Accurate prediction of RNA translation with a deep learning architecture

Xiaojuan Fan[1, 2, *, #], Tiangen Chang[4, *], Chuyun Chen[1, 3], Markus Hafner[2], Zefeng Wang[1, 3, #]

[1] Bio-med Big Data Center, CAS Key Laboratory of Computational Biology, CAS Center for Excellence in Molecular Cell Science, Shanghai Institute of Nutrition and Health,

[2] RNA Molecular Biology Laboratory, National Institute of Arthritis and Musculoskeletal and Skin Disease, Bethesda, MD, USA.

[3] University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

[4] Laboratory of Cancer Data Science, National Cancer Institute, Bethesda, MD, USA.

* These authors contributed equally to this work

[#] Corresponding to: wangzefeng@picb.an.cn, fanx3@nih.gov

Abstract

Accurate annotation of coding regions in RNAs is essential for understanding gene translation. We developed a deep neural network to directly predict and analyze translation initiation and termination sites from RNA sequences. Trained with human transcripts, our model learned hidden rules of translation control and achieved a near perfect prediction of translation sites across entire transcriptome. Our model revealed a surprising role of codon usage in regulating translation termination, which was experimentally validated. We also identified thousands of new open reading frames in mRNAs or annotated lncRNAs, some of which were confirmed experimentally. Remarkably, the model trained with human mRNAs achieved high prediction accuracy in all eukaryotes and good prediction in polycistronic transcripts from prokaryotes or RNA viruses, suggesting a high degree of conservation in translation control. Collectively, this study presents a general and efficient deep learning model for RNA translation, providing new insights into the complexity of translation regulation.

Introduction

Efforts to annotate protein-coding open reading frames (ORFs) have traditionally relied on empirical rules such as sequence conservation pattern (Frankish et al., 2019), the longest ORFs in one transcript (Mouilleron et al., 2016) and initiation from the first AUG (Hinnebusch, 2011). However, these rules are over-simplified without considering the complex regulation by other factors, such as RNA structure and various regulatory *cis*-elements. In addition, RNA translation has recently been found in previously annotated non-coding RNAs (e.g., lncRNAs and circRNAs) (Jackson et al., 2018; Matsumoto et al., 2017; Yang et al., 2017), and can be initiated from alternative sites or non-canonical codons (Chen et al., 2020; Raj et al., 2016), resulting in non-annotated translation products that play critical and diverse cellular roles (Chen *et al.*, 2020; Matsumoto *et al.*, 2017). Considering the complexities of translation, the prediction of translation landmarks, such as initiation and termination sites with high accuracy remains a challenging task, especially for poorly annotated genomes or in the context of complex regulatory networks. Rapid advances in artificial intelligence (AI) provide new hope for dissecting complex biological systems, such as alternative splicing (Jaganathan et al., 2019) and protein folding (Jumper et al., 2021), by uncovering hidden associations and unknown rules (Bogard et al., 2019; Eraslan et al., 2019; Vamathevan et al., 2019; Zhou et al., 2018).

The complete synthesis of a protein requires accurate initiation and termination of translation by ribosomes at specific positions. In eukaryotes, translation initiation is generally mediated by the 5′-cap, which recruits various translation initiation factors

62 and small ribosomal subunit to form preinitiation complex (Sonenberg and
63 Hinnebusch, 2009). This complex subsequently scans through the 5′-untranslated
64 region (5′-UTR) to recognize the optimal start codon, which is determined by both
65 codon identity and its immediate nucleotide context known as Kozak sequence
66 GCCRCCAUGG (R represents purine) (Hinnebusch, 2011). However, in certain
67 cases, the ribosome may skip some weak start codons, resulting in leaky scanning
68 (Hinnebusch, 2011). In addition, many genes may contain several alternative
69 translation initiation sites (TISs), and it is unclear how the ribosomes select one of
70 these sites to initiate translation. Evaluating the strength of each TIS within one
71 transcript and the general rules of selecting the authentic ORFs from possible decoys
72 remain challenging. Therefore the systematic identification of translation sites
73 (translation initiation and termination sites, or TIS/TTS) and the quantitative
74 evaluation of their strength become a prerequisite for exploring the complexity of
75 translation regulation.
76 New experimental techniques, including GTI-seq (global translation initiation
77 sequencing) (Lee et al., 2012) and QTI-seq (quantitative translation initiation
78 sequencing) (Gao et al., 2015), have been developed to identify TISs by directly
79 capturing the stalled initiating ribosomes at single nucleotide resolution across the
80 transcriptome. However, these methods were typically conducted in specific cell
81 types, which limits the transcriptome-wide TIS identification. Several computational
82 methods have also been developed to predict translation initiation sites from mRNA
83 sequences. For example, a physicochemical-property based predictor using pseudo-

trinucleotide composition was constructed to identify TIS in human genes (Chen et al., 2014), and a variety of machine learning methods have also been developed to predict the TIS by analyzing a short sequence fragment (e.g. 200nt) around the start codon of mRNA (Goel et al., 2020; Kalkatawi et al., 2019; Zhang et al., 2017). However, such methods are biased because they are limited to the immediate sequence context surrounding the TISs. Nevertheless, non-coding regions of mRNAs (i.e. 5′-UTR and 3′-UTR), which are not captured in these approaches play an important regulatory role in selecting the start codon (Leppek et al., 2018; Mayr, 2019), which in turn affects the accuracy of TIS predictions.

Here, we developed a deep learning method based on a multilevel dilated convolution network, named TranslationAI, to independently predict TISs and TTSs from full-length mRNA sequences alone. This method can accurately learn the triplicity in the genetic code, and achieve >99% accuracy in predicting known TISs and TTSs in human transcriptome. Our model uncovered regulatory sequences in the UTRs that determine the strength of translation start and stop codons, and identified thousands of new ORFs in mRNAs or annotated lncRNAs. In addition, this model can be extended to other eukaryotes, prokaryote and several human viruses, suggesting a strong conservation in defining coding sequences across different domains of life.

Results

Modeling translation from full length mRNA with deep learning

We constructed a deep residual convolutional neural network, TranslationAI, using the full-length mRNA sequences as input to independently predict the TIS and TTS (Fig. 1A, see methods). Previous approaches only considered short sequence stretches around a potential start codon (Goel *et al.*, 2020; Kalkatawi *et al.*, 2019; Zhang *et al.*, 2017), or reported all potential ORFs in six frames with specified start codon (ATG), stop codon (TAG/TAA/TGA), and length cutoffs (ORFfinder in NCBI). In contrast, our deep learning model evaluates the potential of each position in a given mRNA to function as a TIS or TTS (or neither) without using any prior knowledge of translation (e.g., the triplet or identity of genetic code). Such *ab initio* prediction enables the neural network to learn the hidden rules of translation from sequence context alone.

The deep residual neural network was constructed using the one-hot encoding of mRNA sequences as input (Fig. 1A). We built a 32-layer dilated convolution neural network architecture that produces an output matrix of the probability for the given position being a TIS, TTS or neither (Fig. S1, see methods). To obtain a comprehensive and well-curated dataset, we utilized the RefSeq gene annotation and extracted 47,098 protein-coding transcripts (47,098 TIS-TTS pairs). We trained the model using the transcripts on chromosomes 2, 4, 6, 8, 10-22, X, and Y, and tested the model with the the transcripts from chromosomes 1, 3, 5, 7, and 9. The stringent top-1 accuracy was used for TIS/TTS prediction. Specifically, the score cutoff was

126 set at the point where the number of predicted sites is equal to the real number (i.e.,
127 the false positive rate equal to false negative rate, the score cutoff at this point is 0.5),
128 which ensured that only the high-confidence predicted sites were considered for
129 downstream analysis.

130 To explore the effect of different input windows on prediction accuracy, we
131 considered four input windows that include 60, 200, 600, and 2,000 (2k) nucleotides
132 on both sides of a given position. Interestingly, the prediction accuracy improved
133 dramatically with the increasing input window size, reaching >99% accuracy with a
134 2k nt input window (Fig. 1B). Considering that >70% of transcripts are longer that 2k
135 nt (Fig. S2A), such high accuracy of TranslationAI prediction is particularly
136 remarkable, suggesting that the identity of TIS/TTS could be influenced by the
137 sequences at thousands of nucleotides away, potentially through long-range
138 interactions such as RNA structure or RNA binding proteins. Our findings further
139 suggest that TranslationAI has learned to capture these interactions using sequence
140 information alone.

141 To assess the impact of sequence context on the accuracy of TIS/TTS prediction,
142 we compared the performance of the models trained on 200 and 2k nt sequence
143 context. The 2k model achieved high prediction accuracy (>0.95) in TIS/TTS
144 prediction across a wide range of transcript lengths, however the accuracy of the 200
145 nt model decreases dramatically with increase of transcript length (Fig. 1C). This
146 observation is in agreement with earlier report that distal regions far away from the
147 TIS/TTS can affect the translation decision (Leppek *et al.*, 2018). With the 2k

148 model, the score distribution showed a clear separation between the positive and
149 negative predictions (Fig. 1D), suggesting that the model is robust under different
150 cutoff selections. To further examine the false discovery rate (FDR), we used the 2k
151 model to score potential TISs/TTSs in the shuffled mRNA sequences. At the
152 probability score cutoff of 0.5 for both TIS and TTS, less than 5% of the ORFs in
153 shuffled dataset passed the cutoff, yielding an empirical FDR < 0.05 (Fig. 1E).
154

155 **Key features learned by the AI model for accurate prediction**

156 Encouraged by the near-perfect performance of our deep learning network in
157 predicting translation sites, we sought to examine the features learned by this "black
158 box" model. We performed systematic *in silico* perturbations on different regions of
159 mRNAs to measure their effects on reducing the probability scores of the authentic
160 TISs/TTSs. Such perturbations can reflect how the AI model make the accurate
161 translation prediction in a computational sense, however the a different molecular
162 mechanism may be used by cells to define translation initiation and termination sites.
163 Nevertheless, it would be very interesting to compare how a biological process like
164 mRNA translation is perceived differently by the AI model and empirical tests.

165 We found that the replacement of annotated TISs or TTSs with random codons
166 dramatically reduced the predicted scores of cognate sites (Fig. 1F, Table S1),
167 suggesting this model had successfully learned the identity of start and stop codons
168 from the sequence alone. Consistently, introducing a known stop codon into the
169 coding sequences (CDS) led to a drastic reduction in predicted scores of both TISs

170 and TTSs. In addition, the shuffling of 5′-UTRs significantly affected corresponding
171 TIS prediction but not TTS, whereas the shuffling of 3′-UTR had little effect on
172 TIS/TTS prediction (Fig. 1G, Table S1), suggesting that the sequence of 5′-UTR (but
173 not 3′-UTR) play a critical role in ORF selection. Surprisingly, the deletions of 5′- or
174 3′-UTRs had little effect on TISs/TTSs prediction, with near perfect prediction for
175 transcripts with only 50 nt left in the UTRs (Fig. 1G, Table S1), however the
176 underlying reason is unclear. Finally, random deletions of one or two nucleotides
177 within CDS dramatically reduced the prediction scores, whereas random deletion of
178 three nucleotides in CDS had little effect (Fig. 1H, Table S1), suggesting that the
179 model was capable of learning the triplet rule of the genetic code by itself.
180 To further examine if our prediction was affected by codon usage, we introduced
181 a series of perturbations to the genetic codons of CDS. We found that synonymous
182 substitutions or the in-frame triplet shuffling (maintaining the same amino acid
183 compositions) had little effect on the prediction, whereas nonsynonymous
184 substitutions caused a small but statistically significant reduction in prediction scores
185 (Fig. 1I, Table S1). These results suggest that, while the choice of codons has little
186 effect on TranslationAI, the amino acid compositions of proteins might be used by the
187 AI model to make prediction although such information was hidden behind the input
188 of RNA sequences. Collectively, these results suggest that the TranslationAI
189 network has learned the rules of RNA translation from a subset of the transcriptome
190 without any prior knowledge. However this model did not use all the available
191 information, as shuffling of 3′-UTR had little effect on the prediction, despite reports

192 that the sequences in 3′-UTR play key roles in translation regulation (Mayr, 2019;
193 Vislovukh et al., 2014).
194 In the RefSeq dataset, ~100% of the annotated ORFs follow the triplet codon rule
195 (length=3N), 98% are the longest ORFs in the transcripts, and ~40% use the first
196 AUG triplet as their start codon, raising the possibility that only using these simple
197 rules may be sufficient to make prediction. To determine the extent to which
198 TranslationAI relies on these simple rules, we generated *in silico* permutations in the
199 test dataset and measured their effects on prediction accuracy. The introduction of
200 frameshift permutations significantly reduced the prediction accuracy (Fig. S2B),
201 however >50% of transcripts were still predicted to use original TIS/TTS despite
202 breaking the 3N rule (i.e. ORF length ≠ 3N), suggesting that the rule of triplet codon
203 is important but not unconditional for TIS/TTS prediction by TranslationAI. In
204 addition, when ORF length was reduced by introducing premature stop codons or
205 frameshifts, TranslationAI predicted the longest ORF in ~60% cases (Fig. S2C), again
206 suggesting that such a simple rule was partially followed by this model. Finally, the
207 preference for the first AUG was also affected by several permutations throughout the
208 transcript, including changes in TIS and the introduction of early stop or frameshift
209 (Fig. S2D). Interestingly, shortening the length of the 5′-UTR increased the ratio of
210 first AUG usage, whereas shuffling of the 5′-UTR reduced it, suggesting that the 5′-
211 UTR sequence plays a critical role in TIS selection (Leppek *et al.*, 2018; Mayr, 2019).
212 Collectively, these results indicate that TranslationAI has learned additional features
213 beyond simple traditional rules (i.e., the longest ORF, triplet code, and preference of

214 the first AUG) for accurate prediction, implying that the adjacent sequences near
215 TIS/TTS also play crucial roles in selecting translation start and stop sites.
216 We found no correlation (Spearman's correlation coefficient close to zero)
217 between the predicted score of TIS/TTS and the number of transcript isoforms
218 containing the same TIS/TTS sequences (Fig. S2E), suggesting that the scores of this
219 model were not biased by the training data. Additionally, the TIS and TTS scores in
220 the same ORF had a weak but statistically significant positive correlation (Spearman's
221 correlation coefficient $r = 0.2$, $p < 1*10^{-16}$) (Fig. S2F), although they were predicted
222 independently. Interestingly, the TIS/TTS scores by TranslationAI showed a positive
223 correlation with the translation efficiency estimated by the ratio of read density from
224 Ribosome profiling (Ribo-seq) normalized by mRNA abundance (Chen *et al.*, 2020)
225 (Fig. S2G and S2H), suggesting that these predicted scores partially reflect their
226 endogenous translational activities.
227
228 **Examine the sequences surrounding the TISs/TTSs of different strengths**
229 We next sought to compare the key features between the strong and weak
230 TIS/TTS sites defined by the TranslationAI scores (the top *vs.* the bottom 5th
231 percentile). Interestingly, the sequences surrounding strong TISs/TTSs are more
232 conserved than those around weak TISs/TTSs (Fig. 2A), implying additional selective
233 pressures for translation. Additionally, we examined the ribosome occupancy on the
234 mRNAs with different TIS/TTS scores using Ribo-seq data from iPSC and iPSC-
235 derived cardiomyocytes (Chen *et al.*, 2020; Fields et al., 2015). Our analysis
236 revealed that the transcripts with strong TISs or TTSs exhibit a higher ribosome peak

237 at the start or stop positions (Fig. 2B-2C, Fig. S3A-S3B), which is consistent with
238 their higher activity in mediating translation initiation or termination. These results
239 suggest that the predicted scores of TISs/TTSs reflect their activities in regulating
240 translational initiation/termination. In addition, the transcripts with weak TIS/TTS
241 generally have a longer 5′-UTR than those with strong TIS/TTS (Fig. S3C), but such
242 length difference was observed in the 3′-UTRs only for TTS with different strengths.
243 However the implication behind such length difference is unclear.

244 We further examined the enriched sequence motifs around the strong *vs.* weak
245 TISs/TTSs (from -30 nt to +33 nt) (Fig. 2D). As expected, we found an enriched
246 motif RCC<u>ATG</u>GC (R = A or G, with start codon underlined) at the strong TISs,
247 resembling the Kozak sequence that typically enhances translation initiation (Kozak,
248 1986; 1987). We also discovered a novel motif S<u>TGA</u>G at the strong TTSs but not
249 the weak TTSs (S = C or G, with stop codon underlined, Fig. 2D). Surprisingly, the
250 fragments surrounding the strong TTS are generally biased towards CG-rich
251 sequences, implying that the C/G-rich region may mediate translation termination.
252 Moreover, the C/G enrichment showed a triplet periodicity before the strong TTSs
253 (Fig. 2D), with the third position of a codon being most C/G biased. Such codon
254 bias probably reflects the evolutionary selection under the constrain of protein
255 sequences, suggesting an unappreciated role of codon usage in translation termination.

256 To confirm this codon bias, we analyzed the 30-nt region immediately upstream
257 of the stop codon of human mRNA (equivalent of 10 codons). Consistent with the
258 motif enrichment, the synonymous codons with C/G at the third position were

259 preferably used before the strong TTSs (Fig. 2E). Remarkably, such codon bias was
260 statistically significant for all 18 amino acids with multiple codons (Fig. 2E),
261 particularly for the amino acids with 4-6 codons (e.g., Thr, Val and Ser etc, Fig. S3D).
262 Together, these results reveal a complicated new role of TTS in translational
263 regulation, and indicate that the TranslationAI has learned this codon bias to make a
264 TTS prediction.

265 To experimentally validate the unexpected role of codon bias in translation
266 termination, we designed a series of translation readthrough reporters with different
267 TTSs (Fig. 2F). This reporter contains two independent transcription units, one
268 encoding the firefly luciferase (Fluc) as a transfection control, and the other encoding
269 a fusion protein of GFP and Nano-luciferase (Nluc) separated by a set of short
270 variable sequences (45 nt) containing candidate TTSs of different strengths (Fig. 2F).
271 As a control, we inserted the sequence around the terminal codon of the human
272 vitamin D receptor (VDR) gene, which is known to undergo translation readthrough
273 (Loughran et al., 2018). We further introduced different synonymous mutations
274 before the termination codon of VDR to either strengthen or weaken the TTS, and
275 assayed the levels of mRNA and translation products from these reporters (Fig. 2F
276 and S3E). The results showed that introducing the C/G-rich synonymous mutations
277 in VDR (i.e., generating a strong TTS) eliminated the product from translation
278 readthrough (indicated by arrows, Fig. 2F). Conversely, a weak TTS mutation with
279 A/U-rich synonymous codons before the stop codon increased the translation
280 readthrough. This experiment directly supported our finding on the role of codon

281 selection in translation termination. As an independent validation, we re-analyzed
282 the Ribo-seq data from two distinct cell lines to examine potential translation read
283 through (Chen *et al.*, 2020). As expected, the ribosomal density in the 3′-UTRs
284 following weak TTSs was significantly higher than that following strong TTSs (Fig.
285 2G), suggesting a translation "leakage" following the weak termination sites.
286

287 **Alternative TIS predicted by TranslationAI**

288 Certain mRNAs may use alternative TIS to produce additional protein isoforms
289 with extended N terminals (de Klerk and t Hoen, 2015), therefore we further examined
290 if the transaltionAI can predict alternative TISs that may compete with the annotated
291 sites (see methods). Based on the distribution of translationsAI scores (Fig. 1D), we
292 used the score cutoff at 0.1 to increase the sensitivity of the alternative TISs prediction.
293 In total, 5336 alternative TISs have been identified. Interestingly, the numbers of
294 alternative TIS identified by TranslationAI is negatively correlated with the strength of
295 the annotated TISs (Fig. S3F), suggesting that alternative translation initiation may
296 happen in the mRNA without dominant TIS. Specifically, ~55% of mRNAs with
297 weak annotated TISs contain an alternative TIS (1340 out of 2416 mRNAs), whereas
298 the alternative TISs were found in less than 3% of mRNAs with strong annotated TISs
299 (68 out of 2352 mRNAs). In addition, the score difference between the annotated TISs
300 and the alternative TIS was much smaller in the mRNAs containing weak TISs (Fig.
301 S3G, left), whereas the mRNAs with strong annotated TISs generally lack alternative
302 TISs with a competitive score (Fig. S3G, right). A consistent result was also found

303 when we only considered the alternative TISs in the same reading frame with the
304 annotated TISs (Fig. S3G). These results collectively suggested that many mRNAs
305 with weak TISs may also be translated from an alternative start site.

306 Further analysis suggested that the sequences at predicted alternative TISs are less
307 conserved than the annotated TISs (Fig. S3H), suggesting that these new TISs may be
308 emerged more recently during evolution. The predictions of alternative TISs were
309 further supported by the Ribo-seq data obtained in two different cell lines (Chen *et al.*,
310 2020), where the ribosome density in the 5′-UTRs could reflect the noncanonical
311 translation initiation driven by alternative TIS. Consistent with our prediction, the
312 transcripts with a predicted alternative TIS showed significantly higher ribosomal read
313 density in their 5′-UTRs compared to the mRNAs without alternative TIS (Fig. S3I),
314 suggesting a leaky translation in the UTRs due to alternative translation initiation.

315

316 **Identify non-canonical ORFs in human transcriptome**

317 It was recently reported that non-coding RNAs in the human transcriptome or the
318 non-coding regions of mRNAs may contain unannotated ORFs that are translated into
319 new proteins or small peptides (Chen *et al.*, 2020; Jackson *et al.*, 2018). Therefore,
320 we next used TranslationAI to examine the noncanonical ORFs in human
321 transcriptome. Using the stringent cutoff for positive TISs/TTSs (i.e., same cutoff in
322 mRNAs), we were able to identify 4618 noncanonical ORFs defined by new TIS/TTS
323 pairs (Fig. 3A). These newly predicted ORFs included 501 upstream ORFs (uORFs,
324 exemplified in Fig. 3B), 115 downstream ORFs (dORFs, exemplified in Fig. 3B), 208

325 overlapping ORFs with different reading frames from the annotated ORFs (Table S3).
326 In addition, the model also identified 3794 new ORFs from annotated noncoding
327 RNAs (Fig. 3A, Table S4). These results demonstrate the potential for deep learning
328 approaches to uncover previously unrecognized translation events and shed light on
329 new regulatory mechanisms in gene expression.

330 We next validated the prediction by comparing the predicted TISs in new ORFs
331 with all TISs experimentally identified *via* ribosome profiling (Chen *et al.*, 2020).
332 We found that TranslationAI accurately predicted 99.6% of the annotated ORFs,
333 however a smaller fraction of the noncanonical ORFs supported by Ribo-seq were
334 successfully predicted by TranslationAI (39% of uORFs, 37% of dORFs, 24% of dual
335 reading frame, and 43% new ORFs in lncRNAs were identified, Fig. 3C), suggesting
336 that this model is biased towards canonical TISs/TTSs. Since the training set only
337 contain canonical TIS/TTS information, the TranslationAI model may have learned
338 limited features of non-canonical TISs/TTSs (see discussion).

339 We further focused on the thousands of predicted new ORFs from non-coding
340 RNAs (ncRNAs), which are annotated in ENSEMBL as lncRNA, antisense RNA,
341 miRNA precursors, processed transcripts, etc. (Fig. 3D). The shuffled sequences of
342 all lncRNAs were analyzed using the same deep learning model as a background
343 control. We found that the "non-coding" RNAs indeed contains significantly more
344 ORFs than their shuffled counterparts (Fig. 3E), suggesting that some of the predicted
345 ORFs are authentic. While previous reports suggested that some lncRNAs can be
346 translated into small peptides (Chen *et al.*, 2020; Jackson *et al.*, 2018; Matsumoto *et*

347 *al.*, 2017; Yang *et al.*, 2017), detailed analysis of these noncanonical ORFs was
348 inadequate. Our results showed that the newly predicted ORFs in lncRNA are
349 significantly longer (Fig. S4A) and more conserved in 46 vertebrates (Fig. S4B)
350 compared to the control ORFs defined by the longest fragments between ATG and
351 TAA/TAG/TGA in all lncRNAs, suggesting a potential functionality. However, the
352 TISs/TTSs from translatable lncRNAs have lower scores and are less conserved
353 compared to canonical TISs/TTSs from mRNAs (Fig. S4C), implying that the
354 translation efficiency of these newly identified ORFs may be lower than those from
355 annotated mRNAs. This could partially explain why these 'lncRNAs' were
356 originally defined as non-coding RNAs.
357 To validate the prediction of these new ORFs, we re-analyzed the Ribo-seq data
358 from iPSC and cardiomyocytes (Chen *et al.*, 2020). In total 64 newly predicted
359 ORFs by TranslationAI were supported by Ribo-seq signals in either cell line (Table
360 S4). The ribosome density of translatable lncRNAs closely resembled the footprints
361 from control mRNAs, with strong trinucleotide periodicity that is indictive of active
362 translation (Fig. 3F and Fig. S4D). In addition, mass spectrometry (MS)-based
363 proteomics confirmed the stable expression of 191 non-canonical peptides (Table S4,
364 with two examples shown in Fig. 3G). Among them, the ORFs in 10 lncRNAs were
365 supported by both Ribo-seq and MS. The low number of experimentally validated
366 ORFs may be contributed by several factors, including the cell line specific
367 expression of most lncRNAs, the generally low expression level of lncRNAs, and the
368 limited recover rate in both Ribo-seq and MS experiments.

369 Notably, we have correctly predicted ten out of the thirteen translated lncRNAs
370 validated by independent functional studies, including the HOXB-AS3 that was
371 reported to encode a 53aa peptide (Huang et al., 2017), the LINC00961 that encodes a
372 90aa peptide involved in the mTOR activation (Matsumoto *et al.*, 2017), and the
373 RP11-132A1.4 that encodes a functional peptide of 124aa (Chen *et al.*, 2020). All of
374 the three lncRNAs were mis-annotated in RefSeq, probably because their expression
375 and translation are limited to specific cell types. It is also worth noted that the
376 predicted ORFs in annotated lncRNAs generally have low scores in our model (Table
377 S4), which could be attributed to the short ORF length or the presence of a long 5′-
378 UTR. Nevertheless, these results highlight the potential application of AI model in the
379 future study of translatable lncRNA.
380

381 **TranslationAI accurately predicts TISs of other eukaryotes and viruses**

382 An "universal genetic code" is used from bacteria to mammals, suggesting that
383 translation machinery is highly conserved across different organisms. Therefore, we
384 sought to examine if the model trained on human transcriptome can also predict
385 translation sites in other organisms. The accuracies of translation predictions for all
386 eukaryotes tested (Human, Mouse, Zebrafish, Drosophila, Arabidopsis and budding
387 yeast *S. cerevisiae*) were remarkably high (> 90% for all predictions, except that the
388 yeast TIS prediction accuracy is 89%, Fig. 4A), suggesting that our model has learned
389 the essential features generally required for translation initiation and termination in all
390 eukaryotes.

391 We further assess the performance of TranslationAI model trained with human
392 transcripts on the polycistronic transcription units of E. *coli*, which may reflect how
393 much of the coding sequence might be conserved between eukaryotes and
394 prokaryotes. Surprisingly, we obtained a 65% accuracy in *E. coli* with the deep
395 learning model using all parameters trained on the human transcriptome, even though
396 most *E. coli* genes are transcribed as polycistronic units with very short UTRs and
397 different set of genetic codons. Considering the lack of transcriptomic and
398 proteomic information for most bacteria, this model could be useful in predicting their
399 ORFs based solely on their genome. Furthermore, we extended this model on the
400 transcriptomes of *Arabidopsis* chloroplast and human mitochondrion, and the
401 resulting prediction accuracies were ~30% and <1%, respectively (Fig. S5). Given
402 that the translation mechanism and the genetic codon in chloroplast and
403 mitochondrion are drastically different from the human cells, such drop of prediction
404 accuracy is not unreasonable.
405 Finally, we tested whether this model can predict the TISs/TTSs of human viruses
406 that rely on the host cell machinery for protein synthesis, albeit through different
407 translational mechanisms (Stern-Ginossar et al., 2019). Taking two RNA viruses
408 (Ebola and SARS-CoV-2) as examples, we found that the model trained with human
409 data can accurately predict all TISs and TTSs of Ebola virus even when we used the
410 entire genomic RNA sequence as the input (Fig. 4B and Fig. S6A). In addition to
411 the annotated TISs and TTSs, we also predicted a new internal translation start site in
412 the GP gene in Ebola (Fig. 4B), suggesting an additional translation isoform for GP

with truncated N-terminus.

Compared to Ebola virus, TranslationAI achieved a lower prediction accuracy for SARS-CoV-2, with 8 out of 12 annotated TISs and 10 out of 12 annotated TTSs being predicted with high confidence by sub-genomic RNAs (Wang et al., 2021) (Fig. 4C and Fig. S6B). In addition, our model also made several false positive predictions of TISs/TTSs in SARS-CoV-2 genome (Fig. S6B). Similar performance was observed with either genomic or sub-genomic RNAs as input. A notable difference between these two viruses is that the Ebola genome is transcribed into monocistronic mRNAs with both UTRs, whereas the SARS-CoV-2 mRNAs are mainly polycistronic genomic or subgenomic RNAs with short spacer sequences between each ORF (Banerjee et al., 2020; Wang *et al.*, 2021). This may partially explain the different performances of TranslationAI for different viruses.

Taken together, our results suggest that the model containing parameters trained on human transcriptome can accurately predict ORFs in a wide range of organisms. However, in certain cases where the translation regulation is drastically different (e.g., short UTRs or different codon system), the application of this model should be carried out with caution.

Discussion

As one of the most effective and powerful machine learning methods, the deep learning network has been widely used to study complex biological questions ranging from population genetics to precision medicine (Bogard *et al.*, 2019; Eraslan *et al.*, 2019; Vamathevan *et al.*, 2019; Zhou *et al.*, 2018). Here we developed a deep learning architecture, TranslationAI, for *ab initio* prediction and systematic analysis of translation initiation and termination in entire transcriptome. Remarkably, the model trained with ~70% of human transcripts achieved a near-perfect prediction accuracy in not only human, but also a wide range of eukaryotic organisms. The model can also identify new ORFs in annotated noncoding RNAs, some of which were confirmed by independent Ribo-seq and mass spectrometry experiments. Further analysis of this predictive model also identified a new regulation rule for translation termination, providing mechanistic insights into the complex regulation of RNA translation.

The remarkable accuracy of TranslationAI suggests that this “black box” model have learned useful information with biological relevance. We found that the model has successfully learned the identity of canonical start and stop codons (Fig. 1F), which account for the vast majority (99.4% start codon and 99.6% stop codon) of the training set. The model did not identify any TIS/TTS with noncanonical start or stop (i.e, non-AUG start or non-UAG/UAA/UGA stop), probably because they are extremely rare in the training data. Additional features, including triplet reading frameshift, 5′-UTR sequence, and amino acid composition, also contribute to the prediction accuracy. However, all these features contributed partially to the model in terms of TIS/TTS prediction, as deleting these features did not completely invalidate

454 the prediction (Fig. 1G-I).

455 Surprisingly, several known features associated with translation regulation,
456 including the 3′-UTR sequence, synonymous mutation, and sequences of translation
457 product, showed small effect on the accuracy of TIS/TTS prediction, implying that
458 they may not play a dominant role in selection of translation sites. More
459 importantly, we found several new features can influence the prediction of TIS/TTS,
460 such as the sequence motif around the strong TTS, suggesting additional determinants
461 for translation termination. These features may open up new avenues for
462 investigating the molecular mechanisms underlying translation control. Future
463 exploration of the predictive features learned by the AI model may provide
464 mechanistic insights into the functional interactions between the mRNA and
465 translation machine.

466 We also used this model to assess the single nucleotide variants (SNVs) that may
467 alter the TISs/TTSs and cause functional consequences. Surprisingly, we did not
468 identify any disease-associated SNVs in the ORF that can alter the TIS/TTS
469 prediction, unless the mutation directly changed the start or stop codon. This is
470 consistent with the finding that the synonymous/non-synonymous substitutions of
471 amino acids had little effect on TIS/TTS prediction (Fig. 1I). These results suggest
472 that small changes in coding sequence do not significantly impact ORF selection by
473 the translation machinery, *i.e.,* the translation initiation and termination appear to be
474 unaffected by the elongation steps. Alternatively, it is also possible that the model
475 does not rely on the nucleotide composition within ORFs for its prediction.

476 This model also allowed us to identify novel regulatory motifs associated with
477 TIS and TTS. We discovered, for the first time, that the strong stop codons are
478 surrounded by novel motifs that are generally CG-rich, especially in the third position
479 of the codons immediately before the stop codon. Consistently, there is a codon bias
480 associated with the strong termination site (Fig. 2E), and the synonymous mutations
481 with low CG-contents will introduce translation readthrough in a reporter assay (Fig.
482 2F). Because the GC-rich codons were found to be associated with slow translation
483 (Tunney et al., 2018), we speculate that the strength of the stop codon is associated
484 with the decreased translation elongation rate prior to reaching the stop codon. This
485 regulation may be attributed to the changes in the strength of ribosome's interaction
486 with the mRNA before the ribosomes reach the termination site, which “decelerate”
487 the translation before an efficient termination. Alternatively, the motif could alter the
488 ribosome’s conformation, thereby influencing translation termination. This finding
489 suggests that the translation regulation is more intricate than previously thought, and
490 further exploration on this phenomenon may shed light on the mechanism of
491 translation termination and readthrough.
492 The deep learning network trained with human transcripts achieved a surprisingly
493 high accuracy for the prediction of ORFs in bacteria, yeast, plants and certain viruses
494 using only their genome sequences as input (Fig. 4). This result is particularly
495 interesting given that some of these organisms have polycistronic transcription units
496 or genomes, suggesting that the TranslationAI can predict not only uORFs and dORFs
497 in human, but also ORFs in polycistronic sequences from other organisms. However

498 this model failed to predict mitochondrial ORFs and showed mediocre performance
499 for SARS-CoV-2, probably due to the presence of frameshift regions, overlapping
500 ORFs, and a lack of sufficient intergenic sequences to separate different ORFs.
501 These results suggest that additional training may be required to improve the model's
502 performance on more complex or noncanonical genomes.
503 Despite its high accuracy in predicting canonical TISs, TranslationAI has some
504 limitations. First, our model currently struggles in predicting non-canonical TISs,
505 especially for uTISs with non-AUG start codons. This is probably due to the bias in
506 the training dataset, which mainly consists of mono-cistronic translation units with
507 AUG start codon. We attempted to train this model using a subset of mRNAs with
508 multiple annotated ORFs, however the accuracy (~75%) was not as high as that of
509 canonical transcripts, possibly due to the small size of the training data. Furthermore,
510 although this model can predict the strengths of different TISs in the presence of
511 multiple ORFs, it is still challenging to predict which TIS will be used in specific cell
512 lines or tissues. Expanding the ribosome profiling data in different cell types may
513 help refine the tissue-specific translation prediction model. Finally, the model uses a
514 stringent threshold under the assumption that there is a dominant ORF in each
515 transcript. While this assumption holds well for most transcripts, it may be incorrect
516 in some cases. With more non-canonical ORFs being discovered, the score
517 threshold may need to be modified to better fit different type of transcripts. In
518 summary, with increasing translation datasets and more refined parameters, we
519 believe that the TranslationAI framework will improve ORF prediction and deepen

our understanding of the complexity in translation control.

522 **Methods**

523 **Architecture of the TranslationAI deep learning model**

524 We constructed a deep residual neural network using one-hot encoding to
525 represent mRNA sequences as input (Fig. 1A). Our approach employed a 32-layer
526 dilated convolutional neural network architecture (Fig. S1) to generate an output
527 matrix representing the probabilities of a given position being a TIS, TTS, or neither
528 (NS).

529 Four models were developed with distinct architectures, incorporating 60, 200,
530 600, or 2,000 nucleotides on both sides of a position of interest within the full-length
531 mRNA sequence as input (Fig. S1). Each model produces probability scores for TIS,
532 TTS, and NS, with the sum of these probabilities equaling one.

533 The TranslationAI consists of several stacked residual blocks that connect the
534 input layer to the penultimate layer, a framework previously developed for image
535 recognition (He et al., 2016). A convolutional unit with softmax activation links the
536 penultimate layer to the output layer. To enhance convergence speed during training,
537 the output of every fourth residual block is added to the input of the penultimate layer
538 (Fig. S1).

539 The fundamental unit of the TranslationAI model is a residual block, which
540 contains two batch-normalization layers, two rectified linear units (ReLU), and two
541 convolutional units arranged in a specific sequence (Fig. S1). Each residual block
542 includes three hyper-parameters: N, W, and D. N represents the number of
543 convolutional kernels, W signifies the window size, and D indicates the dilation rate

of each convolutional kernel. Given that a convolutional kernel with window size W and dilation rate D processes features across (W-1)×D neighboring positions, a residual block with two convolutional units can handle features across 2(W-1)×D neighboring positions. As a result, a TranslationAI model with K stacked residual blocks is capable of extracting features spanning $\sum_{i=1}^{K} 2(W_i - 1)D_i$ neighboring positions, where N_i, W_i, and D_i correspond to the hyper-parameters of the i-th residual block.

Model training and testing

To obtain a comprehensive, well-curated, and non-redundant set of sequences, we utilized the RefSeq gene annotation and extracted 47,098 protein-coding transcripts with 47,098 TIS-TTS pairs. Different splicing isoforms of the same gene are treated as different transcripts. For model training, we selected transcripts located on chromosomes 2, 4, 6, 8, 10-22, X, and Y, which contain 13,707 genes with 34,292 TIS-TTS pairs. For model evaluation, the transcripts from chromosomes 1, 3, 5, 7, and 9, containing 5,524 genes with 12,806 TIS-TTS pairs, were used. To mitigate overfitting, 10% of the training set were randomly chosen for early-stopping determination during training, while the remainder were used for model training.

We employed a sequence-to-sequence approach with a chunk size of 6,000 for model training and testing. The full-length mature mRNA sequences were one-hot encoded, with A, C, G, and U mapped to [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], and [0, 0, 0, 1], respectively. We then zero-padded the one-hot encoded nucleotide sequences

until their lengths were multiples of 6,000 and further padded them with flanking sequences of length S/2 at the beginning and end, where S corresponds to 60, 200, 600, or 2,000 for the TranslationAI-60, TranslationAI-200, TranslationAI-600, and TranslationAI-2k models, respectively. The padded nucleotide sequences were divided into blocks of length S/2 + 6,000 + S/2, with the i-th block consisting of nucleotide positions from 6,000(i-1) – S/2 + 1 to 6,000i + S/2. In parallel, the output label sequences were one-hot encoded as TIS ([0, 1, 0]), TTS ([0, 0, 1]), or neither ([1, 0, 0]). The one-hot encoded TIS/TTS output label sequences were also zero-padded until their lengths reached multiples of 6,000, and then split into 6,000-length blocks, with the i-th block containing positions from 6,000(i-1) + 1 to 6,000i. These one-hot encoded nucleotide sequences and their corresponding label sequences served as the inputs and outputs for the models, respectively.

The models were trained for 10 epochs using a batch size of 18 on two NVIDIA Tesla K80 GPUs. During training, the categorical cross-entropy loss between target and predicted outputs was minimized using the Adam optimizer. The optimizer's learning rate was set to 0.001 for the first six epochs and halved in each subsequent epoch. For each model, the training procedure was repeated 10 times, and the top five models based on performance were selected. During testing, each input was assessed using all five trained models, and the average of their outputs was utilized as the predicted output. This rigorous training and testing process aimed to ensure optimal performance and reliability in the resulting models.

The performance of the TIS/TTS prediction models was measured with top-k

accuracy. This metric measures the fraction of correctly predicted positions among the top-k predicted sites, where k represents the number of true TIS/TTS sites in the dataset. For example, if the dataset contains k1 TISs and k2 TTSs, we set the score cutoff to predict exactly the top-k1 and top-k2 positions as TIS and TTS sites, respectively. It is worth noting that, unless specifically mentioned otherwise, only the top-k predicted sites were used for downstream analysis in our study.

Role of transcript lengths in mRNA translation

To study the effect of transcript length on translation sites prediction, we sorted the mRNAs in order of increasing length. Using the TranslationAI-200 and TranslationAI-2k models, we calculated the positive rate of TIS/TTS for each transcript by comparing the number of positive TISs to the total number of transcripts with lengths shorter than the given transcript. We plotted these positive rates against the transcript length, as shown in Fig. 1C. Additionally, we included a histogram of transcript lengths from the 47,098 transcripts analyzed in the background of the figure.

Codon distribution at upstream of stop codon

To examine the codon distribution upstream of stop codon, we quantified the frequency of each codon at the -30nt position relative to both strong and weak stop codons. Specifically, we calculated the frequency of each codon as the number of occurrences of that codon at the -30nt position divided by the total number of RNAs

610 analyzed at that position.

611

612 **Conservation analysis**

613 To evaluate the conservation levels of different RNAs, we obtained the
614 PhastCons46 scores from the PhastCons46 or PhyloP46 track in the UCSC Genome
615 Browser. These scores were generated from genome alignments of 46 vertebrates,
616 including human (hg19), and provide a measure of evolutionary conservation across
617 species. Each box plot displays the median as the central line, while the upper and
618 lower edges of the box represent the first and third quartiles, respectively.

619

620 **Non-canonical TISs/TTSs identification**

621 To systematically identify non-canonical TISs and TTSs in transcriptomes, we
622 input human transcriptome into TranslationAI model, and retained only those RNAs
623 that met the threshold criteria (prediction score > 0.5) for both TIS and TTS as
624 translatable RNAs. The ncRNA dataset were retrieved from Ensembl build 75, which
625 was previously used in the published study (Chen *et al.*, 2020) that identified
626 hundreds of noncanonical ORFs using ribosome profiling and mass-spectrometry.

627 To compare the TISs predicted by TranslationAI with TISs identified by ribosome
628 profiling, we expanded our predicted TIS library by retaining the top 10 TISs for each
629 gene predicted by the TranslationAI. This approach was necessary as ribosome
630 profiling often identifies multiple ORFs from one gene.

631 The alternative TISs were defined as the TIS with the highest score (>0.1),

632 excluding the annotated TIS. In addition, in-frame alternative TISs were defined as
633 TISs where the distance between the annotated TIS and alternative TIS is a multiple
634 of three.
635
636 **Ribosome footprint data analysis**
637 The ribosome profiling data analysis was performed on a previously published
638 ribosome footprint data that were retrieved from no drug treatment iPSC and iPSC-
639 induced iPSC-derived cardiomyocytes cells being deposited in GEO database under
640 accession number GSE131650. The genome assembly used throughout this
641 manuscript is hg19/GRCh37 annotated by GENCODE v27. We first filtered out reads
642 aligning to rRNAs using Bowtie v1.0.0 and aligned the remaining reads to the
643 annotated transcriptome with STAR v2.7.5a, using the --outSJfilterReads Unique --
644 outFilterMultimapNmax 1 --outFilterMismatchNmax 999 --
645 outFilterMismatchNoverReadLmax 0.04 to filter multiple aligned and redundant
646 reads. These alignments were assigned a specific P-site nucleotide using a 12-nt offset
647 from the 3′ end of reads. Metagene plots were generated by normalizing read density
648 around the start and stop codons of each gene (i.e, read depth at each position divided
649 by the average reads of each position in one transcript and the number of the covered
650 transcripts).
651 To estimate the correlation between predicted TIS/TTS scores and mRNA
652 translation efficiency, we ranked and divided the scores into percentiles (n~1580
653 transcripts/percentile, Fig. S2G-S2H). We calculated the translation efficiency of

654 transcripts from ribosome profiling data in two human cell lines as the ratio of Ribo-
655 seq read density to RNA-seq read density.
656
657 **Motif identification of strong and weak TISs/TTSs**
658 Motif discovery and analysis were performed using the Meme suite (v5.0.5). The
659 strong and weak TIS/TTS containing sequences (-30nt : + 33nt) were collected as
660 primary sequences for motif prediction. Additionally, all TIS/TTS containing
661 sequences (-30nt : +33nt) were used as control sequences. Meme was employed to
662 identify enriched motifs, with a maximum length of 63nt, that occur any number of
663 times within the given sequence.
664
665 **MS Proteomics and HLA Peptidomics data analysis**
666 MS-based proteomics data are deposited to the ProteomeXchange Consortium via
667 the Proteomics Identifications Database (PRIDE) partner repository with the dataset
668 identifier PXD014031 and PXD000394. Using an open search engine pFind (v3.1.3),
669 we searched the two previously published human proteome datasets against a combined
670 database (25692 proteins) containing all UniProt human proteins and the potential non-
671 canonical ORFs-coded peptides (uORFs, dORFs, dual frame ORFs, and new ORFs).
672 We selected positive mass spectra using following thresholds: $q < 0.01$, peptides length
673 $\geqslant 8$, missed cleavage sites $\leqslant 3$, allowing only common modifications (cysteine
674 carbamidomethylation, oxidation of methionine, protein N-terminal acetylation, pyro-
675 glutamate formation from glutamine, and phosphorylation of serine, threonine, and

676 tyrosine residues).

677 The resulting peptides were searched against non-redundant human protein
678 database using blastp-short and the peptides with less than two mismatches from known
679 proteins were removed.

680

681 **Transcriptomes of other species**

682 The genome sequence and gene annotations of *Mus musculus* (mm10), *Danio*
683 *rerio* (danRer11), *Drosophila melanogaster* (dm6), Ebola virus (eboVir3), SARS-
684 CoV-2 (wuhCor1), and human mitochondrion are obtained from UCSC database. The
685 sub-genomic sequences of SARS-CoV2 were obtained from a published study (Wang
686 *et al.*, 2021). The genome sequence and gene annotation of Arabidopsis thaliana
687 (TAIR10) are obtained from TAIR database. Regarding E. coli (Escherichia coli K-12
688 substr. MG1655), its transcription unit is retrieved from EcoCyc. The mRNA
689 information (5′-UTR, CDS, and 3′-UTR) for yeast was obtained from *Saccharomyces*
690 Genome Database (SGD). The genome sequence and gene annotation of Chloroplast
691 (Arabidopsis thaliana) was obtained from NCBI database. The genome sequence of
692 human mitochondria was obtained from NCBI database and its gene annotation file
693 was obtained from GENCODE database.

694

695 **Plasmid construction and Transfection**

696 To construct the reporter (pcDNA5-HA-GFP-stop_motif-Nluc-Flag) for testing
697 the translation readthrough products, the Fluc-Nluc luciferase sequence (a gift from

698 prof. Rachel Green) was cloned into pcDNA5 backbone digested by SpeI and XbaI
699 (Wangen and Green, 2020). The GFP sequence was amplified (fused with two BsmBI
700 restriction fragments) and inserted before Nluc. The modified VDR sequences were
701 synthesized (GENEWIZ) and cloned into BsmBI digested vector.
702 HEK293T cell line was cultured in DMEM (high glucose) medium containing
703 10% fetal bovine serum (FBS, Hyclone). To transient transfect plasmids into cells,
704 2 μg of reporters were transfected into cells in 6 well plate using lipofectamine 3000
705 (Invitrogen) according to the manufacturer's instruction. After 48 h, cells were
706 collected for further analysis of protein level.
707
708 **Western blot**
709 Cells were lysed in laemmli buffer, and the total cell lysates were resolved with 4-
710 20% ExpressPlus™ PAGE Gel (GeneScript). The following antibodies were used:
711 HA-Tag antibody (CST: 3724S) was diluted by 1:2000, V5 antibody (CST: 13202S)
712 was diluted by 1:2000, Flag antibody (Sigma: F1804-1MG) was diluted by 1:2000.
713 The HRP-linked secondary antibodies (CST: 7076S) were used by 1:4000 dilution
714 and the blots were visualized with the ECL reagents (Bio-Rad).
715
716 **Data and software availability**
717 Training and testing data, prediction scores for all possible single nucleotide
718 substitutions in the reference genome and source code are publicly hosted at GitHub
719 (https://github.com/rnasys/TranslationAI). Prediction scores and source code are
720 publicly released under GPL v3 and are free for use for academic and non-commercial

applications. Raw data from Figures 2G and S3E were deposited on Mendeley at https://data.mendeley.com/v1/datasets/rsssc3hnc4/draft?preview=1.

Acknowledgments

We thank Dr. Rachel Green for kindly sharing the plasmids of translation read-through. This work is supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38040100 to Z.W.), the National Key Research and Development Program of China (2021YFA1300503 and 2018YFA0107602 to Z.W.), the National Natural Science Foundation of China (32030064, 32250013, 91940303 and 31730110 to Z.W.; 32100430 to X.F.), the Starry Night Science Fund at Shanghai Institute for Advanced Study of Zhejiang University (SN-ZJU-SIAS-009) and the NIH Basic Research in Cancer Health Disparities Award to S.R.P. and J.A.F. (R01CA220314). Z.W. is also supported by CAS Interdisciplinary Innovation Team and the type A CAS Pioneer 100-Talent program.

Author Contributions

Conceptualization, Z.W., X.F. and T.C.; Methodology, X.F., T.C., and Z.W.; Software, X.F. and T.C.; Experiments, C.C.; Writing and discussion, X.F., Z.W., T.C., and M.H.; Funding acquisition, X.F., and Z.W.

Declaration of Interests

The authors declare no competing financial interests.

References

Banerjee, A.K., Blanco, M.R., Bruce, E.A., Honson, D.D., Chen, L.M., Chow, A., Bhat, P., Ollikainen, N., Quinodoz, S.A., Loney, C., et al. (2020). SARS-CoV-2 Disrupts Splicing, Translation, and Protein Trafficking to Suppress Host Defenses. Cell *183*, 1325-1339 e1321. 10.1016/j.cell.2020.10.004.

Bogard, N., Linder, J., Rosenberg, A.B., and Seelig, G. (2019). A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. Cell *178*, 91-106 e123. 10.1016/j.cell.2019.04.046.

Chen, J., Brunner, A.D., Cogan, J.Z., Nunez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., and Weissman, J.S. (2020). Pervasive functional translation of noncanonical human open reading frames. Science *367*, 1140-1146. 10.1126/science.aay0262.

Chen, W., Feng, P.M., Deng, E.Z., Lin, H., and Chou, K.C. (2014). iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem *462*, 76-83. 10.1016/j.ab.2014.06.022.

de Klerk, E., and t Hoen, P.A. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends Genet *31*, 128-139. 10.1016/j.tig.2015.01.001.

Eraslan, G., Avsec, Z., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nat Rev Genet *20*, 389-403. 10.1038/s41576-019-0122-6.

Fields, A.P., Rodriguez, E.H., Jovanovic, M., Stern-Ginossar, N., Haas, B.J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S.A., Ingolia, N.T., et al. (2015). A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Mol Cell *60*, 816-827. 10.1016/j.molcel.2015.11.013.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res *47*, D766-D773. 10.1093/nar/gky955.

Gao, X., Wan, J., Liu, B., Ma, M., Shen, B., and Qian, S.B. (2015). Quantitative profiling of initiating ribosomes in vivo. Nat Methods *12*, 147-153. 10.1038/nmeth.3208.

Goel, N., Singh, S., and Aseri, T.C. (2020). Global sequence features based translation initiation site prediction in human genomic sequences. Heliyon *6*, e04825. 10.1016/j.heliyon.2020.e04825.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778. https://doi.org/10.1109/CVPR.2016.90.

Hinnebusch, A.G. (2011). Molecular mechanism of scanning and start codon selection in eukaryotes. Microbiol Mol Biol Rev *75*, 434-467, first page of table of contents. 10.1128/MMBR.00008-11.

Huang, J.Z., Chen, M., Chen, D., Gao, X.C., Zhu, S., Huang, H., Hu, M., Zhu, H., and Yan, G.R. (2017). A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. Mol Cell *68*, 171-184 e176. 10.1016/j.molcel.2017.09.015.

Jackson, R., Kroehling, L., Khitun, A., Bailis, W., Jarret, A., York, A.G., Khan, O.M., Brewer, J.R., Skadow, M.H., Duizer, C., et al. (2018). The translation of non-canonical open reading frames controls mucosal immunity. Nature *564*, 434-438. 10.1038/s41586-018-0794-7.

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535-548 e524. 10.1016/j.cell.2018.12.015.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold.

Nature *596*, 583-589. 10.1038/s41586-021-03819-2.
Kalkatawi, M., Magana-Mora, A., Jankovic, B., and Bajic, V.B. (2019). DeepGSR: an optimized deep-learning structure for the recognition of genomic signals and regions. Bioinformatics *35*, 1125-1132. 10.1093/bioinformatics/bty752.
Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell *44*, 283-292. 10.1016/0092-8674(86)90762-2.
Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res *15*, 8125-8148. 10.1093/nar/15.20.8125.
Lee, S., Liu, B., Lee, S., Huang, S.X., Shen, B., and Qian, S.B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A *109*, E2424-2432. 10.1073/pnas.1207846109.
Leppek, K., Das, R., and Barna, M. (2018). Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat Rev Mol Cell Biol *19*, 158-174. 10.1038/nrm.2017.103.
Loughran, G., Jungreis, I., Tzani, I., Power, M., Dmitriev, R.I., Ivanov, I.P., Kellis, M., and Atkins, J.F. (2018). Stop codon readthrough generates a C-terminally extended variant of the human vitamin D receptor with reduced calcitriol response. J Biol Chem *293*, 4434-4444. 10.1074/jbc.M117.818526.
Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A., Nakayama, K.I., Clohessy, J.G., and Pandolfi, P.P. (2017). mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. Nature *541*, 228-232. 10.1038/nature21034.
Mayr, C. (2019). What Are 3' UTRs Doing? Cold Spring Harb Perspect Biol *11*. 10.1101/cshperspect.a034728.
Mouilleron, H., Delcourt, V., and Roucou, X. (2016). Death of a dogma: eukaryotic mRNAs can code for more than one protein. Nucleic Acids Res *44*, 14-23. 10.1093/nar/gkv1218.
Raj, A., Wang, S.H., Shim, H., Harpak, A., Li, Y.I., Engelmann, B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. Elife *5*. 10.7554/eLife.13328.
Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell *136*, 731-745. 10.1016/j.cell.2009.01.042.
Stern-Ginossar, N., Thompson, S.R., Mathews, M.B., and Mohr, I. (2019). Translational Control in Virus-Infected Cells. Cold Spring Harb Perspect Biol *11*. 10.1101/cshperspect.a033001.
Tunney, R., McGlincy, N.J., Graham, M.E., Naddaf, N., Pachter, L., and Lareau, L.F. (2018). Accurate design of translational output by a neural network model of ribosome distribution. Nat Struct Mol Biol *25*, 577-582. 10.1038/s41594-018-0080-2.
Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nat Rev Drug Discov *18*, 463-477. 10.1038/s41573-019-0024-5.
Vislovukh, A., Vargas, T.R., Polesskaya, A., and Groisman, I. (2014). Role of 3'-untranslated region translational control in cancer development, diagnostics and treatment. World J Biol Chem *5*, 40-57. 10.4331/wjbc.v5.i1.40.
Wang, D., Jiang, A., Feng, J., Li, G., Guo, D., Sajid, M., Wu, K., Zhang, Q., Ponty, Y., Will, S., et al. (2021). The SARS-CoV-2 subgenome landscape and its novel regulatory features. Mol Cell *81*, 2135-2147 e2135. 10.1016/j.molcel.2021.02.036.
Wangen, J.R., and Green, R. (2020). Stop codon context influences genome-wide stimulation of termination codon readthrough by aminoglycosides. Elife *9*. 10.7554/eLife.52611.

Yang, Y., Fan, X., Mao, M., Song, X., Wu, P., Zhang, Y., Jin, Y., Yang, Y., Chen, L.L., Wang, Y., et al. (2017). Extensive translation of circular RNAs driven by N(6)-methyladenosine. Cell Res *27*, 626-641. 10.1038/cr.2017.31.

Zhang, S., Hu, H., Jiang, T., Zhang, L., and Zeng, J. (2017). TITER: predicting translation initiation sites by deep learning. Bioinformatics *33*, i234-i242. 10.1093/bioinformatics/btx247.

Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet *50*, 1171-1179. 10.1038/s41588-018-0160-6.

Figures and legends

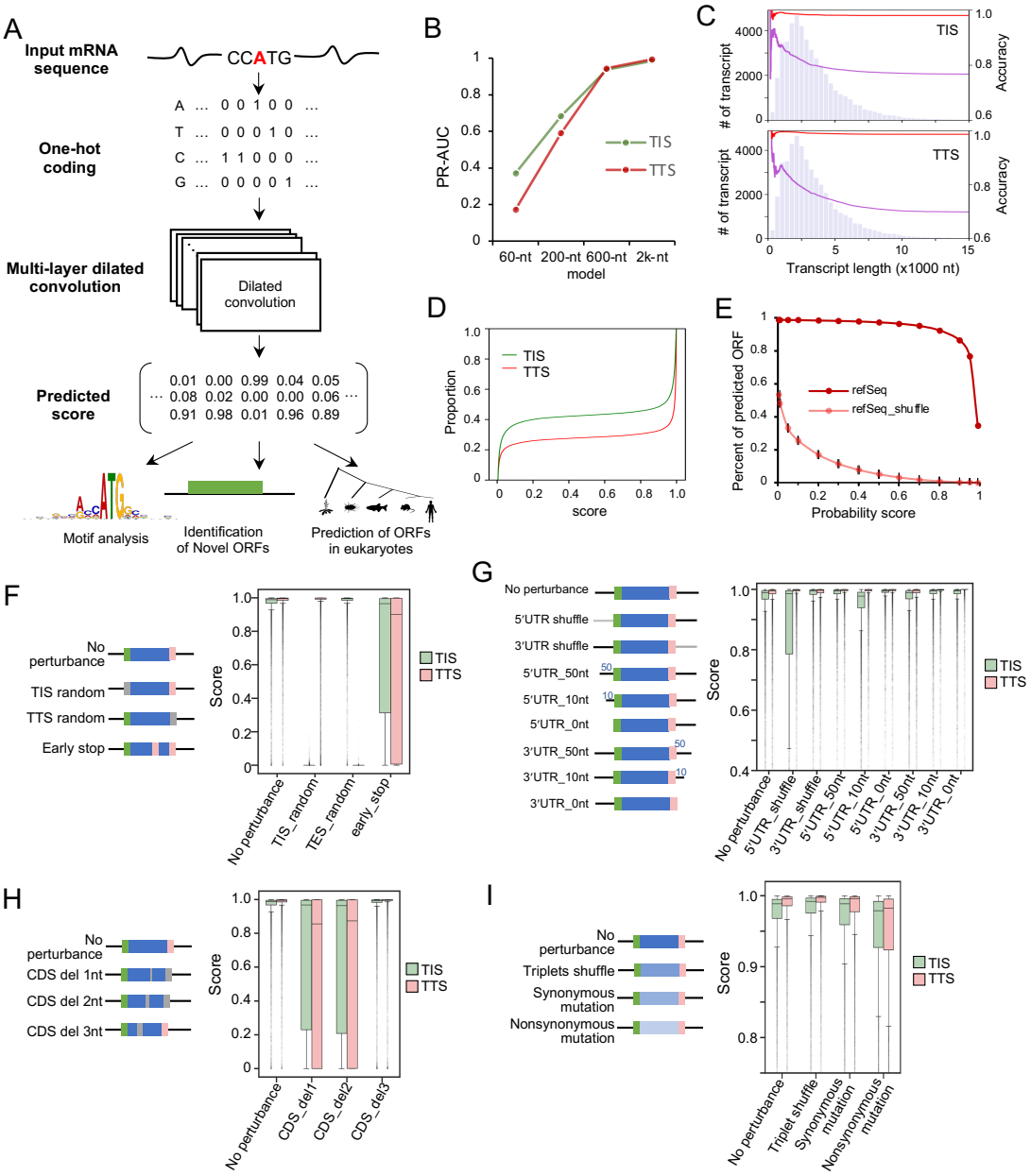

Figure 1. Construction of deep learning network for translation prediction. A. Flowchart of TranslationAI, a computational model for predicting translation initiation and termination sites with full length mRNA. For each position in the full length mRNA, TranslationAI-2k takes 60, 200, 600, and 2,000 nucleotides of flanking sequence as input, and predicts whether that position corresponds to a translation initiation site (TIS), translation termination site (TTS), or neither. **B.** Effect of input sequence context size on network accuracy. PR-AUC is the area under the precision-recall curve. The figure shows the performance of the network at four different input sequence context sizes. **C.**

Relationship between transcript length and the positive rate of TIS/TTS for each transcript by comparing the number of positive TISs from TranslationAI-200 (purple line) or TranslationAI-2k (red line) to the total number of transcripts with lengths shorter than the given transcript. The distribution of transcript length is shown in the background as a histogram. **D.** Accumulative distribution of TIS/TTS score among all mRNAs. **E**. The number of ORFs predicted by TranslationAI-2k using mRNA and shuffled mRNA as input and different cutoff of TIS scores. The average number of ORFs predicted from multiple shuffling, along with error bars representing three times the standard deviation calculated from shuffled sequences. **F-I.** The features learned by TranslationAI. Systematic *in silico* perturbations on different regions of the mRNA, measured as reductions in probability scores for the authentic TISs/TTSs. The perturbations include: **F**, replacement of TIS/TTS identity; **G**, changes in UTR length and sequence; **H**, frameshifts by deleting one, two, or three nucleotides; **I**, mutations in coding sequences. Blue box: ORF (open reading frame), black line: 5′-UTR and 3′-UTR, green line: TIS, pink line: TTS.

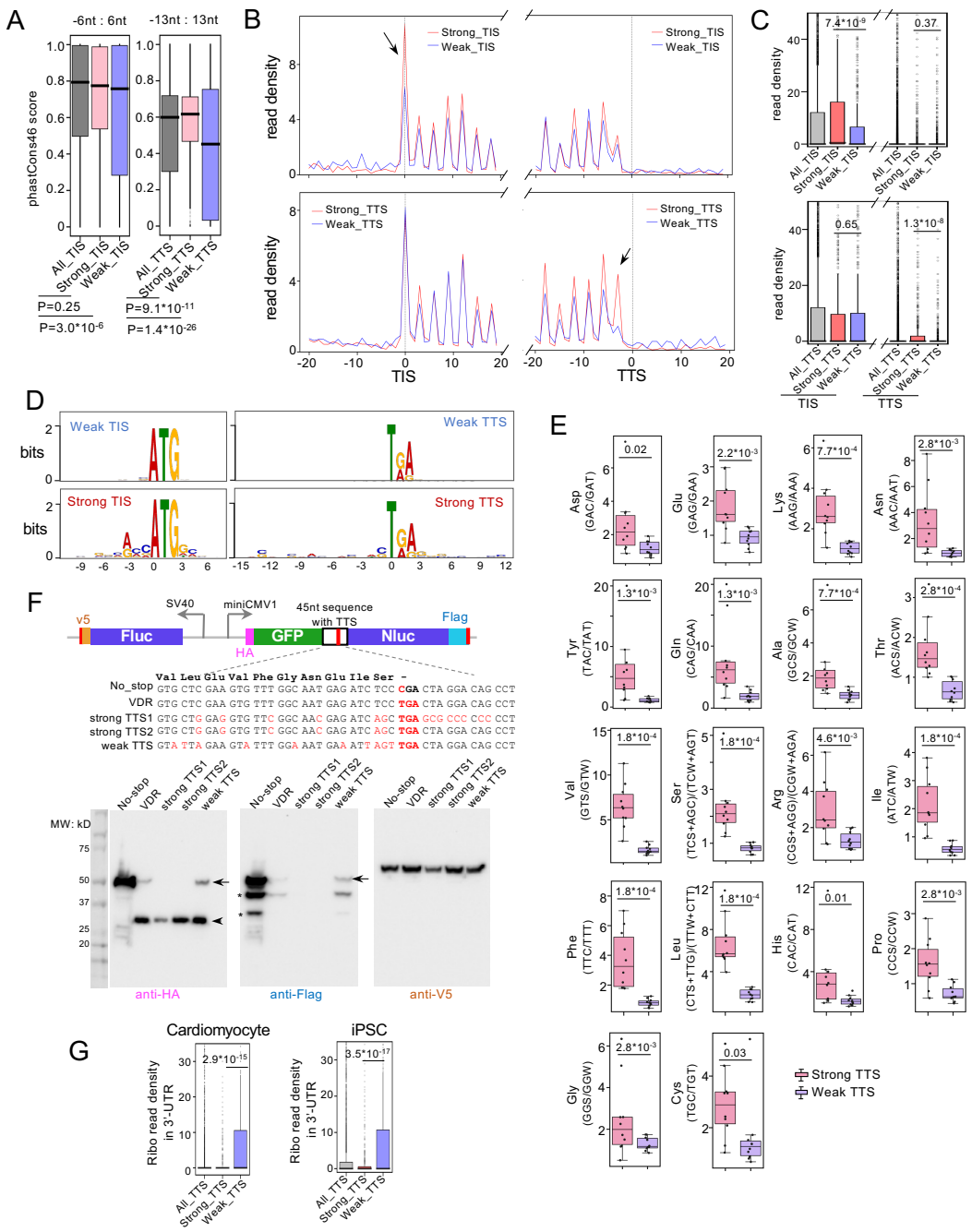

Figure 2. Evaluation of predictive features in TranslationAI. A. Conservation analysis of all TIS/TTS, strong TIS/TTS, and weak TIS/TTS regions. P-values were calculated using Mann–Whitney U test. **B.** Metagene analysis and **C.** boxplot of reads density on transcripts containing strong (red) and weak (blue) TIS/TTS (-20nt : +20nt) in iPSC cell line. P-values were calculated using Mann–Whitney U test. **D.** Motif analysis of strong and weak TIS/TTS. **E.** The codon distribution at the -30nt position of stop codon. The ratios of the same amino acid with C/G and A/T at the third position of each codon before strong and weak stop codons was quantified. P-values were

879 calculated using standard T-test. **F.** Validation of motifs around strong TTS. Known
880 readthrough stop codon from VDR and their mutations (without change of amino acids)
881 were tested for translation readthrough. No_stop: stop codon of VDR (TGA) was
882 mutated to CGA; strong TTS1: stop codon of VDR was mutated to a strong TTS by
883 changing the upstream 27nt and downstream 12nt sequences; strong TTS2: stop codon
884 of VDR was mutated to a strong TTS by changing the upstream 27nt sequence; weak
885 TTS: stop codon of VDR was mutated to a weak TTS by changing the upstream 27nt.
886 See supplementary information in Table S2. **G.** Boxplot analysis of reads density on 3′-
887 UTRs from transcripts with strong/weak TTS in iPSC cell line and iPSC-induced
888 Cardiomyocyte cell line, respectively.
889

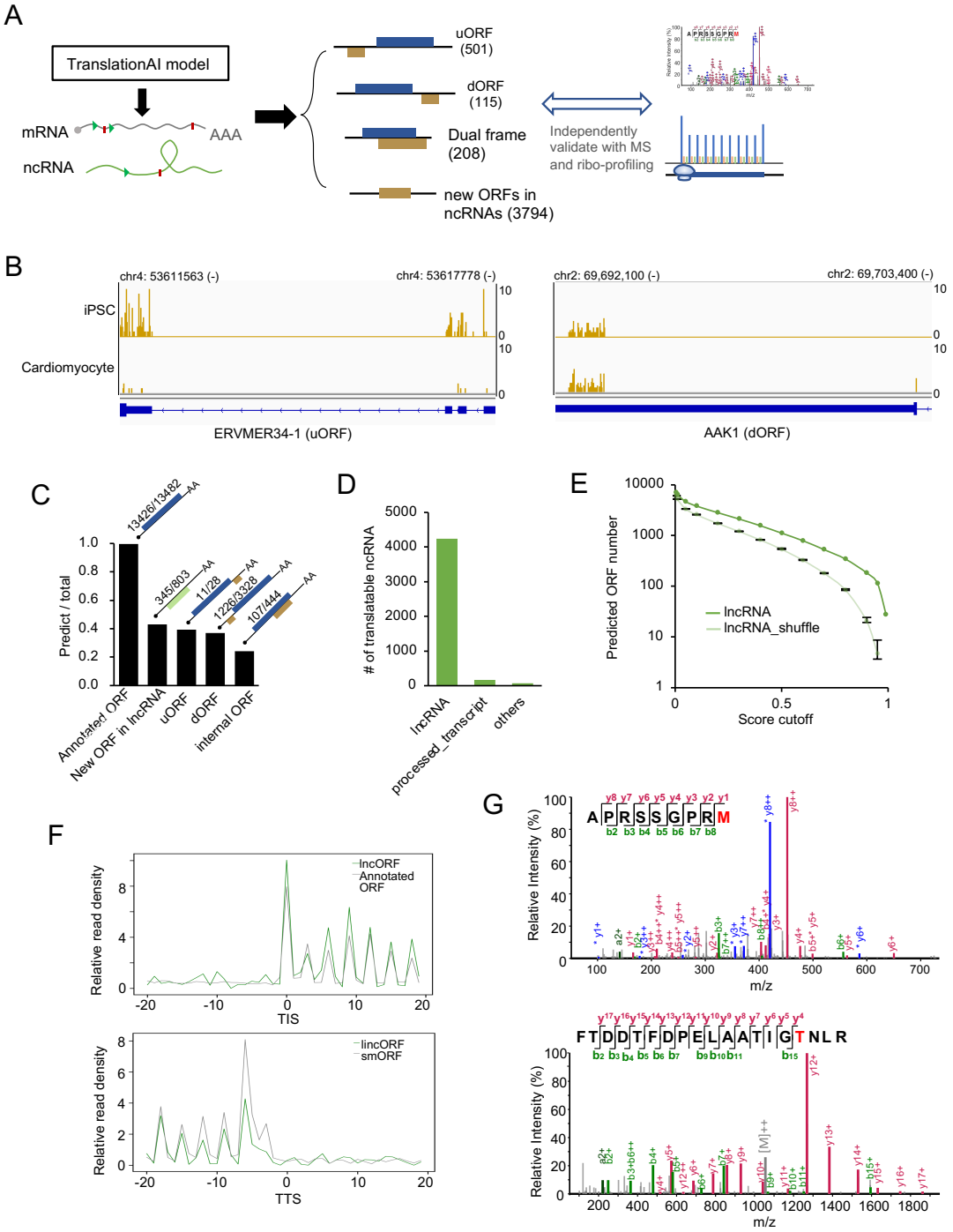

Figure 3. Identification of non-canonical ORFs in human transcriptome. A. Flowchart for identifying non-canonical ORFs, including upstream ORFs, downstream ORFs, dual coding ORFs, and new ORFs from non-coding RNAs. **B.** Example ribosome footprints of a uORF from ERVMER-1 and dORF from AAK1. **C.** The ratio of various types of predicted TISs identified by another published dataset derived from ribosome profiling assays. **D.** The number of newly identified translatable ncRNA in

annotated lncRNAs, processed transcripts and other transcripts without known ORFs. **E.** The number of ORFs predicted by TranslationAI-2k using non-coding RNA and shuffled non-coding RNA sequences as input. The average numbers of ORFs predicted from lncRNAs and shuffled lncRNAs (as control) were shown, with error bars representing 3×standard deviation calculated from shuffled sequences. **F.** Metagene analysis of translatable lncRNAs (green line) and control (grey line, mRNAs with the same ORF length distribution of predicted ORFs from lncRNAs). **G.** The MS/MS spectra of peptides from two ncRNAs: lncRNA ENST00000609975 (APRSSGPRM) and antisense RNA ENST00000413405 (FTDDTFDPELAATIGTNLR). The annotated b- and y-ions are marked in red and green, respectively.

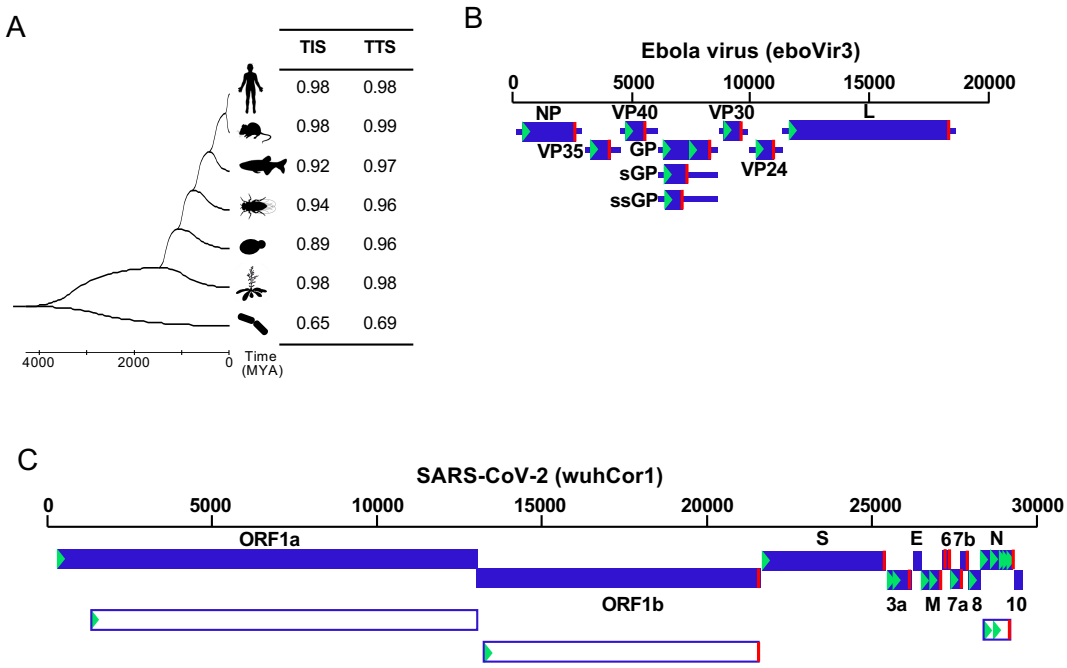

Figure 4. TranslationAI accurately predicts TIS/TTS of eukaryotes, prokaryotes, and viruses. A. The prediction accuracy of TranslationAI-2k applied to eukaryotes (Human, Mouse, Zebrafish, Drosophila, Arabidopsis and budding yeast *S. cerevisiae*) and prokaryotes (*E.* coli). Prediction of TISs/TTSs via **B.** Ebola genomic or RNA sequences and **C.** SARS-CoV-2 genomic or subgenomic sequences. The upper scales represent the genomic sequence position, the blue boxes indicate the annotated ORFs, and the white boxes indicate the newly predicted out-of-frame ORFs. The green triangles and red lines indicate the predicted in-frame TISs and TTSs, respectively.