science Foundation of China [31471232, 31471252]; National Basic Research Program (973 project) [2013CB933900 and 2012CB911201]; Guangdong Natural Science Foundation [S20120011335 and 2014A030313181]; Program of International S&T Cooperation [2014DFB30020]. Funding for open access charge: Recruitment Program of Global Experts and National Natural Science Foundation of China [31471232, 31471252]; National Basic Research Program (973 project) [2013CB933900 and 2012CB911201]; Guangdong Natural Science Foundation [S20120011335 and 2014A030313181]; Program of International S&T Cooperation [2014DFB30020].
Conflict of interest statement. None declared.

REFERENCES

1. Calkhoven,C.F., Muller,C. and Leutz,A. (2002) Translational control of gene expression and disease. *Trends Mol. Med.*, **8**, 577–583.
2. Millonig,G., Muckenthaler,M.U. and Mueller,S. (2010) Hyperferritinaemia-cataract syndrome: worldwide mutations and phenotype of an increasingly diagnosed genetic disorder. *Hum. Genomics*, **4**, 250–262.
3. Stern-Ginossar,N. (2015) Decoding viral infection by ribosome profiling. *J. Virol.*, **89**, 6164–6166.
4. 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.
5. 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.
6. Oh,E., Becker,A.H., Sandikci,A., Huber,D., Chaba,R., Gloge,F., Nichols,R.J., Typas,A., Gross,C.A. and Kramer,G. (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell*, **147**, 1295–1308.
7. Guo,H., Ingolia,N.T., Weissman,J.S. and Bartel,D.P. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835–840.
8. Stadler,M., Artiles,K., Pak,J. and Fire,A. (2012) Contributions of mRNA abundance, ribosome loading, and post-or peri-translational effects to temporal repression of C. elegans heterochronic miRNA targets. *Genome Res.*, **22**, 2418–2426.
9. Michel,A.M., Fox,G., M Kiran,A., De Bo,C., O'Connor,P.B.F., Heaphy,S.M., Mullan,J.P.A., Donohue,C.A., Higgins,D.G. and Baranov,P.V. (2014) GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res.*, **42**, D859–D864.
10. Wan,J. and Qian,S.B. (2014) TISdb: a database for alternative translation initiation in mammalian cells. *Nucleic Acids Res.*, **42**, D845–D850.
11. 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.
12. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
13. DeLuca,D.S., Levin,J.Z., Sivachenko,A., Fennell,T., Nazaire,M.-D., Williams,C., Reich,M., Winckler,W. and Getz,G. (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, **28**, 1530–1532.
14. 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.
15. Anders,S., Pyl,P.T. and Huber,W. (2014) HTSeq–A Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
16. Lawrence,M., Huber,W., Pages,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.
17. 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.
18. Stumpf,C.R., Moreno,M.V., Olshen,A.B., Taylor,B.S. and Ruggero,D. (2013) The translational landscape of the mammalian cell cycle. *Mol. Cell*, **52**, 574–582.

Downloaded from http://nar.oxfordjournals.org/ at Sun Yat-Sen University on January 8, 2017