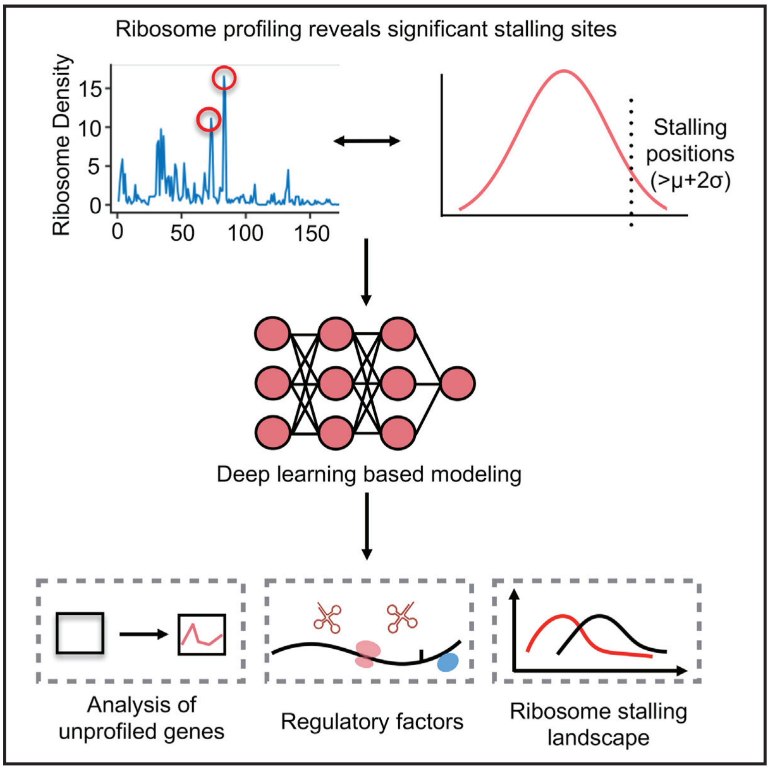
Cell Systems

Report

Analysis of Ribosome Stalling and Translation Elongation Dynamics by Deep Learning

Graphical Abstract


Ribosome profiling reveals significant stalling sites
Ribosome Density
15
10
5
0
0
50
100
150
Stalling positions (>μ+2σ)
Deep learning based modeling
Analysis of unprofiled genes
Regulatory factors
Ribosome stalling landscape

Authors

Sai Zhang, Hailin Hu, Jingtian Zhou, Xuan He, Tao Jiang, Jianyang Zeng

Correspondence

zengjy321@tsinghua.edu.cn

In Brief

ROSE provides a deep learning-based framework for estimating the likelihood of ribosome stalling from the mRNA sequence. Cotranslational events and regulatory factors related to ribosome stalling can be deciphered by ROSE trained on ribosome profiling data. ROSE will facilitate further studies on mRNA translation and protein biogenesis.

Highlights

- ROSE learns sequence features of ribosome stalling from ribosome profiling data
- ROSE accurately estimates the likelihood of ribosome stalling
- ROSE reveals interplays between cotranslational events and ribosome stalling
- ROSE provides a novel and valuable tool to complement ribosome profiling

Zhang et al., 2017, Cell Systems *5*, 212–220
September 27, 2017 © 2017 Elsevier Inc.
http://dx.doi.org/10.1016/j.cels.2017.08.004

CellPress

Cell Systems
Focus on RECOMB
Report

Analysis of Ribosome Stalling and Translation Elongation Dynamics by Deep Learning

Sai Zhang,[1,6] Hailin Hu,[2,6] Jingtian Zhou,[2,6] Xuan He,[1] Tao Jiang,[3,4,5] and Jianyang Zeng[1,7,*]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2]School of Medicine, Tsinghua University, Beijing, China
[3]Department of Computer Science and Engineering, University of California, Riverside, CA, USA
[4]MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST/Department of Computer Science and Technology, Tsinghua University, Beijing, China
[5]Institute of Integrative Genome Biology, University of California, Riverside, CA, USA
[6]These authors contributed equally
[7]Lead Contact
*Correspondence: zengjy321@tsinghua.edu.cn
http://dx.doi.org/10.1016/j.cels.2017.08.004

SUMMARY

Ribosome stalling is manifested by the local accumulation of ribosomes at specific codon positions of mRNAs. Here, we present ROSE, a deep learning framework to analyze high-throughput ribosome profiling data and estimate the probability of a ribosome stalling event occurring at each genomic location. Extensive validation tests on independent data demonstrated that ROSE possessed higher prediction accuracy than conventional prediction models, with an increase in the area under the receiver operating characteristic curve by up to 18.4%. In addition, genome-wide statistical analyses showed that ROSE predictions can be well correlated with diverse putative regulatory factors of ribosome stalling. Moreover, the genome-wide ribosome stalling landscapes of both human and yeast computed by ROSE recovered the functional interplays between ribosome stalling and cotranslational events in protein biogenesis, including protein targeting by the signal recognition particles and protein secondary structure formation. Overall, our study provides a novel method to complement the ribosome profiling techniques and further decipher the complex regulatory mechanisms underlying translation elongation dynamics encoded in the mRNA sequence.

INTRODUCTION

Translation elongation is a crucial step of mRNA translation, in which the ribosome scans the mRNA sequence and gradually grows the nascent peptide chain by appending new amino acids (Figure S1). Although numerous studies have shown that the local elongation rate along an mRNA sequence varies a lot, the underlying regulatory mechanisms of this phenomenon still remain unclear (Brar and Weissman, 2015; Chaney and Clark, 2015; Ingolia, 2014, 2016; Quax et al., 2015). On the other hand, translation elongation plays essential roles in diverse aspects of protein biogenesis, such as differential expression, cotranslational folding, covalent modification, and secretion (Chaney and Clark, 2015; Ingolia, 2016; Quax et al., 2015). In particular, ribosome stalling, which is described as ribosomes piling up at specific positions on mRNAs, can lead to various biological consequences, e.g., mRNA degradation (Buchan and Stansfield, 2007), modulation of protein expression (Tuller et al., 2010), alteration of protein conformations (Tsai et al., 2008), and pathological conditions (Ishimura et al., 2014). In addition, the connection between the local elongation rate and human health is increasingly emerging, which further underscores the necessity of a good understanding of the regulatory mechanisms and functions of elongation dynamics (Sauna and Kimchi-Sarfaty, 2011; Chaney and Clark, 2015).

In recent years, ribosome profiling has emerged as a high-throughput sequencing-based approach to measure the ribosome occupancy on mRNAs at a translatome-wide level *in vivo* (Ingolia et al., 2009, 2012; Brar and Weissman, 2015; Ingolia, 2014, 2016). With an accurate inference of the ribosome A site (i.e., the entry position of aminoacyl-tRNA) in a ribosome-protected fragment (also referred to as the ribosome footprint, ∼30 nucleotides), ribosome profiling provides a genome-wide snapshot of translation elongation dynamics and offers a new angle to estimate translation efficiency. Based on the currently available large-scale studies involving ribosome profiling experiments, several databases, e.g., GWIPS-viz (Michel et al., 2014) and RPFdb (Xie et al., 2015), have been established to store these profiling data.

Although a large amount of sequencing data have been produced by ribosome profiling, researchers are still challenged by the complexity, heterogeneity, and insufficient coverage of these data during the data analysis process (Brar and Weissman, 2015; Ingolia, 2014, 2016; Wang et al., 2016a, 2016b). Recently, deep learning has become one of the most popular and powerful techniques in the machine learning field (Hinton et al., 2006; Hinton and Salakhutdinov, 2006). Its superiority over traditional machine learning models has been demonstrated in a wide range of applications, such as speech recognition (Hinton et al., 2012), image classification (Hinton and Salakhutdinov, 2006), and natural language processing (Collobert et al., 2011). Specifically, deep learning has also been successfully applied to analyze large-scale genomic data and uncover notable biological patterns (Zhang et al., 2015; Alipanahi et al., 2015; Xiong et al., 2015; Zhou and Troyanskaya, 2015), such as the prediction of protein-nucleotide binding (Alipanahi et al., 2015; Zhang et al., 2015) and the effects of noncoding sequence variants (Zhou and Troyanskaya, 2015). In this work, we propose a deep learning-based framework, called ROSE (RibosOme Stalling Estimator), to address the aforementioned challenges and model translation elongation dynamics based on high-throughput ribosome profiling data.

In ribosome profiling experiments, we expect to observe a high ribosome density if ribosome stalling occurs reliably and repeatedly at a specific codon position and causes a ‘‘traffic jam’’ in translation there. Thus, in general, ribosome stalling events can be inferred from ribosome footprint density after normalization, and they have been widely believed to negatively correlate with the local elongation rates (Ingolia, 2014, 2016; Brar and Weissman, 2015). Our framework ROSE cast the ribosome stalling modeling problem into a classification task and predicted ribosome stalling using a deep convolutional neural network (CNN) with encoded sequence features. ROSE was trained in a supervised manner based on both human and yeast ribosome profiling data to revisit evolutionarily conserved observations about ribosome stalling.

RESULTS

Designing and Training ROSE

We proposed a deep learning-based framework, called RibosOme Stalling Estimator (ROSE), to analyze large-scale ribosome profiling data and study the contextual regulation of ribosome stalling and its potential functions in protein biogenesis (Figure 1A). Unlike previous work that characterized translation elongation dynamics using stochastic simulation approaches (Gritsenko et al., 2015; Pop et al., 2014; Reuveni et al., 2011) and density estimation (Liu and Song, 2016; O’Connor et al., 2016), ROSE formalized the modeling problem as a classification task, in which the resulting prediction score can be used to measure the probability of a ribosome stalling event. In this classification framework, codon positions with normalized ribosome footprint densities beyond two standard deviations (SDs) of the density distribution were defined as positive samples (foreground), which represented the occurrences of ribosome stalling, while the remaining sites were regarded as negative samples (background; Figure S2). This threshold was selected to best correlate the normalized reads with the model predictions in a separate validation dataset (Figure S3 and STAR Methods).

We assumed that a ribosome stalling event is primarily determined by its surrounding sequence. The codon position of interest, i.e., the ribosome A site, was first extended both upward and downward by 30 codons, which yielded the codon sequence profile of a putative stalling event. We then encoded this sequence and fed it into a deep convolutional neural network (CNN) to learn the complex relations between ribosome stalling and its contextual features (Figure 1B and STAR Methods). We called the prediction score directly output by the CNN the *intergenic ribosome stalling score*, also termed interRSS (STAR Methods). The name ‘‘interRSS’’ came from the fact that all the scores along the genome were calculated by a universal model and can be compared intergenetically/globally under the same criterion. To further eliminate the possible bias among different genes and facilitate the study on the interplay between intragenic/local factors (e.g., the binding of the signal recognition particle [SRP] on transmembrane segments) and elongation dynamics, we also normalized interRSS within each gene and obtained a local index, called the *intragenic ribosome stalling score*, also termed intraRSS (STAR Methods). Here we followed the same terminologies ‘‘intergenic’’ and ‘‘intragenic’’ from Quax et al. (2015). We collectively called both interRSS and intraRSS the *ribosome stalling score* (RSS). In principle, the RSS can be considered as an estimate of the likelihood of ribosome stalling. A higher RSS generally indicates a higher predicted probability of ribosome stalling at the corresponding codon position.

ROSE relied on a number of motif detectors (i.e., convolution operators) to scan the input sequence and integrated those stalling-relevant motifs to capture the intrinsic contextual features of ribosome stalling (Figure 1B and STAR Methods). Unlike previous CNN architectures used for analyzing biological data (Zhou and Troyanskaya, 2015; Alipanahi et al., 2015), our new CNN framework included multiple parallel convolution-pooling modules, which can not only significantly reduce the model complexity but also alleviate the potential overfitting problem (STAR Methods). After tuning the model hyperparameters using an efficient automatic strategy (STAR Methods), the error back-propagation algorithm was used to learn the network parameters of the CNN model (Rumelhart et al., 1986). We also deployed several optimization techniques, including L_2-regularization (Bengio, 2012), dropout (Srivastava et al., 2014; Bengio, 2012), and early stopping (Bengio, 2012), to further overcome the overfitting problem. To further boost the prediction performance, we also implemented an ensemble version of ROSE (termed eROSE), in which 64 CNNs were initialized and trained independently, and then the average result was used as the final prediction score (STAR Methods).

ROSE Accurately Predicts Ribosome Stalling

In this study, we mainly focused on eukaryotic cells, including human and yeast cells. We first used two datasets downloaded from GWIPS-viz (Michel et al., 2014), including a human dataset of lymphoblastoid cell lines (LCLs) (denoted by Battle15) (Battle et al., 2015) and a yeast dataset of *Saccharomyces cerevisiae* (denoted by Pop14) (Pop et al., 2014) to train our deep learning model and evaluate its prediction performance. In particular, after normalizing the ribosome profiling data and determining the threshold (i.e., $\mu + 2\sigma$; see STAR Methods for more details), the codon sites with normalized footprint densities beyond the threshold were labeled as positive samples, while an equal number of codon sites randomly chosen from the remaining were labeled as negative samples, which resulted in 109,770 and 20,902 samples for Battle15 and Pop14, respectively. For each dataset, we randomly selected 90% of the samples as training data and the remaining 10% as test data. The final performance of our model was mainly reported based on the test data.

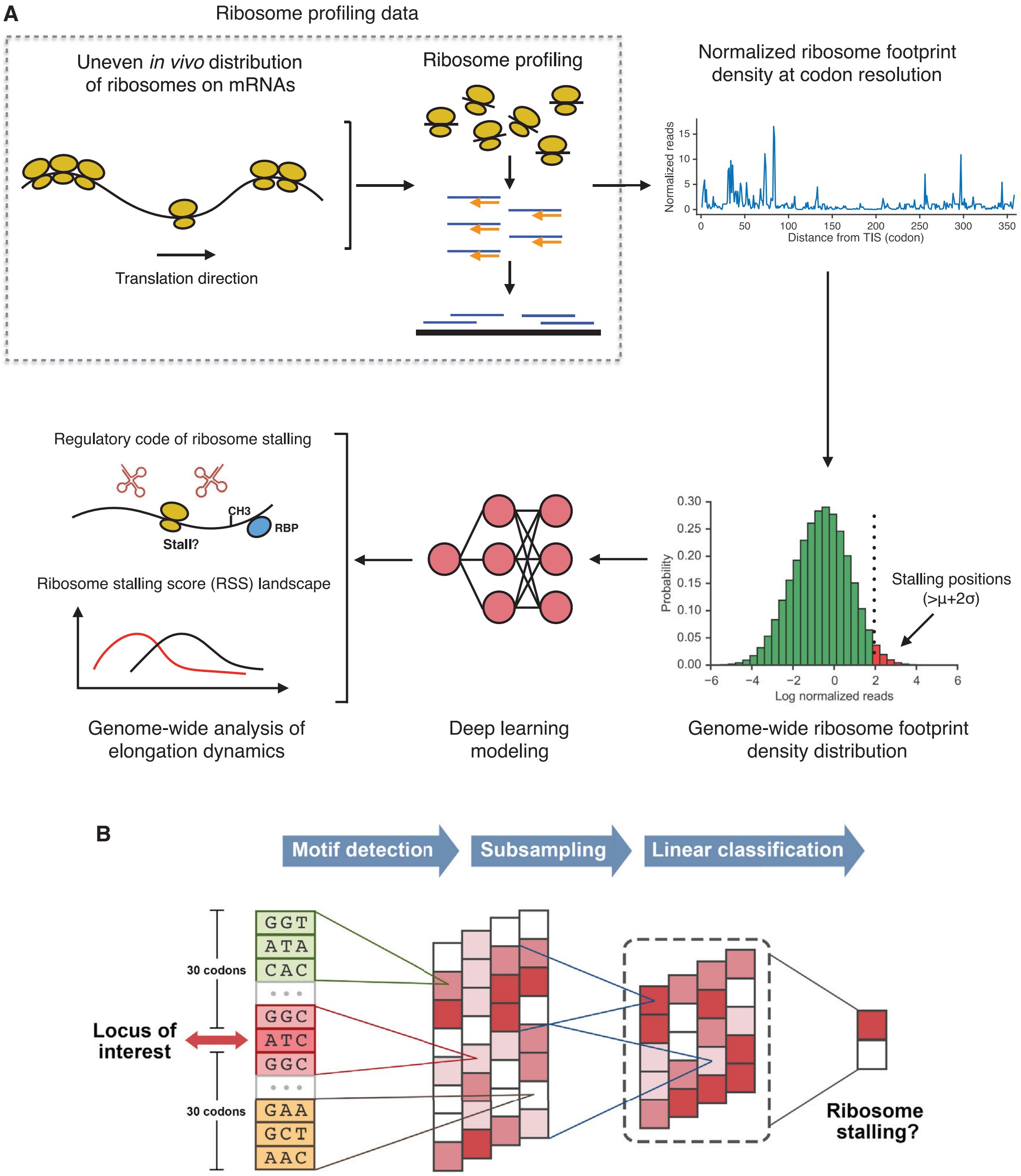


Figure 1. The ROSE Pipeline and the Convolutional Neural Network (CNN) Model

(A) Schematic overview of the ROSE pipeline. The codon sites with normalized ribosome footprint densities beyond two SDs are regarded as positive samples, which represent the ribosome stalling positions, to train a deep CNN model. Then the sequence profiles of individual codon sites along the genome are fed into the trained CNN to compute the distribution of ribosome stalling, which can be further used to study the potential factors affecting ribosome stalling and analyze the genome-wide landscape of translation elongation dynamics.

(B) Schematic illustration of the CNN model used in the ROSE pipeline. More details can be found in the main text.

When compared with a conventional prediction model, called gkm-SVM (Lee et al., 2015; Ghandi et al., 2014), ROSE showed superior performances on both human and yeast datasets with an increase in the area under the receiver operating characteristic curve (AUROC) by up to 18.4% (Figures 1A and 2B). In particular, the ensemble version of ROSE (i.e., eROSE) consistently had superior performance compared with the single version (i.e., sROSE). To validate the effectiveness of our parallel CNN architecture, we also implemented three sequential architectures that stacked two convolution-pooling modules with different kernel sizes in the convolutional layers before the output layer and found that sROSE greatly outperformed those sequential CNNs (Figure 2C).

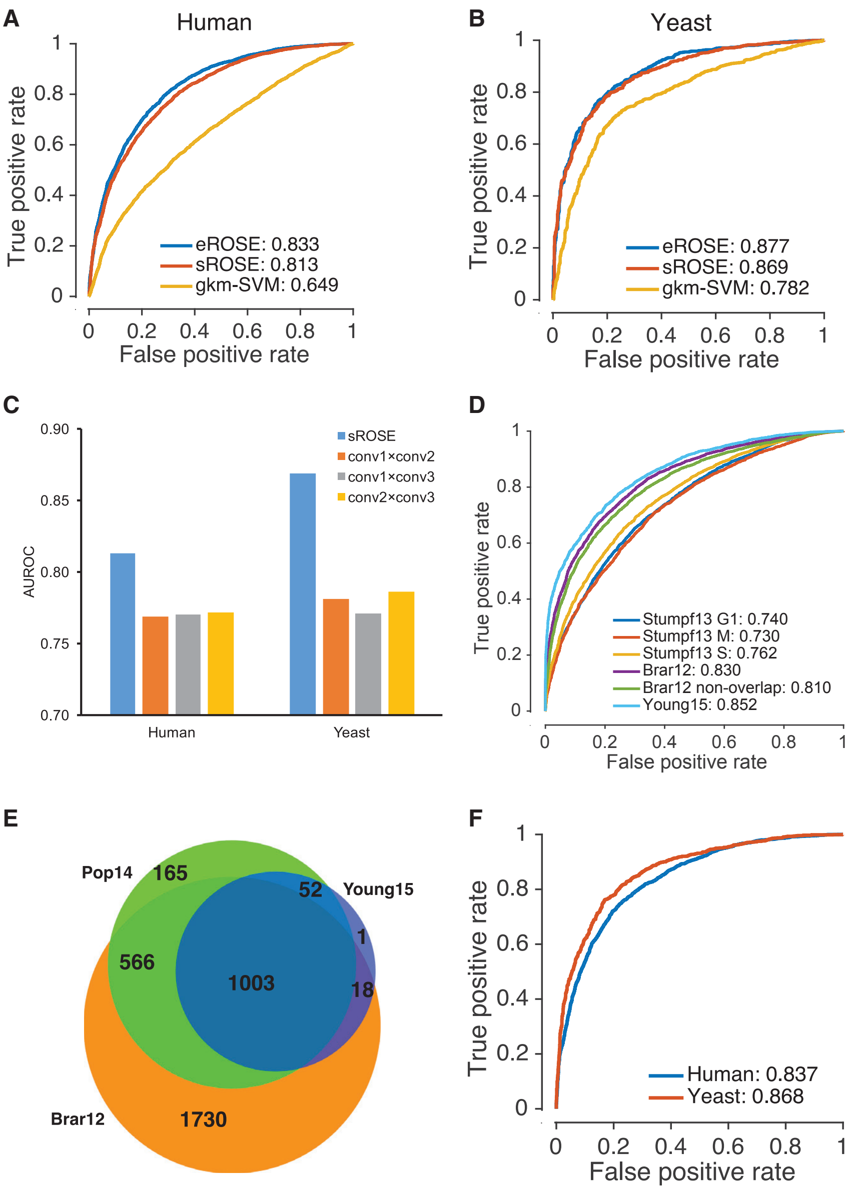


Figure 2. Performance Evaluation of ROSE

(A and B) The receiver operating characteristic (ROC) curves and the area under the corresponding ROC curve (AUROC) scores on the human (Battle15) and yeast (Pop14) test datasets, respectively.

(C) Comparison of AUROCs between parallel and sequential CNN architectures. "sROSE" and "eROSE" stand for the ROSE frameworks with one (single) and 64 (ensemble) CNNs, respectively. conv1, conv2, and conv3 represent the convolutional-pooling modules with kernel sizes of the convolutional layers corresponding to the short, medium, and long ranges used, respectively. × denotes the stacking operation in the sequential architecture.

(D) The ROC curves and the corresponding AUROC scores of the cross-study tests on additional human (Stumpf13 G1, M, and S) and yeast (Brar12 and Young15) datasets. The "Brar12 non-overlapping dataset" includes 1,748 genes with sufficient (over 60%) ribosome profiling coverage in the Brar12 dataset but not in the Pop14 dataset.

(E) The Venn diagram of the three yeast datasets regarding sufficiently covered genes.

(F) The ROC curves and the corresponding AUROC scores on the ramp regions.

We further performed multiple cross-study analyses to examine the generalizability of ROSE over five other ribosome profiling datasets with different experimental conditions, e.g., cell lines/strains and cycloheximide treatment (STAR Methods). Notably, ROSE showed only a moderate decrease in AUROC scores for both human and yeast, when using test datasets from other studies (Figure 2D), or even when the test samples came from genes that were not sufficiently covered (<60%) in the training dataset (Figures 2D and 2E).

It has been widely observed that the first 30–50 codons of a coding sequence are often enriched with rare codons, and create a "ramp" to reduce the elongation rate during the initial translation elongation process (Tuller et al., 2010; Tuller and Zur, 2014). To focus solely on the elongation process and remove the possible biases introduced by the ramp regions, here we excluded all the reads of these regions, i.e., the first 50 codons at the 5′ ends of the coding sequences, from our training data. Intriguingly, even without any training data from ramp regions, ROSE can still successfully predict ribosome stalling in these regions, with the AUROC scores above 83.0% (Figure 2F).

To better evaluate the performance of ROSE, we also provided precision-recall curves, the area under the precision-recall curve, and the accuracy/precision/recall scores for all the tests based on different thresholds (Figure S4 and Table S1).

RSS Associates with Putative Regulatory Factors of Ribosome Stalling

With stringent normalization procedures as well as superior prediction performance, ROSE enables one to systematically investigate diverse factors that may associate with ribosome stalling (STAR Methods). Here, we mainly focused on codon usage bias, tRNA adaptation, codon co-occurrence bias, proline codons, mRNA N^6-methyladenosine modification, mRNA secondary structure, protein-nucleotide binding, and positively charged amino acids, and studied how they correlate with the intraRSS (STAR Methods, Figures 3, S5, S6, S7, S8 and Table S2) as well as their predictive power for ribosome stalling (Figure S9). In addition to revisiting previously accepted conclusions, we also proposed several novel hypotheses based on the prediction results of ROSE, including a dose-dependent stalling tendency associated with proline residues (Figure 3C) and a negative correlation between ribosome stalling and codon co-occurrence (Figures 3A and 3B). We also carried out two negative control tests on a set of randomly selected codons and another set of

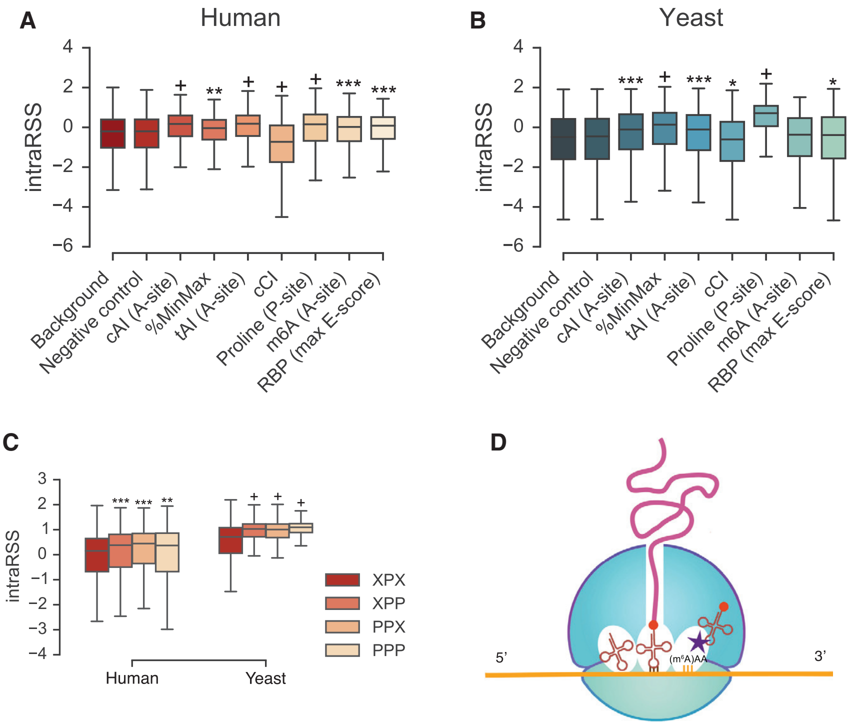


Figure 3. A Comprehensive Reexamination on the Relations between Diverse Putative Regulatory Factors and Ribosome Stalling Using ROSE

(A and B) Comparisons of intraRSS between the codon sites enriched with individual factors and the background for human and yeast, respectively. Data labels: negative control (randomly selected 10,000 codon sites), cAI (codon adaptation index), tAI (tRNA adaptation index), cCI (codon co-occurrence index), and %MinMax score (the codon rareness measurement proposed in Clarke and Clark, 2008).

(C) Comparisons of intraRSS between the single-peptide pattern of proline (i.e., XPX) and the multiple-peptide patterns of proline, including dipeptide (i.e., XPP and PPX) and tripeptide (i.e., PPP), where "P" and "X" stand for proline and any non-proline amino acid, respectively.

(D) A schematic illustration of the m^6A modification of a codon (e.g., AAA) to delay tRNA accommodation (denoted by the purple star) during translation elongation.

$^*5 \times 10^{-25} < p < 1 \times 10^{-2}$; $^{**}5 \times 10^{-50} < p \leq 5 \times 10^{-25}$; $^{***}5 \times 10^{-100} < p \leq 5 \times 10^{-50}$; $^{+}p \leq 5 \times 10^{-100}$; one-sided Wilcoxon rank-sum test. Note that here it is not necessary that the median was zero for the "Background" column, as the intraRSS was normalized within individual genes rather than along the genome. Also, the extremely small p values were partially due to the large sample size (N = 10,000).

randomly selected adenine-containing codon sites, in which the statistical comparisons with the background were insignificant (Figures 3A, 3B, and S5C), to further support the biological relevance of our investigations.

RSS Correlates with Protein Secondary Structure

Here, we sought to probe the relations between the protein secondary structure elements (SSEs) and ribosome stalling based on the RSS computed by ROSE. In particular, we first derived a set of non-redundant protein chains across human and yeast genomes from the PDB (Madej et al., 2012) (STAR Methods). We then investigated intraRSS landscapes of different SSE patterns, including a single chain of α helix (H), β strand (B), or random coil (C), and transitions between different SSEs.

We first obtained the average position-specific intraRSSes of each SSE pattern of interest with a specific window size (STAR Methods). Overall, we found that with a window size of six, all the tendencies of the intraRSS change for individual SSE patterns were species independent (Figures 4A and 4B; Spearman correlation coefficient $R > 0.6$). The conclusions were also confirmed with a window size of ten to eliminate the possible bias caused by the variation of window size. We further compared the intraRSSes of the structured (i.e., α helix or β strand) and random coil residues at the ribosome P sites. Consistent with the previous report that frequent codons were usually enriched in the structured regions while depleted in the random coils (Pechmann and Frydman, 2013), our results showed a significantly higher stalling probability in the coils than in the α helix or β strand regions (Figure 4C; $p < 10^{-25}$ by one-sided Wilcoxon rank-sum test). Furthermore, we examined the tendency of the intraRSS change along a protein secondary structure fragment. As expected, the intraRSS landscape showed a lower chance of stalling in the middle of a structured region but a higher chance in the middle of a coil region compared with the corresponding flanking regions on both sides (Figures 4A and S10A). This behavior was reminiscent of another previous study on the relations between codon frequency and protein secondary structure, in which the tRNA adaptation index (i.e., tAI) was mainly used as an indicator of the elongation rate (Saunders and Deane, 2010). Our intraRSS landscape showed a similar but more symmetrical trend to the previous finding that the transitions from structured to coil regions generally accompanied an increase in the stalling probability on the transition boundaries (Figures 4B and S10B). In addition, the opposite transitions (i.e., from coil to structured regions) exhibited roughly symmetrical trends in the change of intraRSS (Figures 4B and S10B).

RSS Associates with SRP Recognition

Next, we investigated whether the RSS landscape can reflect the elongation process that regulates the coupling between the protein translation and translocation activities. We were particularly interested in the interplay between the predicted likelihood of translational pause and the SRP binding of transmembrane (TM) segments (Figure 4D). We expected that our model would effectively capture the ribosome stalling events encoded by the heterogeneity of the amino acid composition in and around the TM domains.

We downloaded all the available TM protein sequences of human and yeast as well as the corresponding TM domain

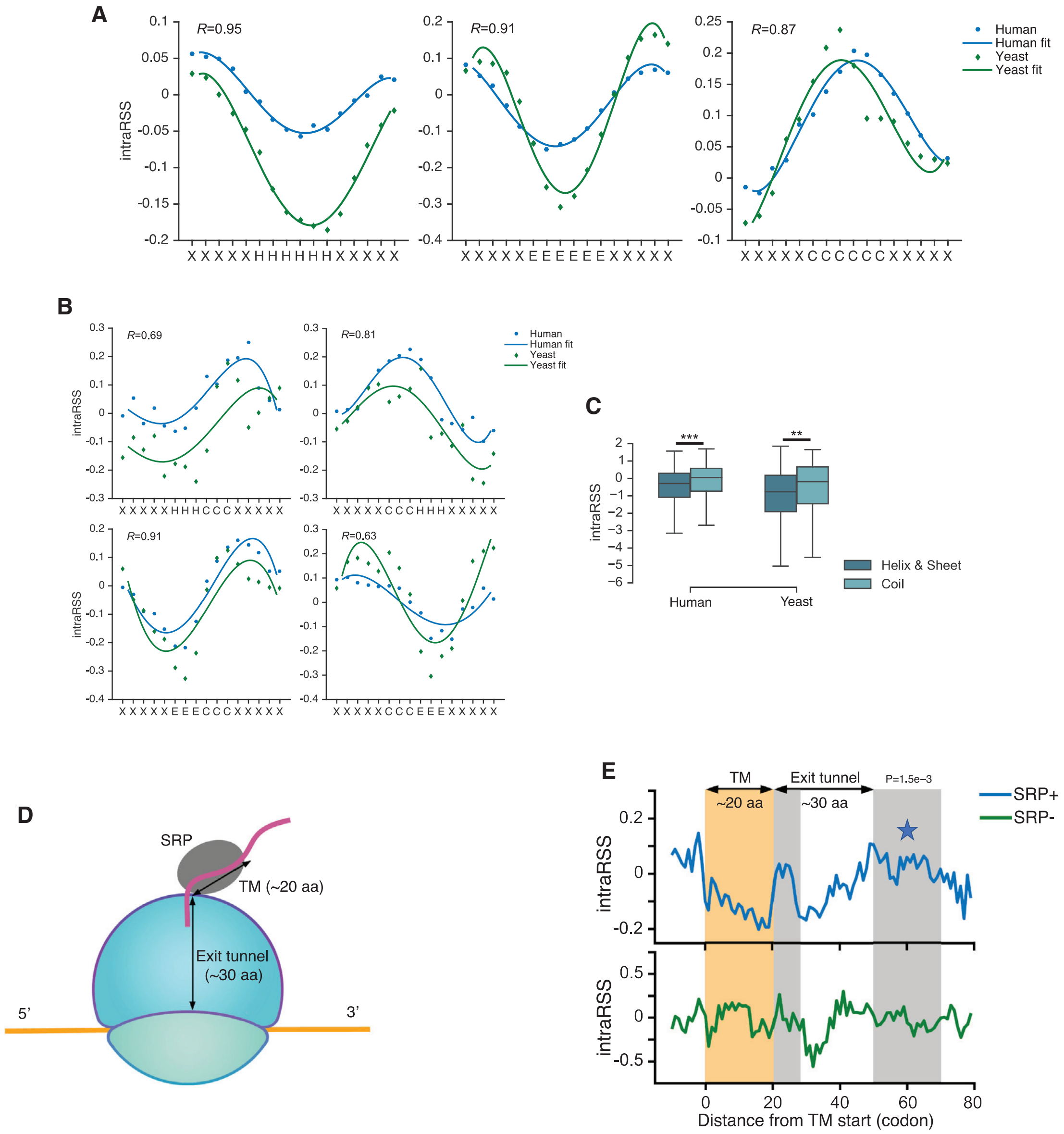


Figure 4. The Intragenic RSS Landscapes Reveal that Ribosome Stalling Associates with Protein Secondary Structure and the SRP Binding of Transmembrane (TM) Segments

(A) The intraRSS landscapes of α helix, β strand, and random coil regions.

(B) The intraRSS landscapes of the secondary structure element (SSE) transition regions. "H", "E" and "C" stand for α helix, β strand, and random coil, respectively, while "X" represents any SSE type in the flanking regions on both sides. Polynomial curve fitting of degree four was used to show the general intraRSS tendency. The Spearman correlation coefficients between human and yeast intraRSS tendencies were calculated.

(C) The overall comparisons of intraRSS between the structured (i.e., α helix and β strand) and random coil residues. $**5 \times 10^{-50} < p \leq 5 \times 10^{-25}$; $***5 \times 10^{-100} < p \leq 5 \times 10^{-50}$; one-sided Wilcoxon rank-sum test.

(D) A schematic illustration of the SRP binding of a TM segment during translation elongation.

(E) Comparison of intraRSS tendency between the TM segments with (SRP+) and without SRP binding (SRP−) in yeast, in which all the protein sequences were aligned with regard to the start of the TM segment whose position was indexed as zero. The yellow rectangle covers the TM segment, while the gray rectangles represent two intraRSS peaks downstream of the TM segment. The intraRSS peak marked with the blue star (i.e., positions from +50 to +70) was significantly diminished ($p = 1.5 \times 10^{-3}$ by one-sided Wilcoxon rank-sum test).

information from the Uniprot database (UniProt Consortium, 2015) (STAR Methods). We first focused on the yeast TM proteins, whose translation had been previously characterized both computationally and experimentally (Pechmann et al., 2014). The intraRSS landscape computed by ROSE captured two major stalling events during the TM protein translation process (Figure 4E). The first stalling event after the TM start occurred right at the end of the TM segment, where the structured TM segment (mostly α helix) transits to a more flexible intracellular region. This result agreed well with our previous conclusion about the relations between RSS and protein secondary structure (Figures 4B and S10B). The other intraRSS peak, spanning positions from +50 to +70, probably represented intrinsic stalling to promote nascent-chain recognition by SRP, which was consistent with a previous report (Pechmann et al., 2014). Indeed, a TM segment generally contains ~20 residues, and the length of the ribosome exit tunnel is ~30 residues. Thus, position +50 is approximately the place where the translated TM segment emerges from the exit tunnel and is bound by SRP. We also observed that the second peak was significantly diminished (Figure 4E; $p = 1.5 \times 10^{-3}$ by one-sided Wilcoxon rank-sum test) in the TM segments that were not associated with SRP binding (termed SRP−) (Alamo et al., 2011), which further validated our findings. The corresponding analyses for the end-aligned yeast TM proteins and human TM proteins can also be found in Figure S11.

DISCUSSION

In this study, we proposed a deep learning-based framework to predict the likelihood of ribosome stalling by integrating the underlying sequence features. To the best of our knowledge, our work is the *first* attempt to exploit the deep learning technique to predict ribosome stalling and model translation elongation dynamics based on large-scale ribosome profiling data. The detailed rationale for the application of deep learning in our problem setting can be found in STAR Methods.

Similar to many other high-throughput sequencing techniques, the current analysis of ribosome profiling data is also faced with several technical challenges, e.g., aligning reads across exon-exon junctions, ambiguous mapping, and sequencing bias (Ingolia et al., 2012). In this study, we relied on several widely accepted data preprocessing approaches, e.g., RNA-seq unified mapper (RUM) (Grant et al., 2011) in the alignment of splicing junction reads in GWIPS-viz (Michel et al., 2014) and the stringent normalization procedure proposed in Artieri and Fraser (2014), to at least partially remove the bias caused by these problems. Together with the accurate and robust prediction performance of ROSE as well as the physiologically relevant phenomena it detected, it is unlikely that the prediction of ROSE will suffer from the technical bias problem.

Our current study is a demonstration of applying ROSE in several specific scenarios in which ribosome stalling has been known to lead to significant physiological consequences. We believe that our ROSE framework will offer more insights into other important translation-related phenomena with the incorporation of more ribosome profiling data and more sophisticated problem formulation in the future.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
    - ○ Data Preprocessing and Normalization
    - ○ Model Design
    - ○ Model Training and Model Selection
    - ○ Rationales for the Application of Deep Learning in Our Problem Setting
    - ○ Implemrentation of gkm-SVM
- QUANTIFICATION AND STATISTICAL ANALYSIS
    - ○ Statistical Analysis on the Associations between Diverse Putative Factors and RSS
    - ○ Quantification of Diverse Putative Factors Related to Ribosome Stalling
    - ○ Statistical Analysis on the Associations between Protein Secondary Structure and RSS
    - ○ Statistical Analysis on the Association between Transmembrane Domains and RSS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes 12 figures and 4 tables and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2017.08.004.

AUTHOR CONTRIBUTIONS

S.Z., H.H., J. Zhou, and J. Zeng conceived the research project. J. Zeng supervised the research project. S.Z. preprocessed raw data, designed and implemented ROSE, and carried out model training and validation tasks. X.H. prepared the sequencing data of m^6A modification. S.Z., H.H., J. Zhou, T.J., and J. Zeng performed the computational and statistical analyses. S.Z., H.H., and J. Zeng wrote the manuscript. All the authors discussed the test results and commented on the manuscript.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China grant 61472205, the US National Science Foundation grant DBI-1262107 and IIS-1646333, China's Youth 1000-Talent Program, and the Beijing Advanced Innovation Center for Structural Biology. The authors are grateful to Q. Zhang and W. Chen for their helpful discussions about this work. They thank Ms. T. Chu for her help in preparing the figures in this paper. An early version of this paper was submitted to and peer reviewed at the 21st Annual International Conference on Research in Computational Molecular Biology (RECOMB 2017). The manuscript was revised and then independently further reviewed at *Cell Systems*.

Received: March 7, 2017
Revised: June 27, 2017
Accepted: August 4, 2017
Published: September 27, 2017

SUPPORTING CITATIONS

The following references appear in the Supplemental Information: Chen et al. (2014); Darnell et al. (2011).

REFERENCES

Alamo, M.D., Hogan, D.J., Pechmann, S., Albanese, V., Brown, P.O., and Frydman, J. (2011). Defining the specificity of cotranslationally acting chaperones by systematic analysis of mRNAs associated with ribosome-nascent chain complexes. PLoS Biol. *9*, e1001100.

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Artieri, C.G., and Fraser, H.B. (2014). Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. Genome Res. *24*, 2011–2021.

Ascano, M., Mukherjee, N., Bandaru, P., Miller, J.B., Nusbaum, J.D., Corcoran, D.L., Langlois, C., Munschauer, M., Dewell, S., Hafner, M., et al. (2012). FMRP targets distinct mRNA sequence elements to regulate protein expression. Nature *492*, 382–386.

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015). Impact of regulatory variation from RNA to protein. Science *347*, 664–667.

Bengio, Y. (2009). Learning deep architectures for AI. Foundations Trends Machine Learn. *2*, 1–127.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In Neural Networks: Tricks of the Trade, Second Edition, G. Montavon, G.B. Orr, and K.-R. Müller, eds. (Springer), pp. 437–478.

Bengio, Y., Courville, A.C., and Vincent, P. (2013). Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. *35*, 1798–1828.

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.

Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., and Weissman, J.S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science *335*, 552–557.

Buchan, J.R., and Stansfield, I. (2007). Halting a cellular production line: responses to ribosomal pausing during translation. Biol. Cell *99*, 475–487.

Cannarozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C., Gonnet, P., Gonnet, G., and Barral, Y. (2010). A role for codon order in translation dynamics. Cell *141*, 355–367.

Chaney, J.L., and Clark, P.L. (2015). Roles for synonymous codon usage in protein biogenesis. Annu. Rev. Biophys. *44*, 143–166.

Chen, E., Sharma, M.R., Shi, X., Agrawal, R.K., and Joseph, S. (2014). Fragile X mental retardation protein regulates translation by binding directly to the ribosome. Mol. Cell *54*, 407–417.

Choi, J., Ieong, K.W., Demirci, H., Chen, J., Petrov, A., Prabhakar, A., O'Leary, S.E., Dominissini, D., Rechavi, G., Soltis, S.M., et al. (2016). N6-methyladenosine in mRNA disrupts tRNA selection and translation-elongation dynamics. Nat. Struct. Mol. Biol. *23*, 110–115.

Clarke, T.F., 4th, and Clark, P.L. (2008). Rare codons cluster. PLoS One *3*, e3412.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. J. Mach. Learn. Res. *12*, 2493–2537.

Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K.Y., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W., et al. (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell *146*, 247–261.

Doerfel, L.K., Wohlgemuth, I., Kothe, C., Peske, F., Urlaub, H., and Rodnina, M.V. (2013). EF-P is essential for rapid synthesis of proteins containing consecutive proline residues. Science *339*, 85–88.

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., and Futcher, B. (2014). Measurement of average decoding rates of the 61 sense codons in vivo. Elife *3*, e03735.

Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M.A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput. Biol. *10*, 1–15.

Glorot, X., and Bengio, Y.. (2010), Understanding the difficulty of training deep feedforward neural networks. Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10), AUSTATS 249–256.

Grant, G.R., Farkas, M.H., Pizarro, A.D., Lahens, N.F., Schug, J., Brunk, B.P., Stoeckert, C.J., Hogenesch, J.B., and Pierce, E.A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics *27*, 2518–2528.

Gritsenko, A.A., Hulsman, M., Reinders, M.J.T., and de Ridder, D. (2015). Unbiased quantitative models of protein translation derived from ribosome profiling data. PLoS Comput. Biol. *11*, 1–26.

Gutierrez, E., Shin, B.S., Woolstenhulme, C.J., Kim, J.R., Saini, P., Buskirk, A.R., and Dever, T.E. (2013). eIF5A promotes translation of polyproline motifs. Mol. Cell *51*, 35–45.

Hinton, G.E., and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. Science *313*, 504–507.

Hinton, G.E., Osindero, S., and Teh, Y.W. (2006). A fast learning algorithm for deep belief nets. Neural Comput. *18*, 1527–1554.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal. Proc. Mag. *29*, 82–97.

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

Ingolia, N.T. (2016). Ribosome footprint profiling of translation throughout the genome. Cell *165*, 22–33.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

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.

Ishimura, R., Nagy, G., Dotu, I., Zhou, H., Yang, X.L., Schimmel, P., Senju, S., Nishimura, Y., Chuang, J.H., and Ackerman, S.L. (2014). Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. Science *345*, 455–459.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R.B., Guadarrama, S., and Darrell, T.. (2014). Caffe: convolutional architecture for fast feature embedding. arXiv:1408.5093.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637.

Ke, S., Alemu, E.A., Mertens, C., Gantman, E.C., Fak, J.J., Mele, A., Haripal, B., Zucker-Scharff, I., Moore, M.J., Park, C.Y., et al. (2015). A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation. Genes Dev. *29*, 2037–2053.

Kim, Y.. (2014). Convolutional neural networks for sentence classification. arXiv:1408.5882.

Kingma, D.P., and Ba, J.. (2014). Adam: a method for stochastic optimization. arXiv:1412.6980.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proc. IEEE *86*, 2278–2324.

Lee, D., Gorkin, D.U., Baker, M., Strober, B.J., Asoni, A.L., McCallion, A.S., and Beer, M.A. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nat. Genet. *47*, 955–961.

Linder, B., Grozhik, A.V., Olarerin-George, A.O., Meydan, C., Mason, C.E., and Jaffrey, S.R. (2015). Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat. Methods *12*, 767–772.

Liu, T.Y., and Song, Y.S. (2016). Prediction of ribosome footprint profile shapes from transcript sequences. Bioinformatics *32*, i183–i191.

Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA package 2.0. Algorithms Mol. Biol. *6*, 1–14.

Madej, T., Addess, K.J., Fong, J.H., Geer, L.Y., Geer, R.C., Lanczycki, C.J., Liu, C., Lu, S., Marchler-Bauer, A., Panchenko, A.R., et al. (2012). MMDB: 3D structures and macromolecular interactions. Nucleic Acids Res. *40*, D461–D464.

Michel, A.M., Fox, G., M Kiran, A., De Bo, C., O'Connor, P.B., Heaphy, S.M., Mullan, J.P., 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.

O'Connor, P.B.F., Andreev, D.E., and Baranov, P.V. (2016). Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. Nat. Commun. *7*, 12915.

Pechmann, S., and Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat. Struct. Mol. Biol. *20*, 237–243.

Pechmann, S., Chartron, J.W., and Frydman, J. (2014). Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. Nat. Struct. Mol. Biol. *21*, 1100–1105.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Peil, L., Starosta, A.L., Lassak, J., Atkinson, G.C., Virumäe, K., Spitzer, M., Tenson, T., Jung, K., Remme, J., and Wilson, D.N. (2013). Distinct XPPX sequence motifs induce ribosome stalling, which is rescued by the translation elongation factor EF-P. Proc. Natl. Acad. Sci. USA *110*, 15265–15270.

Pop, C., Rouskin, S., Ingolia, N.T., Han, L., Phizicky, E.M., Weissman, J.S., and Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Mol. Syst. Biol. *10*, 770.

Quax, T.E., Claassens, N.J., Söll, D., and van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. Mol. Cell *59*, 149–161.

Ray, D., Kazan, H., Cook, K., Weirauch, M., Najafabadi, H., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177.

Reis, M.D., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. *32*, 5036–5044.

Reuveni, S., Meilijson, I., Kupiec, M., Ruppin, E., and Tuller, T. (2011). Genome-scale analysis of translation elongation with a ribosome flow model. PLoS Comput. Biol. *7*, 1–18.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature *323*, 533–536.

Sabi, R., and Tuller, T. (2015). A comparative genomics study on the effect of individual amino acids on ribosome stalling. BMC Genomics *16*, S5.

Sauna, Z.E., and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. Nat. Rev. Genet. *12*, 683–691.

Saunders, R., and Deane, C.M. (2010). Synonymous codon usage influences the local protein structure observed. Nucleic Acids Res. *38*, 6719–6728.

Schwartz, S., Agarwala, S.D., Mumbach, M.R., Jovanovic, M., Mertins, P., Shishkin, A., Tabach, Y., Mikkelsen, T.S., Satija, R., Ruvkun, G., et al. (2013). High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. Cell *155*, 1409–1421.

Sharp, P.M., and Li, W.H. (1987). The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. *15*, 1281–1295.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. *15*, 1929–1958.

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.

Touw, W.G., Baakman, C., Black, J., te Beek, T.A., Krieger, E., Joosten, R.P., and Vriend, G. (2014). A series of PDB-related databanks for everyday needs. Nucleic Acids Res. *43*, D364–D368.

Tsai, C.J., Sauna, Z.E., Kimchi-Sarfaty, C., Ambudkar, S.V., Gottesman, M.M., and Nussinov, R. (2008). Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. J. Mol. Biol. *383*, 281–291.

Tuller, T., and Zur, H. (2014). Multiple roles of the coding sequence 5′ end in gene expression regulation. Nucleic Acids Res. *43*, 13–28.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell *141*, 344–354.

Ude, S., Lassak, J., Starosta, A.L., Kraxenberger, T., Wilson, D.N., and Jung, K. (2013). Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches. Science *339*, 82–85.

UniProt Consortium (2015). Uniprot: a hub for protein information. Nucleic Acids Res. *43*, D204–D212.

Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., et al. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. Nature *505*, 117–120.

Wang, X., Zhao, B.S., Roundtree, I.A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K., Shi, H., and He, C. (2015). N6-methyladenosine modulates messenger RNA translation efficiency. Cell *161*, 1388–1399.

Wang, H., McManus, J., and Kingsford, C. (2016a). Accurate recovery of ribosome positions reveals slow translation of wobble-pairing codons in yeast. In Proceedings of the 20th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2016), M. Singh, ed. (Springer), pp. 37–52.

Wang, H., McManus, J., and Kingsford, C. (2016b). Isoform-level ribosome occupancy estimation guided by transcript abundance with Ribomap. Bioinformatics *32*, 1880–1882.

Wohlgemuth, I., Brenner, S., Beringer, M., and Rodnina, M.V. (2008). Modulation of the rate of peptidyl transfer on the ribosome by the nature of substrates. J. Biol. Chem. *283*, 32229–32235.

Woolstenhulme, C.J., Parajuli, S., Healey, D.W., Valverde, D.P., Petersen, E.N., Starosta, A.L., Guydosh, N.R., Johnson, W.E., Wilson, D.N., and Buskirk, A.R. (2013). Nascent peptides that block protein synthesis in bacteria. Proc. Natl. Acad. Sci. USA *110*, E878–E887.

Xie, S.Q., Nie, P., Wang, Y., Wang, H., Li, H., Yang, Z., Liu, Y., Ren, J., and Xie, Z. (2015). RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. Nucleic Acids Res. *44*, D254–D258.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Young, D.J., Guydosh, N.R., Zhang, F., Hinnebusch, A.G., and Green, R. (2015). Rli1/ABCE1 recycles terminating ribosomes and controls translation reinitiation in 3' UTRs in vivo. Cell *162*, 872–884.

Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., and Zeng, J. (2015). A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Res. *44*, e32.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Human ribo-seq data (LCLs)	Battle et al., 2015	GEO GSE61742
Human ribo-seq data (HeLa)	Stumpf et al., 2013	NCBI SRA099816
Yeast ribo-seq data (288C)	Pop et al., 2014	GEO GSE63789
Yeast ribo-seq data (SK1)	Brar et al., 2012	GEO GSE34082
Yeast ribo-seq data (BY4741)	Young et al., 2015	GEO GSE69414
Software and Algorithms		
ROSE	This paper	v1.0
gkm-SVM	Ghandi et al., 2014	v1.3
Other		
GWIPS-viz Database	Michel et al., 2014	http://gwips.ucc.ie

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resource sharing may be directed to and will be fulfilled by Lead Contact Jianyang Zeng (zengjy321@tsinghua.edu.cn).

METHOD DETAILS

Data Preprocessing and Normalization

All ribosome profiling datasets in this study were downloaded from GWIPS-viz (Michel et al., 2014), in which abundant ribosome profiling data have been maintained and preprocessed as in other widely accepted pipelines (Ingolia et al., 2012). In addition to a human dataset of lymphoblastoid cell lines (LCLs) (denoted by Battle15) (Battle et al., 2015) and a yeast dataset of *S. cerevisiae* (denoted by Pop14) (Pop et al., 2014), we used other five additional datasets for cross-study validation, including three human datasets from different cell cycle stages (i.e., G1, S and M phases) of HeLa cells (Stumpf et al., 2013) (denoted by Stumpf13 G1, S and M, respectively) and two yeast datasets of strain SK1 (Brar et al., 2012) and starin BY4741 (Young et al., 2015) (denoted by Brar12 and Young15, respectively).

Here we applied the normalization method introduced in (Artieri and Fraser, 2014) to remove the technical and experimental biases from the ribosome profiling data. More specifically, after mapping the ribosome profiling and mRNA-seq reads to the reference genome, their codon-level reads were first scaled by the mean coverage level within each gene, which canceled out the coverage differences among genes. Next, the scaled ribosome profiling reads were divided by the scaled mRNA-seq reads in the corresponding locations to eliminate the shared biases between these two fractions. After that, a logarithm operation was further performed to yield the final normalized ribosome footprint density (Figure S2). Since some protein-coding genes can be poorly sequenced due to the issue of sequencing depth and the influence of differential expression, which may introduce unexpected biases to our analysis, here those normalized ribosome footprint densities from genes with sequencing coverage (i.e., the number of codon sites with both non-zero ribosome profiling and mRNA-seq reads divided by the total number of codons in the gene) less than **60%** were excluded from our training and test datasets. Note that such a coverage cutoff was also used in (Sabi and Tuller, 2015), in which robustness of this cutoff has been demonstrated in the analysis of ribosome profiling data.

To label samples for training a binary classifier to detect ribosome stalling events, we first tested several labeling thresholds based on the standard deviation of the normalized ribosome footprint density distribution. In particular, we considered four possible thresholds, including μ, $\mu+\sigma$, $\mu+2\sigma$ and $\mu+3\sigma$, where μ and σ represented the mean and the standard deviation of the normalized footprint density distribution, respectively. For each possible choice of threshold, those codon positions with normalized densities beyond the threshold were labeled as the ribosome stalling positions (i.e., positive samples), while the remaining were regarded as the background (i.e., negative samples). Then we trained four preliminary CNN models based on the datasets derived from these four thresholds, respectively. We also constructed a separate validation dataset that contained an equal number of samples randomly selected from six bins, including $(-\infty,\mu-2\sigma)$, $[\mu-2\sigma,\mu-\sigma)$, $[\mu-\sigma,\mu)$, $[\mu,\mu+\sigma)$, $[\mu+\sigma,\mu+2\sigma)$ and $[\mu+2\sigma,\infty)$. The basic principle of choosing the optimal threshold was the expectation that our model with the best threshold should yield predictions best correlated with their corresponding experimentally observed values in the independent validation dataset. We performed such a test for both human and yeast datasets, from which $\mu+2\sigma$ was determined as our final threshold (Figure S3).

After the above operations, the determined threshold was used to label samples, i.e., the codon sites with normalized footprint densities beyond the threshold were labeled as positive (i.e., foreground) samples, while the same number of codon sites randomly chosen from the remaining were labeled as negative (i.e., background) samples, which results in 109,770 and 20,902 samples for Battle15 and Pop14, respectively. For each dataset, we randomly selected **90%** of the samples as training data and the remaining **10%** as test data. The final performance of our model was mainly reported based on the test data. Note that here we excluded all the reads of the ramp regions (i.e., the first 50 codons at the 5′ ends of coding sequences) from the training data. For cross study validation, the ROC curves were obtained by applying the trained eROSE of the specific species to the validation data.

Model Design

A convolutional neural network (CNN) is a specific type of neural network in deep learning, which has been widely used in common data science fields, such as computer vision (LeCun et al., 1998) and natural language processing (Kim, 2014). In particular, CNNs have also been used to model biological sequence data, e.g., the predictions of protein-nucleotide binding (Alipanahi et al., 2015) and effects of noncoding variants (Zhou and Troyanskaya, 2015). Generally speaking, a CNN is comprised of multiple local motif detectors (i.e., convolution operators) that are invariant with certain transformations, such as translation and rotation, and subsampling (i.e., pooling operators) for dimension reduction and efficient training. To further increase the learning capacity of the network, many layers of these operators are often stacked together, and then followed by several fully-connected layers, and finally the output layer.

In our framework, we first encode the input codon sequence using the one-hot encoding technique (Pedregosa et al., 2011), that is, the *m*th codon type is encoded as a binary vector of length 64, in which the *m*th position is one while the others are zeros, after indexing all 64 codon types. Then the encoded information is fed into one convolutional layer and one pooling layer to learn the hidden features. In the convolutional layer, several one-dimensional convolution operations are performed over the 64-channel input data, in which each channel corresponds to one dimension of the input vector, and the weight matrix (i.e., kernel) can be regarded as the position weight matrix (PWM). More specifically, given a codon sequence $s=(c_1,\ldots,c_n)$ and the corresponding one-hot representation S, where n stands for the input length (here $n=61$ as we extend the codon site of interest on both sides by 30 codons) and $\boldsymbol{c_i}$ represents the ith codon in the sequence, the convolutional layer computes $X=\text{conv}(S)$, i.e.,

$$X_{i,k}=\sum_{j=0}^{m-1}\sum_{l=1}^{64}W_{k,j,l}S_{i+j,l},$$

where $1\leq i\leq n-m+1$, $1\leq k\leq d$, m is the kernel size, and d is the kernel number. Next, the rectified linear activation function (ReLU) is used to imitate the neuron activation, that is, the output of the convolutional layer is further processed by the activation function $Y=\text{ReLU}(X)$, where

$$\text{ReLU}(x)=\begin{cases}x & \text{if } x\geq 0,\\ 0 & \text{if } x<0.\end{cases}$$

After convolution and rectification, we reduce the dimension of matrix Y using the max pooling operation, which computes the maximum value within a scanning window of size three and step size two. More specifically, given the upstream input Y, the max pooling operation computes $Z=\text{pool}(Y)$, i.e.,

$$Z_{i,k}=\max\left(Y_{j,k},Y_{(j+1),k},\ldots,Y_{(j+m-1),k}\right),$$

where i is the index of the output position, j is the index of the start input position, k is the index of the kernel, and m is the size of the scanning window during the pooling operation (here we choose $\boldsymbol{m=3}$).

To enable the local motif detectors to scan sequence motifs in different ranges synchronously, while not increasing the model complexity too much, here we propose a *parallel* architecture, which includes three kernels of different sizes, corresponding to short (5–7), mediate (8–9) and long (10–13) ranges, respectively. The outputs of these three kinds of convolution operators are further rectified and then subsampled independently and in parallel, and finally concatenated into a unified representation U. To calculate the final probability of a ribosome stalling event, the unified representation is directly fed to a sigmoid layer, which computes

$$\Pr\{\text{Ribosome stalling}\}=\text{sigm}(U)=\frac{1}{1+exp(-WU)},$$

where W is the weight matrix of the sigmoid layer.

Note that the sequential (i.e., layer-wise) architecture in conventional CNNs, in which several convolutional and pooling layers are stacked together, can also detect motifs in different ranges. The reason that our parallel architecture can significantly reduce the model complexity comes from the fact that the parallelism simulates the SUM operation, e.g., $(a_1+a_2)+(b_1+b_2)$, while the sequentiality mimics the PRODUCT operation, e.g., $(a_1+a_2)\times(b_1+b_2)$. Obviously, the computational complexity of the latter is much higher than that of the former. Our network reduction can be useful for relieving the potential overfitting problem during the training process. We note that a similar idea has also been proposed in (Kim, 2014). However, the pooling operation in (Kim, 2014) is carried out over the whole convolutional layer without any window restriction, which is quite different from ours. In summary, a complete CNN in our deep learning framework can be formulated as

$$p(s) = \mathrm{sigm}\left(\mathrm{concat}_{i=1,2,3}\left(\mathrm{pool}^i\left(\mathrm{ReLU}^i\left(\mathrm{conv}^i(\mathrm{encode}(s))\right)\right)\right)\right),$$

where i represents the kernel index in the parallel architecture, and encode($\cdot$), conv($\cdot$), ReLU($\cdot$), pool($\cdot$), concat($\cdot$) and sigm($\cdot$) represent the one-hot encoding, convolution, ReLU, max pooling, concatenation and sigmoid operations, respectively.

The above calculated probability $p(s)$ is defined as the *intergenic ribosome stalling score* (also termed interRSS), which measures the likelihood of ribosome stalling at a codon position. To eliminate the interRSS bias among different genes, we further define the *intragenic ribosome stalling score* (also termed intraRSS) as follows,

$$\mathrm{intraRSS}(position|gene) = \log\left(\frac{\mathrm{interRSS}(position)}{\mathrm{mean}(gene)}\right),$$

where interRSS(*position*) represents the interRSS of the codon position of interest and mean(*gene*) stands for the mean interRSS of the corresponding gene. When computing mean(*gene*), we exclude those codon positions in the ramp regions (i.e., the first 50 codons at the 5′ ends of coding sequences).

Model Training and Model Selection

Given the training samples $\{(s_i,y_i)\}_i$, the loss function of our model is defined as the sum of the negative log likelihoods (NLLs), i.e.,

$$\sum_i \mathrm{NLL}_i = -\sum_i \log(y_i p(s_i) + (1-y_i)(1-p(s_i))),$$

where s_i is the input codon sequence and y_i is the true label. To train the CNN, the standard batch gradient descent method with the error backpropagation algorithm is performed (Rumelhart et al., 1986). To further optimize the training procedure, we also apply several training strategies, including the mini-batch and momentum techniques (Bengio, 2012). In addition, we use the Adam algorithm for stochastic optimization to achieve an adaptive moment estimation (Kingma and Ba, 2014). To further overcome the overfitting issue, we also apply several regularization techniques, including L_2-regularization-based weight decay (Bengio, 2012), dropout (Srivastava et al., 2014) and early stopping (Bengio, 2012).

The network structure and the aforementioned optimization techniques introduce a number of hyperparameters to our framework, such as the kernel size, kernel number, base learning rate, weight decay coefficient and the max number of training iterations. It is important to perform proper hyperparameter calibration and model selection for accurate modeling. Although we can achieve this goal using the conventional cross-validation strategies, it is generally time-consuming to test all possible combinations of these hyperparameters. To conquer this difficulty, here we propose a *one-way model selection* strategy for automatic and efficient hyperparameter calibration. In this strategy, we first arbitrarily choose the initial values of the hyperparameters from a candidate set. Then, we separate the hyperparameters into two groups, including those describing the network structure (denoted by H_1), such as the kernel size and the kernel number, and those describing the optimization procedure (denoted by H_2), such as the base learning rate and the weight decay coefficient. Next, by fixing the values of the hyperparameters in H_2, we calibrate those hyperparameters in H_1 using a three-fold cross-validation (CV) procedure, and determine their optimal values that achieve the best CV performance. Similarly, the hyperparameters in H_2 are also calibrated via the three-fold CV procedure after fixing the previously determined values of the hyperparameters in $\boldsymbol{H_1}$. The final values of all hyperparameters of ROSE are provided in Table S3. The ROC curves and AUROC scores of the CNNs with calibrated hyperparameters are shown in Figure S12 for the Battle15 and Pop14 datasets, respectively. Though we can carry out this procedure for more iterations (i.e., multi-way), our test results show that the one-way implementation generally yields satisfying prediction performance in this study.

After hyperparameter calibration and model selection, we train the final ROSE model using the whole training dataset. Due to the nature of non-convex optimization, random weight initialization may affect the search result of the gradient descent algorithm. Here, we use the Xavier initialization algorithm to automatically determine the initial scales of weights according to the number of input and output neurons (Glorot and Bengio, 2010). To account for the potential initialization bias and further boost the prediction performance, we also implement an ensemble version of ROSE (termed eROSE), in which 64 CNNs are trained independently and then combined together to compute the final prediction score, i.e.,

$$p(s) = \frac{1}{64}\sum_{i=1}^{64} p_i(s),$$

in which $\boldsymbol{p_i(s)}$ represents the probability calculated by the ith CNN.

Our implementation of ROSE depends on the Caffe library (Jia et al., 2014), and the Tesla K20c GPUs are used to speed up the training process.

Rationales for the Application of Deep Learning in Our Problem Setting

The application of deep learning in our problem setting is mainly based on the following rationales. First, we assumed that ribosome stalling can be characterized by its surrounding context and sequence motifs that encode different factors affecting ribosome stalling (Chaney and Clark, 2015; Ingolia, 2016; Quax et al., 2015). Here, we used the convolution layers, acting as motif detectors, to model and extract the complex nonlinear sequence features. The convolutional neural networks (CNN) have been successfully applied to model various biological sequence features in previous studies, showing superior prediction performance to conventional machine learning methods (Zhang et al., 2015; Alipanahi et al., 2015; Xiong et al., 2015; Zhou and Troyanskaya, 2015). In fact, the multi-layer neural networks have been shown to be a universal estimator of functions (Bengio, 2009), which means that for any function, there exists a neural network that can estimate its value in any accuracy. In addition, the multi-layer convolution can extract the input hierarchical features automatically without any artificial feature engineering and facilitate the binary classification/prediction in the final layer (Bengio, 2009). Although how a deep neural network automatically learns the intermediate features/representations is still an open question in the machine learning field, numerous empirical studies have demonstrated its effectiveness in various learning tasks (Bengio, 2009; Bengio et al., 2013). In general, with the abundant amount of training data, a deep neural network can often yield superior predictive power over conventional learning approaches (e.g., SVM). Based on these reasonings, as well as the fact that translation elongation dynamics is generally affected by a complicated interplay between heterogenous factors, we believe that the deep convolutional neural network is a proper choice to predict ribosome stalling from the large-scale ribosome profiling data.

Implemrentation of gkm-SVM

To conduct a pair comparison between ROSE and gkm-SVM, the length of the input mRNA sequence to gkm-SVM was also set to 183, that is, the codon position of interest was extended both upward and downward by 90 nucleotides (30 codons). Then gkm-SVM computed the gapped *k*-mer features of the input sequences to classify positive and negative samples (Lee et al., 2015; Ghandi et al., 2014). All the parameters of gkm-SVM were set as default values.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical Analysis on the Associations between Diverse Putative Factors and RSS

Diverse factors, such as tRNA adaptation and mRNA secondary structure, can interplay with each other to affect the ribosome stalling tendency. To investigate whether a factor potentially correlates with ribosome stalling, we first identified those codon sites along the genome that were enriched with this factor, and then checked whether the predicted (intra)RSSes of these positions were significantly different from those of the background. In particular, given a factor, such as tRNA adaptation, we first computed its quantity (e.g., tAI) across the genome and then chose those codon sites whose quantities were in the top *N* list (*N* was set to 10,000 in our study). After that, we ran the Wilcoxon rank sum test to compare the (intra)RSSes of the chosen sites to those of a background dataset, which was generated by randomly selecting 10,000 ribosome occupancy sites from the genome. If the (intra)RSSes of the codon sites enriched with the factor and the background were significantly different, we said this factor correlates with ribosome stalling. In addition, we probed the correlations between different factors based on the background dataset, and found little correlation between these factors that we were interested in, except for cAI, %MinMax and tAI (Table S4).

Since ROSE can also output a binary annotation for each codon site of interest, we further analyzed the binary annotations of those codon sites enriched with several main regulatory factors, including cAI, %MinMax, cCI and proline codons. More specifically, given a fixed threshold (which was set to be 0.5 in this study), we computed the distributions of the binary labels "1" (i.e., interRSS$\geq$0.5) and "0" (i.e., interRSS<0.5) output by ROSE for the top *N* enriched sites and *N* randomly selected genomic loci (here *N* was still set to be 10,000), respectively. The comparison between these two distributions of binary outputs was carried out using the chi-square test, yielding consistent conclusions with those derived from the probability outputs (Figure S8).

Quantification of Diverse Putative Factors Related to Ribosome Stalling

In this study, we mainly focused on codon usage bias, tRNA adaptation, codon cooccurrence bias, proline codons, mRNA N$^{\mathbf{6}}$-methyladenosine modification, mRNA secondary structure, protein-nucleotide binding and positively-charged amino acids, and investigated how they associate with the intraRSS.

For codon usage bias, we applied several proposed metrics for codon frequency estimation, such as the codon adaptation index (cAI) (Sharp and Li, 1987) and the %MinMax score (Clarke and Clark, 2008). In particular, we calculated cAI for both ribosome A- and P-sites, and %MinMax for the local region around the ribosome A-site (i.e., five codons both upstream and downstream from the A-site). Also, from a tRNA perspective, the tRNA adaptation index (tAI) has been proposed to consider both the tRNA concentration (approximated by the copy number of the corresponding tRNA gene) and the strength of codon-anticodon pairing (computed according to the Crick wobble rules) (Reis et al., 2004). Again, to facilitate the genome wide statistical analysis, the tAI values for both ribosome A- and P-sites for each codon across the genome were compared.

The codon cooccurrence bias, i.e., the non-uniform distribution of synonymous codon orders, was also reported to affect translation elongation dynamics (Cannarozzi et al., 2010; Quax et al., 2015). To further examine this factor, we first defined a new metric, called the *codon cooccurrence index* (cCI), which measures the autocorrelation (i.e., reuseness) of isoaccepting codons in a local region. Precisely speaking, given the codon of interest at position *i*, we only considered its local region $[i-w,i+w]$, where *w* stands for the window size. For each codon at position $p\in[i-w,i+w]$, we checked whether it had an isoaccepting codon in the upstream region $[i-u,p-1]$. We used notation iso_p to represent this indicator, that is, $\text{iso}_p=1$ if the indicator holds true, and $\text{iso}_p=0$ otherwise. Thus, the cCI at position i was defined as

$$cCI_i=\frac{\sum_{p\in[i-w,i+w]}\text{iso}_p}{2w},$$

in which we set $w=5$ and $u=30$.

The unique structure of proline side chain is generally associated with a relatively low efficiency in its peptide bond formation, which may slow down translation elongation (Woolstenhulme et al., 2013; Artieri and Fraser, 2014; Doerfel et al., 2013; Gardin et al., 2014; Wohlgemuth et al., 2008; Ude et al., 2013; Peil et al., 2013; Gutierrez et al., 2013). Several studies have confirmed the relatively low translation elongation rates at proline codons (Artieri and Fraser, 2014; Gardin et al., 2014). Here we performed an extended study on the relation between proline codons and ribosome stalling using ROSE. In particular, four peptide patterns of proline were investigated, including XPX, XPP, PPX and PPP, in which the three positions correspond to the ribosome E-sites, P-sites and A-sites, respectively, and "P" and "X" represent proline and non-proline amino acids, respectively.

N^6-methyladenosine is probably the most prevalent post-transcriptional modification in mRNAs and plays vital roles in regulating mRNA stability and translation efficiency (Wang et al., 2014, 2015). Recently, Choi *et al.* elucidated that the m^6A-modified codons at the ribosome A-sites can reduce the translation elongation rate in *E.coli* (Choi et al., 2016). We used ROSE to test this hypothesis based on the translatome-wide m^6A mapping obtained from the known single-nucleotide resolution sequencing data, including two human datasets (denoted by Linder15 (Linder et al., 2015) and Ke15 (Ke et al., 2015), respectively) and one yeast dataset (denoted by Schwartz13 (Schwartz et al., 2013)). To ensure that the statistical analysis result did not result from the underlying adenine nucleotides in the codon sites of interest, we also constructed a control dataset which contained **10,000** randomly-selected codon sites covering the adenine nucleotides but without m^6A modification (Figures S5C and S5D).

To evaluate the mRNA structure stability, we first ran RNAfold (Lorenz et al., 2011) to predict the secondary structures of all mRNA sequences in the background dataset, which contained 10,000 randomly-selected ribosome occupancy sites from the genome. Here, the mRNA sequences covering the codon sites of interest were 183 nucleotides long, as we extended each putative ribosome occupancy site by 30 codons both upward and downward as input to ROSE. We then measured the folding level of each sequence by computing its double-stranded ratio (denoted by ds**%**) in the local region of a ribosome A-site, and regarded the top 5,000 mRNA sequences with the highest ds**%** scores as highly folded. Next, we compared the intraRSSes of highly and weakly double-stranded regions for both human and yeast (Figure S5E).

Moreover, we also investigated the relationship between RBP binding and the predicted ribosome stalling scores. Generally, we estimated of the binding affinity of RBPs using the E- and Z-scores provided by the CISBP-RNA database (Ray et al., 2013). In particular, given a region R and an RBP binding motif set M, for any 7-mer m, we defined $\text{aff}-\max(R)=\max_{m\in R}(\max_{m\in M}(m))$ for the max-score estimation, and $\text{aff}-\text{mean}(R)=\text{mean}_{m\in R}(\max_{m\in M}(m))$ for the mean-score estimation, where $\max_{m\in M}(m)$ returns the maximum E- or Z-score of the 7-mer m within the set M. All the four criterias were applied to detect strong RBP binding codon sites for the subsequent statistical analysis (Figures 3A, 3B, S5A, and S5B).

As a specific RBP, the fragile X mental retardation protein (FMRP) (Figure S6A) had been relatively well studied in the literature. Here, we estimated the FMRP binding affinity of the region downstream the ribosome A-site based on the known FMRP binding sites identified by the PAR-CLIP experiment (Ascano et al., 2012). In particular, suppose that we index the codon position at the ribosome A-site as zero. Then the downstream region covering positions from +1 to +3 is still protected by the ribosome (Figure S1). We were particularly interested in estimating the binding affinity of FMRP in the region of next ten codons after the ribosome protected fragment (i.e., codons from +4 to +13), which was denoted by R, and then investigating the correlation between this estimated binding affinity score and RSS. We mainly used the abundance of the mapped reads of FMRP binding sites identified by PAR-CLIP (Ascano et al., 2012) to estimate its binding affinity. Specifically, if there were N reads identified in region $[i,i+x]$, then for any site $s\in[i,i+x]$, its FMRP binding affinity, denoted by $\text{aff}(s)$, was estimated by $\text{aff}(s)=N/x$. After that, the overall binding affinity of the region R right after the ribosome protected fragment was calculated by $\text{aff}(R)=\sum_{s\in R}\text{aff}(s)$. Here we only considered the binding sites whose lengths were within one standard deviation from the mean calculated based on the length distribution of FMRP binding sites, as the extremely long regions may introduce bias to our analysis (Figure S6B).

Finally, we also reexamined the influence of positively charged residues on ribosome stalling. We first tested those codon sites enriched with the positively-charged amino acids upstream (i.e., with the 10,000 highest ratios of the positively-charged amino acids in the upstream 30 codons) in the genome. To probe this problem in more detail, we further separately looked into the specific positively-charged amino acids, including histidine, lysine and arginine (Figure S7).

Statistical Analysis on the Associations between Protein Secondary Structure and RSS

To prepare the protein SSE data, we first derived a set of non-redundant protein chains across human and yeast genomes from the Protein Data Bank (PDB) (Madej et al., 2012), in which BLAST (Altschul et al., 1990) with the sequence-similarity cutoff $P=10^{-7}$ was used to compare two protein sequences. The SSEs of these protein chains (5,054 from human and 766 from yeast) were then determined based on the mapping to the DSSP database (Kabsch and Sander, 1983; Touw et al., 2014), which contains the experimentally-determined secondary structure assignments for the protein sequences in the PDB. To obtain the average position-specific intraRSSes of a certain SSE pattern, all the eligible SSE-aligned sequences with a particular window size were extracted from the genome with five flanking amino acids on both sides, and then the mean intraRSS of each position was calculated. Note that here we mainly considered the intraRSSes of those codons at the ribosome P-sites, where the corresponding amino acids are concatenated to the nascent peptides (Figure S1).

Statistical Analysis on the Association between Transmembrane Domains and RSS

To prepare the TM protein data, we first downloaded all the available TM protein sequences of human and yeast as well as the corresponding TM domain information from the Uniprot database (UniProt Consortium, 2015). To avoid the biases that may be caused by the influence between different TM segments, here we only considered the single-pass integral proteins and the last TM segments of multispan TM proteins, which resulted in 4,235 human and 561 yeast proteins. For yeast proteins, we also excluded 65 TM sequences that are not bound by SRP according to the previous experimental study (Alamo et al., 2011). To characterize the intraRSS landscape along the elongation process, all the protein sequences were aligned with regard to the start of the TM segment whose position was indexed as zero, and then the mean intraRSS of each codon between positions -10 and +80 was calculated.

DATA AND SOFTWARE AVAILABILITY

The source code of ROSE is available at https://github.com/mlcb-thu/rose.