Modeling translation elongation dynamics by deep learning reveals new insights into the landscape of ribosome stalling

Sai Zhang[1,†], Hailin Hu[2,†], Jingtian Zhou[2,†], Xuan He[1], and Jianyang Zeng[1,*]

Abstract

Translation elongation plays a central role in multiple aspects of protein biogenesis, e.g., differential expression, cotranslational folding and secretion. However, our current understanding on the regulatory mechanisms underlying translation elongation dynamics and the functional roles of ribosome stalling in protein synthesis still remains largely limited. Here, we present a deep learning-based framework, called ROSE, to effectively decipher the contextual regulatory code of ribosome stalling and reveal its functional connections to translational control of protein expression from ribosome profiling data. Our validation results on both human and yeast datasets have demonstrated superior performance of ROSE over conventional prediction models. With high prediction accuracy, ROSE provides a precise index to estimate the translation elongation rate at codon resolution. We have demonstrated that ROSE can successfully decode diverse regulatory factors of ribosome stalling, including codon usage bias, tRNA adaptation, codon cooccurrence bias, proline codons, N^6-methyladenosine (m^6A) modification, mRNA secondary structure and protein-nucleotide binding. In addition, our comprehensive genome-wide *in silico* studies based on ROSE have revealed notable functional interplay between elongation dynamics and several cotranslational events in protein biogenesis, including protein targeting by the signal recognition particle (SRP) and protein secondary structure formation. Furthermore, our intergenic analysis suggests that the enriched ribosome stalling events at the 5′ end of the coding sequences (also referred to as the ramp sequences) can be involved in the modulation of translation efficiency. These findings indicate that ROSE can offer a powerful tool to analyze the large-scale ribosome profiling data and provide novel insights into the landscape of ribosome stalling, which will further expand our understanding on translation elongation dynamics.

Keywords: deep learning; ribosome profiling; ribosome stalling; translation elongation dynamics; translational regulation; protein biogenesis

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China.
[2] School of Pharmaceutical Sciences, Tsinghua University, Beijing, China.
[†] These authors contributed equally to this work.
[*] To whom correspondence should be addressed. Email: `zengjy321@tsinghua.edu.cn`.

Introduction

The translation process, including translation initiation, elongation and termination, is a fundamental biological process that delivers genetic information to functional proteins in living cells. The dysregulation of translation has been shown to be associated with a variety of diseases, such as neurological disorder and cancers [1]. Elongation is a crucial step of mRNA translation after initiation, in which the ribosome scans the mRNA sequence and gradually grows the nascent peptide chain by appending new amino acids (Supplementary Figure 1). 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 remains far from clear [2–6]. 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 [3, 4, 6]. In addition, the connection between the elongation rate and human health is increasingly emerging, which further underscores the necessity of a well understanding on the regulatory mechanisms and functions of ribosome stalling and translation elongation dynamics [3, 7].

Gene expression profiling techniques, such as microarrays [8] and RNA-Seq [9], which measure mRNA abundance at a transcriptome-wide level, have been routinely used to study the regulation of gene expression. However, most of the biological questions with respect to the translational control of protein expression cannot be addressed by these methods. 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* [2, 4, 5, 10, 11]. 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, $\sim$30 nucleotides), ribosome profiling provides a genome-wide snapshot of translation elongation dynamics and offers a new angle to estimate translation efficiency. Currently, ribosome profiling has been widely used to study a number of important biological problems related to translation [2, 4, 5], such as the identification of novel or alternatively translated genome regions [12–19], and the discovery of critical regulatory factors in translational control and protein biogenesis [20–36]. Based on the current available large-scale studies involving ribosome profiling experiments, several databases, e.g., GWIPS-viz [37] and RPFdb [38], have been established to collect and curate these profiling data.

Although a large amount of sequencing data have been produced by ribosome profiling experiments, researchers are still challenged by the complexity, heterogeneity and insufficient coverage of these data to derive unbiased and biologically relevant conclusions [2, 4, 5]. Recently, deep learning has become one of the most popular and powerful techniques in the machine learning field [39, 40]. Its superiority over traditional machine learning models has been demonstrated in a wide range of applications, such as speech recognition [41], image classification [40] and natural language processing [42]. Specifically, deep learning has also been successfully applied to analyze large-scale genomic data and uncover notable biological patterns [43–48], such as the predictions of protein-nucleotide binding [44] and effects of noncoding variants [47]. 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 the ribosome profiling data.

Our framework ROSE casts the modeling problem into a classification task and predicts the ribosome stalling events by integrating the encoded sequence features. It is trained based on both human and yeast ribosome profiling data to derive evolutionarily conserved conclusions about

the underlying regulatory mechanisms of ribosome stalling and translation elongation dynamics. Our validation results have demonstrated that ROSE can greatly outperform conventional machine learning models for analyzing genomic sequence data. We have shown that ROSE can successfully decipher diverse regulatory factors of ribosome stalling, including codon usage bias, tRNA adaptation, codon cooccurrence bias, proline codons, N^6-methyladenosine (m^6A) modification, mRNA secondary structure and protein-nucleotide binding, and provide an accurate estimate of the elongation rate at codon resolution. Moreover, our comprehensive *in silico* studies reveal several notable regulatory relations between elongation dynamics and the cotranslational events in protein biogenesis. In particular, our analysis identifies interesting interplay between elongation dynamics and the signal recognition particle (SRP) binding of the transmembrane (TM) segments. Our results indicate that ribosomes tend to stall with high probability at a position $\sim$50 codons downstream from a TM segment, which may promote the molecular recognition by SRP. In addition, our studies show that protein secondary structure elements (SSEs) and their transition patterns tend to be highly correlated with the local elongation rates, which may imply an essential regulatory effect of elongation dynamics on cotranslational folding. Furthermore, our intergenic analysis suggests that the enriched ribosome stalling events in the ramp sequences (i.e., the 5′ end of the coding sequences) may be functionally important in the modulation of translation efficiency. These results demonstrate that ROSE can offer an effective and powerful tool for analyzing large-scale ribosome profiling data and provide new insights for understanding the landscape of ribosome stalling, which is crucial for revealing the contextual regulation and the functional roles of translation elongation dynamics.

Results

Designing, training and validating ROSE

We propose a deep learning-based framework, called RibosOme Stalling Estimator (ROSE), to analyze the large-scale ribosome profiling data and study the contextual regulation of ribosome stalling and its functions in protein biogenesis (Fig. 1a). Unlike previous work that characterized translation elongation dynamics mainly using the stochastic simulation approaches [22, 27, 28], ROSE formalizes the modeling problem into 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, those codon positions in which the ribosome footprint densities exceed 95% of the profiled regions are defined as the positive samples representing the occurrences of ribosome stalling, while the remaining sites are regarded as the negative samples (Online Methods).

We assume that a ribosome stalling event is primarily determined by its surrounding sequence. The codon position of interest, i.e., the ribosome A-site (Supplementary Fig. 1), is first extended both upward and downward by 30 codons, which yields the codon sequence profile of a putative stalling event. We then encode this sequence and feed the encoded features into a deep convolutional neural network (CNN) to learn the complex relations between ribosome stalling and its contextual features (Fig. 1b). We call the prediction score directly outputted by the CNN the *intergenic ribosome stalling score*, which is also termed interRSS (Online Methods). The name "interRSS" comes from the fact that all the scores along the genome are calculated by a universal model and can be compared intergenetically under the same criterion. To further eliminate the interRSS bias among different genes and facilitate the study on the interplay between the intragenic factors (e.g.,

the binding of the signal recognition particle (SRP) on the transmembrane segments) and elongation dynamics, we also normalize the interRSS within each gene and obtain the *intragenic ribosome stalling score*, which is also termed intraRSS (Online Methods). We collectively call interRSS and intraRSS the *ribosome stalling score* (RSS). In principle, RSS can be statistically considered as an estimate of the local translation elongation rate. A higher RSS generally indicates a lower elongation rate and vice versa.

ROSE relies on a number of motif detectors (i.e., convolution operators) to scan the input sequence and integrate those stalling relevant motifs to capture the intrinsic contextual features of ribosome stalling (Fig. 1b, Online Methods). Unlike previous CNN architectures used for analyzing biological data [44, 47], our new CNN framework includes multiple parallel convolution-pooling modules, which can not only significantly reduce the model complexity, but also alleviate the potential overfitting problem (Online Methods). The standard error back-propagation algorithm is used to learn the network parameters of the CNN model [49]. We also deploy several optimization techniques, including regularization [50], dropout [50, 51] and early stopping [50], to further overcome the overfitting problem. In addition, we propose an efficient strategy for large-scale automatic model selection, i.e., calibrating the model hyperparameters without any human intervension (Online Methods). To further boost prediction performance, we also implement an ensemble version of ROSE which consists of 64 CNN classifiers.

Based on our screening principles (Online Methods), we selected two ribosome profiling datasets from GWIPS-viz [37] as our training and test data, including one human dataset [52] (denoted by Battle15) and one yeast dataset [27] (denoted by Pop14). We first performed a normalization procedure similar to that in [21] to remove the technical and experimental biases from the ribosome profiling data (Fig. 1a, Online Methods). 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 we mainly focused on those genes with sequencing coverage above 60%.

We compared the performance of ROSE to that of a classical model, called gkm-SVM, which is developed mainly based on the support vector machine (SVM) and has been successfully applied in various analyses of genomic sequence data [53, 54]. Our tests on both human and yeast datasets showed that ROSE can greatly outperform gkm-SVM with an increase in the area under the receiver operating characteristic curve (AUROC) by up to 18.4% (Figs. 2a and 2b). In addition, the ensemble version of ROSE (also termed eROSE) consistently had superior performance compared to the single version (also termed sROSE).

It has been widely observed that the first 30–50 codons of a coding sequence (CDS) are often enriched with rare codons, and create a "ramp" to reduce the elongation rate during the initial translation elongation process [55, 56]. Such a ramp sequence has been proposed to be universal in prokaryotic and eukaryotic genes, and has been suggested to serve the purpose of reducing the likelihood of downstream ribosomal traffic jams and thus increasing the overall translation efficiency [6, 55, 56]. To focus solely on the elongation process and remove the possible measurement biases introduced by the ramp regions, here we excluded all the reads of these regions, i.e., the first 50 codons at the 5′ end of the coding sequences, from training data. On the other hand, we showed that even without using any training sample from ramps, ROSE can still successfully predict the ribosome stalling events in these regions, with the AUROC scores above 83.0% (Fig. 2c). Although the existence of a ramp region and its effects on the downstream elongation process still remain controversial [29, 57], here we observed a significantly increasing trend of ribosome stalling in the

ramp regions based on our method (Fig. 2d; $P = 4.64 \times 10^{-33}$ for human and $P = 3.80 \times 10^{-3}$ for yeast, one-sided Wilcoxon rank-sum test). As ROSE was trained independently from the ribosome profiling reads of the 5′ end of the coding sequences, our observation reached an unbiased conclusion that the translation elongation rate in the ramp sequences is relatively low compared to that in the downstream regions, which basically reconfirmed the existence of the ramp sequences.

ROSE deciphers the regulatory code of ribosome stalling

Next, we analyzed several factors that may play important roles in regulating ribosome stalling during translation elongation. Previous studies [4, 10, 12, 21, 22, 24, 27–29, 58, 59] have provided various mechanistic insights into translation elongation dynamics based on statistical analysis. Nevertheless, due to unexpected experimental distortions (e.g., the use of cycloheximide in ribosome stabilization [24]) and insufficient modeling power, these studies may derive insignificant or even contradictory conclusions [2, 21, 58]. With stringent screening and normalization procedures as well as superior prediction performance, ROSE enables one to systematically investigate and effectively decode diverse factors that regulate ribosome stalling. Here, we mainly focused on codon usage bias, tRNA adaptation, codon cooccurrence bias, proline codons, N^6-methyladenosine (m^6A) modification, mRNA secondary structure and protein-nucleotide binding, and studied how they affect ribosome stalling and elongation dynamics.

Codon usage bias

Systematic variation has been observed in codon usage across species, among genes in a genome, or even within a gene [60]. Such a codon usage bias has been demonstrated to be crucial in various cellular functions, such as splicing control, stabilization of mRNA structure, maintenance of translation fidelity and regulation of protein folding [3, 7]. Particularly, rare codons are normally highly associated with low translation elongation rates, which has been widely believed to be responsible for proper protein folding [3, 6, 7, 60, 61]. Several metrics have been proposed to measure the codon usage frequency, such as the codon adaptation index (cAI) [62] and the %MinMax score [63]. Here, we used ROSE to reexamine how the codon rareness can affect ribosome stalling based on both cAI and %MinMax profiles. Specifically, we scanned the ribosome occupancy sites along a genome, and then compared the intraRSSes of those sites that were enriched with rare codons to those of the background (Online Methods). 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). Consistent results were observed for both human and yeast (Figs. 3a, 3b and Supplementary Figs. 2a, 2b), i.e., those sites enriched with rare codons displayed significantly higher intraRSSes than the background ($P < 10^{-65}$ for cAI and $P < 10^{-25}$ for %MinMax, one-sided Wilcoxon rank-sum test). These results implied that the codon bias is an important factor to modulate translation elongation dynamics.

tRNA adaptation

Another codon related feature affecting the translation elongation rate is the tRNA concentration. In general, the codons recognized by abundant cognate tRNAs have short decoding time [64]. However, previous analyses of ribosome profiling data often led to inconsistent conclusions on the effects of this feature, which may be attributed to experimental bias, methodological difference or

unknown coregulation factors [5, 24, 27, 57, 61]. In addition to tRNA concentration, the strength of a wobble pairing interaction may also influence the elongation rate [3]. 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) [65]. In fact, it has been found that codon bias is often correlated with tRNA abundance [55, 66–68]. Our analysis also observed a certain correlation between cAI (also %MinMax) and tAI for both human and yeast datasets (Supplementary Table 1). We reexamined the effects of tRNA concentration and the strength of wobble pairing on the elongation rate using ROSE, in which the ribosome occupancy sites were quantified by their tAI scores. In particular, we compared the intraRSSes of those ribosome A- and P-sites enriched with low tAI scores to those of the background (Online Methods), and our comparative analysis on both yeast and human datasets supported the conclusion that lower tRNA concentrations and weaker wobble pairing interactions tend to induce ribosome stalling and associate with lower elongation rates (Figs. 3a, 3b and Supplementary Figs. 2a and 2b; $P < 10^{-60}$ by one-sided Wilcoxon rank-sum test).

Codon cooccurrence bias

In addition to the aforementioned codon usage bias, the codon cooccurrence bias, i.e., the non-uniform distribution of synonymous codon orders, can also affect translation elongation dynamics [6, 69]. Previous studies have suggested that after being recharged, tRNAs may still stay around the ribosome, and tRNA recycling can modulate the translation elongation rate [6, 69]. Under this hypothesis, we would expect the highly isoaccepting-codon-reused regions to be depleted of ribosome stalling events. To examine this problem, 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 (Online Methods). Our studies on both human and yeast showed that those regions enriched with autocorrelated codons (i.e., with high cCI scores) displayed significantly lower intraRSSes than the background (Figs. 3a and 3b; $P < 10^{-9}$ by one-sided Wilcoxon rank-sum test). This result provided a novel evidence to support the argument that the codon cooccurrence bias, probably coordinated with tRNA recycling, can modulate the ribosome stalling tendency and control the local elongation rate.

Proline codons

The unique structure of the proline side chain is generally associated with a relatively low efficiency in its peptide bond formation, which may slow down translation elongation [21, 61, 70–75]. Several studies have confirmed the relatively low translation elongation rates at proline codons [21, 61]. Here we performed an extended study on the influence of proline codons on 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-, P- and A-sites, respectively, and "P" and "X" represent proline and non-proline amino acids, respectively. The comparative analysis on both human and yeast datasets showed that these peptide patterns of proline displayed significantly higher intraRSSes than the background (Figs. 3a and 3b; $P < 10^{-100}$ by one-sided Wilcoxon rank-sum test). In addition, the dipeptide and tripeptide patterns of proline, including XPP, PPX and PPP, showed significantly higher intraRSSes than XPX (Fig. 3c; $P < 10^{-40}$ by one-sided Wilcoxon rank-sum test). This result implied that a codon sequence with more prolines may have a higher chance to induce ribosome stalling and yield a

lower translation elongation rate.

N^6-methyladenosine modification

Notably, we found that the mRNA modification N^6-methyladenosine (m^6A) within codons can also influence ribosome stalling (Fig. 3d). N^6-methyladenosine is probably the most prevalent post-transcriptional modification in mRNAs and plays vital roles in regulating mRNA stability and translation efficiency. Most m^6A sites are distributed in mRNA regulatory regions such as 3'-UTRs and the locations around the stop codons, where the m^6A "readers" such as YTHDF1 and YTHDF2 can bind to dynamically regulate gene expression [76]. Moreover, it has been reported that m^6A in the CDS can modulate the translation initiation to promote translation efficiency [77]. During translation, the additional methyl group of an m^6A site may act as a stumbling block to ribosomes and thus result in a translation pause. 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* [78]. Thus, it is reasonable to hypothesize that the m^6A marks in eukaryotic cells may also be closely correlated to the ribosome stalling tendency. 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 [79] and Ke15 [80], respectively) and one yeast dataset (denoted by Schwartz13 [81]). The analysis results showed that in human, those codons modified by m^6A at the ribosome A-sites had a significantly higher ribosome stalling tendency than the background (Fig. 3a and Supplementary Fig. 2c; $P = 9.61 \times 10^{-13}$ for Linder15 and $P = 2.20 \times 10^{-64}$ for Ke15, one-sided Wilcoxon rank-sum test). To ensure that such an increase of intraRSS 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. In the control test, those m^6A codon sites from the human translatome still exhibited higher intraRSSes than the control dataset (Supplementary Fig. 2d; $P = 1.76 \times 10^{-9}$ for Linder15 and $P = 4.52 \times 10^{-54}$ for Ke15, one-sided Wilcoxon rank-sum test). For yeast, the difference of intraRSS between the m^6A codon sites and the background was insignificant (Fig. 3b and Supplementary Fig. 2d; $P > 0.05$ by two-sided Wilcoxon rank-sum test). This result may be due to the fact that the number of codons modified by m^6A in the yeast dataset was quite limited (only 278 samples after read mapping), and the m^6A modification was only observed in yeast meiosis [81]. Overall, our analysis provided a new insight into the relation between the m^6A modification and translation elongation in human and yeast, which can further expand the previous conclusion drawn from the *E. coli* data [78].

mRNA secondary structure

During the translation elongation process, the ribosome should first unwind the locally folded mRNA secondary structures (e.g., stem-hairpin or stem-internal loops) to move forward [82, 83]. This indicates that in a highly double-stranded region, translation elongation can be slowed down, which thus increases the probability of ribosome stalling [3, 5, 6, 27, 58, 84–87]. To verify this hypothesis, we first ran RNAfold [88] 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 com-

puting 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 those highly-folded mRNAs to those of the remaining sequences. For both human and yeast, we observed an expected increase of intraRSS for those mRNAs with highly-folded structures (Figs. 3a and 3b; $P < 10^{-8}$ by one-sided Wilcoxon rank-sum test), which provided a novel evidence on the regulatory effect of mRNA secondary structure on translation elongation.

RNA-binding proteins

Recently, RNA-binding proteins (RBPs) have received broad interests for their crucial roles in post-transcriptional and translational regulation [89, 90]. By affecting the stability and the translation process of their target mRNAs, RBPs can act as important regulatory factors to control gene expression [91, 92]. As a specific RBP, the fragile X mental retardation protein (FRMP) is essential for neuronal translation regulation, whose transcriptional inactivation has been known to be involved in many diseases, such as fragile X syndrome and autism [93]. It has been found that FMRP can associate with polyribosomes and impede the elongation of a peptide chain [94]. Structural studies have also indicated that FRMP may directly bind to both RNAs and ribosomal proteins to prohibit ribosome movement (Fig. 3e) [95]. Therefore, a region with downstream FMRP binding in the CDS is expected to have a higher chance of ribosome stalling. Here, we estimated the FMRP binding affinity of a region downstream the ribosome A-site based on the known FMRP binding sites identified by the PAR-CLIP experiment [93] (Online Methods). We calculated the intraRSSes of the codon sites enriched with FMRP binding downstream, and found that these sites had significantly higher intraRSSes than the background (Fig. 3f; $P = 5.49 \times 10^{-15}$ by one-sided Wilcoxon rank-sum test), which confirmed the effectiveness of our model to capture the ribosome stalling events regulated by this specific RNA-binding protein.

For general RBPs, we hypothesized that it would be highly probable for ribosomes to stall in an mRNA region enriched with RBP binding motifs downstream, as the bound RBPs may obstruct ribosome movement. To verify this hypothesis, we analyzed the intraRSSes of those codon sites with strong RBP binding propensity downstream (Online Methods), in which the RBP binding motifs and the corresponding binding affinities were downloaded from the CISBP-RNA database [90]. We found that indeed these sites exhibited significantly higher intraRSSes than the background (Figs. 3a and 3b; max E-score estimation, $P = 9.36 \times 10^{-82}$ for human and $P = 1.90 \times 10^{-4}$ for yeast, one-sided Wilcoxon rank-sum test). The increases were significant for human with respect to all four criteria used to estimate the RBP binding affinities, i.e., max/average E- and Z-scores (Fig. 3a and Supplementary Fig. 2a, Online Methods). However, for yeast, the increase was only significant for the max E-score (Fig. 3b and Supplementary Fig. 2b), which may be attributed to the limited number of RBP binding motifs available in yeast (3 motifs in yeast vs. 91 motifs in human). Overall, our analysis showed that RBP binding may act as another factor in influencing ribosome stalling. Of course, we cannot rule out other unknown factors associated with RBP binding motifs that may virtually control ribosome stalling, which will certainly require additional experimental studies and further investigation.

ROSE unveils intragenic RSS landscapes

As the RSS outputted by ROSE encodes diverse factors affecting the ribosome stalling events, it can be regarded as a useful indicator of the local elongation rate. On the other hand, due to the uneven nature of translation elongation dynamics, each gene can have a specific intragenic landscape of ribosome stalling, which may be important for many cellular functions, e.g., modulating the distribution of ribosomes along an mRNA [55, 56], regulating the protein cotranslational folding process [96], and assisting protein translocation across membranes [97]. Here, we analyzed the intraRSS landscape by relating it to several cotranslational events in protein biogenesis, including protein secondary structure formation and protein targeting by the signal recognition particle (SRP).

RSS correlates with protein secondary structure

Although codon bias has been shown to associate with protein secondary structure elements (SSEs) and cotranslational folding [3, 6, 97, 98], it lacks more direct evidence to verify that such an association is imparted from the local elongation dynamics. Here, we sought to probe the relation between the protein SSEs and the local elongation rates based on the intraRSS computed by ROSE. In particular, we first derived a set of non-redundant protein chains across human and yeast genomes from the Protein Data Bank (PDB) [99], in which BLAST [100] with the sequence-similarity cutoff $P = 10^{-7}$ was used to compare two protein sequences. The SSEs of these protein chains were then determined based on the mapping to the DSSP database [101, 102], which contains the experimentally-determined secondary structure assignments for the protein sequences in the PDB. We then investigated the intraRSS landscapes of different SSE patterns, including a single chain of alpha helix (H), beta stand (B) or random coil (C), and transitions between different SSEs.

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 (Supplementary Fig. 1). Overall, we found that with the window size of six, all the tendencies of the intraRSS change for individual SSE patterns were species independent, showing consistent trends for both human and yeast datasets (Figs. 4a and 4b; Spearman correlation coefficient $R > 0.6$). To further eliminate the bias that may be caused by the window size, we also repeated the same analysis procedure with the window size of ten, and found that five out of seven SSE patterns still showed similar trends for both human and yeast (Supplementary Fig. 3; Spearman correlation coefficient $R > 0.5$). The remaining two SSE patterns displayed relatively weak agreement in the intraRSS tendencies (Supplementary Fig. 3; Spearman correlation coefficient $R < 0.1$), which was probably attributed to the dramatic drop of the sample size in both datasets (i.e., from thousands to tens).

We further compared the intraRSSes of the structured (i.e., alpha helix or beta 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 [97], our results showed a significantly higher stalling probability in the coils than in the alpha helix or beta strand regions (Fig. 4c). 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 while a higher chance in the middle of a coil region, when compared to the corresponding flanking regions on both sides (Fig. 4a and Supplementary Fig. 3a). 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 [98]. Our intraRSS landscape showed a similar trend to the previous finding [98] that the transitions from structured to coil regions generally accompanied an increase in the stalling probability on the transition boundaries (Fig. 4b and Supplementary Fig. 3b). In addition, the opposite transitions (i.e., from coil to structured regions) exhibited roughly symmetrical trends in the change of intraRSS (Fig. 4b and Supplementary Fig. 3b). Notably, the ribosome stalling positions revealed by intraRSS here were slightly different from those reported in [98]. In particular, our results displayed more symmetrical stalling positions than the previous results [98], suggesting that ribosomes are prone to stall at the coil residues near the transition boundaries. Admittedly, we cannot exclude the influence of other factors on these findings, such as the database upgrade or the discrepancy of SSE assignment between different databases (e.g., JOY [103] vs. DSSP [101, 102]). Nevertheless, our results indicated the necessity of incorporating other controlling factors in addition to tRNA adaptation to better estimate the translation elongation rate.

RSS associates with SRP recognition

Next, we investigated whether the intraRSS 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 translational pause and the signal recognition particle (SRP) binding of the transmembrane (TM) segments (Fig. 5a). We expected that our model would effectively capture the ribosome stalling events encoded by the heterogeneity of amino acid composition in and around the TM domains.

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 [104]. 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 the 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 [105]. 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.

We first focused on the yeast TM proteins, whose translation has been previously characterized both computationally and experimentally [106]. The intraRSS landscape computed by ROSE captured two major stalling events during the TM protein translation process (Fig. 5b). The first stalling event after the TM start occurred right at the end of the TM segment, where the structured TM segment (majorly alpha helix) is transited to a more flexible intracellular region. This result agreed well with our previous conclusion about the relation between intraRSS and protein secondary structure (Fig. 4b and Supplementary Fig. 3b). The other intraRSS peak, spanning positions from +50 to +70, probably represented the intrinsic stalling to promote the nascent-chain recognition by SRP, which was consistent with the previous report [106]. Indeed, a TM segment generally contains $\sim$20 residues and the length of the ribosome exit tunnel is $\sim$30 residues. Thus position +50 is approximately the place where the translated TM segment emerges from the exit

tunnel and is bound by SRP. To further verify whether this intraRSS peak was truly relevant to SRP binding, we also specifically examined the intraRSS landscape of the TM segments that were not associated with SRP binding (termed SRP-), which were obtained from the previous experimental study [105]. Compared to the above result on the sequences with SRP binding (termed SRP+), the first stalling event remained at position +20, while the intraRSS peak in the putative SRP-associated stalling region (i.e., positions from +50 to +70) was significantly diminished (Fig. 5b; $P = 1.5 \times 10^{-3}$ by one-sided Wilcoxon rank-sum test), which indicated that intraRSS is an excellent indicator of the ribosome stalling event that regulates the synthesis of a TM domain. Notably, the TM segments at positions between 0 and +20 of the SRP- proteins showed a remarkable increase in intraRSS compared to the result of the SRP+ proteins, suggesting that an alternative membrane targeting may occur in the absence of SRP binding. To probe whether our position alignment scheme can induce bias to our analysis, we also performed a similar study in which the TM ends were aligned together, from which we drew the same conclusion (Supplementary Fig. 4a).

For human, as we lacked the SRP recognition data, it was difficult to separate the TM domains bound and unbound by SRP. Here, we only analyzed the intraRSS landscape of a mixed human dataset that did not distinguish the TM segments with and without SRP binding. In this analysis, although we still observed a similar intraRSS peak near the end of the TM segment, the peak signal around position +50 was relatively weak compared to that in yeast (Supplementary Fig. 4b). Such a discrepancy may be caused by either the mixed effect of the TM domains with and without SRP binding in the human dataset or the intrinsic mechanistic difference in cotranslational protein targeting mediated by SRP recognition between human and yeast.

ROSE uncovers intergenic RSS landscapes

As discussed previously (see Section "Designing, training and validating ROSE"), ROSE was able to capture the stalling features of the ramp region (i.e., the first 30–50 codons downstream from the start codon) of a gene. The ribosome stalling in the ramp was expected to enable the fluency of the downstream translation elongation process by reducing the chance of ribosome jamming. In other words, the aggregation of ribosomes in the ramp regions may facilitate the downstream translation [10, 55, 56, 60]. Thus, we expected a gene with a higher interRSS in the ramp region to be more efficiently translated. To verify this hypothesis, we performed a large-scale study to investigate the relation between the interRSS of the ramp region and the translation efficiency of the corresponding gene.

We used the logarithm of the protein expression level divided by the corresponding mRNA expression level to measure the translation efficiency (TE) of each gene. In our analysis, the mRNA and protein expression data of human (lymphoblastoid cell lines, LCLs) and yeast (*S. cerevisiae*) were obtained from [52] and [107, 108], respectively. As short genes may introduce bias to our analysis of the ramp regions, here we only focused on those genes with more than 300 codons, which in total resulted in 12,734 and 3,590 genes for human and yeast, respectively. We then divided these obtained genes into two classes: those with the highest 10% mean interRSSes of the ramp regions were defined as the *ramp genes*, while the remaining ones were regarded as the *non-ramp genes*. Our comparison showed that the ramp genes owned stringently higher translation efficiency than the non-ramp genes (Figs. 6a and 6b; $P = 3.85 \times 10^{-7}$ for human and $P = 2.7 \times 10^{-3}$ for yeast, one-sided Wilcoxon rank-sum test). In particular, for human, the ramp genes displayed dominantly higher translation efficiency than the non-ramp genes approximately starting from

the TE median, while for yeast, the same phenomenon was observed for those genes roughly with top 10% TE (Figs. 6a and 6b).

Furthermore, we compared the enriched functional categories between ramp and non-ramp genes. Specifically, we only focused on those genes with high translation efficiency (i.e., with top 50% TE) as they may be crucial for supporting the fundamental cellular activities, and used all the expressed genes in the dataset as the background for the gene ontology (GO) analysis (Supplementary Table 2). Interestingly, we found that among these genes in yeast, the ramp genes were significantly enriched with many housekeeping GO terms, such as translation regulation and amino acid biogenesis, while the non-ramp genes showed much weaker enrichment or even depletion of these GO categories (Fig. 6c and Supplementary Table 3). For human, such a GO enrichment was not significantly observed after correcting the P values (Supplementary Table 3; $P > 0.05$). This may be due to the intrinsic expression property of LCLs, as we also failed to observe a significant GO enrichment even when simply focusing on those genes of high TE without differentiating their interRSS levels of the ramps (Supplementary Table 3; $P > 0.05$). Nevertheless, for those cell cycle related GO terms, the corrected P values of the ramp genes with high TE were expectedly smaller than those of the non-ramp genes (Supplementary Table 3). Taken together, these results suggested that interRSS can provide a good indicator for studying the regulatory functions of the ramp sequences, which may modulate the translation efficiency of important genes at the elongation level.

Discussion

In addition to those factors investigated in the previous sections, we also studied the correlation between the amino acid charge and ribosome stalling, which still remains controversial and unclear. Several studies claimed that the positively-charged amino acids can slow down the formation of a peptide chain by interacting with the negatively-charged ribosomal exist tunnel [27, 58, 86, 111, 112]. However, others found no such correlation by arguing the quality of experimental data and the methodological limitations in the previous studies [21]. Here, we re-examined this problem based on our method. Indeed, for 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, we did not observe a significant increase of intraRSS compared to the result of the background. To probe this problem in more detail, we further looked into the spesific positively-charged amino acids, including histidine, lysine and arginine, and asked whether any particular amino acid can contribute to ribosome stalling. However, although significant difference of intraRSS was observed between those sites enriched with a particular amino acid upstream and the background, we cannot reach a consistent conclusion over human and yeast datasets (Supplementary Fig. 5).

In this study, we have proposed a deep learning-based method, called ROSE, to model translation elongation dynamics by integrating the underlying sequence features. To our best knowledge, our work is the first attempt to exploit the machine learning/deep learning technique to model the elongation process and predict ribosome stalling based on the large-scale ribosome profiling data. Through comprehensive analyses on both human and yeast datasets, we have shown that ROSE can effectively decipher diverse factors that control the elongation rate and the ribosome stalling tendency. Moreover, ROSE provides a genome-wide estimate on the local elongation rate and enables us to investigate both intragenic and intergenic landscapes of ribosome stalling. Based

on ROSE, we have validated the existence of the ramp sequences in an unbiased manner, and uncovered interesting interplay between elongation dynamics and the cotranslational events in protein biogenesis. For instance, our results have revealed that translation elongation dynamics may modulate the TM recognition by SRP and also associate with the protein secondary structure formation during the cotranslational folding process. In addition, the interRSS landscape indicates that the ramp sequences may be involved in the modulation of translation efficiency. Although translation initiation has been confirmed to be crucial for regulating translation efficiency and differential expression [6, 56], our studies have suggested a potential regulatory function of translation elongation to determine protein expression. Overall, our method can provide an effective and powerful tool to analyze the large-scale ribosome profiling data, and further expand our understanding on the regulatory mechanisms and functions of ribosome stalling and translation elongation dynamics.

Methods

Datasets and data preprocessing

The training and test datasets were downloaded from GWIPS-viz [37], in which abundant ribosome profiling data have been maintained to facilitate the downstream analysis. Here we applied the normalization method introduced by Artieri and Fraser [21] 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 values of the ribosome profiling data. To exclude the unexpected biases that may be introduced from the reads of poorly-sequenced genes, here we mainly considered the reads of those genes whose read coverage ratios were larger than 60%. After normalization, we selected those codon sites whose read abundance was in the top 5% as positive (i.e., foreground) samples, while randomly chose the same number of codon sites from the remaining 95% as negative (i.e., background) samples. These labeled samples were then used as training and test data in a binary classification task. Note that here we excluded all the reads of the ramp regions (i.e., the first 50 codons at the 5′ end of the coding sequences) from the training data.

In this study, we mainly focused on eukaryotic cells, including human and yeast cells. As a previous study [24] has shown that the use of cycloheximide (CHX) for stabilizing ribosomes before sequencing can induce unexpected bias to the measurement (i.e., the ribosomes can continue moving forward in the present of CHX), here we only considered the profiling data with the use of the flash-freezing method [11] instead of CHX. Moreover, after applying the aforementioned normalization technique on the original human and yeast ribosome profiling datasets downloaded from GWIPS-viz, we only kept those datasets with more than $10,000$ positive samples, which were generally sufficient enough to train a reliable machine learning model. In the end, such a normalization and screening procedure yielded two qualified datasets (Supplementary Fig. 6), i.e., the human dataset of lymphoblastoid cell lines (LCLs) [52] (denoted by Battle15) and the yeast dataset of *S. cerevisiae* [27] (denoted by Pop14), which included 109,770 and 20,902 samples,

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 (Figs. 2a and 2b). Note that as the isoform-level analysis of ribosome profiling data has not been completely solved [10], here we mainly chose the largest transcript of each gene as the reference genome to construct the gene-level ribosome profiles.

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 [113] and natural language processing [114]. In particular, CNNs have also been used to model biological sequence data, e.g., the predictions of protein-nucleotide binding [44] and effects of noncoding variants [47]. 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 [115], that is, the mth codon is encoded as a binary vector of length 64, in which the mth position is one while the others are zeros, after indexing all 64 codons. 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 $s = (c_1, ..., c_n)$ 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 c_i represents the ith codon in the sequence, the convolutional layer computes $X = \mathrm{conv}(S)$, 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 $Y = \mathrm{ReLU}(X)$, where

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

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 $Z = \mathrm{pool}(Y)$, i.e.,

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

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 $m = 3$).

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

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

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 [114]. However, the pooling operation in [114] is carried out over the whole convolutional layer without any window restriction, which is quit different from ours. In summary, a complete CNN in our deep learning framework can be formulated as

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

where i represents the kernel index in the parallel architecture, and $\text{encode}(\cdot)$, $\text{conv}(\cdot)$, $\text{ReLU}(\cdot)$, $\text{pool}(\cdot)$, $\text{concat}(\cdot)$ and $\text{sigm}(\cdot)$ 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,

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

where $\text{interRSS}(position)$ represents the interRSS of the codon position of interest and $\text{mean}(gene)$ stands for the mean interRSS of the corresponding gene. When computing $\text{mean}(gene)$, we exclude those codon positions in the ramp region (i.e., the first 50 codons at the 5′ end of the coding sequence). In addition, as limited by the input length of our model (i.e., 61 codons), we mainly use the mean interRSS of codon positions 31–50 to estimate the interRSS level of the ramp region.

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 stochastic gradient descent with the error backpropagation algorithm is performed [49]. To further optimize the training procedure, we also apply several training strategies, including the mini-batch and momentum techniques [50]. In addition, we use the Adam algorithm for stochastic optimization to achieve an adaptive moment estimation [116]. To further overcome the overfitting issue, we also apply several regularization techniques, including L_2-regularization-based weight decay [50], dropout [51] and early stopping [50].

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 H_1. The final values of all hyperparameters of ROSE are provided in Supplementary Table 4. The ROC curves and AUROC scores of the CNNs with calibrated hyperparameters are shown in Supplementary Figs. 7a and 7b 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 [117]. 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 $p_i(s)$ represents the probability calculated by the ith CNN.

Our implementation of ROSE depends on the Caffe library [118], and the Tesla K20c GPUs are used to speed up the training process.

Statistical analysis on diverse factors influencing ribosome stalling

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 influences 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 (here 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 with those of a background dataset, which was generated by randomly selecting 10,000 ribosome occupancy sites from the genome. If the (intra)RSSes of those codon sites enriched with the factor and the background were significantly different, we said this factor can influence ribosome stalling. In addition, we probed the correlations between different factors based on the background dataset, and found little correlation between different factors that we were interested in, except cAI, %MinMax and tAI (Supplementary Table 1). This enhanced our conclusion about the regulatory function of a factor in controlling ribosome stalling, as it basically excluded the possibility that the observed effect was indirect and caused by the influence from other known factors.

Codon cooccurrence index

Cannarozzi *et al.* [69] have proposed the tRNA pairing index (TPI) to characterize the codon order within each gene. However, the TPI measurement is not suitable for our study as here we mainly focused on the codon distribution within a local region rather than over the whole transcript. Here, we proposed a novel index, called the *codon cooccurrence index* (cCI), to estimate the codon cooccurrence level 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 $u = 30$ in our study.

Estimating the RBP binding affinity

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 (Supplementary Figure 1). We were particularly interested in estimating the binding affinity of FMRP or other RBPs in the region of the next ten codons after the ribosome protected fragment (i.e., codons from +4 to +13), which was denoted by R, and then investigating the influence of this estimated binding affinity

score on ribosome stalling. For FMRP, we mainly used the abundance of the mapped reads of the FMRP binding sites identified by PAR-CLIP [93] to estimate its binding affinity. In particular, if there are 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 those binding sites whose lengths are within one standard derivation from the mean calculated based on the length distribution of FMRP binding sites, as the extremely long regions may introduce bias to our analysis.

For general RBPs, the estimation of their binding affinity was similar except that now we used the E- and Z-scores provided by [90] instead of the mapped PAR-CLIP reads. 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 $\text{aff-mean}(R) = \text{mean}_{m \in R}(\max_{m \in M}(m))$ 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.

Acknowledgments

This work was supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003 and 61472205, and China's Youth 1000-Talent Program, the Beijing Advanced Innovation Center for Structural Biology. The authors are grateful to Drs. T. Jiang, Q. Zhang and W. Chen for their helpful discussions about this work. They thank Ms. T. Chu for her help on preparing the figures in this paper.

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 and J.Zeng performed all the computational and statistical analyses. S.Z., H.H., J.Zeng, J.Zhou and X.H. wrote the manuscript. All the authors discussed the test results and commented on the manuscript.

Competing financial interests

The authors declare no competing financial interests.

References

[1] G. C. Scheper, M. S. van der Knaap, and C. G. Proud, "Translation matters: Protein synthesis defects in inherited disease," *Nat Rev Genet*, vol. 8, pp. 711–723, Sept. 2007.

[2] G. A. Brar and J. S. Weissman, "Ribosome profiling reveals the what, when, where and how of protein synthesis," *Nat Rev Mol Cell Biol*, vol. 16, pp. 651–664, Nov. 2015.

[3] J. L. Chaney and P. L. Clark, "Roles for synonymous codon usage in protein biogenesis," *Annual Review of Biophysics*, vol. 44, no. 1, pp. 143–166, 2015.

[4] N. Ingolia, "Ribosome footprint profiling of translation throughout the genome," *Cell*, vol. 165, no. 1, pp. 22–33, 2016.

[5] N. T. Ingolia, "Ribosome profiling: New views of translation, from single codons to genome scale," *Nat Rev Genet*, vol. 15, pp. 205–213, Mar. 2014.

[6] T. Quax, N. Claassens, D. Söll, and J. van der Oost, "Codon bias as a means to fine-tune gene expression," *Molecular Cell*, vol. 59, no. 2, pp. 149–161, 2015.

[7] Z. E. Sauna and C. Kimchi-Sarfaty, "Understanding the contribution of synonymous mutations to human disease," *Nat Rev Genet*, vol. 12, pp. 683–691, Oct. 2011.

[8] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA microarrays," *Nat Genet*, vol. 21, pp. 33–37, 1999.

[9] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: A revolutionary tool for transcriptomics," *Nat Rev Genet*, vol. 10, pp. 57–63, Jan. 2009.

[10] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, "Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling," *Science*, vol. 324, no. 5924, pp. 218–223, 2009.

[11] N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, and J. S. Weissman, "The ribosome profiling strategy for monitoring translation *in vivo* by deep sequencing of ribosome-protected mRNA fragments," *Nat. Protocols*, vol. 7, pp. 1534–1550, Aug. 2012.

[12] N. Ingolia, L. Lareau, and J. Weissman, "Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes," *Cell*, vol. 147, no. 4, pp. 789–802, 2011.

[13] A. A. Bazzini, T. G. Johnstone, R. Christiano, S. D. Mackowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, and A. J. Giraldez, "Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation," *The EMBO Journal*, 2014.

[14] L. Calviello, N. Mukherjee, E. Wyler, H. Zauber, A. Hirsekorn, M. Selbach, M. Landthaler, B. Obermayer, and U. Ohler, "Detecting actively translated open reading frames in ribosome profiling data," *Nat Meth*, vol. 13, pp. 165–170, Feb. 2016.

[15] G.-L. Chew, A. Pauli, J. L. Rinn, A. Regev, A. F. Schier, and E. Valen, "Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs," *Development*, vol. 140, no. 13, pp. 2828–2834, 2013.

[16] J. Crappé, E. Ndah, A. Koch, S. Steyaert, D. Gawron, S. De Keulenaer, E. De Meester, T. De Meyer, W. Van Criekinge, P. Van Damme, and G. Menschaert, "PROTEOFORMER: Deep proteome coverage through ribosome profiling and MS integration," *Nucleic Acids Research*, 2014.

[17] J. G. Dunn, C. K. Foo, N. G. Belletier, E. R. Gavis, and J. S. Weissman, "Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*," *eLife*, vol. 2, p. e01179, Dec. 2013.

[18] A. Fields, E. Rodriguez, M. Jovanovic, N. Stern-Ginossar, B. Haas, P. Mertins, R. Raychowdhury, N. Hacohen, S. Carr, N. Ingolia, A. Regev, and J. Weissman, "A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation," *Molecular Cell*, vol. 60, no. 5, pp. 816–827, 2015.

[19] Z. Ji, R. Song, A. Regev, and K. Struhl, "Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins," *eLife*, vol. 4, p. e08890, Dec. 2015.

[20] L. F. Lareau, D. H. Hite, G. J. Hogan, and P. O. Brown, "Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments," *eLife*, vol. 3, p. e01257, May 2014.

[21] C. G. Artieri and H. B. Fraser, "Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation," *Genome Research*, 2014.

[22] A. A. Gritsenko, M. Hulsman, M. J. T. Reinders, and D. de Ridder, "Unbiased quantitative models of protein translation derived from ribosome profiling data," *PLoS Comput Biol*, vol. 11, pp. 1–26, Aug. 2015.

[23] N. Guydosh and R. Green, "Dom34 rescues ribosomes in 3′ untranslated regions," *Cell*, vol. 156, no. 5, pp. 950–962, 2014.

[24] J. A. Hussmann, S. Patchett, A. Johnson, S. Sawyer, and W. H. Press, "Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast," *PLoS Genet*, vol. 11, pp. 1–25, Dec. 2015.

[25] F. Loayza-Puch, K. Rooijers, L. C. M. Buil, J. Zijlstra, J. F. Oude Vrielink, R. Lopes, A. P. Ugalde, P. van Breugel, I. Hofland, J. Wesseling, O. van Tellingen, A. Bex, and R. Agami, "Tumour-specific proline vulnerability uncovered by differential ribosome codon reading," *Nature*, vol. 530, pp. 490–494, Feb. 2016.

[26] D. Nedialkova and S. Leidel, "Optimization of codon translation rates via tRNA modifications maintains proteome integrity," *Cell*, vol. 161, no. 7, pp. 1606–1618, 2015.

[27] C. Pop, S. Rouskin, N. T. Ingolia, L. Han, E. M. Phizicky, J. S. Weissman, and D. Koller, "Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation," *Molecular Systems Biology*, vol. 10, no. 12, 2014.

[28] S. Reuveni, I. Meilijson, M. Kupiec, E. Ruppin, and T. Tuller, "Genome-scale analysis of translation elongation with a ribosome flow model," *PLoS Comput Biol*, vol. 7, no. 9, pp. 1–18, 2011.

[29] P. Shah, Y. Ding, M. Niemczyk, G. Kudla, and J. Plotkin, "Rate-limiting steps in yeast protein translation," *Cell*, vol. 153, no. 7, pp. 1589–1601, 2013.

[30] B. Zinshteyn and W. V. Gilbert, "Loss of a conserved tRNA anticodon modification perturbs cellular signaling," *PLoS Genet*, vol. 9, pp. 1–12, Aug. 2013.

[31] C. C. Williams, C. H. Jan, and J. S. Weissman, "Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling," *Science*, vol. 346, no. 6210, pp. 748–751, 2014.

[32] A. H. Becker, E. Oh, J. S. Weissman, G. Kramer, and B. Bukau, "Selective ribosome profiling as a tool for studying the interaction of chaperones and targeting factors with nascent polypeptide chains and ribosomes," *Nat. Protocols*, vol. 8, pp. 2212–2239, Nov. 2013.

[33] Y. Han, A. David, B. Liu, J. G. Magadán, J. R. Bennink, J. W. Yewdell, and S.-B. Qian, "Monitoring cotranslational protein folding in mammalian cells at codon resolution," *Proceedings of the National Academy of Sciences*, vol. 109, no. 31, pp. 12467–12472, 2012.

[34] C. H. Jan, C. C. Williams, and J. S. Weissman, "Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling," *Science*, vol. 346, no. 6210, 2014.

[35] E. Oh, A. Becker, A. Sandikci, D. Huber, R. Chaba, F. Gloge, R. Nichols, A. Typas, C. Gross, G. Kramer, J. Weissman, and B. Bukau, "Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor *in vivo*," *Cell*, vol. 147, no. 6, pp. 1295–1308, 2011.

[36] Z. Xiao, Q. Zou, Y. Liu, and X. Yang, "Genome-wide assessment of differential translations with ribosome profiling data," *Nat Commun*, vol. 7, Apr. 2016.

[37] A. M. Michel, G. Fox, A. M. Kiran, C. De Bo, P. B. F. O'Connor, S. M. Heaphy, J. P. A. Mullan, C. A. Donohue, D. G. Higgins, and P. V. Baranov, "GWIPS-viz: Development of a ribo-seq genome browser," *Nucleic Acids Research*, vol. 42, no. D1, pp. D859–D864, 2014.

[38] S.-Q. Xie, P. Nie, Y. Wang, H. Wang, H. Li, Z. Yang, Y. Liu, J. Ren, and Z. Xie, "RPFdb: A database for genome wide information of translated mRNA generated from ribosome profiling," *Nucleic Acids Research*, 2015.

[39] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, July 2006.

[40] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[41] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, pp. 82–97, Nov. 2012.

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

[43] S. Zhang, J. Zhou, H. Hu, H. Gong, L. Chen, C. Cheng, and J. Zeng, "A deep learning framework for modeling structural features of RNA-binding protein targets," *Nucleic Acids Research*, 2015.

[44] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nat Biotech*, vol. 33, pp. 831–838, Aug. 2015.

[45] M. K. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey, "Deep learning of the tissue-regulated splicing code," *Bioinformatics*, vol. 30, no. 12, pp. i121–i129, 2014.

[46] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, and B. J. Frey, "The human splicing code reveals new insights into the genetic determinants of disease," *Science*, vol. 347, no. 6218, 2015.

[47] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," *Nat Meth*, vol. 12, pp. 931–934, Oct. 2015.

[48] M. Liang, Z. Li, T. Chen, and J. Zeng, "Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach," *IEEE/ACM Trans. Comput. Biology Bioinform.*, vol. 12, no. 4, pp. 928–937, 2015.

[49] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.

[50] Y. Bengio, *Neural Networks: Tricks of the Trade: Second Edition*, ch. Practical Recommendations for Gradient-Based Training of Deep Architectures, pp. 437–478. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jan. 2014.

[52] A. Battle, Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford, J. K. Pritchard, and Y. Gilad, "Impact of regulatory variation from RNA to protein," *Science*, vol. 347, no. 6222, pp. 664–667, 2015.

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

[54] M. Ghandi, D. Lee, M. Mohammad-Noori, and M. A. Beer, "Enhanced regulatory sequence prediction using gapped *k*-mer features," *PLoS Comput Biol*, vol. 10, pp. 1–15, July 2014.

[55] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel, "An evolutionarily conserved mechanism for controlling the efficiency of protein translation," *Cell*, vol. 141, no. 2, pp. 344–354, 2010.

[56] T. Tuller and H. Zur, "Multiple roles of the coding sequence 5′ end in gene expression regulation," *Nucleic Acids Research*, 2014.

[57] A. Dana and T. Tuller, "The effect of tRNA levels on decoding times of mRNA codons," *Nucleic Acids Research*, 2014.

[58] C. A. Charneski and L. D. Hurst, "Positively charged residues are the major determinants of ribosomal velocity," *PLoS Biol*, vol. 11, pp. 1–20, Mar. 2013.

[59] C. A. Charneski and L. D. Hurst, "Positive charge loading at protein termini is due to membrane protein topology, not a translational ramp," *Molecular Biology and Evolution*, vol. 31, no. 1, pp. 70–84, 2014.

[60] J. B. Plotkin and G. Kudla, "Synonymous but not the same: The causes and consequences of codon bias," *Nat Rev Genet*, vol. 12, pp. 32–42, Jan. 2011.

[61] J. Gardin, R. Yeasmin, A. Yurovsky, Y. Cai, S. Skiena, and B. Futcher, "Measurement of average decoding rates of the 61 sense codons *in vivo*," *eLife*, vol. 3, p. e03735, Oct. 2014.

[62] P. M. Sharp and W.-H. Li, "The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications," *Nucleic Acids Research*, vol. 15, no. 3, pp. 1281–1295, 1987.

[63] T. F. Clarke, IV and P. L. Clark, "Rare codons cluster," *PLoS ONE*, vol. 3, pp. 1–5, Oct. 2008.

[64] E. P. Rocha, "Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization," *Genome Research*, vol. 14, no. 11, pp. 2279–2286, 2004.

[65] M. d. Reis, R. Savva, and L. Wernisch, "Solving the riddle of codon usage preferences: A test for translational selection," *Nucleic Acids Research*, vol. 32, no. 17, pp. 5036–5044, 2004.

[66] T. Ikemura, "Codon usage and tRNA content in unicellular and multicellular organisms," *Molecular Biology and Evolution*, vol. 2, no. 1, pp. 13–34, 1985.

[67] E. M. Novoa and L. Ribas de Pouplana, "Speeding with control: Codon usage, tRNAs, and ribosomes," *Trends in Genetics*, vol. 28, pp. 574–581, Nov. 2012.

[68] H. Dong, L. Nilsson, and C. G. Kurland, "Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates," *Journal of Molecular Biology*, vol. 260, pp. 649–663, Aug. 1996.

[69] G. Cannarozzi, N. N. Schraudolph, M. Faty, P. von Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet, and Y. Barral, "A role for codon order in translation dynamics," *Cell*, vol. 141, no. 2, pp. 355–367, 2010.

[70] C. J. Woolstenhulme, S. Parajuli, D. W. Healey, D. P. Valverde, E. N. Petersen, A. L. Starosta, N. R. Guydosh, W. E. Johnson, D. N. Wilson, and A. R. Buskirk, "Nascent peptides that block protein synthesis in bacteria," *Proceedings of the National Academy of Sciences*, vol. 110, no. 10, pp. E878–E887, 2013.

[71] L. K. Doerfel, I. Wohlgemuth, C. Kothe, F. Peske, H. Urlaub, and M. V. Rodnina, "EF-P is essential for rapid synthesis of proteins containing consecutive proline residues," *Science*, vol. 339, no. 6115, pp. 85–88, 2013.

[72] I. Wohlgemuth, S. Brenner, M. Beringer, and M. V. Rodnina, "Modulation of the rate of peptidyl transfer on the ribosome by the nature of substrates," *Journal of Biological Chemistry*, vol. 283, no. 47, pp. 32229–32235, 2008.

[73] S. Ude, J. Lassak, A. L. Starosta, T. Kraxenberger, D. N. Wilson, and K. Jung, "Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches," *Science*, vol. 339, no. 6115, pp. 82–85, 2013.

[74] L. Peil, A. L. Starosta, J. Lassak, G. C. Atkinson, K. Virumäe, M. Spitzer, T. Tenson, K. Jung, J. Remme, and D. N. Wilson, "Distinct XPPX sequence motifs induce ribosome stalling, which is rescued by the translation elongation factor EF-P," *Proceedings of the National Academy of Sciences*, vol. 110, no. 38, pp. 15265–15270, 2013.

[75] E. Gutierrez, B.-S. Shin, C. Woolstenhulme, J.-R. Kim, P. Saini, A. Buskirk, and T. Dever, "eIF5A promotes translation of polyproline motifs," *Molecular Cell*, vol. 51, no. 1, pp. 35–45, 2013.

[76] X. Wang, Z. Lu, A. Gomez, G. C. Hon, Y. Yue, D. Han, Y. Fu, M. Parisien, Q. Dai, G. Jia, B. Ren, T. Pan, and C. He, "N6-methyladenosine-dependent regulation of messenger RNA stability," *Nature*, vol. 505, pp. 117–120, Jan. 2014.

[77] X. Wang, B. Zhao, I. Roundtree, Z. Lu, D. Han, H. Ma, X. Weng, K. Chen, H. Shi, and C. He, "N6-methyladenosine modulates messenger RNA translation efficiency," *Cell*, vol. 161, no. 6, pp. 1388–1399, 2015.

[78] J. Choi, K.-W. Ieong, H. Demirci, J. Chen, A. Petrov, A. Prabhakar, S. E. O'Leary, D. Dominissini, G. Rechavi, S. M. Soltis, M. Ehrenberg, and J. D. Puglisi, "N6-methyladenosine in mRNA disrupts tRNA selection and translation-elongation dynamics," *Nat Struct Mol Biol*, vol. 23, pp. 110–115, Feb. 2016.

[79] B. Linder, A. V. Grozhik, A. O. Olarerin-George, C. Meydan, C. E. Mason, and S. R. Jaffrey, "Single-nucleotide-resolution mapping of m^6A and m^6Am throughout the transcriptome," *Nat Meth*, vol. 12, pp. 767–772, Aug. 2015.

[80] S. Ke, E. A. Alemu, C. Mertens, E. C. Gantman, J. J. Fak, A. Mele, B. Haripal, I. Zucker-Scharff, M. J. Moore, C. Y. Park, C. B. Vågbø, A. Kusśnierczyk, A. Klungland, J. E. Darnell, and R. B. Darnell, "A majority of m^6A residues are in the last exons, allowing the potential for 3′ UTR regulation," *Genes & Development*, vol. 29, no. 19, pp. 2037–2053, 2015.

[81] S. Schwartz, S. Agarwala, M. Mumbach, M. Jovanovic, P. Mertins, A. Shishkin, Y. Tabach, T. Mikkelsen, R. Satija, G. Ruvkun, S. Carr, E. Lander, G. Fink, and A. Regev, "High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis," *Cell*, vol. 155, pp. 1409–1421, Dec. 2013.

[82] C. Chen, H. Zhang, S. L. Broitman, M. Reiche, I. Farrell, B. S. Cooperman, and Y. E. Goldman, "Dynamics of translation by single ribosomes through mRNA secondary structures," *Nat Struct Mol Biol*, vol. 20, pp. 582–588, May 2013.

[83] J.-D. Wen, L. Lancaster, C. Hodges, A.-C. Zeri, S. H. Yoshimura, H. F. Noller, C. Bustamante, and I. Tinoco, "Following translation by single ribosomes one codon at a time," *Nature*, vol. 452, pp. 598–603, Apr. 2008.

[84] T. E. Gorochowski, Z. Ignatova, R. A. Bovenberg, and J. A. Roubos, "Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate," *Nucleic Acids Research*, 2015.

[85] J.-R. Yang, X. Chen, and J. Zhang, "Codon-by-codon modulation of translational speed and accuracy via mRNA folding," *PLoS Biol*, vol. 12, p. e1001910, July 2014.

[86] T. Tuller, I. Veksler-Lublinsky, N. Gazit, M. Kupiec, E. Ruppin, and M. Ziv-Ukelson, "Composite effects of gene determinants on the translation speed and density of ribosomes," *Genome Biology*, vol. 12, no. 11, pp. 1–18, 2011.

[87] T. Tuller, Y. Y. Waldman, M. Kupiec, and E. Ruppin, "Translation efficiency is determined by both codon bias and folding energy," *Proceedings of the National Academy of Sciences*, vol. 107, no. 8, pp. 3645–3650, 2010.

[88] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "ViennaRNA package 2.0," *Algorithms for Molecular Biology*, vol. 6, no. 1, pp. 1–14, 2011.

[89] S. Gerstberger, M. Hafner, and T. Tuschl, "A census of human RNA-binding proteins," *Nature Reviews Genetics*, vol. 15, pp. 829–845, Nov. 2014.

[90] D. Ray, H. Kazan, K. Cook, M. Weirauch, H. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. Matzat, R. Dale, S. Smith, C. Yarosh, S. Kelly, B. Nabet, D. Mecenas, W. Li, R. Laishram, M. Qiao, H. Lipshitz, F. Piano, A. Corbett, R. Carstens, B. Frey, R. Anderson, K. Lynch, L. Penalva, E. Lei, A. Fraser, B. Blencowe, Q. Morris, and T. Hughes, "A compendium of RNA-binding motifs for decoding gene regulation," *Nature*, vol. 499, pp. 172–177, July 2013.

[91] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. Beckmann, C. Strein, N. Davey, D. Humphreys, T. Preiss, L. Steinmetz, J. Krijgsveld, and M. Hentze, "Insights into RNA biology from an atlas of mammalian mRNA-binding proteins," *Cell*, vol. 149, no. 6, pp. 1393–1406, 2012.

[92] A. Baltz, M. Munschauer, B. Schwanhäusser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, and M. Landthaler, "The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts," *Molecular Cell*, vol. 46, no. 5, pp. 674–690, 2012.

[93] M. Ascano, N. Mukherjee, P. Bandaru, J. B. Miller, J. D. Nusbaum, D. L. Corcoran, C. Langlois, M. Munschauer, S. Dewell, M. Hafner, Z. Williams, U. Ohler, and T. Tuschl, "FMRP targets distinct mRNA sequence elements to regulate protein expression," *Nature*, vol. 492, pp. 382–386, Dec. 2012.

[94] J. C. Darnell, S. J. Van Driesche, C. Zhang, K. Y. S. Y. Hung, A. Mele, C. E. Fraser, E. F. Stone, C. Chen, J. J. Fak, S. W. W. Chi, D. D. Licatalosi, J. D. Richter, and R. B. Darnell, "FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism," *Cell*, vol. 146, pp. 247–261, July 2011.

[95] E. Chen, M. R. Sharma, X. Shi, R. K. Agrawal, and S. Joseph, "Fragile X mental retardation protein regulates translation by binding directly to the ribosome," *Molecular cell*, vol. 54, pp. 407–417, May 2014.

[96] G. Zhang, M. Hubalewska, and Z. Ignatova, "Transient ribosomal attenuation coordinates protein synthesis and co-translational folding," *Nat Struct Mol Biol*, vol. 16, pp. 274–280, Mar. 2009.

[97] S. Pechmann and J. Frydman, "Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding," *Nat Struct Mol Biol*, vol. 20, pp. 237–243, Feb. 2013.

[98] R. Saunders and C. M. Deane, "Synonymous codon usage influences the local protein structure observed," *Nucleic Acids Research*, vol. 38, no. 19, pp. 6719–6728, 2010.

[99] T. Madej, K. J. Addess, J. H. Fong, L. Y. Geer, R. C. Geer, C. J. Lanczycki, C. Liu, S. Lu, A. Marchler-Bauer, A. R. Panchenko, J. Chen, P. A. Thiessen, Y. Wang, D. Zhang, and S. H. Bryant, "MMDB: 3D structures and macromolecular interactions," *Nucleic Acids Research*, vol. 40, no. D1, pp. D461–D464, 2012.

[100] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[101] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[102] W. G. Touw, C. Baakman, J. Black, T. A. te Beek, E. Krieger, R. P. Joosten, and G. Vriend, "A series of PDB-related databanks for everyday needs," *Nucleic Acids Research*, 2014.

[103] K. Mizuguchi, C. M. Deane, T. L. Blundell, M. S. Johnson, and J. P. Overington, "JOY: Protein sequence-structure representation and analysis," *Bioinformatics*, vol. 14, no. 7, pp. 617–623, 1998.

[104] T. U. Consortium, "Uniprot: A hub for protein information," *Nucleic Acids Research*, vol. 43, no. D1, pp. D204–D212, 2015.

[105] M. d. Alamo, D. J. Hogan, S. Pechmann, V. Albanese, P. O. Brown, and J. Frydman, "Defining the specificity of cotranslationally acting chaperones by systematic analysis of mRNAs associated with ribosome-nascent chain complexes," *PLoS Biol*, vol. 9, pp. e1001100–, July 2011.

[106] S. Pechmann, J. W. Chartron, and J. Frydman, "Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP *in vivo*," *Nat Struct Mol Biol*, vol. 21, pp. 1100–1105, Dec. 2014.

[107] L. M. F. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frohlich, T. C. Walther, and M. Mann, "Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast," *Nature*, vol. 455, pp. 1251–1254, Oct. 2008.

[108] I. Nookaew, M. Papini, N. Pornputtpong, G. Scalcinati, L. Fagerberg, M. Uhlén, and J. Nielsen, "A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: A case study in *Saccharomyces cerevisiae*," *Nucleic Acids Research*, 2012.

[109] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[110] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat. Protocols*, vol. 4, pp. 44–57, Dec. 2008.

[111] J. Lu and C. Deutsch, "Electrostatics in the ribosomal tunnel modulate chain elongation rates," *Journal of Molecular Biology*, vol. 384, pp. 73–86, Dec. 2008.

[112] J. Lu, W. R. Kobertz, and C. Deutsch, "Mapping the electrostatic potential within the ribosomal exit tunnel," *Journal of Molecular Biology*, vol. 371, pp. 1378–1391, Aug. 2007.

[113] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, 1998.

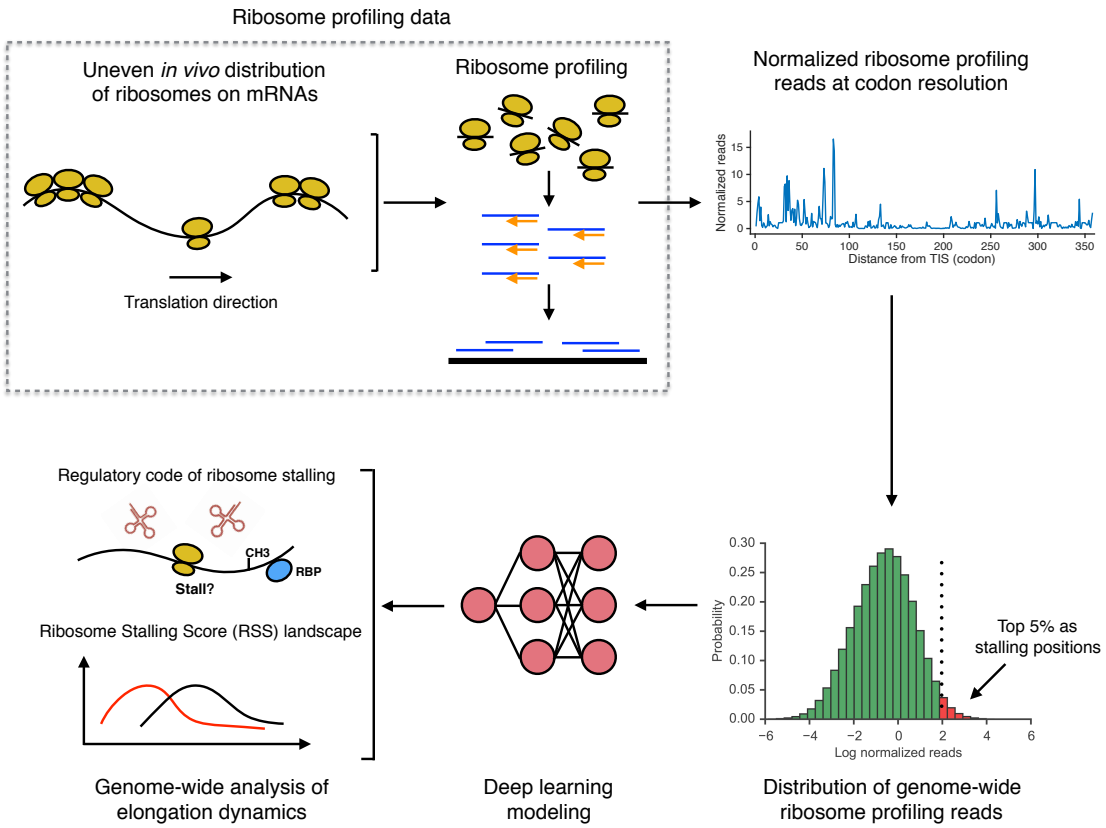
[114] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014.

[115] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

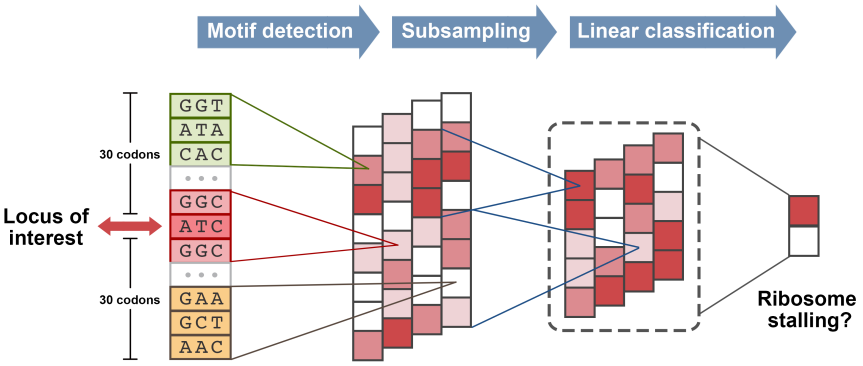
[116] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[117] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*, 2010.

[118] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *CoRR*, vol. abs/1408.5093, 2014.



(a)



(b)

Figure 1: The ROSE pipeline and the CNN model. (a) Schematic overview of the ROSE pipeline. The codon sites with the top 5% ribosome footprint densities 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.

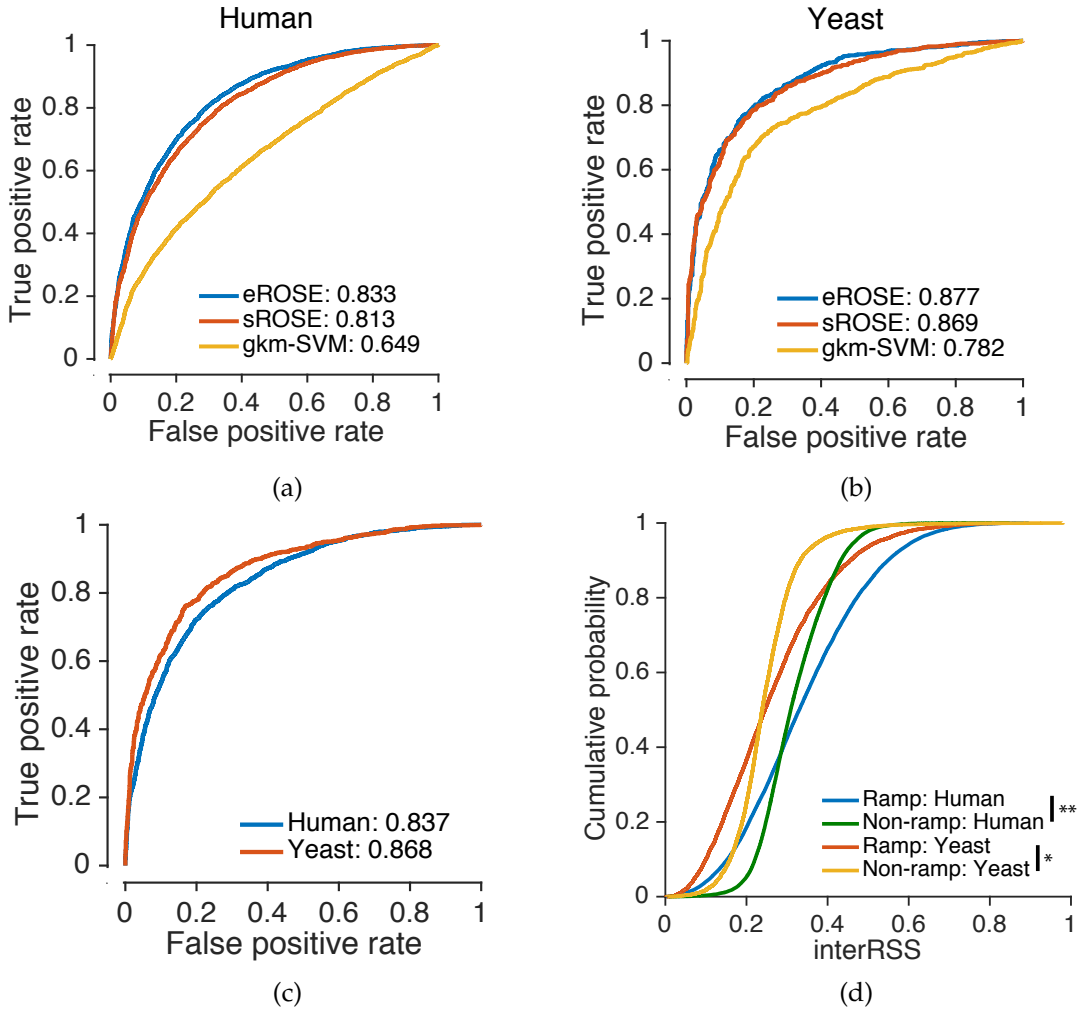
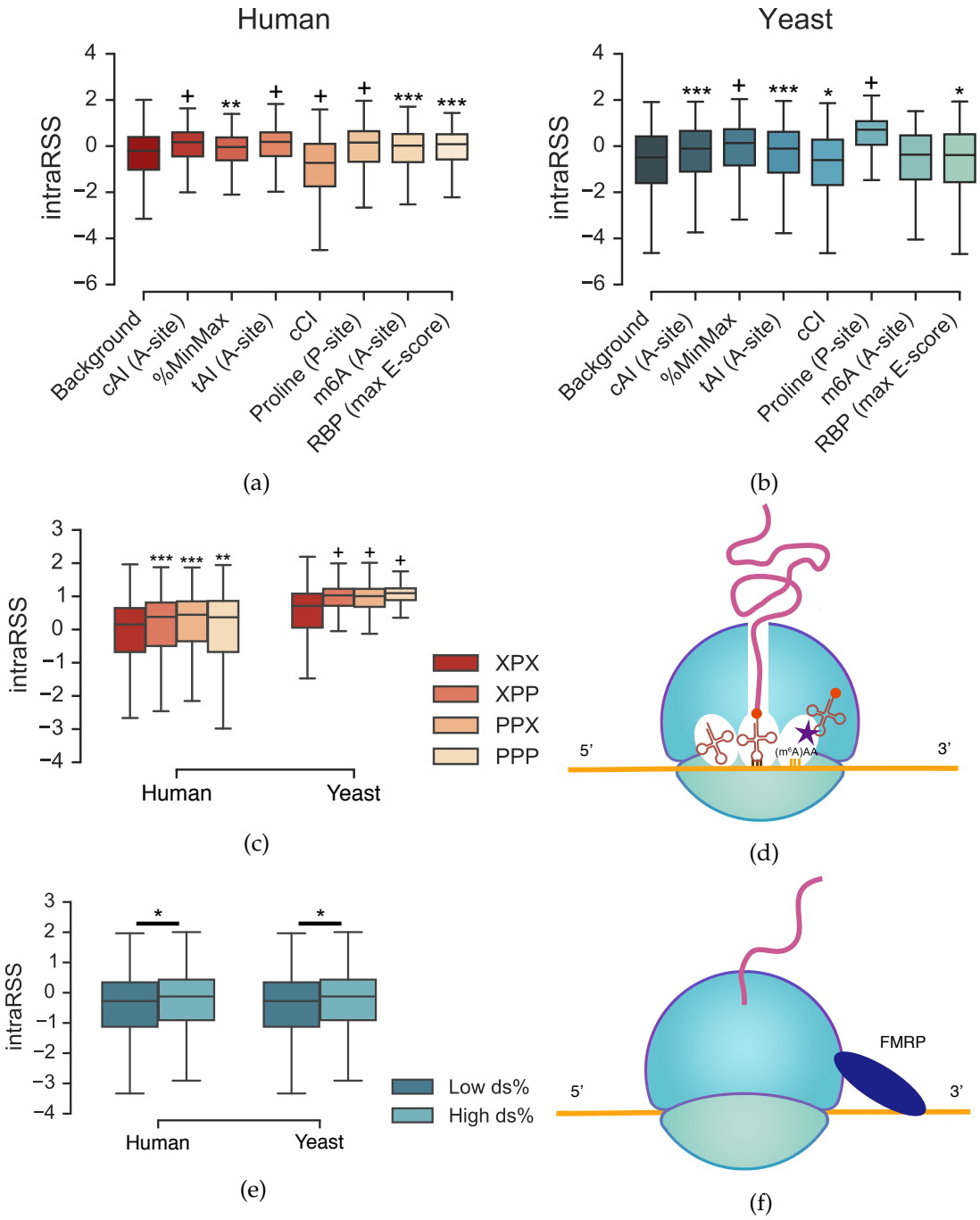
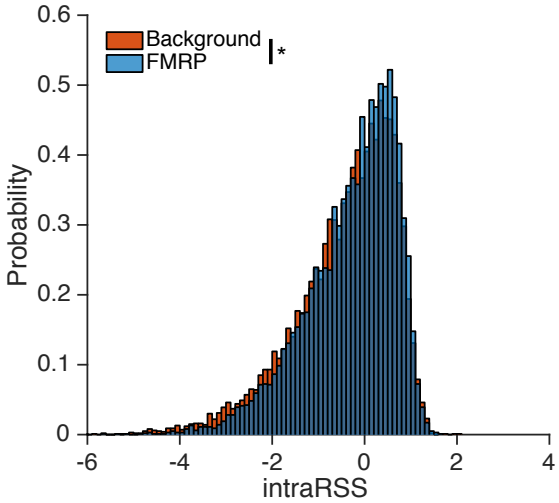


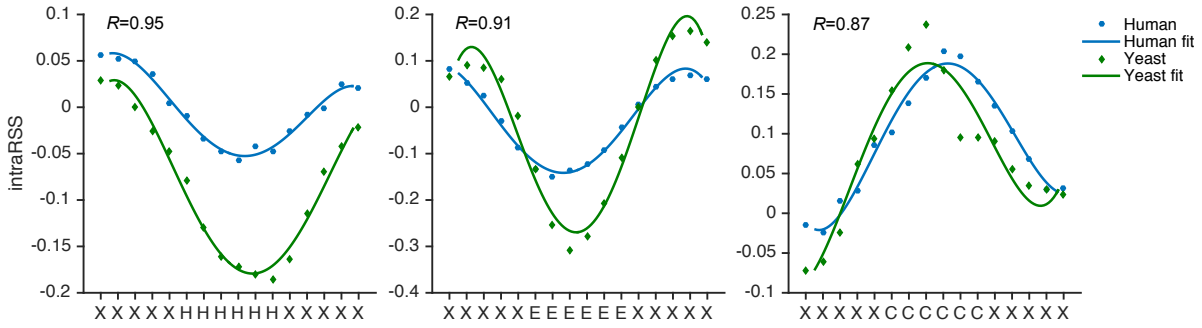
Figure 2: Performance evaluation of ROSE. (a) and (b) The receiver operating characteristic (ROC) curves and the area under the ROC curve (AUROC) scores on the human (Battle15) and yeast (Pop14) test datasets, respectively. "sROSE" and "eROSE" stand for the ROSE frameworks with one (single) and 64 (ensemble) CNNs, respectively. (c) The ROC curves and the corresponding AUROC scores on the ramp regions. (d) The empirical cumulative distribution function (CDF) curves of interRSS on both ramp and non-ramp regions. *: $5 \times 10^{-25} < P < 1 \times 10^{-2}$; **: $5 \times 10^{-50} < P \leq 5 \times 10^{-25}$; one-sided Wilcoxon rank-sum test.


Human
Yeast
intraRSS
Background
cAI (A-site)
%MinMax
tAI (A-site)
cCI
Proline (P-site)
m6A (A-site)
RBP (max E-score)
(a)
(b)
XPX
XPP
PPX
PPP
Human
Yeast
(c)
5'
3'
(m6A)AA
(d)
Low ds%
High ds%
(e)
FMRP
(f)

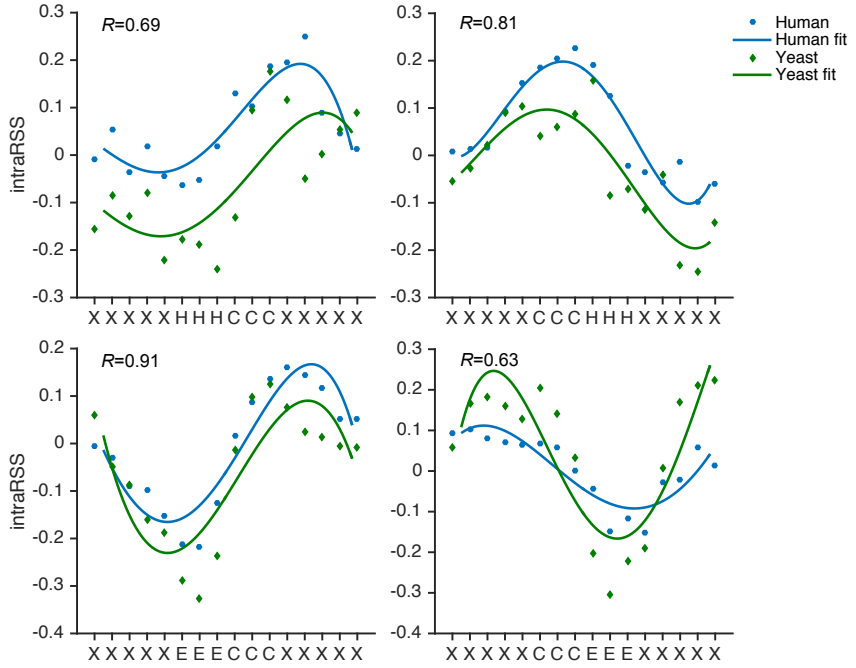


(g)

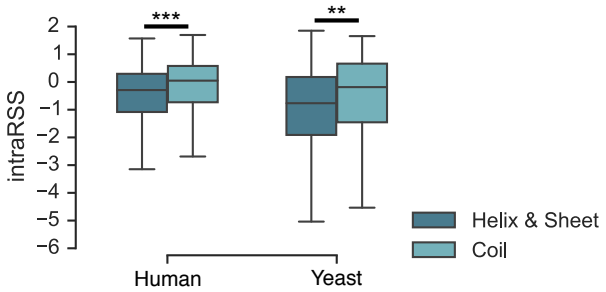
Figure 3: A comprehensive investigation on the regulatory effects of different factors on ribosome stalling using ROSE. (a) and (b) The comparisons of intraRSS between the codon sites enriched with individual factors and the background for human and yeast, respectively. Details of the tests can be found in the main text. (c) The 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 the tRNA accommodation during translation elongation. (e) The comparisons of intraRSS between highly and weakly double-stranded regions for both human and yeast. (f) A schematic illustration of the FMRP binding to impede ribosome movement. (g) The comparison of intraRSS between the FMRP target regions and the background. *: $5 \times 10^{-25} < P < 1 \times 10^{-2}$; **: $5 \times 10^{-50} < P \leq 5 \times 10^{-25}$; ***: $5 \times 10^{-100} < P \leq 5 \times 10^{-50}$; +: $P \leq 5 \times 10^{-100}$; one-sided Wilcoxon rank-sum test.


R=0.95
R=0.91
R=0.87
intraRSS
Human
Human fit
Yeast
Yeast fit
X X X X X H H H H H H X X X X X
X X X X X E E E E E E X X X X X
X X X X X C C C C C C X X X X X

(a)

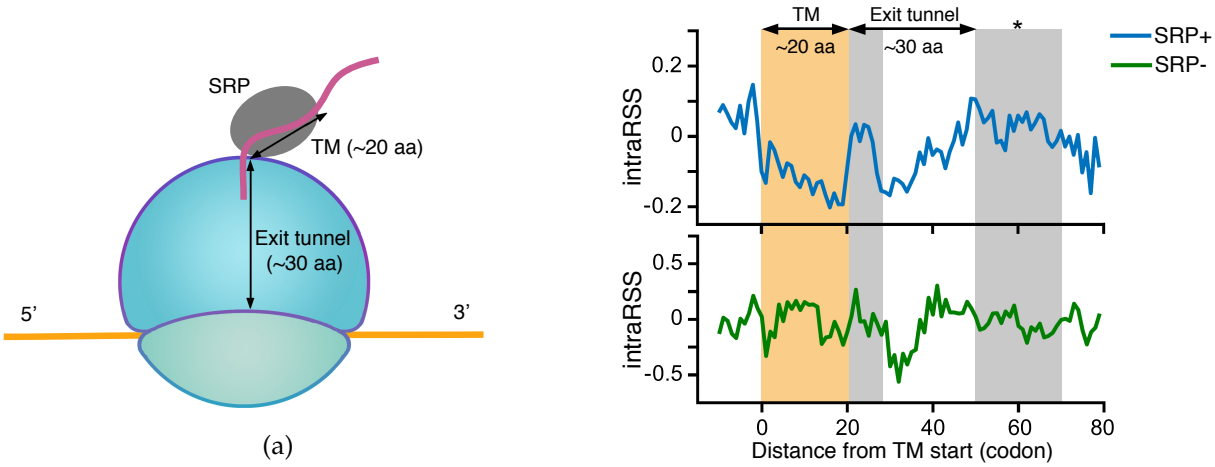

R=0.69
R=0.81
R=0.91
R=0.63
intraRSS
Human
Human fit
Yeast
Yeast fit
X X X X X H H H C C C X X X X X
X X X X X C C C H H H X X X X X
X X X X X E E E C C C X X X X X
X X X X X C C C E E E X X X X X

(b)



(c)

Figure 4: The intragenic RSS landscape unveils that the translation elongation rate associates with the local protein secondary structure. (a) The intraRSS landscapes of the alpha helix, beta strand and random coil regions. (b) The intraRSS landscapes of the SSE transition regions. "H", "E" and "C" stand for alpha helix, beta 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., alpha helix and beta 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.



(a) (b)

Figure 5: The intragenic RSS landscape reveals that translation elongation dynamics promotes the SRP binding of the TM segments. (a) A schematic illustration of the SRP binding of a TM segment during translation elongation. (b) The comparison of intraRSS tendency between the TM segments with and without SRP binding 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 grey rectangles represent two intraRSS peaks downstream the TM segment. *: $5 \times 10^{-25} < P < 1 \times 10^{-2}$; one-sided Wilcoxon rank-sum test.

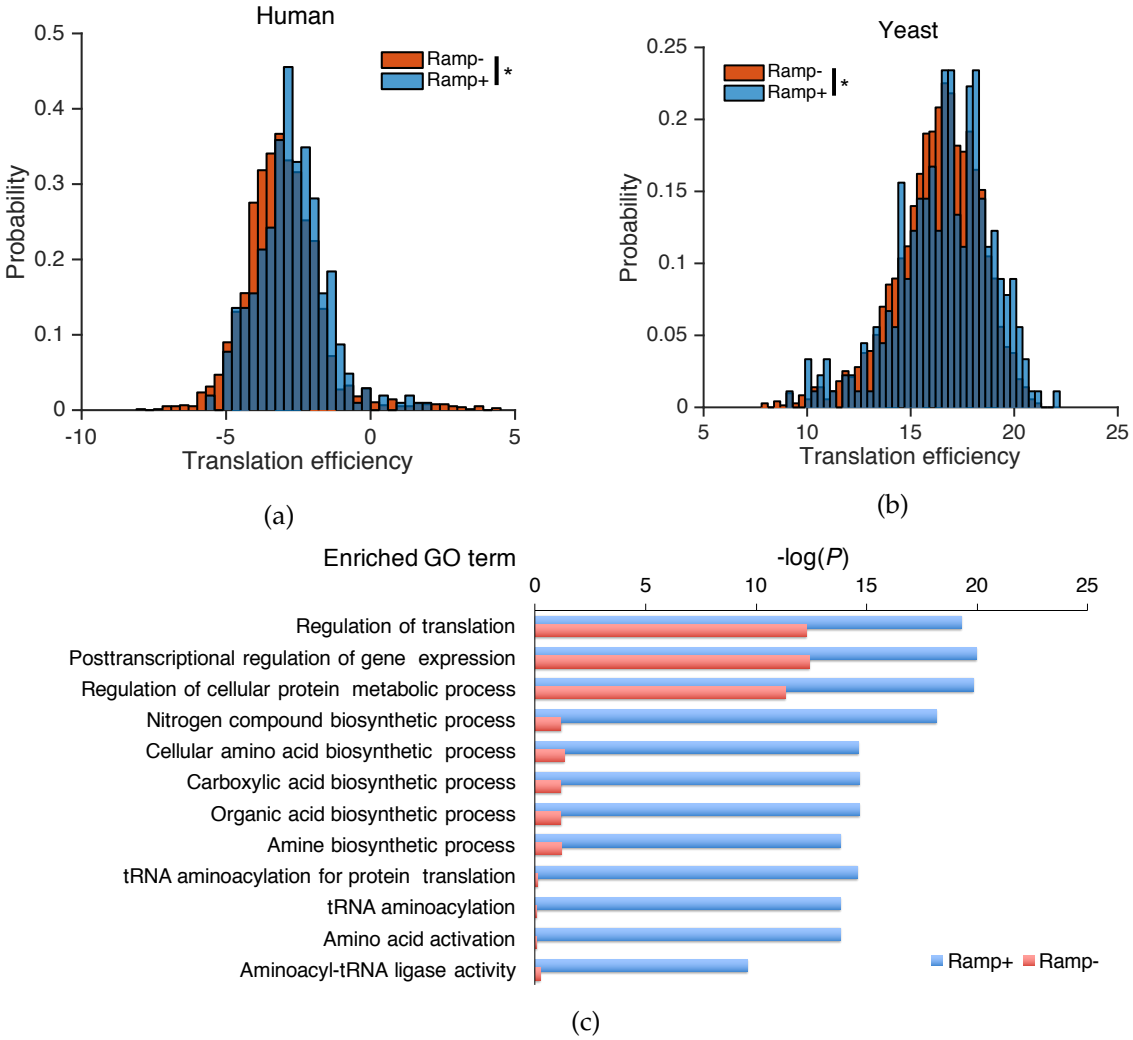


Figure 6: The intergenic RSS landscape reveals the difference of translation efficiency and enriched functional categories between ramp and non-ramp genes, which were defined according to the interRSSes of the ramp regions (see the main text for more details). (a) The interRSS distributions of both ramp and non-ramp genes in human. The mRNA levels measured by RNA-Seq were quantified by reads per kilobase transcript per million mapped reads (RPKM), while the protein levels measured by mass spectrometry were quantified by the SILAC ratio [52]. (b) The interRSS distributions of both ramp and non-ramp genes in yeast. The mRNA levels measured by RNA-Seq were quantified by fragments per kilobase of transcript per million mapped reads (FPKM) [108], while the protein levels measured by mass spectrometry were quantified by the summed ion intensity [107]. The translation efficiency was estimated by the logarithm of the protein expression level divided by the corresponding mRNA expression level. "Ramp+" and "ramp-" stand for the ramp and non-ramp genes, respectively. (c) The comparison of the enriched GO terms between ramp and non-ramp genes in yeast. GO enrichment analysis was performed using DAVID [109, 110]. Here, the P values were computed after multiple testing correction according to the Benjamini-Hochberg procedure. For a full list of the functional annotation clustering, see Supplementary Table 3. *: $5 \times 10^{-25} < P < 5 \times 10^{-2}$; one-sided Wilcoxon rank-sum test.