Deciphering RNA regulation with a foundation language model

Hanwen Zhou [1, 2, †], Yue Hu [3, † *], Yulong Zheng [1, 2], Jiefu Li [1, 2], Jielong Peng [1, 2], Guoqing Zhang [1, 2 *], Zefeng Wang [2, 4 *]

[1] National Genomics Data Center & Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China.

[2] University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing, China.

[3] School of Medicine, Department of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[4] School of Life Science, Southern University of Science and Technology, Shenzhen, Guangdong, China

Keywords: RNA regulation, language model, splicing, translation, degradation

† These authors contributed equally to this work.

* Correspondence to: Yue Hu, Tel: +1 919 600 0000, E-mail: yue_hu@med.unc.edu; Guoqing Zhang, Tel: (+86) 021-54920465, E-mail: gqzhang@picb.ac.cn; Zefeng Wang, Tel: (+86) 021-54920416, E-mail: wangzefeng@picb.ac.cn

Abstract

RNAs are widely regulated in cellular processes by their *cis*-regulatory elements and *trans*-acting RNA binding proteins, the information of which is encoded in RNA sequences spanning evolutionary diversity. Deciphering the rules underlying RNA regulation can provide new insights into molecular mechanisms and RNA therapies. The large language models have demonstrated their efficacy in the analysis of human languages and protein sequences. Considering the similarities in semantic and syntactic features between RNA sequences and human language, we developed LAMAR, a foundation language model for RNA regulation, aimed at capturing the intrinsic characteristics of RNA sequences. The model was first pretrained on approximately 15 million sequences of genes and transcripts from 225 mammals and 1569 viruses. By leveraging LAMAR as a foundational platform, we fine-tuned the pretrained model with labeled datasets across various tasks. The resulting fine-tuned model outperformed the best benchmark by up to 9% for predicting mRNA translation efficiency, and improved by 7% for predicting the RNA half-life. We further applied LAMAR to predict the splice sites of pre-mRNAs and internal ribosome entry sites driving cap-independent translation, the performances of which were comparable with the state-of-the-art methods. The results indicated that our foundation language model is applicable for comprehensive regulatory analysis of RNA, providing new clues into molecular cell biology and disease.

The code is available at https://github.com/rnasys/LAMAR.

Keywords: RNA regulation, language model, splice site, translation, degradation

Introduction

As an important component of the central dogma, RNAs play various roles in multiple biological processes, ranging from the templates for protein synthesis [1] to regulators of gene expression [2]. The cellular processes of RNA are regulated by its *cis*-regulatory elements and *trans*-acting recruited RNA binding proteins [3], the information of which are encoded in RNA sequences spanning evolutionary diversity. Deciphering the regulatory rules of RNA will provide new insights into molecular mechanisms and further design novel therapeutics.

With the advancement of biochemical technologies and bioinformatics, many researchers have developed computational tools to analyze the accumulated regulatory maps of RNA [4-8]. While well-designed algorithms were previously developed for specific question, a unified framework to learn the diverse regulation of RNA is lacking. The Large Language Models have demonstrated that unsupervised pretraining of a vast corpuses of text can significantly advance our comprehension [9] and generative capabilities of language [10]. The pretrained model serves as a foundation platform that can further be fine-tuned to solve multiple downstream tasks. In light of the similarity of semantic and syntactic features between RNA sequences and human language, some studies explored the potential of language models to capture the essence of RNA. RNA-MSA [11] and RNAErnie [12] were developed to predict the secondary structures of non-coding RNAs. RNA-FM was pretrained on non-coding RNAs and fine-tuned to solve multiple tasks including prediction of RNA tertiary structures and binding sites of RNA binding proteins [13]. However, the present foundation language models still lacked insights on RNA regulation. In addition, the current models were all pretrained on non-coding RNA sequences, and thus the extracted sequence features are mainly related to the functions and structures of non-coding RNAs. Therefore, the extracted sequence features may be biased to highly structured non-coding RNAs and insufficient in dissecting downstream tasks related to processing of RNA precursors and functions of mature mRNAs.

To address these challenges, we developed a foundation language model for RNA regulation, named LAMAR, by unsupervised pretraining on approximately 15 million RNA sequences with multiple functions across mammalians. LAMAR can effectively distinguish between regulatory elements or transcripts with distinct functions within the representation space, indicating that it has successfully captured the functional attributes solely from sequences. By leveraging LAMAR as a foundation platform, we were able to fine-tune this pretrained model with labeled datasets across a range of tasks including supervised modeling of splice sites, translation efficiency, internal ribosome entry sites (IRESs) and degradation. Our results showed that the performances of these fine-tuned models were better than or comparable with the state-of-the-art methods that were specially designed for each task, which demonstrated the power of large language model in dissecting the diverse functions determined by the RNA sequences, providing new clues for underlying cellular mechanisms.

Results

The pretrained RNA foundation model

We developed the RNA foundation model with the similar architecture as the transformer encoders. We first pretrained the model with large-scale unlabeled RNA sequences, and then fine-tuned the model with task-specific datasets to solve the downstream tasks. To efficiently learn the sequence features and evolutionary information of RNA, around 15 million nucleotide sequences of multiple functions across species were downloaded and processed as training corpus from RefSeq and RNACentral databases (Figure 1A). The mRNAs, non-coding RNAs and genes of mammals and viruses were included, covering around 270 billion nucleotide tokens (Figure 1B, C). In the pretraining stage, we used the mask learning strategy to train the model. A portion of the nucleotides were masked and then predicted by contextual sequences, enabling the model to learn the interactions between nucleotides (Figure 1E). We pretrained two models with the context length of up to 2048 and 4096 tokens, considering the sequence length of genes and transcripts and training resources (Figure 1D). In the fine-tuning stage, the task-specific predictor was stacked on the encoder for downstream tasks. We will validate our RNA foundation model by four downstream tasks: prediction of splicing sites, translation efficiency and IRES activity and degradation rates of RNA (Figure 1F).

The pretrained model learned the functional features of RNAs

To investigate whether the model learned the biochemical properties of nucleotides within pretraining, we compared the hidden embeddings of four classes of nucleotides. Each nucleotide was represented as one vector by averaging the hidden embeddings of nucleotides in 500 randomly selected human mRNAs and long non-coding RNAs. We used the Principal Component Analysis to project the representative vectors into two-dimensional space. As shown in Figure 2A, the adenine A and guanine G which are

115 dicyclic purines, the cytosine C and thymine T which are monocyclic pyrimidines are
116 the closest pair for the first dimension, indicating that the model learned the
117 biochemical properties of nucleotides from large-scale unlabeled sequences. The
118 distances of hidden embeddings were presented in Supplementary Figure 1.

119 To verify whether the pretrained model extracted the functional features of RNAs,
120 we input translation regulatory elements of different functions into the pretrained model
121 and observed the clusters of outputs. By calculating the mean of the hidden embeddings
122 at each position in sequence, each element was represented as a single vector and further
123 projected into two-dimensional space using t-Distributed Stochastic Neighbor
124 Embedding (tSNE). Compared with the random scatter of the outputs from the
125 untrained model (NMI=0.041), the RNA sequences with same functions were more
126 likely to be clustered together (NMI=0.531) (Figure 2B). The same results were also
127 found for RNAs with different function (NMI=0.060 for untrained model vs
128 NMI=0.327 for pretrained model) (Figure 2C). These results indicate that the model
129 successfully extracted the functional features of RNA sequences in unlabeled self-
130 supervised pretraining.

131 **The fine-tuned model predicted splice sites accurately**

132 In the process of eukaryotic gene expression, the spliceosome precisely identifies
133 splice sites and catalyzes the transformation of pre-mRNAs into mature transcripts [14].
134 Accurately predicting splice sites is crucial for understanding the link between genetic
135 mutations and their phenotypes [15, 16]. Currently, various computational tools such as
136 SpliceAI [5] have been developed to forecast splice sites within pre-mRNA sequences.

137 A large number of studies have shown that the human splice sites have regular
138 sequence motifs in their flanking sequences [17], suggesting that the splicing sites have
139 similar contextual information and the hidden embeddings of which from pretrained
140 model were expected be similar. To investigate the effect of pretraining, we used the
141 tSNE algorithm to project the hidden embeddings of the 5' and 3' splicing sites into the

two-dimensional space. As shown in Figure 3B, the splice sites and non-splicing sites were grouped respectively, indicating that the pretrained model successfully learned the contexts of splicing sites, which could be applied to subsequent fine-tuning.

We further fine-tuned the LAMAR to predict the probabilities of 5' and 3' splice sites for each nucleotide in human pre-mRNA sequences. We used the same datasets consistent with SpliceAI and set the sequence length as 1024 for fine-tuning (Figure 3A). A variety of strategies were tested: a) fine-tuning without pretrained weights, b) fine-tuning with frozen pretrained weights and c) fine-tuning with all parameters. As illustrated in Figure 3C, fine-tuning with all parameters achieved the best Precision-Recall Area Under Curve (PRAUC) and outperformed strategies a and b by up to 0.013 and 0.050, respectively. We defined the best model as LAMAR-SS. It is worth noting that fine-tuning with only a simple prediction head could achieve good result (PRAUC=0.918), indicating the effectiveness of pretraining. We compared LAMAR-SS with benchmark methods including SpliceAI and RNAFM. To control the sequence length for fine-tuning, the sub-models spliceAI-400 and spliceAI-2k were selected which predicted the splicing sites based on flanking 200nt and 1000nt sequences. Figure 3D showed that our method achieves better results than other methods. In particular, LAMAR-SS outperforms spliceAI-400 by up to 0.036, spliceAI-2k by up to 0.02 and RNAFM by up to 0.015 in terms of PRAUC (outperforms spliceAI-400 by up to 0.049, spliceAI-2k by up to 0.035 and RNAFM by up to 0.057 in terms of Top-k accuracy).

The fine-tuned model predicted translation efficiency accurately

The 5' untranslated region (5' UTR) plays a major role in the control of translation efficiency of mRNA [18]. Understanding the relationship between 5' UTR sequences and levels of protein expression can help enhance the protein production to optimize the potency of mRNA therapy [19, 20].

In this section, we fine-tuned LAMAR to predict the translation efficiency of human mRNAs based on 5' UTR sequences (Figure 4A). The 10903 transcripts from

169 human embryonic kidney (HEK) 293T cell line was selected for training [19]. The
170 translation efficiency of these transcripts was measured by Ribo-seq. We fine-tuned
171 LAMAR with all parameters and achieved the best result in all strategies, which we
172 named it as LAMAR-TE (Supplementary Figure 2). Since UTR-LM outperformed than
173 classical methods in prediction of translation efficiency of 5' UTRs [21], we selected
174 UTR-LM and RNAFM as benchmarks. The results of 10-fold cross validation showed
175 that LAMAR-TE outperformed RNAFM and UTR-LM by up to 9% and 14% for the
176 mean squared error respectively (MSE) (Figure 4B), and by up to 7% and 18% for the
177 Spearman correlation coefficient (Figure 4C). UTR-LM was pretrained on sequences
178 and secondary structures of 5' UTRs using approximately 1 million parameters. The
179 superior performance of LAMAR-TE suggested that our model probably learned the
180 structural information of 5' UTRs exclusively from sequences with much more
181 parameters.

182 **The fine-tuned model predicted IRES activity accurately**

183 The internal ribosome entry sites (IRESs) drive the cap-independent translation of
184 some mRNAs under stressful conditions [22]. The translation of circular RNAs
185 (circRNAs) lacking 5' and 3' ends can also be initiated through IRESs [23, 24].
186 Although a small number of IRESs were identified from high-throughput screening
187 experiments [25-27], most of them have not been annotated.

188 In this section, we fine-tuned LAMAR to classify sequences with IRES activity
189 (Figure 5A). The 1,901 experimentally validated IRESs from IRESite [28], IRESbase
190 [29], and RFAM [30] databases were collected as positive sequences (Figure 5B). The
191 structural RNAs were sampled from RFAM database as negative sequences. We fine-
192 tuned the model with all parameters and achieved the best result in all fine-tuned
193 strategies, which was named as LAMAR-IRES (Supplementary Figure 3). Some
194 algorithms have been built to annotate the IRESs, such as IRESPred [31], IRESfinder
195 [32] and IRESpy [33]. Since these methods were trained on noisy experimental libraries,

196 we did not select them and fine-tuned RNAFM as the benchmark. Figure 5C showed
197 that both two fine-tuned models achieved the good AUC around 0.98.
198 To verify the effect of *in silico* screening of IRES, we used LAMAR-IRES to
199 classify the published experimental libraries sourced from viral and cellular mRNAs
200 [25-27]. As illustrated in Figure 5D, the median of predicted probabilities of putative
201 IRESs are significantly higher than those of remaining sequences, revealing the *in silico*
202 screening ability of LAMAR-IRES. We further tested the model with in-house
203 experimental data sourced from single strand viral genomes. The results showed that
204 the activities of predicted IRES were significantly higher than those of predicted not
205 IRES (Figure 5E), suggesting that our model integrated the causal relationship between
206 RNA sequence and IRES activity.

207 **The fine-tuned model predicted RNA degradation rates accurately**

208 As one of the most prominent sequences regulating mRNA decay, the 3'
209 untranslated regions (3' UTRs) recruit RNA binding proteins to promote or inhibit
210 degradation pathways [34].
211 In this section, we fine-tuned our LAMAR to predict the half-life of mRNAs
212 containing different 3' UTRs (Figure 6A). The half-lives of 1967 3 'UTRs were
213 measured by high-throughput experiments in the BEAS-2B cell line in David Erle
214 dataset [35]. Consistent with the previous section, we measured the model performance
215 in predicting the half-lives of 3' UTR with ten-fold cross-validation. We compared the
216 performances of various fine-tuning strategies and found that the full-parameter fine-
217 tuned model with the contextual length of 4096 tokens achieved the best result
218 (Supplementary Figure 4A, B). We named the fine-tuned model as LAMAR-DR and
219 benchmarked with other methods. Comparison with RNAFM, LAMAR-DR improved
220 the MSE by 7.4% (Figure 6B) and the correlation coefficient by 8.0% (Figure 6C).
221 We further tested LAMAR-DR by predicting the mutational effects with the
222 saturation mutation data on CXCL2 3' UTR. For each mutation, we take the difference

223 between the half-life of the mutant sequence and the natural sequence as the mutational
224 effect. The results showed that the mutation effect predicted by LAMAR-DR was
225 consistent with the experimental results, and revealed that there was a strong
226 degradation element, AU-rich element (ARE), in the upstream of 3' UTR (Figure 6D).
227 The Spearman correlation coefficient between predicted mutation effect and
228 experimental results was 0.515 (Supplementary Figure 4C).
229

Discussion

We presented a foundation language model LAMAR for multilayer regulation of RNA. The model learned the semantic representations of RNA by unsupervised pretraining on a large number of cross-species and multi-type RNA sequences. We fine-tuned the model with labeled datasets for several tasks including prediction of splice sites, translation efficiency, IRES activity and degradation rate. Our model outperformed benchmark methods including RNAFM and UTR-LM, reflecting the extraordinary performance of LAMAR in RNA regulation modeling. Interestingly, although RNAFM was pretrained on non-coding RNA sequences from RNACentral database, it could be finetuned to achieve good results in tasks about mRNA regulation, indicative of the powerful generalization of the pretrained models.

Since LAMAR shared the similar architecture with transformer encoders containing the self-attention layers, the number of parameters is proportional to the square of the input length. Therefore, some ultra-long RNA sequences cannot be fully input into the model during pre-training and fine-tuning due to the limitation of graphic memory. Training with split sequence segments leads to the loss of distant coevolutionary information in RNA. In addition, the attributes of RNAs such as translation efficiency and degradation rate are dependent on entire sequences [36]. Training with partial sequence elements will lose the model performance. Fortunately, several frameworks have been designed for ultra-long sequences recently [37, 38], and we will explore these techniques in the next version of the model.

References

1. Brenner S, Jacob F, Meselson M: **An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis.** *Nature* 1961, **190:**576-581.
2. Statello L, Guo C-J, Chen L-L, Huarte M: **Gene regulation by long non-coding RNAs and its biological functions.** *Nature Reviews Molecular Cell Biology* 2021, **22:**96-118.
3. Licatalosi DD, Darnell RB: **RNA processing and its regulation: global insights into biological networks.** *Nature Reviews Genetics* 2010, **11:**75-87.
4. Fan X, Chang T, Chen C, Hafner M, Wang Z: **Accurate prediction of RNA translation with a deep learning architecture.** *bioRxiv* 2023**:**2023.2007.2008.548206.
5. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, et al: **Predicting Splicing from Primary Sequence with Deep Learning.** *Cell* 2019, **176:**535-548.e524.
6. Song Z, Huang D, Song B, Chen K, Song Y, Liu G, Su J, Magalhães JPd, Rigden DJ, Meng J: **Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications.** *Nature Communications* 2021, **12:**4011.
7. Stroup EK, Ji Z: **Deep learning of human polyadenylation sites at nucleotide resolution reveals molecular determinants of site usage and relevance in disease.** *Nature Communications* 2023, **14:**7378.
8. Sun L, Xu K, Huang W, Yang YT, Li P, Tang L, Xiong T, Zhang QC: **Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures.** *Cell Research* 2021, **31:**495-516.
9. Devlin J, Chang M-W, Lee K, Toutanova K: **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** In *North American Chapter of the Association for Computational Linguistics*. 2019
10. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al: **Language Models are Few-Shot Learners.** *ArXiv* 2020, **abs/2005.14165**.
11. Zhang Y, Lang M, Jiang J, Gao Z, Xu F, Litfin T, Chen K, Singh J, Huang X, Song G, et al: **Multiple sequence alignment-based RNA language model and its application to structural inference.** *Nucleic Acids Research* 2023, **52:**e3-e3.
12. Wang N, Bian J, Li Y, Li X, Mumtaz S, Kong L, Xiong H: **Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning.** *Nature Machine Intelligence* 2024, **6:**548-557.

13. Chen J, Hu Z, Sun S, Tan Q, Wang Y, Yu Q, Zong L, Hong L, Xiao J, King I, Li Y: **Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions.** *bioRxiv* 2022.
14. Matera AG, Wang Z: **A day in the life of the spliceosome.** *Nature Reviews Molecular Cell Biology* 2014, **15:**108-121.
15. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, et al: **The human splicing code reveals new insights into the genetic determinants of disease.** *Science* 2015, **347:**1254806.
16. Soemedi R, Cygan KJ, Rhine CL, Wang J, Bulacan C, Yang J, Bayrak-Toydemir P, McDonald J, Fairbrother WG: **Pathogenic variants that alter protein code often disrupt splicing.** *Nature Genetics* 2017, **49:**848-855.
17. Wang ZF, Burge CB: **Splicing regulation: From a parts list of regulatory elements to an integrated splicing code.** *RNA* 2008, **14:**802-813.
18. Hinnebusch AG, Ivanov IP, Sonenberg N: **Translational control by 5′-untranslated regions of eukaryotic mRNAs.** *Science* 2016, **352:**1413-1416.
19. Cao J, Novoa EM, Zhang Z, Chen WCW, Liu D, Choi GCG, Wong ASL, Wehrspaun C, Kellis M, Lu TK: **High-throughput 5′ UTR engineering for enhanced protein production in non-viral gene therapies.** *Nature Communications* 2021, **12:**4138.
20. Castillo-Hair S, Fedak S, Wang B, Linder J, Havens K, Certo M, Seelig G: **Optimizing 5'UTRs for mRNA-delivered gene editing using deep learning.** *Nature Communications* 2024, **15:**5284.
21. Chu Y, Yu D, Li Y, Huang K, Shen Y, Cong L, Zhang J, Wang M: **A 5′ UTR language model for decoding untranslated regions of mRNA and function predictions.** *Nature Machine Intelligence* 2024, **6:**449-460.
22. Yang Y, Wang Z: **IRES-mediated cap-independent translation, a path leading to hidden proteome.** *Journal of Molecular Cell Biology* 2019, **11:**911-919.
23. Fan X, Yang Y, Chen C, Wang Z: **Pervasive translation of circular RNAs driven by short IRES-like elements.** *Nature Communications* 2022, **13:**3751.
24. Pamudurti NR, Bartok O, Jens M, Ashwal-Fluss R, Stottmeister C, Ruhe L, Hanan M, Wyler E, Perez-Hernandez D, Ramberger E, et al: **Translation of CircRNAs.** *Molecular Cell* 2017, **66:**9-21.e27.
25. Chen C-K, Cheng R, Demeter J, Chen J, Weingarten-Gabbay S, Jiang L, Snyder MP, Weissman JS, Segal E, Jackson PK, Chang HY:

Structured elements drive extensive circular RNA translation. *Molecular Cell* 2021, **81:**4300-4318.e4313.

26. Weingarten-Gabbay S, Elias-Kirma S, Nir R, Gritsenko AA, Stern-Ginossar N, Yakhini Z, Weinberger A, Segal E: **Systematic discovery of cap-independent translation sequences in human and viral genomes.** *Science* 2016, **351:**aad4939.
27. Yang Y, Fan X, Ye Y, Chen C, Ludwig SEJ, Zhang S, Lu Q, Will CL, Urlaub H, Sun J, et al: **Unbiased screen of human transcriptome reveals an unexpected role of 3′ UTRs in translation initiation.** *bioRxiv* 2021**:**2021.2005.2028.446092.
28. Mokrejš M, Mašek T, Vopálenský V, Hlubuček P, Delbos P, Pospíšek M: **IRESite—a tool for the examination of viral and cellular internal ribosome entry sites.** *Nucleic Acids Research* 2010, **38:**D131-D136.
29. Zhao J, Li Y, Wang C, Zhang H, Zhang H, Jiang B, Guo X, Song X: **IRESbase: A Comprehensive Database of Experimentally Validated Internal Ribosome Entry Sites.** *Genomics, Proteomics & Bioinformatics* 2020, **18:**129-139.
30. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, et al: **Rfam 14: expanded coverage of metagenomic, viral and microRNA families.** *Nucleic Acids Research* 2021, **49:**D192-D200.
31. Gritsenko AA, Weingarten-Gabbay S, Elias-Kirma S, Nir R, de Ridder D, Segal E: **Sequence features of viral and human Internal Ribosome Entry Sites predictive of their activity.** *PLOS Computational Biology* 2017, **13:**e1005734.
32. Zhao J, Wu J, Xu T, Yang Q, He J, Song X: **IRESfinder: Identifying RNA internal ribosome entry site in eukaryotic cell using framed k-mer features.** *Journal of Genetics and Genomics* 2018, **45:**403-406.
33. Wang J, Gribskov M: **IRESpy: an XGBoost model for prediction of internal ribosome entry sites.** *BMC Bioinformatics* 2019, **20:**409.
34. Mayr C: **Regulation by 3′ -Untranslated Regions.** *Annual Review of Genetics* 2017, **51:**171-194.
35. Zhao W, Pollack JL, Blagev DP, Zaitlen N, McManus MT, Erle DJ: **Massively parallel functional annotation of 3′ untranslated regions.** *Nature Biotechnology* 2014, **32:**387-391.
36. Agarwal V, Kelley DR: **The genetic and biochemical determinants of mRNA degradation rates in mammals.** *Genome Biology* 2022, **23:**245.
37. Nguyen E, Poli M, Faizi M, Thomas AW, Sykes CB, Wornow M, Patel A, Rabideau C, Massaroli S, Bengio Y, et al: **HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution.** In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. pp. Article 1872. New Orleans, LA, USA: Curran

Associates Inc.; 2024:Article 1872.

38. Nguyen E, Poli M, Durrant MG, Thomas AW, Kang B, Sullivan J, Ng MY, Lewis A, Patel A, Lou A, et al: **Sequence modeling and design from molecular to genome scale with Evo.** *bioRxiv* 2024**:**2024.2002.2027.582234.

39. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al: **Evolutionary-scale prediction of atomic-level protein structure with a language model.** *Science* 2023, **379:**1123-1130.

40. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.** *Nucleic Acids Research* 2016, **44:**D733-D745.

41. Consortium RN: **RNAcentral 2021: secondary structure integration, improved sequence search and new member databases.** *Nucleic Acids Research* 2021, **49:**D212-D220.

42. Dao T, Fu DY, Ermon S, Rudra A, Ré C: **FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness.** *arXiv e-prints* 2022**:**arXiv:2205.14135.

Methods

1. Model architecture

We developed a foundation language model for studying RNA regulation, named LAMAR. The model shares the similar architecture with ESM-2 [39] and consists of the embedding layer, encoder block and predictor. The encoder block contains 12 encoder layers, each of which has 12 self-attention heads. The specific predictor was designed for different tasks.

When RNA sequences are input into LAMAR, sequence x was tokenized into the list of tokens $(x_1, x_2, \ldots, x_L)$, and was added with [cls] and [eos] token at the beginning and end of the list, where $x_i \in$ ['A', 'T', 'C', 'G', 'N'] and L represents the length of RNA sequence. The tokens were converted into vectors with 768 dimensions by the embedding layer and finally the RNA sequence was represented as a (L+2, 768) dimensional embedding matrix. The Embedding matrix becomes the final embedding of the RNA sequence through the encoder block.

In the pretraining stage, we trained the model with masked language modeling. The 15% of the nucleotides in RNA sequences were randomly masked and then predicted with the remaining unmasked nucleotides. The predictor was consisted of two linear layers and a layer-normalization layer. The Loss function can be expressed as:

$$Loss = E_x E_M \sum_{i \in M} -logP(x_i | x_{\sim M}; \Theta) \quad (1)$$

In formula (1), M represents the indices of masked nucleotides, and $\sim M$ represents the complement of M, that is, the indices of unmasked nucleotides. Θ represents the trainable parameters of the model.

The downstream tasks were categorized into two types: token classification and sequence regression. For token classification, the functional class of each nucleotide was predicted through the two stacked linear layers. The classification loss was defined as the cross-entropy between the predicted and true labels of tokens. For sequence regression, the functional attributes of RNAs were predicted by the embedding of the

426 [cls] token followed by two stacked linear layers. The regression loss was calculated by
427 the squared error between the predicted and actual values.

428 **2. Pretrained datasets**

429 The pretrained datasets of LAMAR derived from four sources: 1) Protein coding
430 and non-coding transcripts of mammals from RefSeq [40]; 2) Non-coding transcripts
431 of mammals from RNACentral [41]; 3) Referenced genes of mammals from NCBI
432 Genome; 4) Referenced genes of viruses that infect mammals from NCBI Genome.
433 RefSeq (Release 221) contained 13,319,105 transcripts of 223 mammalian species
434 in 15 suborders. The sequences were downloaded as fna.gz format from website
435 https://ftp.ncbi.nlm.nih.gov/refseq/release/vertebrate_mammalian/. RNACentral
436 (Release 23) contained 2,649,775 non-coding transcripts from 4,669 mammals, 214 of
437 which were included to be consistent with species in RefSeq datasets. The sequences
438 could be downloaded from https://ftp.ebi.ac.uk/pub/databases/RNAcentral/releases/23
439 .0/sequences/. As of Dec 14 2023, NCBI Genome contained 222 mammalian reference
440 genomes annotated by RefSeq. The sequences and annotations were respectively
441 downloaded as fna.gz and gff.gz from the links on the metadata file
442 (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/assembly_summ
443 ary.txt). The 6,033,179 reference genes were extracted from the reference genomes by
444 bedtools with annotations. As of Jan 4 2024, NCBI Virus contained 15,037 viral
445 reference genomes, 1579 of which infected mammals were included into pretrain
446 datasets (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral/assembly_summary.txt).
447 The 19,007 reference genes were extracted by bedtools with annotations. The above
448 data were merged into the initial corpus, 15,455,049 transcripts and 5,832,792 genes in
449 total.
450 For the initial pretrained corpus, the sequences less than 15nt in length were filtered.
451 In order to remove the redundant sequences, we used MMseq2 tool to respectively
452 cluster transcripts and genes (cluster mode=2, identity=0.8, coverage=0.8), and

obtained 6,673,915 and 4,557,965 clusters, respectively.

3. Training procedure

3.1 Pretraining stage

To access our pretrained model, we built a validation set by randomly leaving out 1% of gene and transcript clusters, and pretrained the model with the remaining clusters. The representative sequences from each cluster and 50% of the member sequences were included into pretraining corpus, with an upper limit of 200 members per cluster. In addition, the testing corpus was composed of the representative sequences of testing cluster. In all, the training corpus encompassed a total of 14,873,638 sequences, which accounted for approximately 269 billion tokens. The testing corpus comprised 112,808 sequences, providing representative samples for validation.

We pretrained two models with contextual lengths of 2048 and 4096 tokens. When the length of the pretraining sequences exceeded these limits, they were segmented into fragments shorter than the contextual length. The overlap of 256 and 512 tokens were ensured between adjacent fragments to maintain contextual information. For the evaluation phase, we randomly extracted fragments with contextual length from testing sequences that exceeded the pretraining length. Any testing sequences with more than 10% Ns were excluded. The 10,000 sequences were randomly sampled for validation.

Our model underwent 250,000 training steps to reach convergence, with batch sizes set to 512 and 256 for context lengths of 2048 and 4096 tokens. We utilized the AdamW optimizer which was regulated by a learning rate schedule involving warm up and decay. The learning rate was warmed up during the first 10,000 steps to the maximum value 1e-4 and then decayed in remaining steps. The model was pretrained with FP16 mixed precision and speeded up with FlashAttention library [42]. The pretraining was conducted on 8 A800 80GB graphic process units, taking around 260 hours to finish training.

3.2 Fine-tuning stage

We used 10-fold cross-validation to evaluate the performance of fine-tuned models except for prediction of splice sites. The optimal combination of hyperparameter was identified using a grid search of batch size, learning rate, and training steps. Consistent with the pretraining stage, the AdamW optimizer with linear decay of learning rate was selected. The model was fine-tuned with FP32 precision on the Sugon Z-100 16GB and Tesla V100 32GB clusters of graphic process units.

4. Analysis of RNA embedding

The encoder block transforms a sequence of length L into an embedding matrix of size (L+2, 768). The vector (i+1, 768) is set as the embedding for the i_{th} nucleotide, while the average of the embedding matrix is served as the embedding for the entire sequence.

5. Downstream tasks

5.1. Prediction of splice sites

We used the same dataset as SpliceAI [5] to predict the splice sites of human pre-mRNAs. The dataset consisted of 15,036 protein coding transcripts annotated by Gencode V24lift37. The transcripts from chromosome 2, 4, 6, 8, 10-22, X, Y were used to fine-tune the model (13,384 transcripts, 130,796 pairs of splice sites), and the other transcripts from chromosome 1, 3, 5, 7, 9 were included in validation set (1,652 transcripts, 14,289 pairs of splice sites). We used bedtools to extract the transcript sequences from hg19 reference genome with annotation files. The fine-tuning length was set to 1024 tokens due to limited graphic memory. The pre-mRNAs longer than 1024 tokens were segmented into fragments of 1024 tokens, with an overlap of 256 tokens. The labels of sequences were assigned using the annotation file, where the 5' splice site was encoded as 1, the 3' splice site as 2, and neither as 0. The middle regions

of 768 tokens were predicted by the entire fragments during fine-tuning.

5.2. Prediction of translation efficiency

We selected Cao dataset [19] to predict the translation efficiencies of mRNAs based on 5' UTRs. The expression and translation efficiency of 10903 human transcripts were measured by mRNA-seq and Ribo-seq in HEK293 cell line. The 5' UTRs were extracted by bedtools from hg19 reference genome with GENCODE V17 annotation. The fine-tuning length was set to 1024 tokens for this downstream task. When the 5' UTR sequence exceeded 1024 tokens, the segment of 1024 tokens closest to the 3' end was included in the fine-tuned set.

5.3. Annotation of IRES

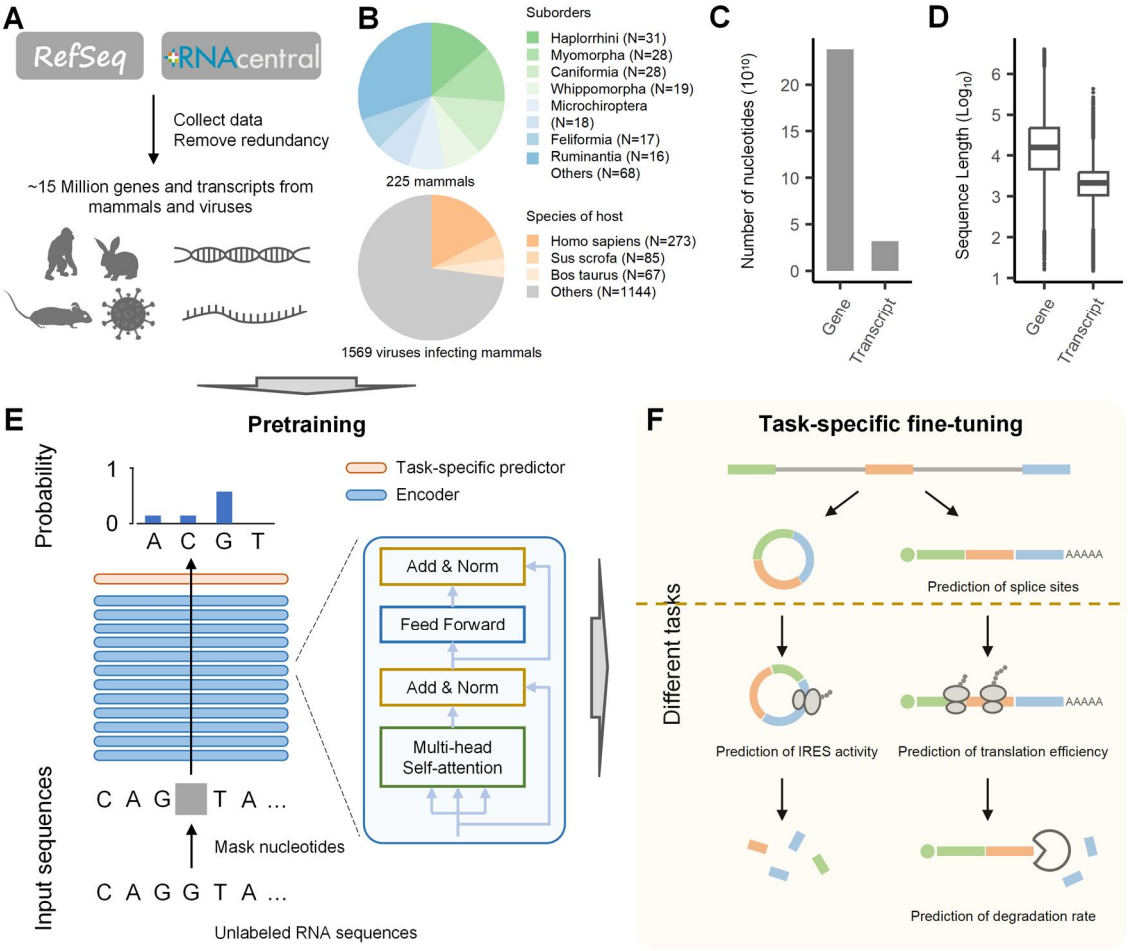
The training set for IRES annotation were derived from IRESite [28], IRESbase [29] and RFAM [30] databases. The 1,901 experimentally validated IRES sequences and 2,026,606 non-IRES sequences were collected. To avoid the data leakage caused by sequence similarity, we used MMseqs2 to cluster positive sequence and negative sequence respectively (identity=0.9, coverage=0.9), and obtained 903 positive clusters and 810,096 negative clusters. We sampled the representative sequences from the negative clusters so that the number of positive and negative sequences is the same and the length distribution is similar. We randomly selected 50% of the positive and negative clusters as the training set and the rest as the evaluation set. We repeated the experiment 10 times to measure the performance of the model.

5.4. Prediction of degradation rates

We selected David Erle dataset [35] to predict the half-lives of mRNAs based on 3' UTRs. The half-lives of 2790 3' UTRs were measured by Massive Parallel Reporter Assays in BEAS-2B cell line. The 1967 3'UTRs of 160 nucleotides with detected half-lives were included in fine-tuned dataset. The AU-rich element and context sequence

529 of CXCL2 3' UTR was saturated mutated [35]. The variant effect was assessed by
530 comparing the steady-state mRNA levels between control and mutated sequences.
531

532 **Figures and figure legends**



533

534 **Fig 1. Pretraining an RNA foundation model for multilayer regulation of RNA.**

535 **(A)** Schematic diagram of pretraining data processing. **(B)** The taxonomic makeup of
536 the pretraining species consisting of the mammals (top) and the viruses (bottom). **(C)**
537 The number of the nucleotide in the genes and transcripts. **(D)** The sequence length of
538 the genes and transcripts in the pretraining dataset. **(E)** Schematic diagram of the
539 pretraining model which predicts the likelihoods of the masked tokens given a sequence.
540 The encoder comprised the self-attention layers and fully-connected layers. **(F)**
541 Schematic diagram of the downstream tasks including supervised prediction of splice
542 site, translation efficiency, internal ribosome entry site (IRES) and degradation rate.
543

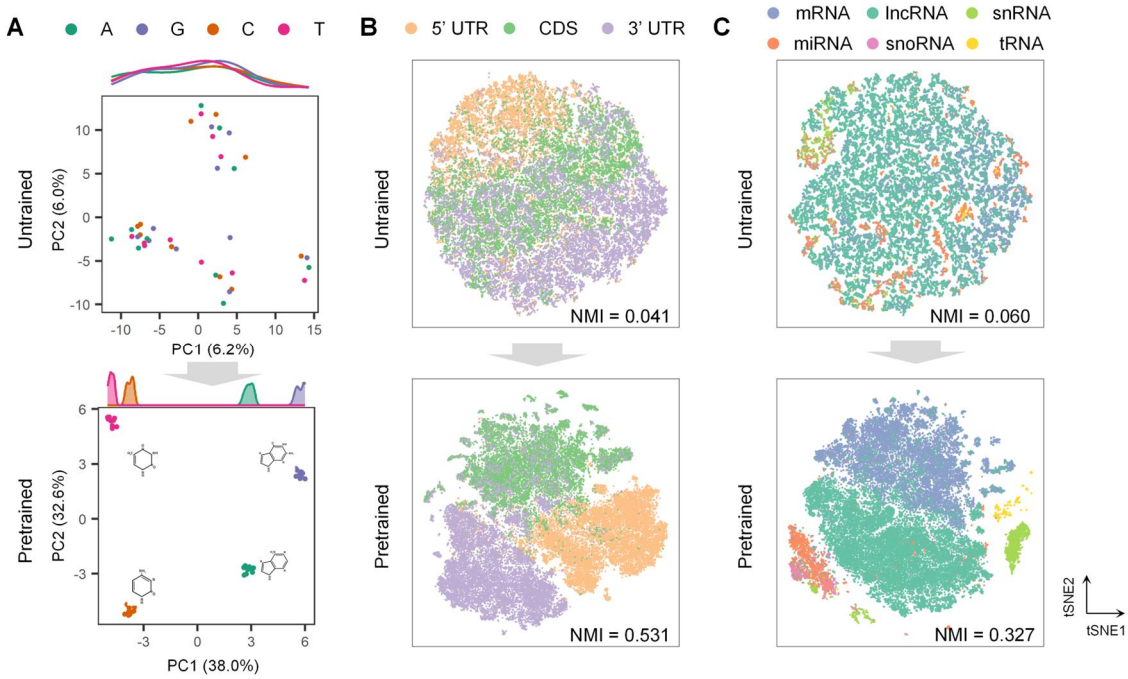


Fig 2. LAMAR learned the biological properties of nucleotides and the function features of RNAs.

Top, embeddings from pretrained model. Bottom, embeddings from untrained model. **(A)** The PCA plot visualizing the embeddings of four classes of nucleotides. **(B)** The tSNE plot visualizing the embeddings of translation regulated elements: 5' UTR, CDS and 3' UTR. **(C)** The tSNE plot visualizing the embeddings of RNAs: mRNA, lncRNA, snRNA, miRNA, snoRNA and tRNA.

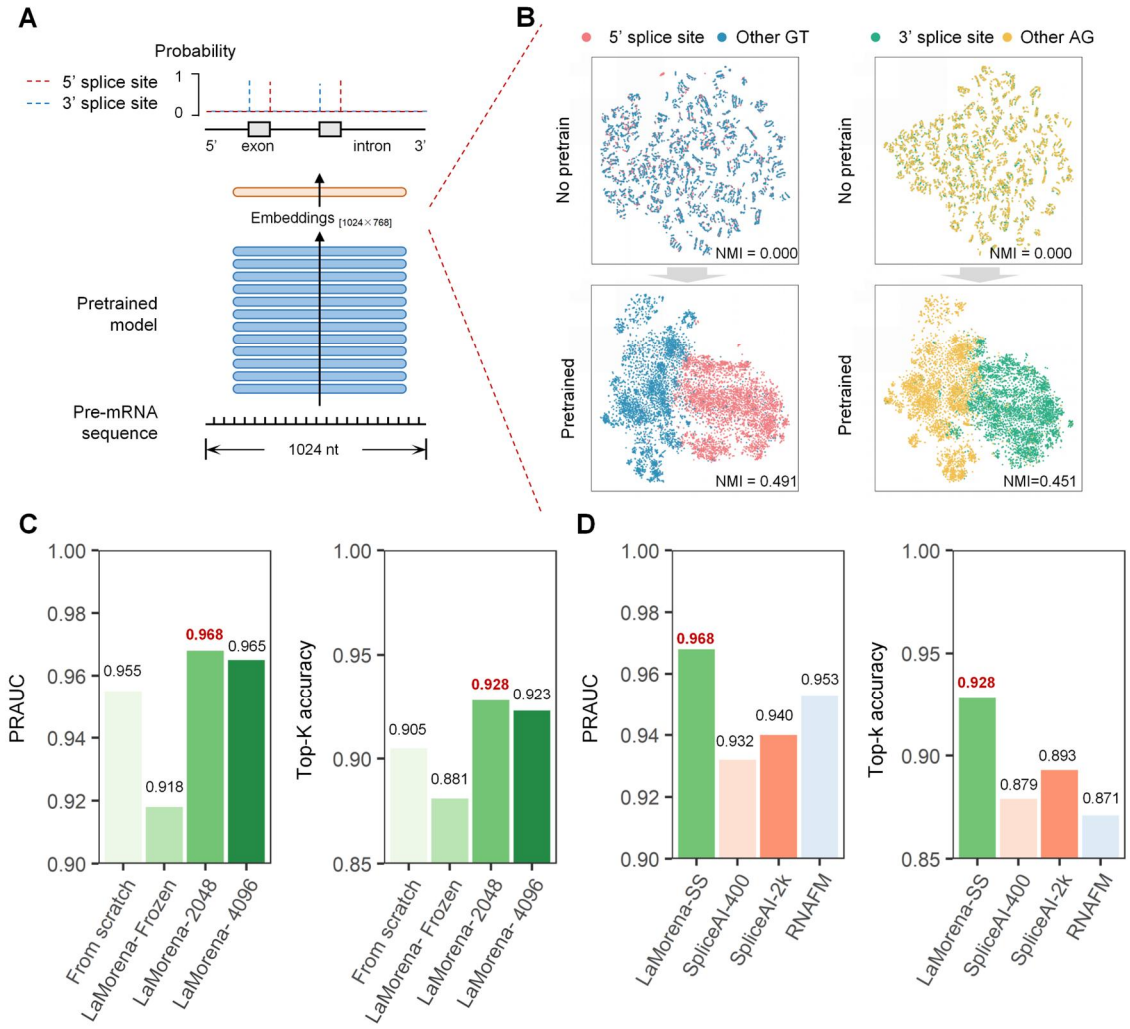


Fig 3. Prediction of splice sites of pre-mRNAs.

(A) Schematic diagram of the fine-tuned model predicting splice sites. The model predicts the probability that a nucleotide is a 5' splice site, a 3' splice site or neither given the pre-mRNA segments that are 1024 nucleotides long. **(B)** The tSNE plots visualizing the embeddings of the splice sites and control sites output from the untrained model (top) and the pretrained model (bottom). **(C)** The Top-K accuracy and PRAUC of the models finetuned with different strategies. **(D)** The Top-K accuracy and PRAUC of LAMAR-SS and benchmarks including SpliceAI submodels and RNAFM.

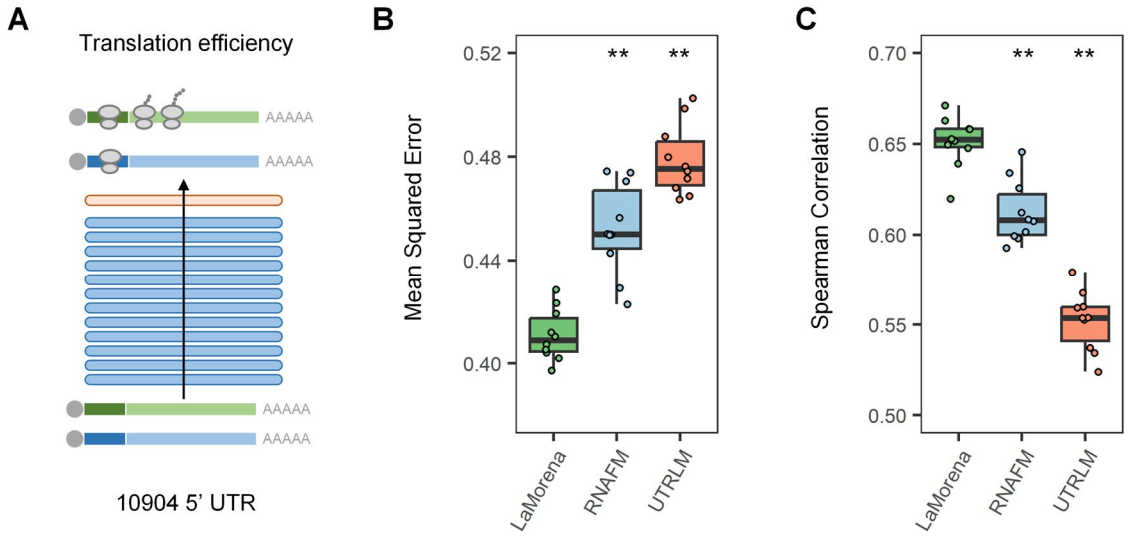
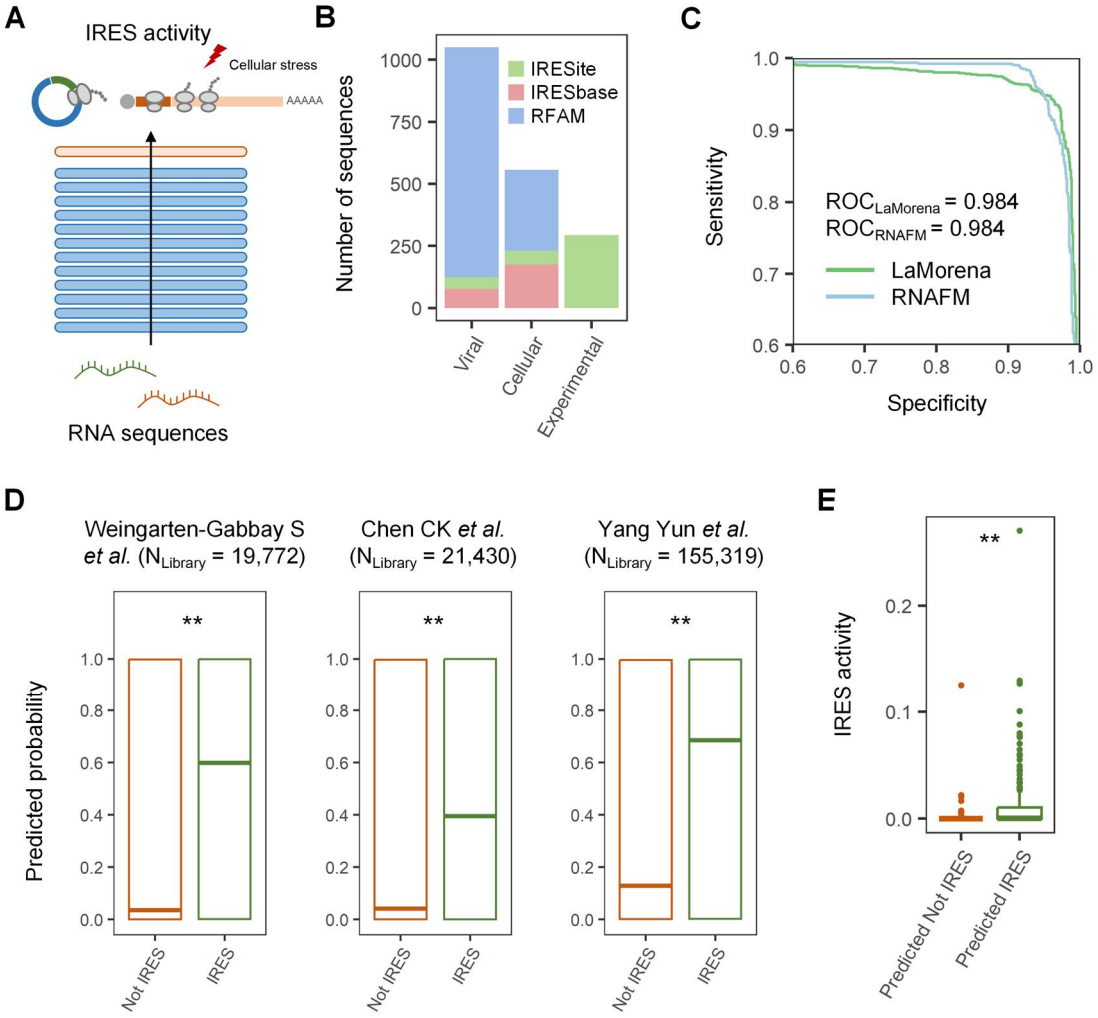


Fig 4. Prediction of translation efficiencies of mRNAs base on 5' UTRs.

(A) Schematic diagram of the fine-tuned model predicting translation efficiencies of mRNAs based on 5' UTRs. **(B)** The mean squared error between observed translation efficiency and predicted values of LAMAR-TE and benchmarks including RNAFM and UTRLM. **(C)** The spearman correlation coefficient between observed translation efficiency and predicted values of LAMAR-TE and benchmarks. The P values were calculated with Wilcoxon signed rank test. *, P value < 0.05; **, P value < 0.01.



572

573 **Fig 5. Annotation of IRES.**

574 **(A)** Schematic diagram of the fine-tuned model annotating IRES. **(B)** The number of
575 collected experimentally validated IRESs sourced from IRESite, IRESbase and RFAM
576 databases. **(C)** The AUC curves of LAMAR-IRES and RNAFM. **(D)** The predicted
577 probabilities of sequences in three experimental libraries screening IRES. **(E)** The
578 activities of predicted IRESs and not-IRESs by LAMAR-IRES in HepG2 cell lines.
579

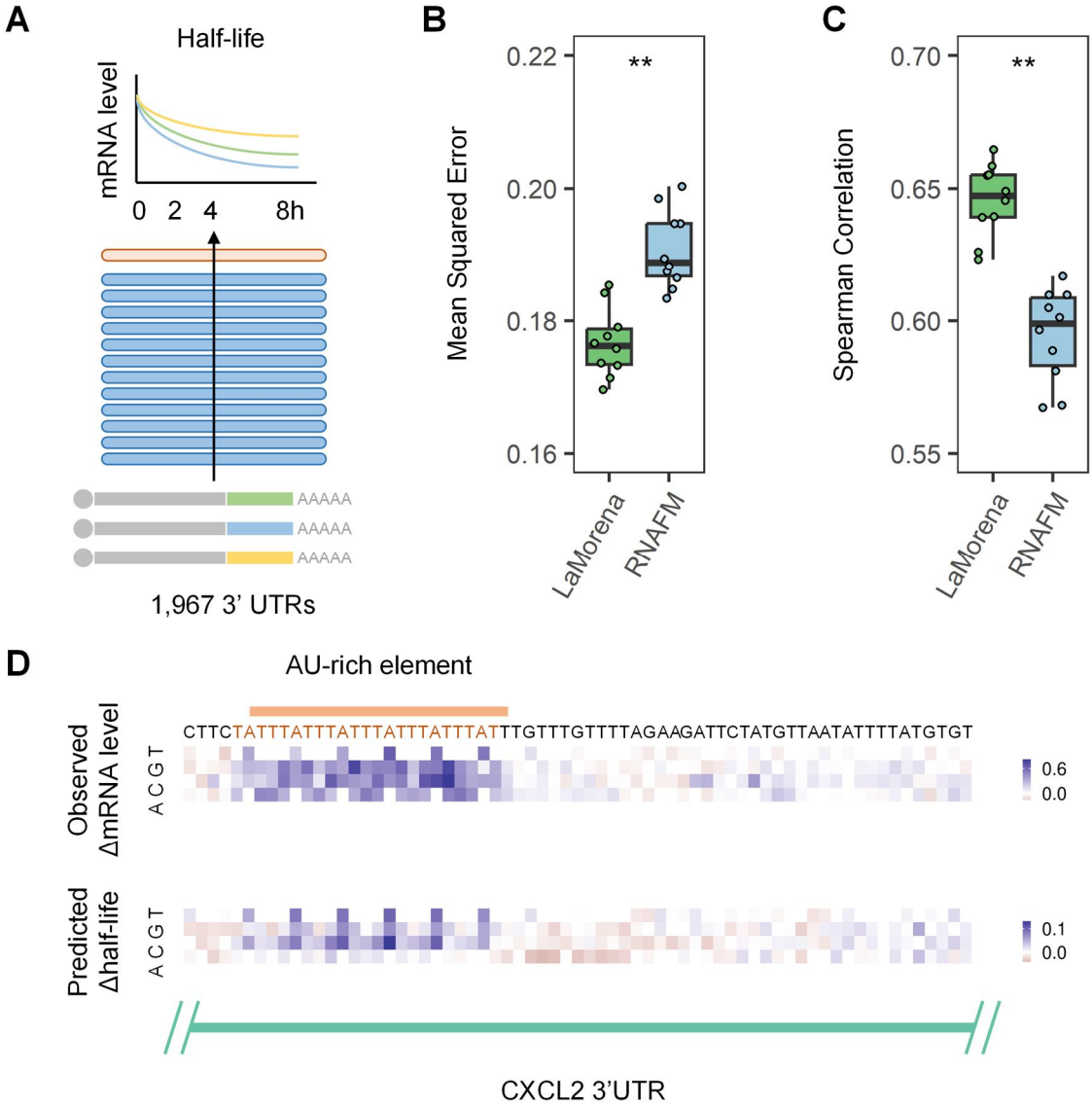


Fig 6. Prediction of degradation rates of 3' UTRs.

(A) Schematic diagram of the fine-tuned model predicting the half-times of mRNAs containing different 3' UTRs. **(B)** The mean squared error between observed half-times and predicted values of LAMAR-DR and benchmarks. **(C)** The spearman correlation coefficient between observed half-times and predicted values of LAMAR-DR and benchmarks. **(D)** The observed (Top) and predicted (Bottom) variant effects of CXCL2 3' UTR. The nucleotides of the AU-rich element were colored in orange. *P* values were calculated with Wilcoxon signed rank test. *, *P* value < 0.05; **, *P* value < 0.01.