PAL-AI reveals genetic determinants that control poly(A)-tail length during oocyte maturation, with relevance to human fertility

Kehui Xiang[1,2,3*], David P. Bartel[1,2,3*]

[1]Howard Hughes Medical Institute, Cambridge, MA 02142, USA

[2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA

[3]Department of Biology, Massachusetts Institute of Technology, Cambridge MA 02139, USA

[*]For correspondence: dbartel@wi.mit.edu, kxiang@wi.mit.edu

Abstract

In oocytes of mammals and other animals, gene regulation is mediated primarily through changes in poly(A)-tail length[1–9]. Here, we introduce PAL-AI, an integrated neural network machine-learning model that accurately predicts tail-length changes in maturing oocytes of frogs and mammals. We show that PAL-AI learned known and previously unknown sequence elements and their contextual features that control poly(A)-tail length, enabling it to predict tail-length changes resulting from 3'-untranslated region (3'-UTR) single-nucleotide substitutions. It also predicted tail-length-mediated translational changes, allowing us to nominate genes important for oocyte maturation. When comparing predicted tail-length changes in human oocytes with genomic datasets of the All of Us Research Program[10] and gnomAD[11], we found that genetic variants predicted to disrupt tail lengthening have been under negative selection in the human population, thereby linking mRNA tail lengthening to human female fertility.

Introduction

Fully grown vertebrate oocytes, which are arrested at prophase I of meiosis, have minimal transcriptional activity due to the condensation of chromosomes. With little potential for transcriptional control, post-transcriptional control, particularly translational control, plays a major role in regulating gene expression during the subsequent processes of oocyte maturation and early embryonic development, after which the zygotic genome is transcriptionally activated[12–14]. Proper progression through this period of oocyte maturation and early development requires translational upregulation of select maternally deposited mRNAs, such as *MOS* and *CCNB1*[15,16], which encode essential proteins for meiosis, and *TPRX1/2/L*[7], which encode transcription factors necessary for zygotic genome activation. Leveraging the strong coupling between poly(A)-tail length and translational efficiency during these developmental stages[3,4,6,9,17,18], translation activation of these mRNAs is achieved by extension of their poly(A) tails[15,19,20], which enables them to compete better for the binding of limited poly(A)-binding proteins and subsequent recruitment of translation initiation factors[4]. Thus, disruption of either translational or poly(A) tail-length changes in these maturing oocytes or early embryos can cause developmental arrest in frogs, mice, and humans[7,21,22]. Genetic alterations to these processes have also been linked to human female infertility[23–25].

Control of poly(A)-tail length during oocyte maturation of vertebrate animals is a tug-of-war between the default transcriptome-wide deadenylation and the mRNA-specific cytoplasmic polyadenylation[3,8,9,18,26,27]. The specificity of cytoplasmic polyadenylation relies on two sequence elements within mRNA 3' untranslated regions (UTRs): a cytoplasmic polyadenylation element (CPE, motif UUUUA), which is recognized by the CPEB1 protein, and a polyadenylation signal (PAS, motif AWUAAA, where W is either A or U), which is recognized by the cleavage and polyadenylation specificity factor (CPSF) complex[18,28–31]. Contextual features of these two sequence elements, including their relative positions, flanking nucleotides, and structural accessibility, as well as the number of each element within the 3' UTR, influence the extent of cytoplasmic polyadenylation[18]. Although several

other sequence elements have been proposed to influence poly(A)-tail length, their effects are usually small, indirect, and do not appear to impact more than a handful of mRNAs[18,30,32,33].

Despite advances in our knowledge of cis-acting elements that control cytoplasmic polyadenylation, a quantitative understanding of poly(A) tail-length changes for different mRNAs during oocyte maturation is still lacking. A sequence-based predictive model would provide further insights into known functional sequence elements and enable the discovery and evaluation of additional motifs or regulatory principles. For example, such a model would enable large-scale assessment of the consequences of genetic mutations on tail-length control as well as translational regulation, thus providing an opportunity to nominate disease-contributing variants that are otherwise difficult to identify in genome-wide association studies. Therefore, we set out to develop a machine-learning model that accurately predicts poly(A) tail-length changes in humans and other vertebrate animals.

Results

An integrated neural network model predicts tail-length changes during frog oocyte maturation

To predict tail-length changes during frog oocyte maturation directly from mRNA sequences, we initially developed a multiple linear regression model based on *k*-mer (sequence motif length of *k*) compositions within 3' UTRs. Given the importance of both the number and position of CPE and PAS elements in influencing tail-length changes[18], we defined two features for each *k*-mer of varying lengths: 1) a count metric, specifying the total number of a *k*-mer within the 3' UTR; 2) a positional metric, calculated as the sum of the inverse distances between each *k*-mer and the 3' end. We performed training and testing in a 10-fold cross-validation procedure on frog endogenous mRNA tail-length changes, comparing between 0 h and 7 h post progesterone treatment[18] (Fig. 1a). Test-set predictions from each fold were concatenated and compared to the measured values (Fig. 1b). We tested different *k*-mer and 3' UTR lengths as inputs and conducted extensive hyperparameter tuning (Supplementary Fig. 1a). The best model performed moderately well and explained approximately 40% of the variance of measured tail-length changes (Fig. 1c; Spearman correlation coefficient $R_s = 0.64$; Pearson correlation coefficient $R_p = 0.63$). Notably, the CPE and the PAS, two known motifs mediating cytoplasmic polyadenylation, were among the top-ranked features in both the count and the positional metrics (Supplementary Fig. 1b, c).

Although the linear model identified the essential motifs for poly(A)-tail lengthening, it under-predicted tail-length changes for most mRNAs whose tails were extended by > 30 nt. Inspired by recent advances in deep learning, we developed a neural network model, termed poly(A)-tail length AI (PAL-AI) (Fig. 1d, Supplementary Fig. 1d, e), which integrates both convolutional and recurrent neural network components, using an architecture similar to but simpler than that of Saluki[34]. PAL-AI was trained and tested using the sequences of endogenous mRNAs, and the same 10-fold cross-validation strategy as used for the linear model. Combinations of different hyperparameters were tested extensively for the model to achieve the best performance (Supplementary Fig. 2a, b) without

overfitting (Supplementary Fig. 2c, d). The final trained model predicted tail-length changes with high accuracy, significantly out-performing the linear model and explaining more than 67% of the variance of measured values (Fig. 1e; R_s= 0.82, R_p = 0.82).

To examine the impact of input sequence length on prediction accuracy, we trained PAL-AI models on various regions of the 3' UTR and compared Pearson correlation coefficients between predicted and measured values. A model trained exclusively on the last 100 nt of 3' UTRs already explained over 40% of the variance (Fig. 1f; average R_p = 0.64). Model performance improved with increasing 3' UTR length up to 2000 nt, beyond which no significant gains were observed (Fig. 1f, g). In contrast, when the last 100 nt were excluded from the last 1000 nt of 3' UTRs for the input, the predictive power declined markedly (Fig. 1f; average R_p = 0.40), supporting the idea that the 3'-most regions of 3' UTRs impart the strongest effect on tail-length changes[18].

We further explored whether incorporating additional features, such as sequences from coding regions and base-pairing probabilities predicted by RNAfold[35], could improve the overall model performance. However, models including these features showed no significant improvement over those trained on equally-sized 3' UTR-only inputs (Fig. 1f, g), implying that these features, at least as parameterized in our model, contribute minimally to tail-length changes. Nevertheless, among all configurations, the model trained on the last 2000 nt of the 3' UTR plus the coding region performed the best with the shortest input length (Fig. 1f, g) and was therefore used for subsequent analyses unless otherwise specified. We also tested a variant model architecture, in which the convolutional module was replaced with a ResNet block, a design that has demonstrated superior performance in many deep-learning applications[36,37] . This alternative, although at times achieving comparable accuracy, was less consistent (Supplementary Fig. 2e) and required substantially more parameters and computational resources. For this reason, we did not further optimize this ResNet-based model.

PAL-AI learned motifs and contextual features that regulate poly(A)-tail length

To elucidate the principles learned in the training of PAL-AI, we performed in silico mutagenesis and examined the impact of single-nucleotide mutations on predicted poly(A) tail-length changes. Among all 6-mers, the loss of the PAS (AAUAAA) and CPE (UUUUA)-containing 6-mers led to the largest decrease in tail-length changes (Fig. 2a). Sequence logos generated from the most impacted 8-mers also matched the CPE and the PAS (Fig. 2b). Similarly, when examining the consequence of gaining of 6-mer motifs, those motifs containing the PAS and CPE resulted in the greatest predicted tail-length increase (Supplementary Fig. 3a, b). To prevent the signal from spilling from a stronger motif to a weaker one, we reexamined tail-length changes associated with the loss of 6-mers. We identified the 6-mer associated with the largest mean tail-length change and then re-calculated the average tail-length changes for all remaining 6-mers after excluding all point substitutions that disrupted this top 6-mer. This process was repeated 14 times, each time excluding point substitutions that disrupted an additional 6-mer, thereby generating a ranked list of 15 different 6-mers most associated with predicted tail lengthening (Fig. 2c). Likewise, we used an analogous procedure to reexamine tail-length changes associated

149 with the gain of 6-mers (Supplementary Fig. 3c). As expected, the top 6-mers identified in
150 these analyses contained a PAS or CPE.
151 In addition to the PAS and CPE, our analyses also showed that PAL-AI recognized a
152 different class of 6-mers, which were associated with modest but statistically significant
153 tail lengthening (Fig. 2c, Supplementary Fig. 3c). Interestingly, these motifs each contained
154 a UGU or GUU trimer, which might be recognized by other RNA-binding proteins such as
155 CELF or DAZL in frog oocytes[38–40]. Another possibility is that some of these motifs, such as
156 UGUUUU), which resembled the CPE, might be weakly bound by CPEB1. A third possibility
157 is that the presence of these motifs correlates with tail lengthening but is not causal. To
158 explore this latter possibility, we examined the efficacy of these motifs in arbitrary
159 sequences used in a massively parallel reporter 3' UTR library[18]. Because these arbitrary
160 sequences were not biological, they were not susceptible to noncausal correlative
161 associations, and thus any association of a motif with a tail-length change was presumably
162 causal. Analysis of these reporters revealed an association between these UGU/GUU-
163 containing 6 mers and tail lengthening, but it was small — only a 0.64 nt increase on
164 average for the top UGU/GUU-containing 6 mer UGUUUU in the absence of a CPE, which
165 was substantially less than that of a CPE (17.7 nt on average)[18]. In the same library, a
166 somewhat larger increase in tail length was observed for UGU and GUU (1.8 and 1.4 nt
167 above an average effect of a 3-mer, respectively) when a CPE was also present in the 3' UTR
168 (Supplementary Fig. 3d). This added tail-lengthening effect was observed in all CPE-
169 neighboring positions more than 2 nt from the CPE (Fig. 2d), suggesting that these motifs
170 might play a small auxiliary role in tail lengthening beyond merely modulating the context of
171 the CPE. Moreover, for some 3' UTRs, the co-presence of multiple copies of these
172 UGU/GUU-containing motifs might still contribute to detectable tail lengthening[18], perhaps
173 through binding to CPEB1-interacting proteins, such as DAZL[41,42], which in turn could
174 facilitate recruitment of CPEB1 to adjacent CPE-like motifs.
175 We carried out additional in silico mutagenesis by inserting either the CPE or the
176 PAS at each position along the 3' UTRs. As expected[18], PAL-AI predicted the strongest
177 impact on polyadenylation when these elements were introduced near the 3' ends of the 3'
178 UTRs (Fig. 2e). An added CPE was most favorable when located near the PAS on either its 5'
179 or 3' side (Fig. 2f), but not in the position most likely to disrupt the PAS, which is typically
180 ~17 nt from the end of the 3' UTR (Supplementary Fig. 3e)[43], consistent with the positional
181 effects determined from analyses of millions of reporters[18].
182 Although insertion of the canonical PAS variant AUUAAA moderately promoted
183 polyadenylation (Fig. 2e), noncanonical PAS variants[44] had negligible effects on
184 cytoplasmic polyadenylation (Supplementary Fig. 3f). In fact, mutating the canonical PAS
185 AAUAAA to any PAS variant reduced the predicted tail-length increase by 7.7–11.7 nt
186 (Supplementary Fig. 3g). Additionally, in the CPE^{mos}-N60 reporter 3' UTR library[18], when
187 variants containing either an AAUAAA or AUUAAA were excluded, none of the remaining 6-
188 mers were associated with tail lengthening (Supplementary Fig. 3h). Together, these
189 analyses indicate that although non-canonical PAS variants can direct mRNA cleavage and
190 polyadenylation[45], they do not facilitate mRNA polyadenylation in the cytoplasm—at least
191 not during oocyte maturation.

192 To gain more insights into properties that PAL-AI learned at the transcript level, we
193 examined predicted consequences of all possible point substitutions at the last 100 nt of
194 the *mos.L* mRNA. In silico mutations that perturbed either the PAS at position –21 or the
195 CPE at position –86 caused considerable decreases in predicted tail lengthening, whereas
196 a G-to-U mutation at position –62 created a new CPE, thereby causing an increase in
197 predicted tail lengthening (Fig. 2g). *mos.L* also contains a second CPE falling near the PAS,
198 at position –27. This PAS-proximal CPE was expected to be suboptimal for cytoplasmic
199 polyadenylation because its separation from the PAS of only 1 nt was too short[18]. Notably,
200 PAL-AI recognized this proximity feature and correctly predicted only mild mutational
201 outcomes when this motif was disrupted (Fig. 2g).
202 In another example, the *tpx2.L* mRNA, PAL-AI identified two CPEs predicted to be
203 important for tail lengthening. Furthermore, PAL-AI predicted decreases in tail lengthening
204 when Gs at positions –66 and –36 (both at –2 positions relative to CPEs) were mutated and
205 an increase in tail-length change for a C-to-G substitution at position –12 (+1 relative to the
206 PAS) (Fig. 2h)—all observations consistent with the contextual features of the CPE and the
207 PAS identified from analyses of millions of reporters[18]. Similar results were observed for
208 other examples of known substrates of cytoplasmic polyadenylation (Supplementary Fig.
209 3i, j).
210 Together, these analyses indicated that PAL-AI successfully captured the sequence
211 elements and their contextual features important for promoting cytoplasmic
212 polyadenylation, thereby enabling it to accurately predict poly(A) tail-length changes of frog
213 mRNAs during oocyte maturation.
214
215 **PAL-AI predicts poly(A) tail-length change of injected mRNAs**
216 Next, we asked if PAL-AI was able to predict tail-length changes of injected reporter
217 mRNAs. To this end, we amplified a small subset of a 3'-UTR reporter library (N60-PAS^{mos})[18]
218 to generate a library with only ~35,000 variants (Fig. 3a), which we called N60(LC)-PAS^{mos}.
219 The reduced sequence complexity of this library enabled acquisition of tail-length
220 distributions of individual mRNA sequences. We injected this mRNA library into prophase I-
221 arrested frog oocytes, collected RNA samples at 0 and 7 h post-progesterone treatment,
222 and used high-throughput sequencing[18] to identify the 3' UTR variant and determine its tail-
223 length distribution at each of the two stages of maturation (Fig. 3b).
224 When using tail-length changes measured for these injected mRNAs to test our
225 model trained on endogenous mRNAs, our model performed well (Supplementary Fig. 4a).
226 Indeed, correlation coefficients observed when testing on the injected mRNAs (R_s= 0.77, R_p
227 = 0.83) resembled those observed when testing on endogenous mRNAs (Fig. 1e, R_s= 0.82,
228 R_p = 0.82). However, the model appeared to under-predict the magnitude of the tail-
229 lengthening effects (Supplementary Fig. 4a). A few possibilities could explain the
230 systematic difference in the magnitude of tail-lengthening observed for injected mRNAs
231 compared to that predicted by PAL-AI. First, the sequence lengths of the training and the
232 testing datasets were markedly different. Endogenous mRNAs used for training PAL-AI had
233 few 3' UTRs shorter than 100 nt (420 out of 6054), whereas 3' UTRs of the injected mRNAs
234 were all relatively short (81 nt). Second, due to experimental variability, particularly the
235 variable time it takes for oocytes of different batches to mature, the endogenous mRNAs

and injected mRNAs may have been collected at somewhat different maturation stages. Third, the 3' UTRs of the reporter mRNAs consisted of arbitrary sequences, whose composition may differ substantially from natural mRNA sequences.

To test whether sequence composition contributed to the discrepancy, we modified PAL-AI's last layer to enable dataset-specific multi-output predictions (Fig. 3c). Training on alternating batches of endogenous and injected mRNAs maintained performance on endogenous mRNAs (Fig. 4d, R_s= 0.82, R_p = 0.82) while enhancing performance on injected mRNAs (Fig. 4e, R_s= 0.81, R_p = 0.88). To distinguish this multi-output model from the original PAL-AI model, we termed this model as PAL-AI-m ("m" for multi-output) and the original model as PAL-AI-s ("s" for single-output).

Despite the improvement, the PAL-AI-m model still under-predicted tail lengthening for most long-tailed (≥ 80 nt) mRNAs from the injected library. When we re-trained the model with measurements from only injected mRNAs, PAL-AI performed significantly better when tested on held-out data (Supplementary Fig. 4b, R_s= 0.86, R_p = 0.91). In silico mutagenesis analyses indicated that the new model also learned the expected sequence elements important for tail lengthening and their contextual features (Supplementary Fig. 4c–e). This training approach on synthetic, nonbiological reporter sequences also ensured that these learned elements and features are causative for tail-length changes, rather than correlated to causative features.

Predicted effects of single-nucleotide substitutions correspond with experimentally determined effects.

The ability of PAL-AI to predict the fate of injected mRNAs enabled direct assessment of its ability to predict the consequences of mutations, including all possible single-nucleotide substitutions. To this end, we generated a single-nucleotide mutagenesis library of reporter mRNAs with 3' UTR sequences derived from the last 100 nt of 3' UTRs of ten mRNAs selected from frog oocytes. For each of the ten 100-nt UTR fragments, each nucleotide of the 3' UTR was substituted with one of the three alternative nucleotides, resulting in 3,000 variants (Fig. 4a). These 3,000 variants, together with the non-substituted wild-type molecules, were injected into oocytes and the effect of each substitution on tail-length changes was analyzed over the course of maturation. Of the ten 3'-UTR fragments, eight (*atp1a1.S*, *ccnb1.2.L*, *ccnb2.L*, *lima1.L*, *mad2l1.L*, *magoh.S*, *mos.L*, and *tpx2.L*) directed cytoplasmic polyadenylation upon injection into oocytes, whereas two (*aurkaip1.L* and *dbf4.L*) directed no tail lengthening during frog oocyte maturation[18].

When comparing the predicted difference in tail-length change to that measured between each mutant and the wild-type, the PAL-AI-m model performed significantly better compared to the PAL-AI-s model for eight out of ten mRNAs (Fig. 4b). This improved performance suggested that the PAL-AI-m model successfully learned additional features from the non-natural sequences of the N60(LC)-CPE^{mos} library. The PAL-AI-m model achieved high accuracy (R_p: 0.78–0.90) for six mRNAs (*atp1a1.S*, *lima1.L*, *mad2l1.L*, *magoh.S*, *mos.L*, and *tpx2.L*; Fig. 4b, d–g, Fig. 5a, b, Supplementary Fig. 5a–f) and moderate accuracy (R_p: 0.58–0.62) for three mRNAs (*aurkaip1.L*, *ccnb1.2.L*, and *ccnb2.L*; Fig. 4b, Supplementary Fig. 6a, b, d, e, Supplementary Fig. 7a, b).

279 Concordance between predicted and experimental outcomes was particularly clear
280 for substitutions that disrupted either a CPE or PAS (Fig. 4c, f, g, Fig. 5a, Supplementary Fig.
281 5d–f). Substituting the canonical PAS (AAUAAA) with the alternative variant AUUAAA
282 severely affected tail lengthening, but not as severely as observed for any of the other
283 single-nucleotide substitutions of AAUAAA (Fig. 4c). These experimental findings validated
284 our in silico mutagenesis conclusion that noncanonical PAS variants do not support
285 cytoplasmic polyadenylation during oocyte maturation.
286 At the individual mRNA level, ablation of any of the three motifs (two CPEs and one
287 PAS) within *mos.L* mRNA led to substantial decreases in tail lengthening, as predicted by
288 PAL-AI-m (Fig. 4f). The predicted weaker contribution to cytoplasmic polyadenylation by
289 the CPE proximal to the PAS compared to that by the CPE further upstream was also
290 validated (Fig. 4f). Similar results illustrating PAL-AI-m's ability to predict the importance of
291 these motifs and their contextual features were also observed for other mRNAs (Fig. 4g,
292 Supplementary Fig. 5). Notably, the predicted differences in tail-length changes observed
293 between many variants and the wild-type were smaller than those measured, perhaps due
294 to the same reasons previously put forward to explain the under-prediction of changes for
295 the injected N60(LC)-PAS^{mos} library (Supplementary Fig. 4a).
296 In another example, *mad2l1.L* mRNA had no CPE within the last 100 nt of its 3' UTR
297 but contained seven UGU/UGUUU motifs (Fig. 5a). Mutations within these motifs caused
298 modest but consistent reductions in tail lengthening (Fig. 5a, b), agreeing with their
299 detectable but weak roles in cytoplasmic polyadenylation. In contrast, substitutions that
300 introduced a single CPE led to large increases in tail lengthening, almost all of which were
301 correctly predicted by PAL-AI-m (Fig. 5a, b).
302 Despite achieving good performance for the six aforementioned mRNAs (R_p = 0.78–
303 0.9), PAL-AI-m performed only moderately well on three mRNAs (*aurkaip1.L*, *ccnb2.L*, and
304 *ccnb1.2.L*, R_p = 0.58 , 0.62, and 0.59, Supplementary Fig. 6b, e, 7a) and did not perform well
305 at all on one mRNA (*dbf4.L*, R_p = 0.21, Fig.5c). The wide-type *dbf4.L* mRNA, which had no
306 CPE and a relatively weak PAS (AUUAAA), had little tail lengthening, and thus most single-
307 nucleotide substitutions had negligible opportunity to reduce tail lengthening (Fig. 5d),
308 which largely explained their relatively weak effects for this mRNA. PAL-AI-m did predict
309 moderate increases in tail lengthening for a few substitutions that would create CPEs
310 predicted to lengthen tails, but by no more than 10 nt (Fig. 5c). Although some of these
311 predicted increases were validated (positions –61 and –51), others, particularly those
312 downstream the PAS (positions –35, –31, –7 and –6) were not (Fig. 5d). To understand these
313 discrepancies, we predicted the secondary structure of the last 100 nt of the wild-type
314 *dbf4.L* 3' UTR using EternaFold[46]. The substitutions at positions –61 and –51 both created
315 CPEs in loop regions, whereas the substitutions at positions –35, –31, –7 and –6 all created
316 CPEs with nucleotides paired in a stem loop (Fig. 5d, e), which likely prevented these CPEs
317 from being recognized by the CPEB1 protein.
318 For *aurkaip1.L* and *ccnb2.L* mRNAs, PAL-AI-m had difficulty predicting increased tail
319 lengthening caused by substitutions in consecutive positions within a few UTR segments
320 (*aurkaip1.L*: positions –34 to –30 and –19 to –15, Supplementary Fig. 6a; *ccnb2.L*: positions
321 –60 to –55 and –15 to –10, Supplementary Fig. 6d). These substitutions appeared to
322 destabilize stem loops that were sequestering PAS elements in the wild-type UTRs

323 (Supplementary Fig. 6c, f), thus making these elements more accessible for CPSF binding.
324 In support of this hypothesis, substitutions that further stabilized these stem loops
325 reduced tail lengthening (*aurkaip1.L*: A-to-G at position –35 and C-to-U at position –14,
326 Supplementary Fig. 6a, c; *ccnb2.L*: C-to-G at position –54 and C-to-G at position –16,
327 Supplementary Fig. 6d, f). Together, these analyses indicated that sequences within the 3'
328 UTR can indirectly affect poly(A) tail-length changes by modulating the structural
329 accessibility of the CPE and the PAS, a mechanism that likely applies to select mRNAs[33]
330 and thus had not yet been effectively learned by PAL-AI-m.
331 As for the *ccnb1.2.L* mRNA, PAL-AI-m under-predicted the effects of a CPE that
332 overlapped with the PAS (Supplementary Fig. 7a, b). Binding of CPEB1 to this overlapping
333 CPE presumably prevented CPSF from recognizing the PAS, thus inhibiting cytoplasmic
334 polyadenylation as previously suggested[47]. Supporting this hypothesis, the C-to-U
335 mutation at position –18 of *magoh.S* mRNA created an overlapping CPE, which reduced tail
336 lengthening (Supplementary Fig. 5f).
337 Despite some discrepancies between the predictions and experimental results,
338 these analyses demonstrated that, overall, PAL-AI can accurately predict tail-length
339 changes caused by single-nucleotide mutations within mRNA 3' UTRs.
340

341 **PAL-AI predicts poly(A) tail-length change in mammalian oocytes**
342 Because principles of poly(A) tail-length control during oocyte maturation are highly
343 conserved among frogs, mice, and humans[18], we asked if PAL-AI trained from frog oocyte
344 data could predict tail-length changes during mammalian oocyte maturation. For mouse
345 oocytes, PAL-AI achieved moderate accuracy (Supplementary Fig. 8a; R_s= 0.57, R_p = 0.58).
346 The performance of the model was likely affected by unmatched developmental stages
347 between frogs and mice, and also global deadenylation that was independent of 3'-UTR
348 sequences[9,18]. Similar predictive power for PAL-AI was also observed for human oocytes
349 using published datasets (Supplementary Fig. 8b; R_s= 0.61, R_p = 0.59), even though the
350 human measurements were acquired using a different sequencing protocol and platform
351 (PacBio)[8].
352 To minimize prediction errors arising from species-specific sequence composition
353 and experimental variability, we developed a multi-species PAL-AI model trained on pooled
354 data from frogs, mice, and humans, using the same approach as for our frog endogenous
355 mRNAs and injected mRNAs. This model, termed PAL-AI-ms ("ms" for multi-species),
356 showed improved performance for both mouse (Fig. 6a; R_s= 0.63, R_p = 0.64) and human
357 oocytes (Fig. 6b; R_s= 0.68, R_p = 0.66), while maintaining comparable accuracy for frog
358 oocytes (Supplementary Fig. 8c; R_s= 0.79, R_p = 0.77). In addition, the model's predictions
359 also correlated well with independent tail-length measurements obtained using Oxford
360 Nanopore Technologies (ONT) for mRNAs during mouse oocyte
361 maturation[9] (Supplementary Fig. 8d; R_s= 0.69, R_p = 0.69). Although ONT data showed larger
362 absolute tail-length changes compared to our measurements obtained using Illumina
363 sequencing (HiSeq 2000), both datasets exhibited consistent trends (Supplementary Fig.
364 8e). This systematic difference in magnitude likely explained why the model under-
365 predicted tail-length changes in the ONT data (Supplementary Fig. 8d).

366 In frog, mouse, and human oocytes, tail-length changes strongly correlate with
367 translational efficiency changes[3,9,18]. As a result, our model-predicted tail-length changes
368 were expected to be informative regarding up- and down-regulation of mRNA translation—
369 not only for endogenous mRNAs but also for any mRNAs with known 3'-UTR sequences. To
370 confirm this expectation, we predicted tail-length changes for a series of reporter mRNAs
371 with different 3'-UTR sequences used for assessing translational regulation during mouse
372 oocyte maturation[6]. Our model-predicted tail-length changes corresponded well with the
373 reported translational efficiency changes (Fig. 6c; R_s= 0.91, R_p = 0.89).
374 Although tail-length control during oocyte maturation is highly conserved among
375 vertebrates, tail lengthening and associated translational upregulation are known to be
376 required for mRNAs from only a few genes. Conserved tail lengthening of mRNAs from
377 orthologous genes may shed light on the functional significance of tail lengthening during
378 oocyte maturation and early development, but mRNAs from only 19 genes with conserved
379 tail lengthening have been identified[18]. Part of the reason for this small number is that most
380 mRNAs could not be evaluated because they lacked tail-length measurements in at least
381 one of the three species[18]. To overcome this limitation, we used our model-predicted tail-
382 length changes for cases in which tail-length measurements were unavailable, resulting in
383 the identification of another 264 genes (57, 108, and 99 with one, two, and three predicted
384 values respectively) whose mRNA poly(A) tails were either extended or predicted to be
385 extended ≥ 15 nt in all three species (Fig. 6d, Supplementary Fig. 8f, g). Most of these
386 mRNAs had increased translational efficiency, and for those for which translational
387 efficiency changes had not been measured, PAL-AI-predicted tail-length increases could
388 be informative. A Gene Ontology analysis of these genes revealed enrichment of molecular
389 pathways such as cell cycle, microtubule cytoskeleton organization, cell division, and
390 chromosome organization, which are crucial for meiosis and presumably prepare the
391 zygote for cell divisions of early embryonic development (Fig. 6e).
392
393 **Variants predicted to disrupt poly(A)-tail lengthening are under negative selection in**
394 **humans and throughout vertebrates**
395 Next, we asked if genetic variants predicted to influence tail-length changes during oocyte
396 maturation have been under evolutionary selection. We analyzed human 3' UTRs by using
397 PAL-AI-ms to predict the effects of all possible single-nucleotide mutations within their last
398 100 nt. For each position, we calculated the average difference in predicted tail-length
399 change when mutated to alternative nucleotides and classified the wild-type nucleotide by
400 the direction of the effect and severity when mutated: tail-length reduction(severe [≤ –25
401 nt], moderate [–25 to –15 nt], mild [–15 to –5 nt]), minimal-impact [–5 to 5 nt], and tail-
402 length increase (mild [5 to 15 nt], moderate [15 to 25 nt], severe [> 25 nt]). Comparing these
403 predictions to phyloP scores, a conservation metric based on the alignment of 100
404 vertebrate genomic sequences[48,49], revealed strong evolutionary constraints, with
405 conservation levels tracking predicted functional impact. For example, nucleotides for
406 which mutations would cause severe tail-length reduction showed significantly higher
407 phyloP scores than moderate-impact sites (Mann–Whitney U test), which in turn were more
408 conserved than mild-impact sites, with minimal-impact sites being least conserved (Fig.
409 7a, Supplementary Fig. 9a). Nucleotides for which mutations would cause tail-length

410 increase showed a parallel but weaker trend, with only mild/moderate-impact sites
411 exhibiting significantly elevated conservation relative to minimal-impact sites (Fig. 7a,
412 Supplementary Fig. 9a). This is likely in part because tail-length shortening during oocyte
413 maturation is non-specific[9,18,26,27] and therefore very few sites, when mutated, would result
414 in tail-length increase (Fig. 7a). Overall, these findings indicate that genomic positions at
415 which mutations would substantially alter poly(A) tail-length changes during oocyte
416 maturation have been under purifying selection in the vertebrate lineage.
417 We also asked if genetic variants predicted to influence tail-length control during
418 oocyte maturation might have functional consequences for human health. To this end, we
419 analyzed ~8.4 million 3'-UTR variants from the All of Us Research Program[10] using PAL-AI-
420 ms. For each variant, we calculated its predicted effect size as the difference in predicted
421 tail-length change relative to the reference allele (Fig. 7b). Focusing on the last 100 nt of
422 each 3' UTR (~ 0.85 million variants), we binned variants by predicted effect size the same
423 way as we had done for the conservation analysis and examined their relative fractions in
424 each bin for groups with different allele frequencies. This analysis revealed a striking allele-
425 frequency-dependent depletion of tail-lengthening-disrupting variants. For example,
426 variants predicted to severely disrupt tail-lengthening showed 2-, 3-, 7- and 10-fold
427 depletion in common variants (allele frequencies > 0.01%, > 0.1%, > 1%, and > 10%,
428 respectively) compared to singletons ($P < 0.001$, Fisher's exact test). Depletion of frequent
429 alleles was also significant, albeit to a lesser degree, for variants predicted to impart
430 moderate (1.5–1.9 fold) and mild (1.3–1.7 fold) reduced tail-lengthening (Fig. 7c). This
431 depletion pattern remained significant even after excluding all variants that would
432 introduce or ablate PAS motifs (Supplementary Fig. 9b), which were presumably also under
433 negative selection due to their nuclear role in specifying mRNA cleavage and
434 polyadenylation[50]. In contrast, gain-of-function variants were not consistently depleted in
435 common alleles compared to singleton variants (Fig. 7c). Similar results were observed
436 when PAL-AI was applied to the gnomAD v4.1 datasets[11] (Fig. 7d, Supplementary Fig. 9c, d).
437 Together, these analyses indicated that variants that disrupt poly(A) tail-length control,
438 particularly those that disrupt tail lengthening during oocyte maturation, are under negative
439 selection in humans.
440

441 ## Discussion

442 We developed PAL-AI, a neural network model that predicted tail-length changes in frog,
443 mouse, and human oocytes using only sequence information. PAL-AI captured
444 consequential sequence motifs and contextual effects without any feature generation,
445 thereby allowing us to predict tail-length changes of mRNAs with any sequence. Moreover,
446 PAL-AI helped us identify UGU/GUU-containing motifs contributing to poly(A)-tail
447 lengthening. Although previously implicated in regulating translation[38,41,42,51–53], these motifs
448 had not been associated with tail-length control. Our analyses of the injected mRNA
449 reporter libraries further demonstrated that their tail-lengthening effects were causal rather
450 than correlative with other features of endogenous mRNAs. Individually, these motifs
451 contributed weakly to polyadenylation, but multiple copies did result in considerable tail
452 lengthening, as illustrated by the *mad2l1.L* mRNA. Their precise roles in affecting tail

453 lengths might be context-dependent, as some might serve as CPE-like motifs
454 promiscuously bound by CPEB1, whereas others might be recognized by RNA-binding
455 proteins such as DAZL, which also interacts with CPEB1 and thus might help facilitate
456 recruiting CPEB1 to the 3' UTR.
457 The ability of PAL-AI to predict tail-length changes of both endogenous mRNAs and
458 injected mRNAs enabled us to capture the consequences of most single-nucleotide
459 substitutions. For a few mRNAs, PAL-AI under-performed in predicting the outcomes of
460 substitutions—often those substitutions that appeared to alter the structural accessibility
461 of CPE or PAS elements. However, when we combined base-pairing probability, as
462 predicted by RNAfold[35], with mRNA sequences as the input for our model, the performance
463 did not significantly improve, which suggested that more accurate parameterization of
464 mRNA structure would be needed for the model to account for structural accessibility.
465 Perhaps incorporating data from in-oocyte structural probing of mRNAs will further improve
466 the model predictions[54–57].
467 Because of the strong influence of tail length on translational efficiency in oocytes,
468 we envision that our model can be used to engineer mRNA 3'-UTR sequences for desired
469 protein expression in these systems. Furthermore, by predicting tail-length changes that
470 had not yet been measured in frog, mouse, or human oocytes, our model helped identify a
471 group of genes that appear to undergo conserved mRNA tail-lengthening during oocyte
472 maturation. Although for mRNAs of most of these genes, the importance of translational
473 up-regulation during oocyte maturation and early embryogenesis still needs to be
474 examined, knowledge of these genes can inform efforts to identify mutations that affect
475 human female fertility.
476 Indeed, genetic variants predicted by PAL-AI to disrupt tail-length control during
477 oocyte maturation have been under negative selection in vertebrates including humans.
478 Presumably, some of these variants perturb sequence elements or contextual features
479 important for cytoplasmic polyadenylation, thus leading to failed translational activation of
480 mRNAs encoding proteins important for oocyte maturation or early embryonic
481 development. While not evidently observed in this study, as larger genomic data sets
482 become available and more sensitive analyses are conducted, negative selection might
483 also be detected for mutations predicted to lead to inappropriate activation of
484 translationally repressed mRNAs during these developmental transitions.
485 Although the observed depletion of variants predicted to disrupt tail lengthening
486 persisted after excluding PAS-perturbing substitutions, we cannot rule out that some of this
487 selection pressure may reflect pleiotropic effects on other RNA-processing events. For
488 instance, U-rich sequences resembling the CPE near the mRNA 3' end can serve as binding
489 platforms for Fip1 in nuclear pre-mRNA processing[58,59]. Variants disrupting these motifs
490 could simultaneously impair both cytoplasmic polyadenylation and nuclear 3' end
491 formation. Future integration of PAL-AI with tools capable of predicting 3' processing
492 defects, such as APARENT2[50] or PolyaID[60] , could help distinguish these potentially
493 confounding selective pressures.
494 Past efforts to identify genetic causes of female infertility have primarily focused on
495 variants in coding regions, perhaps in part due to the use of exome sequencing[23]. However,
496 mutations that explain defects in oocytes and early embryos associated with female

infertility fall in fewer than 20 genes and are found in only a small portion of patients[61]. Given the importance of tail-length and translational control in human oocytes and early embryos, as well as our analyses of variant abundancies in the human population, some unexplained cases are presumably due to disruptive mutations occurring in 3' UTRs. With more genomes with female reproductive failures being sequenced, we will be able to start assessing contributions by these variants, and PAL-AI can be a valuable tool complementing current approaches, such as genome-wide association studies, to help nominate causal variants within 3' UTRs.

Acknowledgments

We thank Thy Pham, Peter Wang, and other current and former members of the Bartel lab for helpful discussions, Sumeet Gupta for assistant with data collection, and the Whitehead Institute Genome Technology Core for sequencing. D.P.B. is an investigator of the Howard Hughes Medical Institute.

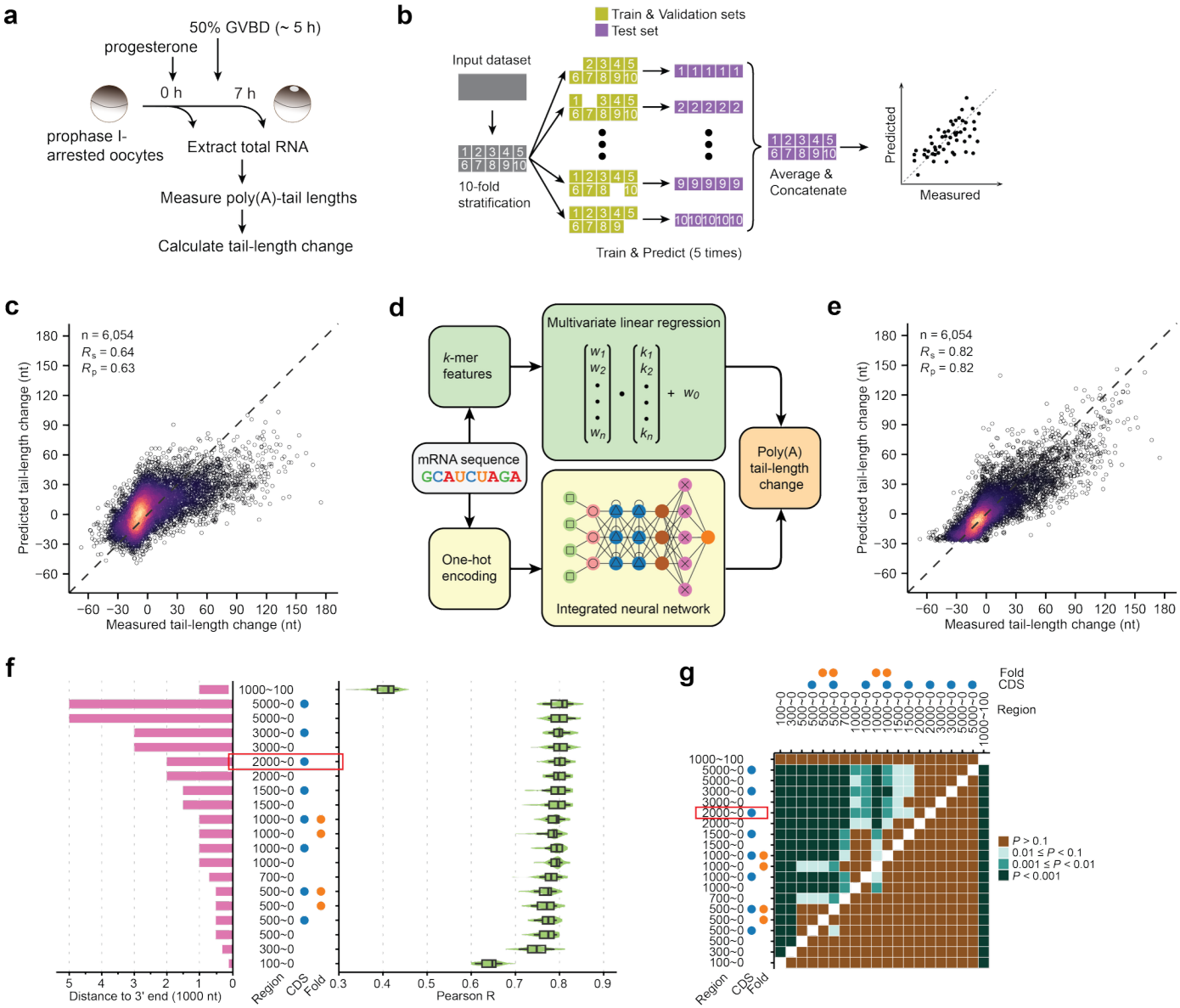
Author contributions

K.X. and D.P.B. designed the study and wrote the manuscript. K.X. conceived the project and performed all the experiments and analyses.

Competing interests

The authors declare no competing interests.

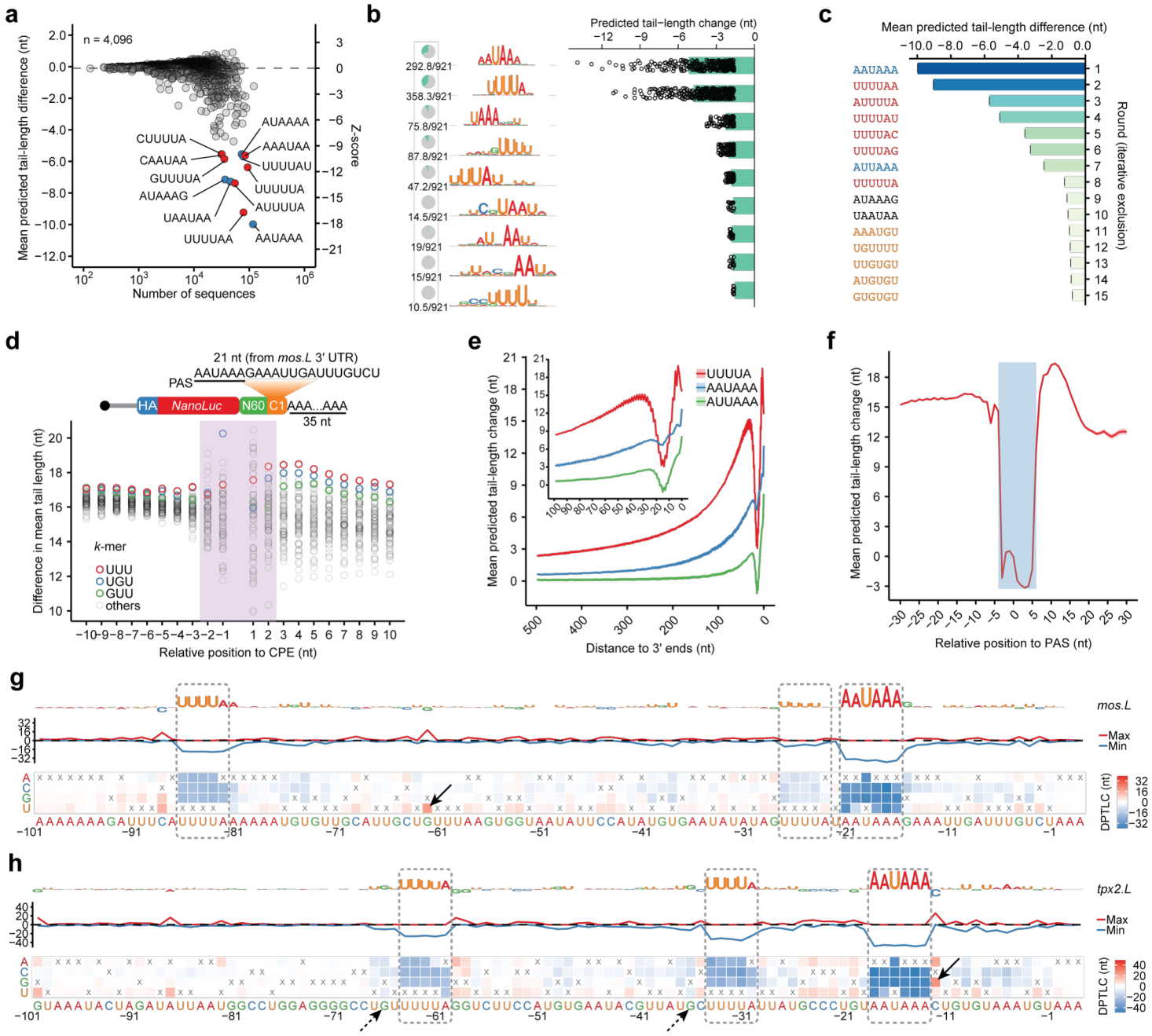
520 **Fig. 1**



521 **Fig. 1: Machine-learning models predict tail-length changes during frog oocyte**
522 **maturation.**
523 **a** Experimental scheme for examining poly(A) tail-length changes of frog mRNAs during
524 oocyte maturation. Total RNA was extracted from oocytes before and after progesterone-
525 induced germinal vesicle breakdown (GVBD), and changes in poly(A) tail lengths were
526 measured. **b** Schematic of the 10-fold cross-validation strategy used to train and test
527 different machine-learning models in this study. Data were partitioned into
528 training/validation and test sets, repeated across 10 different stratified folds. **c**
529 Performance of the multiple linear regression model. Plotted are the tail-length changes
530 predicted by the model as a function of the changes measured in frog oocytes between 0
531 and 7 h post-progesterone treatment. Each point represents a unique poly(A) site of an
532 endogenous mRNA. Colors indicate the density of points. **d** Diagram outlining the two

machine-learning models developed to predict poly(A) tail-length changes from mRNA sequence features: a multiple linear regression model and an integrated neural network (PAL-AI). **e** Performance of PAL-AI; otherwise as in (**c**). **f** Prediction performance of PAL-AI trained on different input regions of mRNA or additional annotation features. Left: input sequence regions (bars) and additional features (CDS or predicted pairing, blue and red dots, respectively). Right: distributions of R_p values observed when comparing predicted and measured tail-length changes for test data held out during training. Ten-fold cross-validation of the model was repeated five times, generating 50 R_p values. The red rectangle indicates the configuration chosen as the final model. CDS, coding sequence. Box and whiskers indicate the 10th, 25th, 50th, 75th, and 90th percentiles. **g** Pairwise comparison of input strategies used for PAL-AI based on prediction performance. Shown are binned *P* values from one-sided t-tests, testing the alternative hypothesis that the mean R_p value of the group indicated on the y axis is greater than that indicated on the x axis. The red rectangle indicates the configuration chosen as the final model.

Fig. 2



549 **Fig. 2: PAL-AI learned sequence elements and contexture features important for**
550 **poly(A) tail-length control during oocyte maturation.**
551 **a** PAL-AI-predicted consequences of motif loss. Shown for each 6-mer is the mean
552 difference between predicted tail-length changes of mutants compared to those of the
553 wild-type. These mean differences are plotted as a function of the mean number of
554 analyzed mutants. Red points represent motifs that contained a CPE and blue points
555 represent those that contained a complete or partial PAS. Only motifs within the last 300 nt
556 of 3' UTRs were included in this analysis. **b** Sequence motifs associated with PAL-AI-
557 predicted changes in poly(A)-tail length. Shown are sequence logos generated from 8-mers

most associated with the largest differences in predicted tail-length change between mutants in which 8-mers were lost and the wild-type. The pie charts on the left indicate the fraction of 8-mers that aligned to the corresponding sequence logo. The bar plots on the right show the mean difference, with points representing individual 8-mers. Only motifs within the last 300 nt of 3' UTRs were included in this analysis. **c** Top 6-mers associated with decreased predicted tail-length change in the in silico mutagenesis analysis. Differences in tail-length change were calculated in an iterative process with exclusion. In each round, differences in tail-length change were calculated for all 6-mers lost in the mutants, and the 6-mer associated with the largest decrease in tail-length changes was selected. In the following rounds, all mutants in which this top 6-mer was disrupted were excluded before this procedure was repeated. Motifs are colored red if they contain a CPE, blue if they contain a PAS, orange if they contain a GUU or UGU, and black if otherwise. Only motifs within the last 300 nt of 3' UTRs were included in this analysis. **d** Positional tail-lengthening effects of 3-mers flanking a CPE. Plotted for each 3-mer are differences in mean tail lengths observed for mRNAs with that 3-mer at the indicated positions relative to a CPE in the 3' UTRs of the N60-PAS^{mos} library[18], comparing between 0 and 7 h post-progesterone treatment. Only variants containing one CPE (UUUUA) were included in this analysis. The shaded area indicates positions where 3-mers impact the flanking nucleotide context of a CPE[18]. **e** Positional effects predicted by PAL-AI for the CPE and PAS. Plotted for each of the motifs are mean differences between predicted tail-length changes conferred upon inserting the motif in silico at the indicated positions along the 3' UTR. The drop observed when inserting a motif near position 17 was attributed to disruption of the PAS (**Supplementary Fig. 3e**). Shaded areas along the lines indicate standard error. The inset shows results for the last 100 nt of the 3' UTR. **f** Effects of the relative position between the CPE and the PAS predicted by PAL-AI. Plotted are the mean differences between predicted tail-length changes conferred upon inserting a CPE in silico as a function of the relative distance between the CPE and the PAS. Shaded areas along the lines indicate standard error. The rectangular shaded area indicates positions where the inserted CPE would overlap with the PAS. **g** Impact of single-nucleotide substitutions on predicted tail-length changes. Shown are results for the last 100 nt of the *mos.L* mRNA 3' UTR. The heatmap indicates the difference in predicted tail-length change (DPTLC) as a result of changing each nucleotide (*x* axis) to each of the three non-wild-type alternatives (*y* axis), with x indicating the non-substituted, wild-type nucleotide. The line plot indicates the maximum (red line) and the minimum (blue line) of the mutational outcomes at each position. The logo plot indicates the importance of each nucleotide, with the height normalized to the negative value of the average outcome of the three possible substitutions. Dashed rectangles indicate model-identified sequence elements (the CPE and PAS) that are important for poly(A)-tail lengthening. The arrow points to an instance of a new CPE, generated by a G-to-U substitution, and its associated increase in tail-length change. **h** Impact of single-nucleotide substitutions in the *tpx2.L* mRNA 3' UTR on predicted tail-length changes, plotted as in (**g**). The arrow with a solid line points to an example of PAS flanking-nucleotide effects, in which a C-to-G substitution at the +1 position increases predicted tail-length change[18]. The arrows with dashed lines point to examples of CPE

flanking nucleotide effects, in which substitutions of G at the –2 position decrease predicted tail-length change[18].

Fig. 3

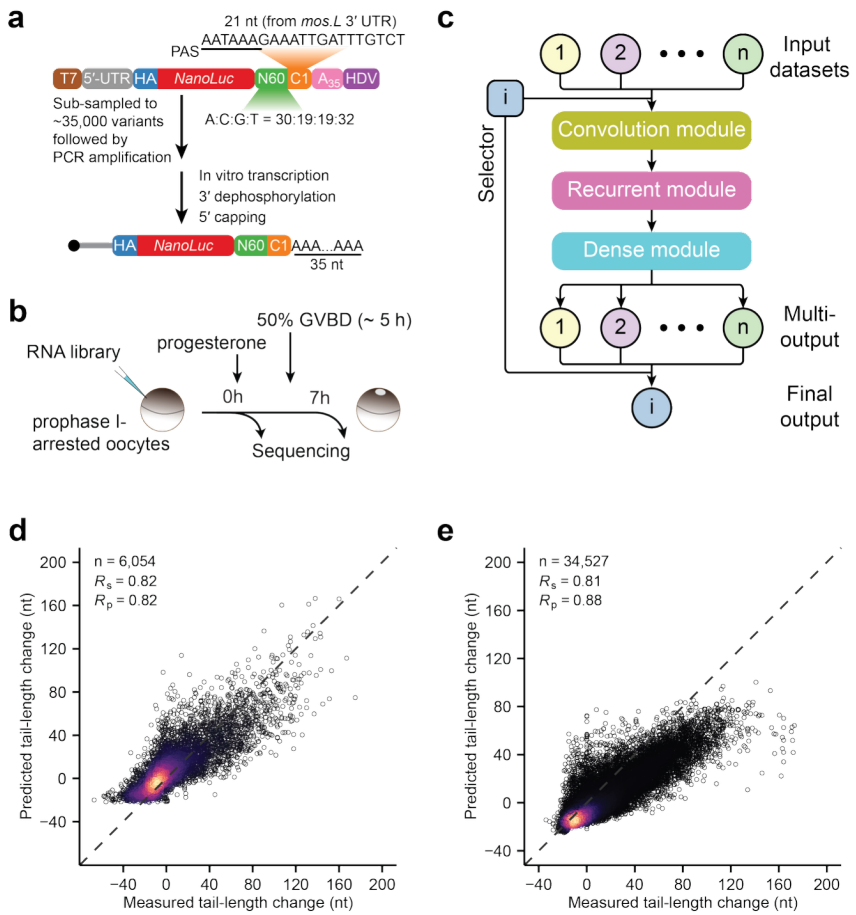


Fig. 3: PAL-AI predicts tail-length changes of synthetic mRNAs in frog oocytes. **a** Schematic of the N60(LC)-PAS^{mos} mRNA library used for injection. **b** Experimental scheme for mRNA library injection and sample collection. **c** Diagram of the architecture of the multi-output PAL-AI model (PAL-AI-m). **d** Performance of the PAL-AI-m model in predicting tail-length changes of frog endogenous mRNAs. Otherwise, this panel is as in **Fig. 1c**. **e** Performance of the PAL-AI-m model in predicting tail-length changes of mRNAs in the N60(LC)- PAS^{mos} library. Otherwise, this is as in **Fig. 1c**.

Fig. 4

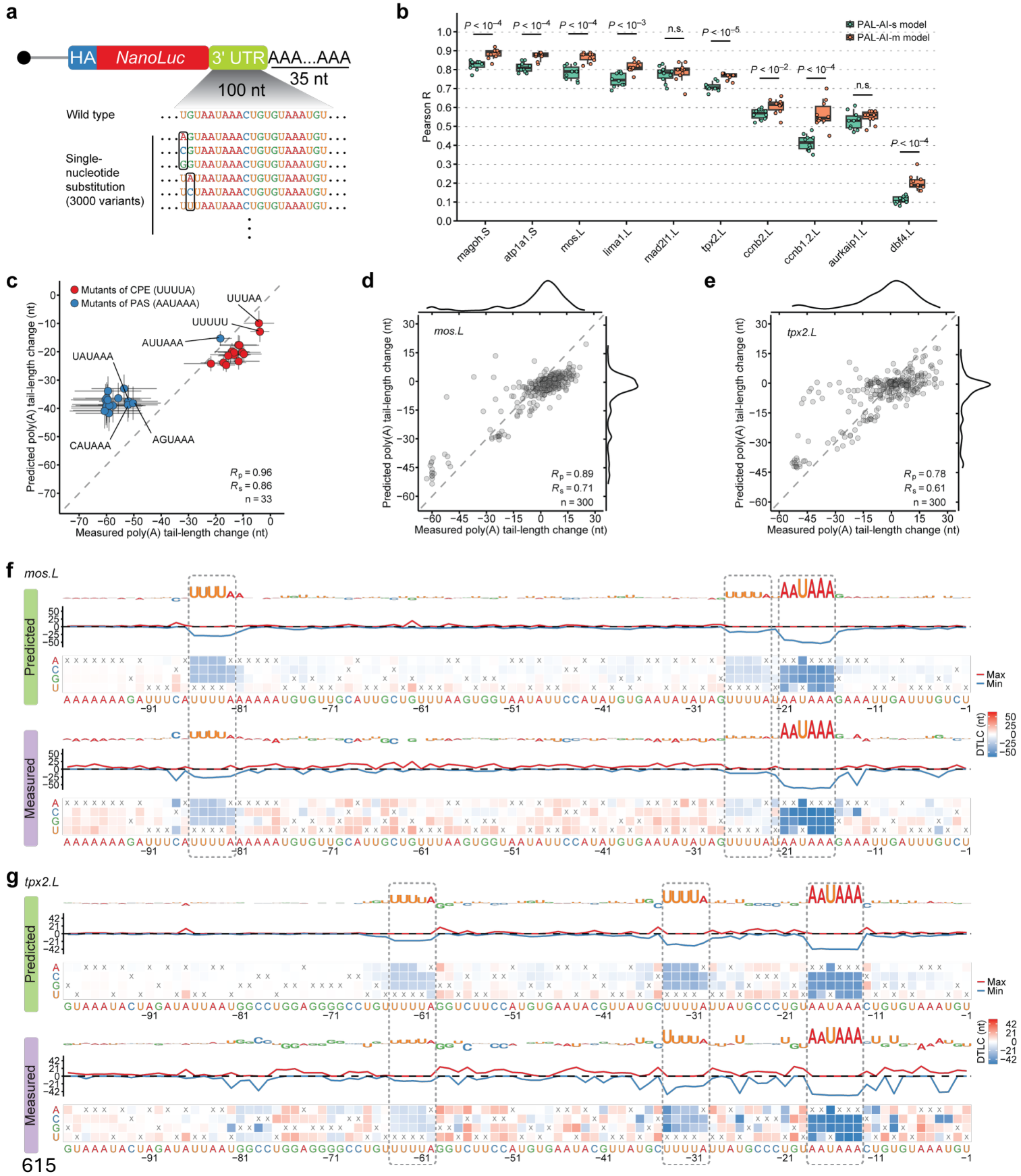

a
HA
NanoLuc
3' UTR
AAA...AAA
35 nt
100 nt
Wild type
...UGUAAUAAACUGUGUAAAUGU...
Single-nucleotide substitution (3000 variants)
b
PAL-AI-s model
PAL-AI-m model
Pearson R
magoh.S
atp1a1.S
mos.L
lima1.L
mad2l1.L
tpx2.L
ccnb2.L
ccnb1.2.L
aurkaip1.L
dbf4.L
n.s.
c
Mutants of CPE (UUUUA)
Mutants of PAS (AAUAAA)
UUUAA
UUUUU
AUUAAA
UAUAAA
CAUAAA
AGUAAA
Predicted poly(A) tail-length change (nt)
Measured poly(A) tail-length change (nt)
Rp = 0.96
Rs = 0.86
n = 33
d
mos.L
Rp = 0.89
Rs = 0.71
n = 300
e
tpx2.L
Rp = 0.78
Rs = 0.61
n = 300
f
mos.L
Predicted
Measured
Max
Min
DTLC (nt)
g
tpx2.L
Predicted
Measured
Max
Min
DTLC (nt)

Fig. 4: Experiments validate predicted effects of single-nucleotide substitutions. **a** Schematic of the library of single-nucleotide substitutions used for injection. In addition to the 3000 variants with single-nucleotide substitutions, mRNAs with the wild-type segments were also included at a proportion of one wild-type for every 300 single-nucleotide variants. **b** Summary of the relationships between the PAL-AI-s- or PAL-AI-m-predicted effects of single-nucleotide substitutions and the measured effects of those substitutions. For each of the 10 mRNAs in the single-nucleotide substitution library, Pearson *R* values were calculated based on the differences in tail-length changes predicted by the indicated model (key) and those measured for each single-nucleotide substitution, comparing between 0 h and 7 h post-progesterone treatment in frog oocytes. The top-performing model trained from each fold of the ten-fold cross-validation was used for predictions (ten models total for each architecture). *P* values are from one-sided t-tests. Box and whiskers indicate the 10th, 25th, 50th, 75th, and 90th percentiles. n.s.: not significant ($P \geq 0.01$). **c** PAL-AI-m prediction of the effects of single-nucleotide substitutions that disrupt either a CPE or a PAS (AAUAAA). For each position of an element, the average difference in tail-length change predicted upon substitution with one of the three non-wild-type possibilities is compared to that measured for mRNA injected into frog oocytes and collected 0 and 7 h post-progesterone treatment. Points for substitutions that disrupt a CPE are in blue, and points for those that disrupt a PAS are in red. Error bars indicate the standard error of the mean of all library variants that had the indicated substitutions. **d** PAL-AI-m prediction of the effects of single-nucleotide substitutions of the *mos.L* mRNA. Plotted for each single-nucleotide substitution is the difference in tail-length change predicted by PAL-AI-m compared to that measured for that mRNA variant injected into frog oocytes and collected 0 and 7 h post-progesterone treatment. On the sides are density distributions for the predicted and measured differences in tail-length changes. **e** PAL-AI-m prediction of the effects of single-nucleotide substitutions of the *tpx2.L* mRNA; otherwise, as in (**d**). **f** Impact of single-nucleotide substitutions on predicted and measured tail-length changes for the *mos.L* mRNA. Shown are results for the last 100 nt of the *mos.L* 3' UTR. The heatmaps indicate the predicted (top) and measured (bottom) differences in tail-length changes (DTLC) as a result of changing each nucleotide (*x* axis) to each of the three non-wild-type alternatives (*y* axis), with x indicating the non-substituted, wild-type nucleotide. The line plots indicate the predicted (top) and measured (bottom) maximum (red line) and minimum (blue line) of the mutational outcomes at each position. The logo plots indicate the predicted (top) and measured (bottom) importance of each nucleotide, with the height normalized to the negative value of the average outcome of the three possible substitutions. Dashed rectangles indicate sequence elements (the CPE and PAS) that are important for poly(A)-tail lengthening. **g** Impact of single-nucleotide substitutions on predicted and measured tail-length changes for the *tpx2.L* mRNA; otherwise as in (**f**).

Fig. 5

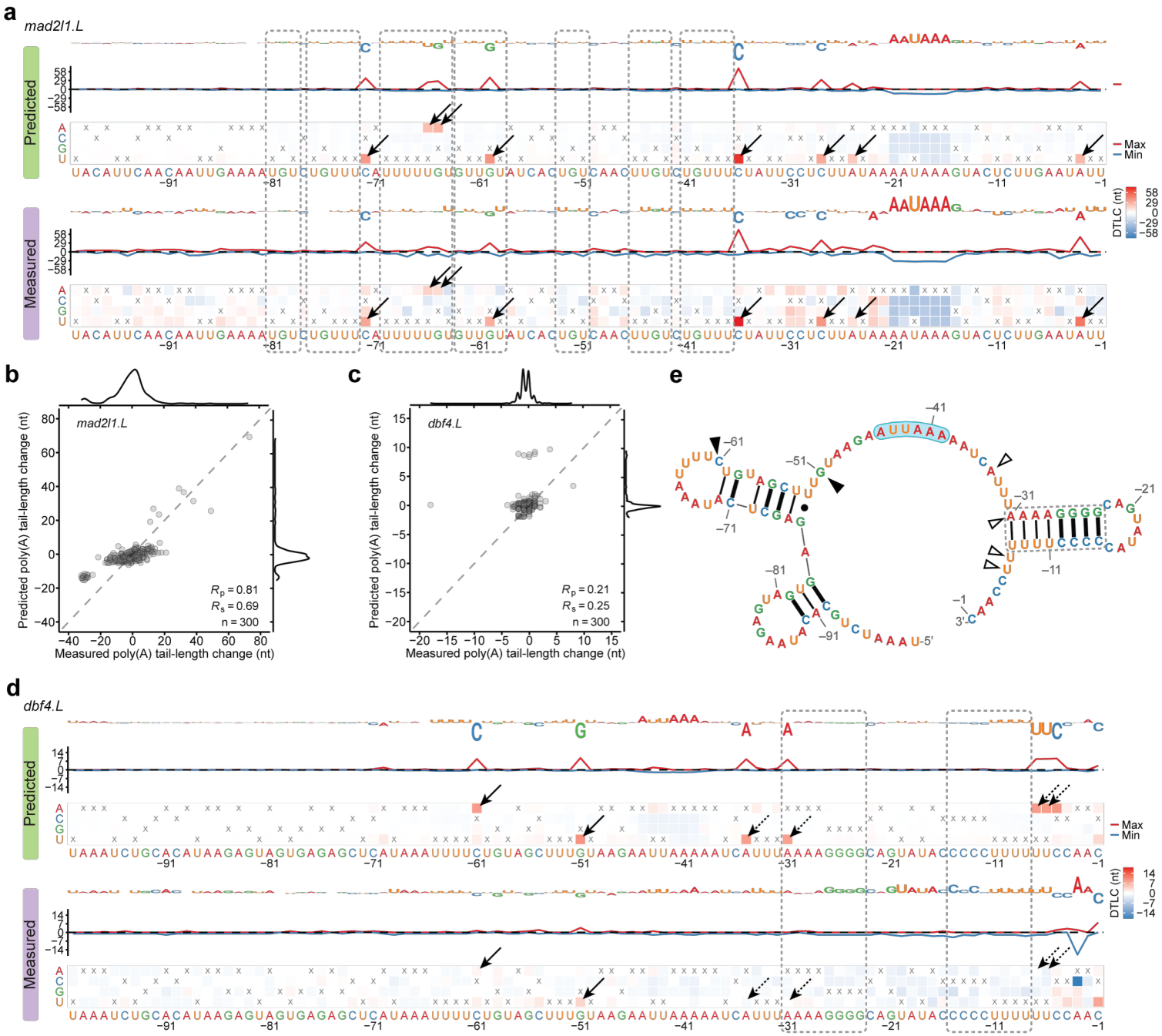


Fig. 5: Auxiliary motifs and structural accessibility influence tail-length changes. **a** Impact of single-nucleotide substitutions on predicted and measured tail-length changes for the *mad2l1.L* mRNA. Dashed rectangles indicate UGU/UGUUU elements that each appears to modestly promote poly(A)-tail lengthening. The arrows point to instances of new CPEs generated by substitutions and their associated increases in tail-length changes. Otherwise, this panel is as in **Fig. 4f**. **b** PAL-AI-m prediction of the effects of single-nucleotide substitutions of the *mad2l1.L* mRNA. Otherwise, this panel is as in **Fig. 4d**. **c** PAL-AI-m prediction of the effects of single-nucleotide substitutions of the *dbf4.L* mRNA.

Otherwise, this panel is as in **Fig. 4d**. **d** Impact of single-nucleotide substitutions on predicted and measured tail-length changes for the *dbf4.L* mRNA. Dashed rectangles indicate two regions predicted to pair with each other, forming the stem of a hairpin. Arrows indicate CPE-creating substitutions with tail-length changes either correctly (solid) or incorrectly (dashed) predicted by PAL-AI-m. **e** EternaFold-predicted maximum expected accuracy secondary structure of the last 100 nt of the *dbf4.L* 3' UTR. The dashed rectangle indicates the stem of the hairpin highlighted in (**d**). Triangles indicate positions where CPE-creating substitutions with tail-length changes either correctly (solid) or incorrectly (hollow) predicted by PAL-AI-m, as shown in (**d**). The PAS is shaded light blue. Base pairings are indicated with thick lines for G–C, thin lines for A–U, and a dot for the G–U wobble.

Fig. 6

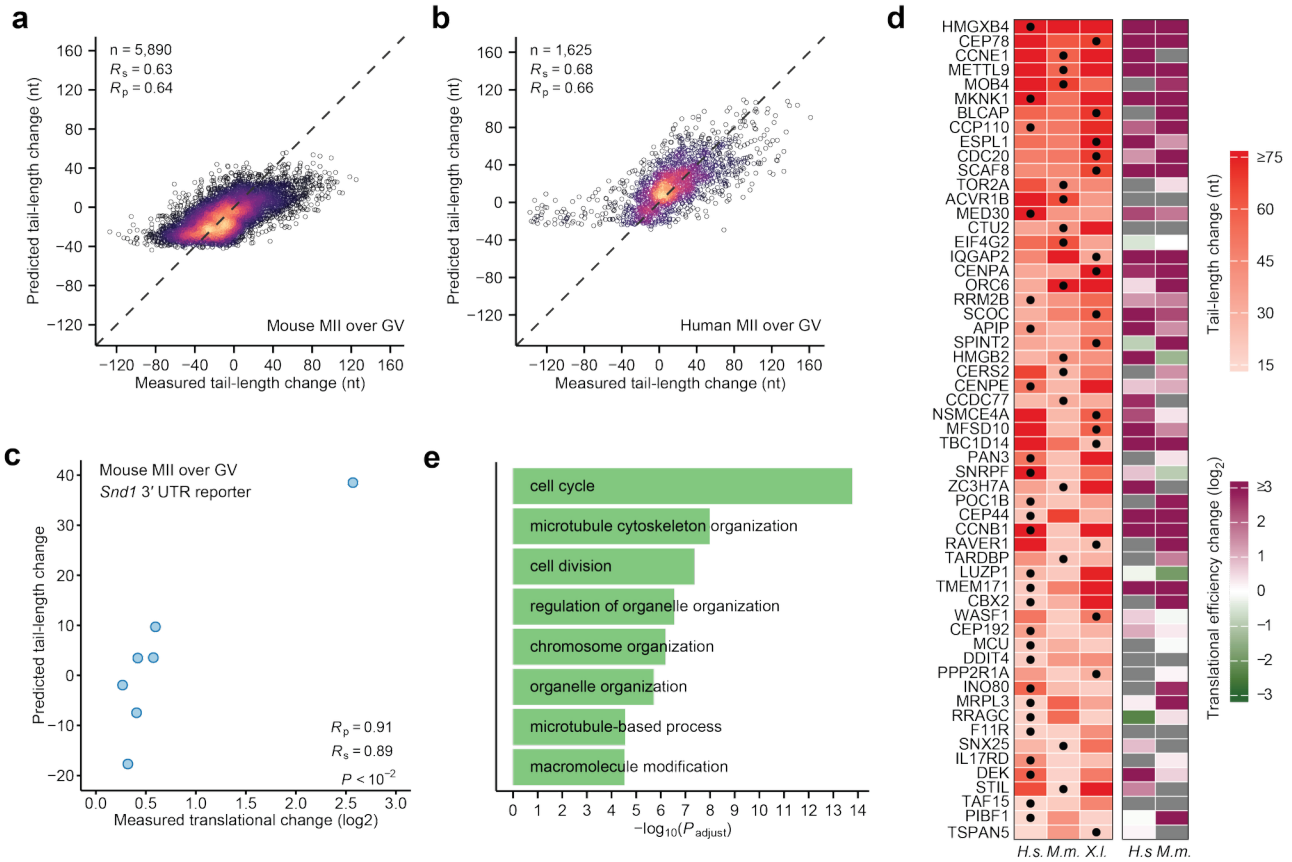


Fig. 6: Multi-species PAL-AI predicts tail-length changes of endogenous mRNAs of mouse and human oocytes.

a Performance of the multi-species PAL-AI model on mouse mRNAs. Plotted are the tail-length changes predicted by PAL-AI-ms to occur upon mouse GV-to-MII oocyte maturation as a function of the measured changes. Each point represents a unique poly(A) site of an endogenous mRNA. Otherwise, this panel is as in **Fig. 1c**. **b** Performance of the multi-species PAL-AI model on human mRNAs. Otherwise, this panel is as in (**a**). **c** Ability of predicted tail-length changes to explain translational efficiency changes. Plotted is the relationship between the multi-species PAL-AI model-predicted tail-length changes and translational efficiency changes measured for seven mRNA reporters upon mouse GV-to-MII oocyte maturation[6]. **d** Genes predicted or shown to have substantial mRNA tail-lengthening (≥ 15 nt) in human (*H.s.*), mouse (*M.m.*), and frog (*X.l.*) oocytes. Heatmaps show measured or predicted (dots) tail-length changes (left) and measured translational efficiency changes (right), comparing between human GV and MII oocytes, mouse GV and MII oocytes, and frog oocytes 0 and 7 h post-progesterone treatment. Only genes with the model-predicted values in one of the three species were shown. Gray indicates values not available. **e** Select ontologies enriched for genes with substantial mRNA tail-lengthening (≥ 15 nt, predicted, if not measured) in human, mouse, and frog oocytes.

Fig. 7

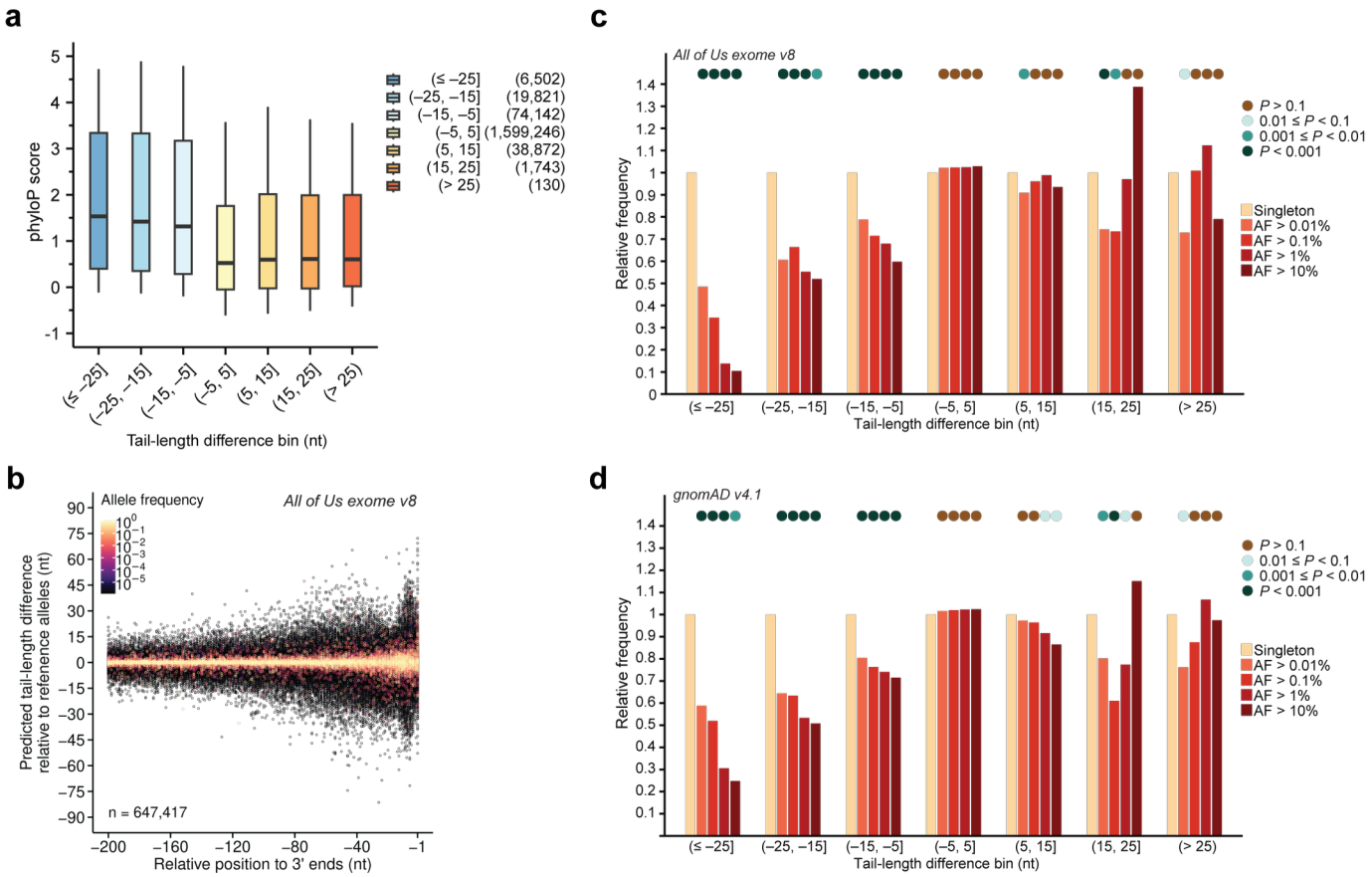
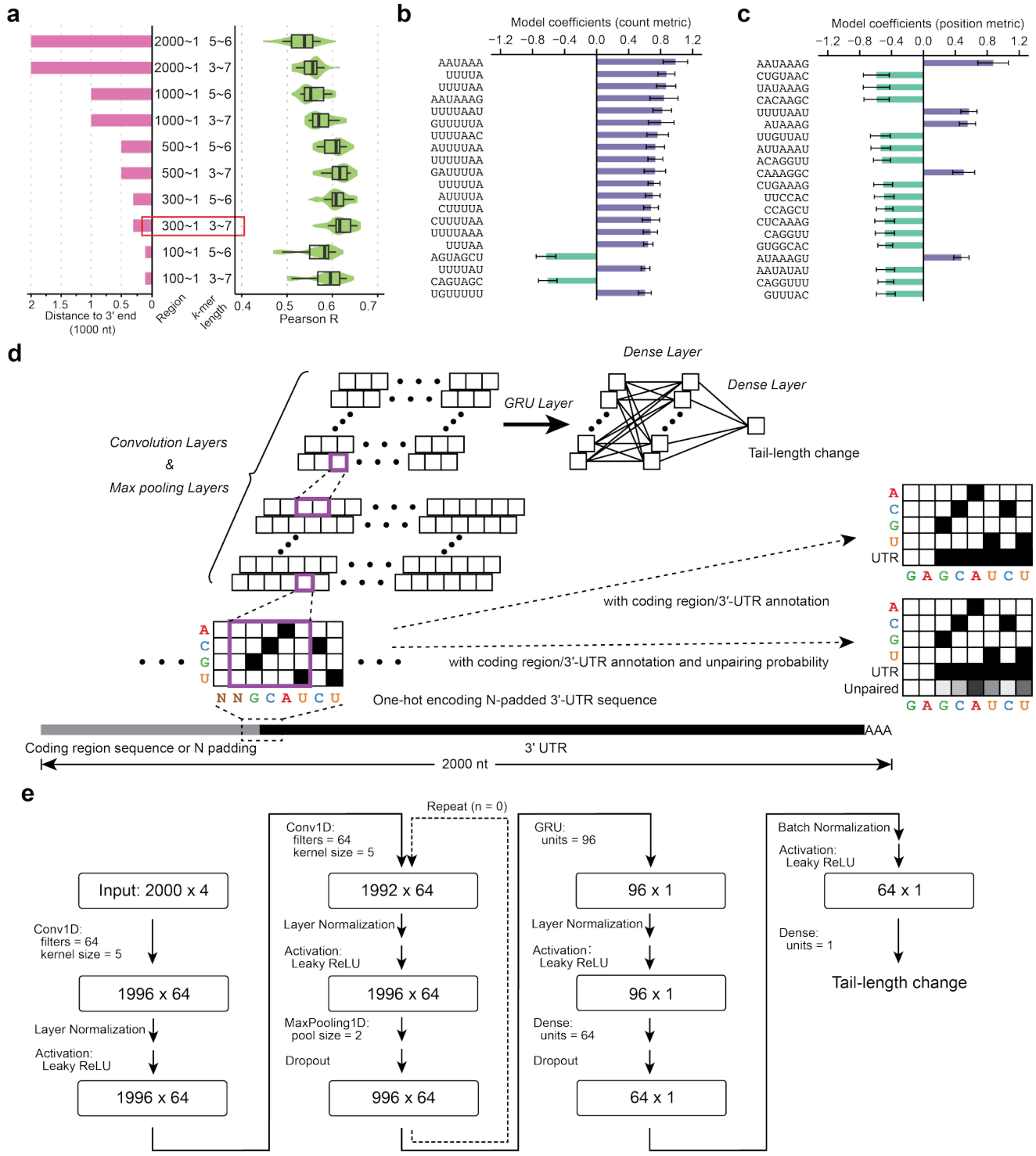


Fig. 7. Variants predicted to disrupt poly(A)-tail lengthening are selected against in the human population and throughout vertebrates.
a Negative selection among sequenced vertebrate species of variants predicted to be more disruptive of tail lengthening. Shown are distributions of phyloP scores[48,49] , grouped by the average PAL-AI-predicted differences in tail-length changes for individual nucleotides when substituted to each of the three alternative nucleotides. Only positions within the last 100 nt of 3' UTRs were included in this analysis. For the tail-length difference bin, parentheses indicate values not included, while square brackets indicate values included. Numbers in the parentheses next to the bins in the legend indicate the number of nucleotides in each bin. Statistical test results are reported in **Supplementary Fig. 9a**. **b** Predicted effects on the tail-length change for human variants. Shown is the PAL-AI-predicted difference in the tail-length change for each variant reported in the *All of Us* Research Program (exome callset v8) and that of the reference allele, plotted as a function of the variant position in its 3' UTR. Colors indicate allele frequencies. **c** Depletion in human 3' UTRs of alleles predicted to disrupt poly(A)-tail lengthening. Shown are relative frequencies of variants, grouped by predicted differences in tail-length change, among cohorts of variants with different allele frequencies reported in the *All of Us* Research Program (exome region v8). Only variants within the last 100 nt of 3' UTRs were included in this analysis. Binned *P* values (circles at the top) were from Fisher's exact tests performed for each cohort against singletons. **d** Same as (**c**), but for variants reported in gnomAD v4.1 dataset.

Supplementary Fig. 1

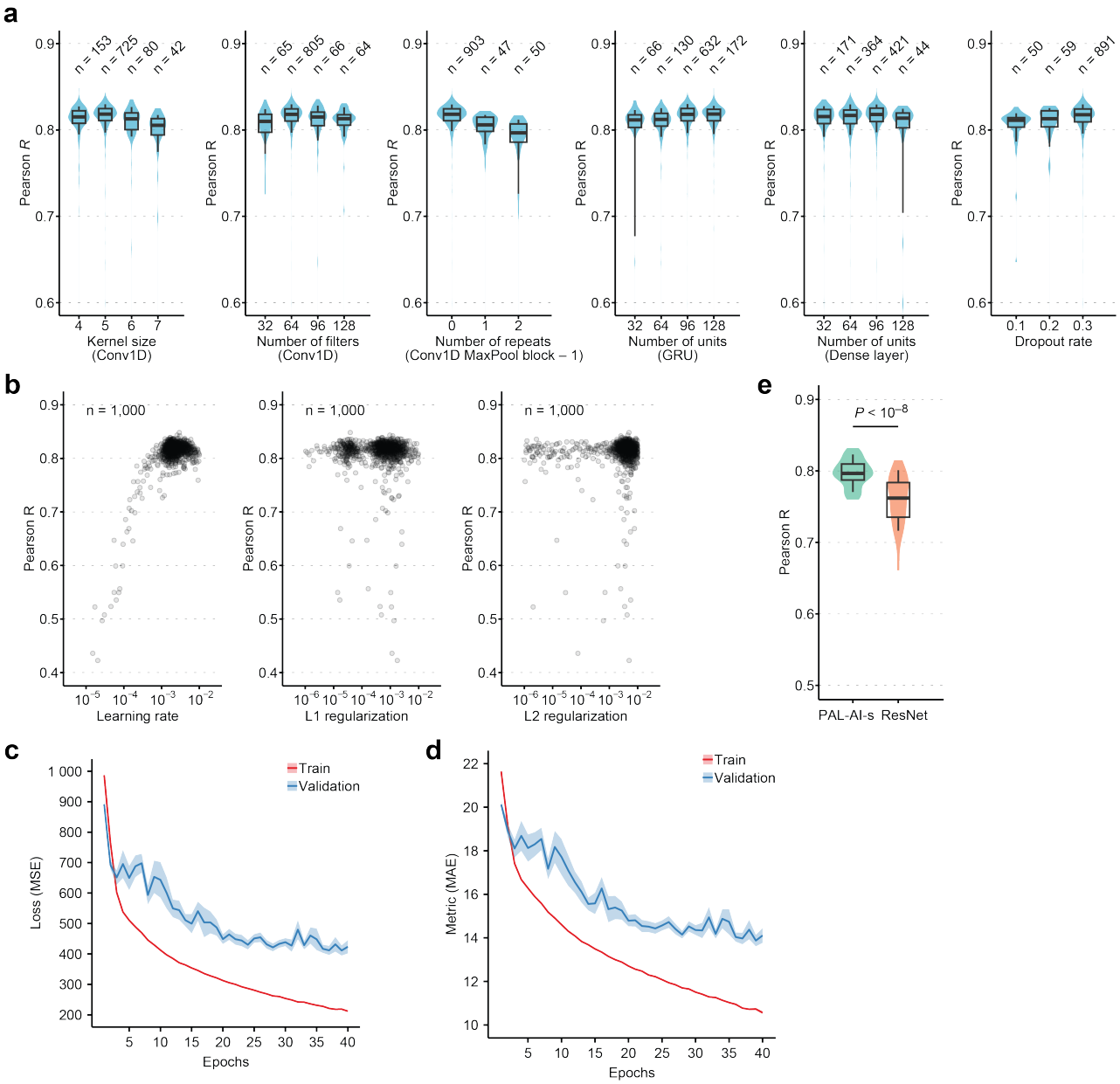


Supplementary Fig. 1: Additional data supporting the prediction of tail-length changes by multiple linear models and the architecture of PAL-AI.

a Prediction performance of the multiple linear regression models trained on different input regions of mRNA and configured with different lengths of *k*-mers; otherwise, as in **Fig. 1f**. **b** Top-ranked features as the number of *k*-mers from the best-performing multiple linear

regression models. Plotted are mean coefficients of the 20 top-ranked (by absolute value) features as the number of *k*-mers from the model. Error bars indicate standard deviation of coefficients of the 50 models from 10-fold cross-validation. **c** Top-ranked features as the position metric of *k*-mers from the best-performing multiple linear regression models. As in (**b**), except plotted are mean coefficients of the 20 top-ranked (by absolute value) features as the position metric (sum of the inverse of the distance between each *k*-mer and the 3' end) of *k*-mers from the model. **d** The architecture of PAL-AI for predicting tail-length changes. **e** Parameters of the architecture of PAL-AI. Shown for each layer are the name and parameters of the layer, and in boxes are the dimensions of the input and output matrices. The dashed arrow indicates repeats of the convolutional module. Different numbers of repeats were tested during hyperparameter searches.

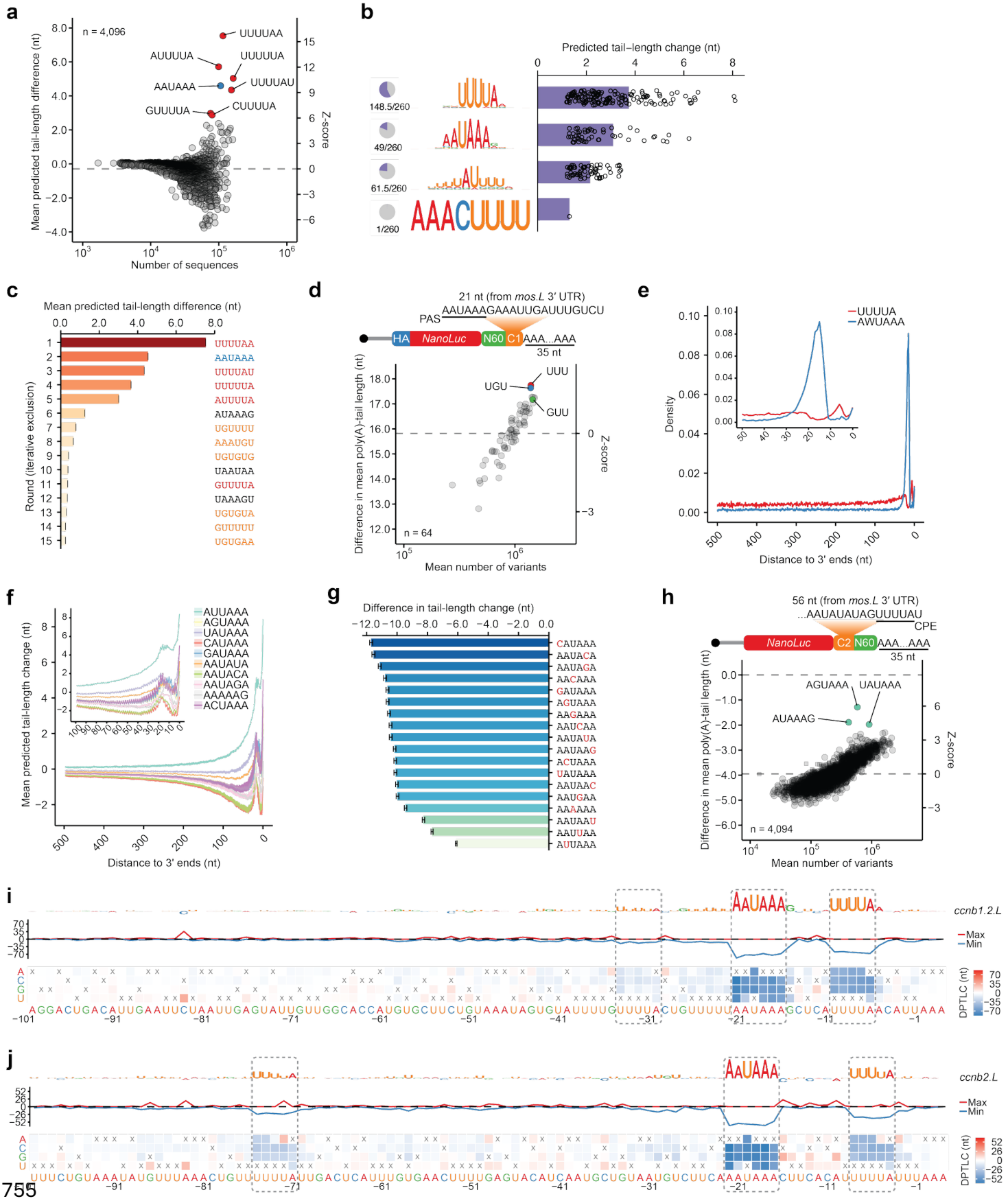
Supplementary Fig. 2



Supplementary Fig. 2: Hyperparameter optimization reveals optimal configuration parameters for PAL-AI.
a Impact of discrete hyperparameters on PAL-AI performance. Shown are distributions of R_p values evaluating the correspondence between predicted and measured tail-length changes for different hyperparameter values in validation data. Numbers at the top indicate sample (trial) sizes for each condition. Box and whiskers indicate the 10th, 25th, 50th, 75th,and 90th percentiles. **b** Relationship between continuous hyperparameters and PAL-AI performance. Scatter plots show R_p values as a function of learning rate (left), L1 regularization (middle), and L2 regularization (right). Each point represents an individual model with specific hyperparameter values. **c** Learning curves showing training and validation loss. Mean squared error (MSE) is plotted against training epochs for both

training (red) and validation (blue) datasets. Results represent means from 50 models (5 repeats for each fold in 10-fold cross-validation). Shaded areas indicate standard error of the mean. **d** Performance metrics during PAL-AI model training. Mean absolute error (MAE) is plotted against training epochs for both training (red) and validation (blue) datasets, analyzed as in (**c**). **e** Comparison of model architectures. Shown are distributions of R_p values evaluating the correspondence between predicted and measured tail-length changes for PAL-AI-s or ResNet implementations (with input sequence length of 2000 nt, 3' UTR only). Ten-fold cross-validation of each model was repeated five times, generating 50 R_p values. Box and whiskers indicate the 10th, 25th, 50th, 75th, and 90th percentiles. *P* value: one-sided t-test.

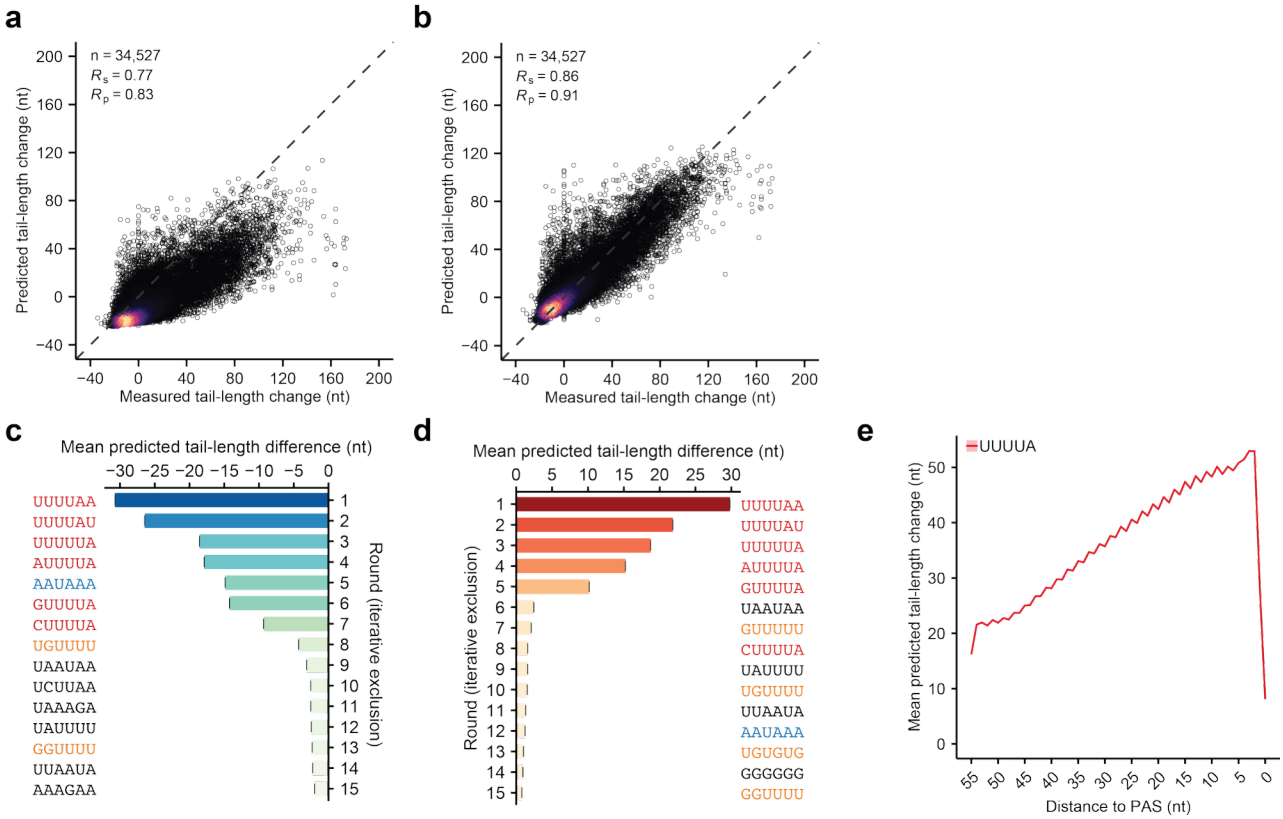
Supplementary Fig. 3



Supplementary Fig. 3: Additional data supporting sequence elements and contextual features learned by PAL-AI.

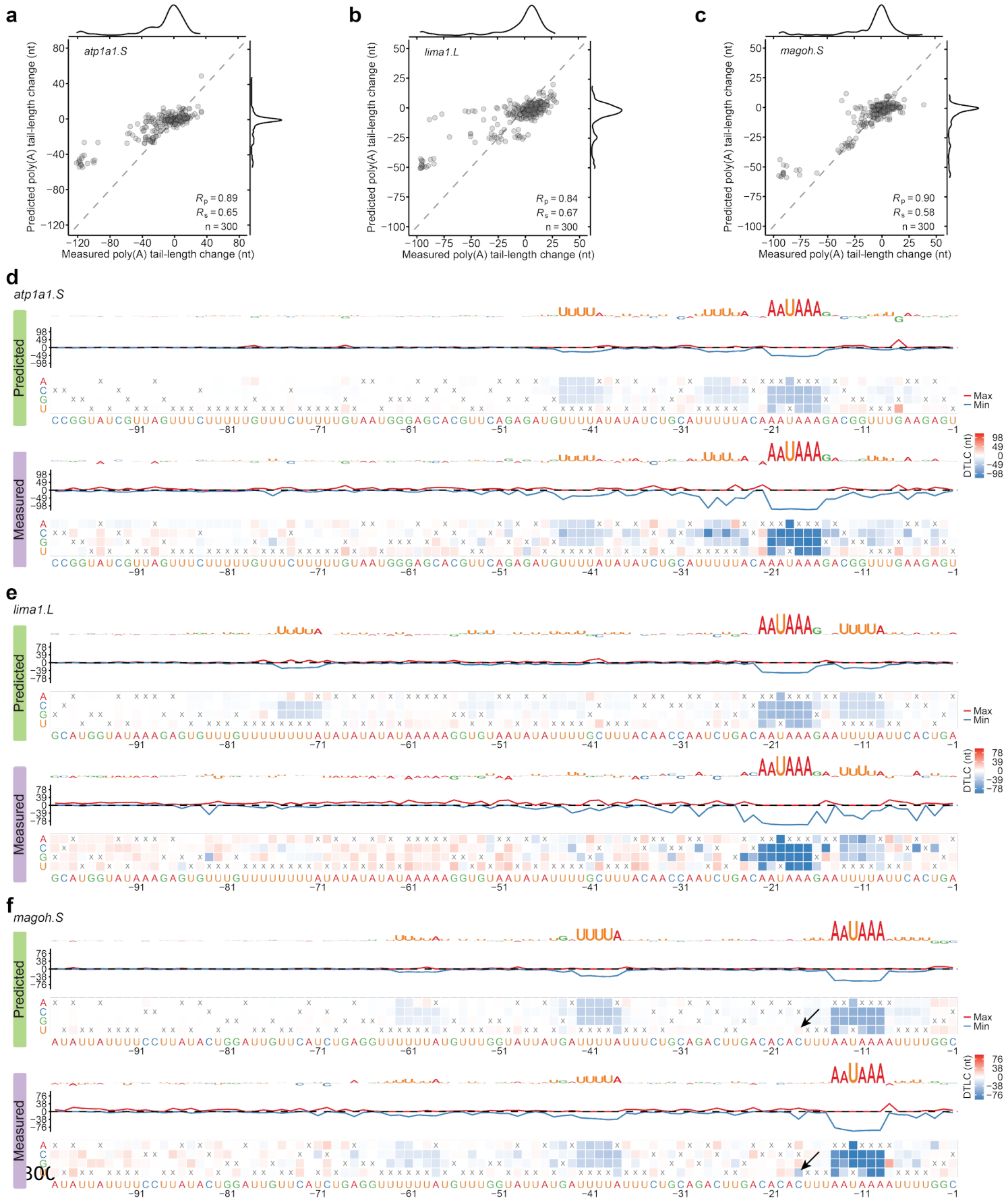
a As in **Fig. 2a**, except for consequences of gain of sequence motifs. **b** As in **Fig. 2b**, except for 8-mers most associated with the largest differences in predicted tail-length change between mutants in which 8-mers were gained and the wild-type. **c** As in **Fig. 2c**, except for top 6-mers associated with increased predicted tail-length change when gained in the in silico mutagenesis analysis. **d** Tail-length changes associated with each 3-mer within the 3' UTRs of the N60-PAS^{mos} library[18], comparing between 0 and 7 h post-progesterone treatment. Plotted for each 3-mer are differences in mean tail lengths observed for mRNAs with that 3-mer. Only variants containing one CPE (UUUUA) were included in this analysis. **e** Densities of the PAS and the CPE along the last 500 nt of frog mRNA 3' UTRs. The inset shows the results for the last 100 nt of the 3' UTR. **f** As in **Fig. 2e**, except for PAS variants. All insertions that would disrupt an AAUAAA were excluded. **g** Reduced tail-length change for PAS variants. Shown are the mean differences in predicted tail-length change when replacing the canonical PAS (AAUAAA) with the indicated PAS variant. Nucleotides that differ from the canonical PAS are colored in red. Only variants within the last 300 nt of 3' UTRs were included in this analysis. Error bars: standard error of the mean. **h** Tail-length changes associated with each 6-mer within the 3' UTRs of the CPE^{mos}-N60 library, comparing between 0 and 5 h post-progesterone treatment[18]. Plotted for each 6-mer are differences in mean tail lengths observed for mRNAs with that 6-mer. Variants containing any canonical PAS (AAUAAA or AUUAAA) were excluded from this analysis. **i** Impact of single-nucleotide substitutions on predicted tail-length changes for the *ccnb1.2.L* mRNA 3' UTR; otherwise, as in **Fig. 2g**. **j** Impact of single-nucleotide substitutions on predicted tail-length changes for the *ccnb2.L* mRNA 3' UTR; otherwise, as in **Fig. 2g**.

Supplementary Fig. 4



Supplementary Fig. 4. Additional data supporting the prediction of tail-length changes for synthetic mRNAs in frog oocytes.
a Performance of the PAL-AI model trained on endogenous frog mRNAs when tested on injected mRNA reporters. Each point represents a variant in the N60(LC)-PAS^{mos} library with at least 50 poly(A) tags in each of the two datasets, which were from oocytes 0 and 7 h post-progesterone treatment; otherwise, as in **Fig. 1c**. **b** Performance of a model trained from data generated from the injected N60(LC)-PAS^{mos} mRNA library and tested using 10-fold cross-validation. The 34,527 variants in the N60(LC)-PAS^{mos} library with at least 50 poly(A) tags in each of the two datasets, which were from oocytes 0 and 7 h post-progesterone treatment, were used in this analysis; otherwise, as in **Fig. 1c**. **c** Top 6-mers associated with decreased predicted tail-length change in the in silico mutagenesis analysis with a model trained with data generated from the injected N60(LC)-PAS^{mos} mRNA library. Otherwise, this panel is as in **Fig. 2c**. **d** As in (**c**), except for top 6-mers associated with increased predicted tail-length change when gained in the in silico mutagenesis analysis. **e** Positional effects for the CPE predicted by a model trained with data generated from the injected N60(LC)-PAS^{mos} mRNA library. Otherwise, this panel is as in **Fig. 2e**.

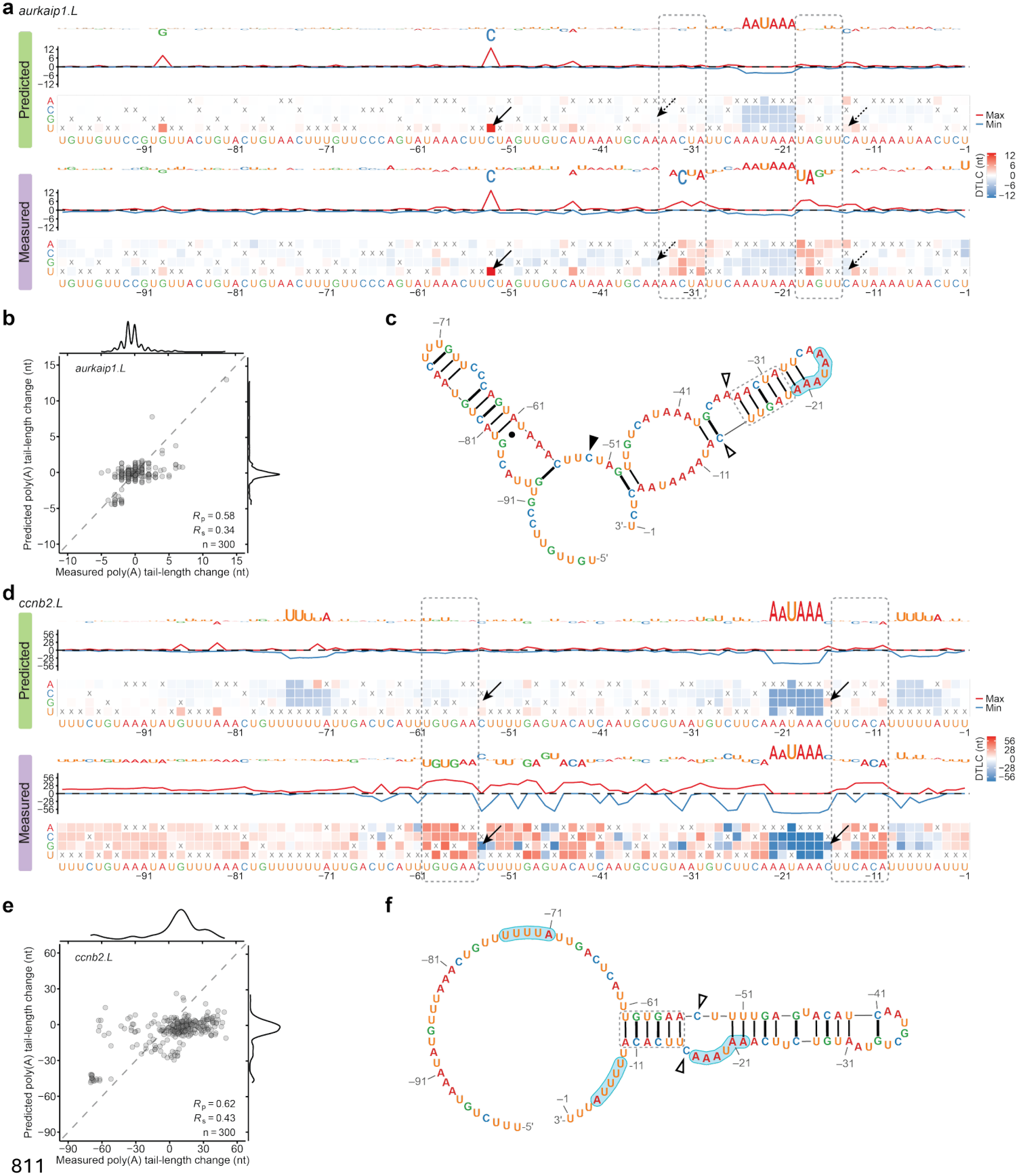
Supplementary Fig. 5



Supplementary Fig. 5: Additional experimental validation of predicted effects of single-nucleotide substitutions.

a PAL-AI prediction of the effects of single-nucleotide substitutions of the *atp1a1.S* mRNA. Otherwise, this panel is as in **Fig. 4d**. **b** As in (**a**), but for the *lima1.L* mRNA. **c** As in (**a**), but for the *magoh.S* mRNA. **d** Impact of single-nucleotide substitutions on predicted and measured tail-length changes for the *atp1a1.S* mRNA. Otherwise, this panel is as in **Fig. 4f**. **e** As in (**d**), but for the *lima1.L* mRNA. **f** As in (**d**), but for the *magoh.S* mRNA. Arrows indicate predicted and measured differences in tail-length changes that result from a C-to-U substitution that creates a CPE that overlaps with the PAS.

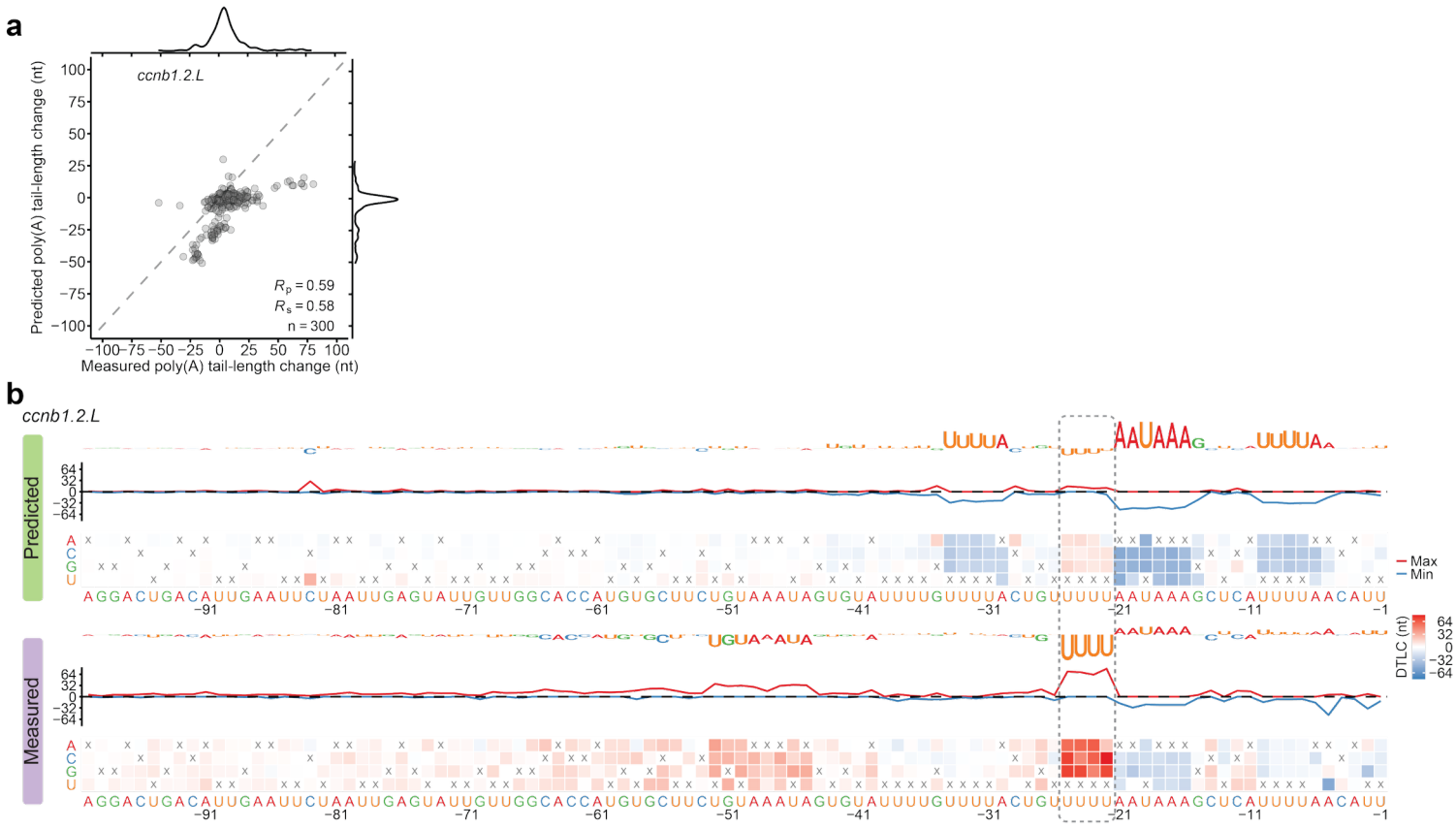
Supplementary Fig. 6



Supplementary Fig. 6: Additional analyses supporting tail-length changes affected by structural accessibility.

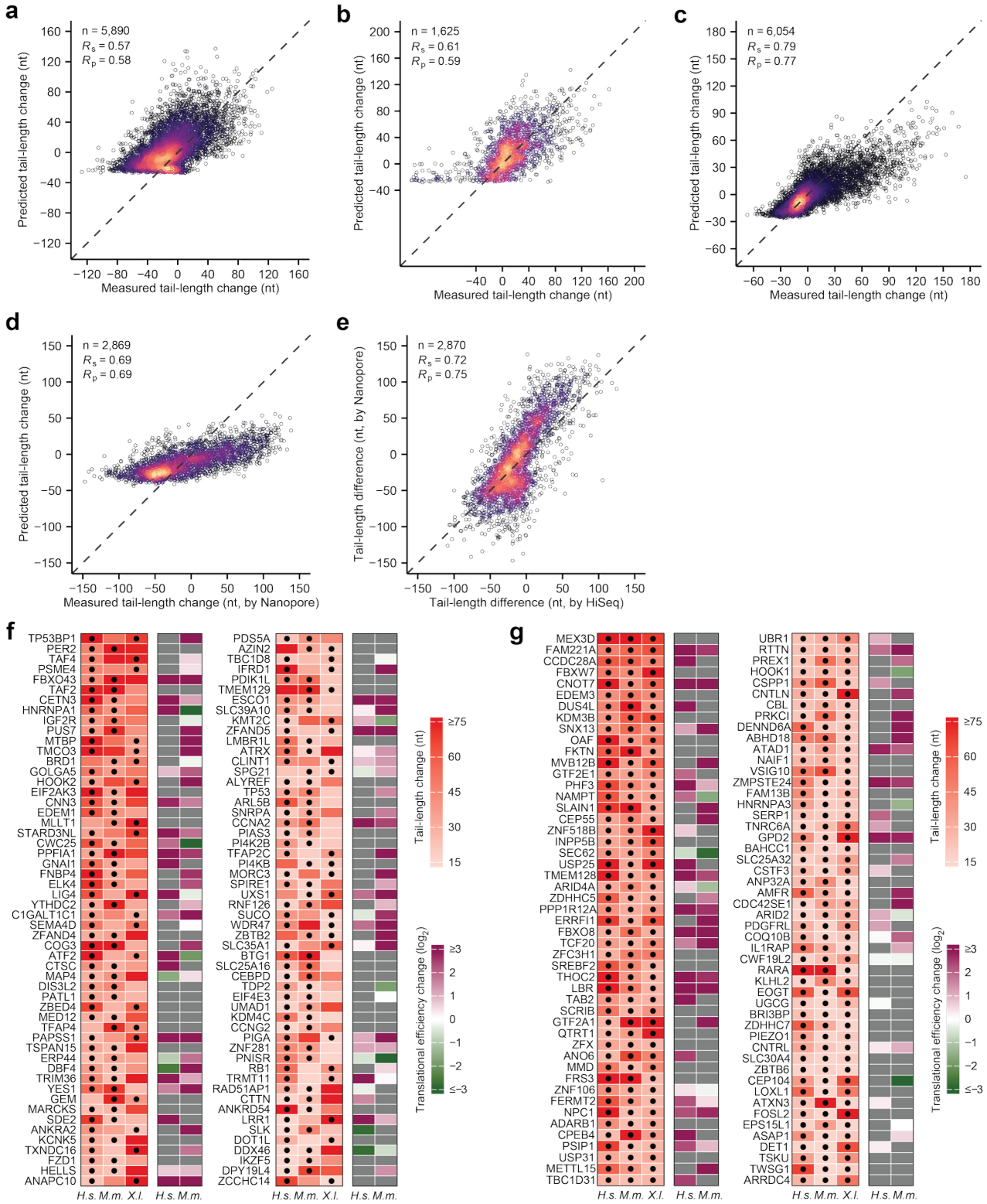
a Impact of single-nucleotide substitutions on predicted and measured tail-length changes for the *aurkaip1.L* mRNA. Arrows indicate predicted and measured differences in tail-length changes of either a C-to-U substitution (solid) that creates a CPE or two substitutions (A-to-G at position –35 and C-to-U at position –14, dashed) that would increase the length of a predicted stem (dashed rectangles) expected to occlude the PAS. Otherwise, this panel is as in **Fig. 4f**. **b** PAL-AI prediction of the effects of single-nucleotide substitutions of the *aurkaip1.L* mRNA. Otherwise, this panel is as in **Fig. 4d**. **c** EternaFold-predicted maximum expected accuracy secondary structure of the last 100 nt of the *aurkaip1.L* 3' UTR. The dashed rectangle indicates the stem of a hairpin, as highlighted in (**a**). The triangles indicate positions of either a C-to-U substitution (solid) that creates a CPE in a predicted loop region with a tail-length change correctly predicted by PAL-AI or two substitutions (A-to-G at position –35 and C-to-U at position –14, hollow) that would increase the length of a predicted stem expected to occlude the PAS. **d** As in (**a**), but for the *ccnb2.L* mRNA. Arrows indicate predicted and measured differences in tail-length changes of C-to-G substitutions at positions –54 and –16, both of which create the potential for a C–G base pair that would increase the length of a predicted stem indicated with dashed rectangles. **e** As in (**b**), but for the *ccnb2.L* mRNA. **f** As in (**c**), but for the *ccnb2.L* mRNA. Shaded in blue are two CPEs and a PAS. The triangles indicate two positions where C-to-G substitutions would extend the predicted stem indicated with the dashed rectangle.

Supplementary Fig. 7



Supplementary Fig. 7: Additional experimental validation of predicted effects of single-nucleotide substitutions, exemplifying consequences of CPE-PAS overlap. **a** PAL-AI prediction of the effects of single-nucleotide substitutions of the *ccnb1.2.L* mRNA. Otherwise, this panel is as in **Fig. 4d**. **b** Impact of single-nucleotide substitutions on predicted and measured tail-length changes for the *ccnb1.2.L* mRNA. The dashed rectangle indicates a CPE that overlaps with a PAS. Otherwise, this panel is as in **Fig. 4f**.

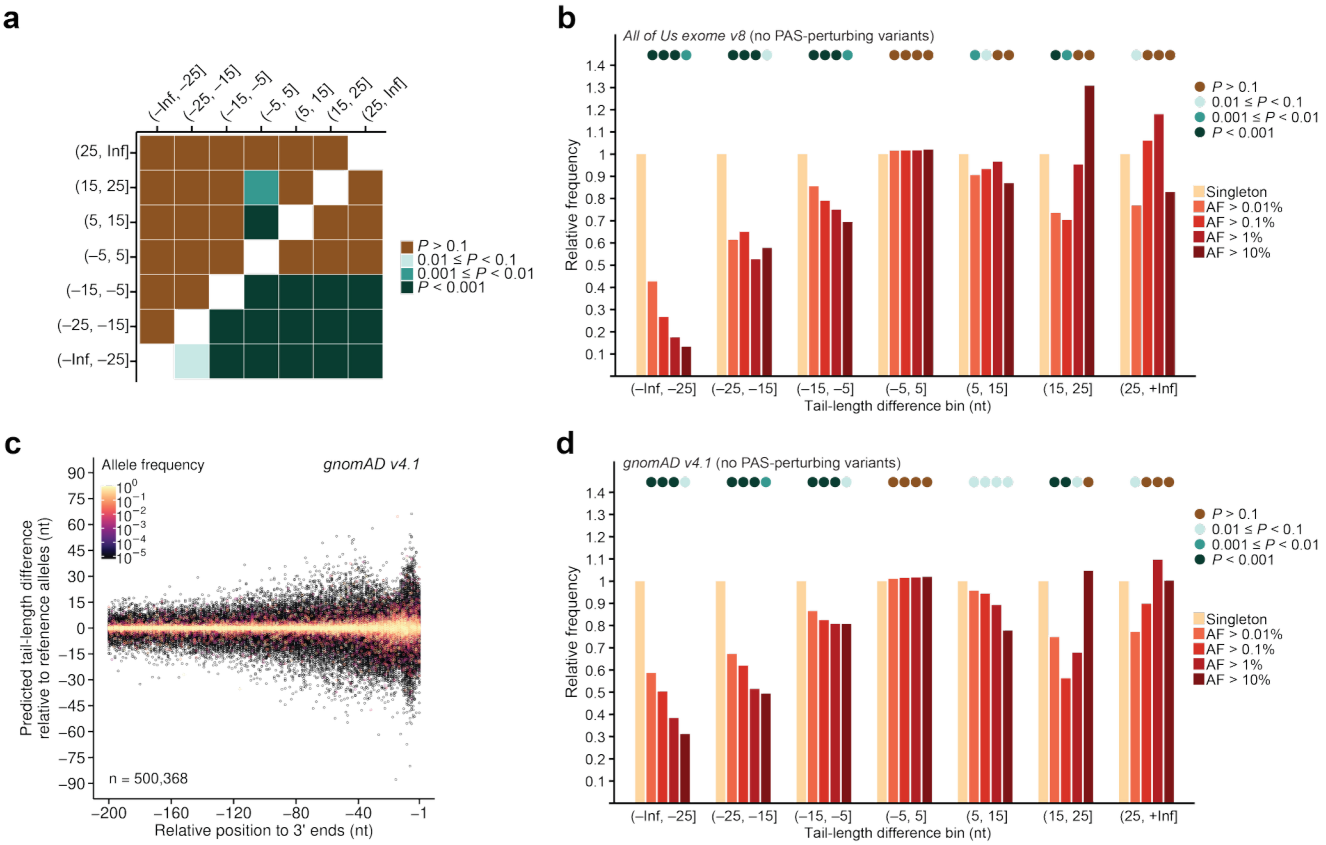
Supplementary Fig. 8



Supplementary Fig. 8: Additional data supporting the prediction of tail-length change for endogenous mRNAs in mouse and human oocytes.

843 **a** Performance of PAL-AI on mouse mRNAs. This panel is as in **Fig. 6a**, except that the
844 predictions were made with the PAL-AI-s model trained solely on frog oocyte endogenous
845 mRNAs. **b** Performance of PAL-AI on human mRNAs. This panel is as in **Fig. 6b**, except that
846 the predictions were made with the PAL-AI-s model trained solely on frog oocyte
847 endogenous mRNAs. **c** Performance of the PAL-AI-ms model on frog mRNAs. Otherwise,
848 this panel is as in **Fig. 6a**. **d** Performance of the PAL-AI-ms model on mouse mRNAs with
849 data generated from ONT in a published study[9]. This panel is as in Fig. 6a, except that the
850 measured tail-length changes during mouse GV-to-MII oocyte maturation were obtained
851 from ONT. **e** Correspondence between HiSeq and ONT measurements. Shown are tail-
852 length differences measured either by HiSeq (x axis) or ONT (y axis) during mouse GV-to-MII
853 oocyte maturation. Otherwise, this panel is as in **Fig. 1c**. **f** Genes predicted or shown to
854 have substantial mRNA tail-lengthening. This panel is as in **Fig. 6d**, except genes with PAL-
855 AI-predicted values in two of the three species are shown. **g** Genes predicted to have
856 substantial mRNA tail-lengthening. This panel is in (**f**), except genes with PAL-AI-predicted
857 values in all three species are shown.
858
859

Supplementary Fig. 9



Supplementary Fig. 9. Additional data supporting negative selection of human variants predicted to disrupt poly(A)-tail lengthening.
a Pairwise comparison of phyloP scores. Shown are binned *P* values from one-sided Mann–Whitney U tests of phyloP scores between groups with indicated differences tail-length changes predicted by PAL-AI, testing the alternative hypothesis that the mean phyloP scores of the group indicated on the y axis is greater than that indicated on the x axis. **b** Depletion among frequent alleles predicted to disrupt poly(A)-tail lengthening, when excluding effects on PAS elements. This panel is as in **Fig. 7c**, except variants that introduced or eliminated PAS elements were excluded. **c** Analysis of gnomAD data, showing predicted effects on tail-length change for human variants. This panel is as in **Fig. 7b**, but for human variants in the gnomAD v4.1 dataset. **d** Analysis of gnomAD data, showing depletion among frequent alleles predicted to disrupt poly(A)-tail lengthening, when excluding effects on PAS elements. This panel is as in (**b**), but for human variants in the gnomAD v4.1 dataset.

Methods

Annotations of poly(A) sites and 3' UTRs. Annotations of poly(A) sites and 3' UTRs of frog, mouse, and human oocytes were obtained from a previous study[18].

Linear regression models for poly(A) tail-length changes. Poly(A) tail-length changes for each isoform of frog oocyte endogenous mRNAs were calculated as the difference between median tail lengths measured at 0 and 7 h post-progesterone treatment, requiring at least 50 poly(A) tags at both time points, as reported previously[18]. Only 3' UTRs not annotated as "uncertain" in previously curated annotations[18] were included.

K-mer features were defined as the positions and counts of all *k*-mers of in the length of either 5–6 nt or 4–7 nt within the specified regions of the 3' UTR of each mRNA isoform were used as features for the linear regression model. For isoforms with 3' UTRs shorter than the length of the specified region, the entire 3' UTR was examined. Isoforms with 3' UTRs shorter than 10 nt were excluded. Each 3' UTR was appended with the tri-nucleotide sequence "AAA" before *k*-mer extraction. This sequence was treated as part of the 3' UTR for both *k*-mer feature calculation and the specified region length (e.g., for a the "last 1000 nt" region, the last 997 nt of the annotated 3' UTR were used before appending "AAA").

The number of each *k*-mer N^k was counted with the *oligonucleotideFrequency* function from the R package *Biostrings*[62]. A composite positional feature value P^k was computed for each *k*-mer to summarize its positional distribution relative to the 3' end:

$$P^k = \begin{cases} \sum_i \frac{1}{\max(d_i, 1)}, & i \notin \emptyset \\ 0, & i \in \emptyset \end{cases}$$

where d_i is distance from the 3' end to the *i*-th occurrence of the *k*-mer. The linear regression model predicting tail-length change L_i for isoform *i* was defined as:

$$L_i = c^0 + \sum_k c_N^k \cdot N_i^k + \sum_k c_P^k \cdot P_i^k$$

where c^0, c_N^k, and c_P^k are the model coefficients to be optimized. All features were scaled using *RobustScaler* from the Python *scikit-learn* package[63]. Data were randomly stratified into 10 folds using *StratifiedKFold*[63], maintaining the proportion of isoforms with 0, 1, 2, 3, or ≥ 4 CPE (UUUUA) elements within their 3' UTRs.

The model was implemented using a single Dense layer in *Tensorflow*[64] and trained with the Adam optimizer. Hyperparameters were tuned using the *Optuna*[65] framework. The negative R^2 (between measured and predicted tail-length changes) for the validation data was used as the objective to minimize for optimization. Final hyperparameters were: learning_rate = 0.0025169, l1_regularizer = 0.000166, l2_regularizer = 0.0001097, epochs =100, batch_size = 50, loss = 'mse', patience = 10. Model performance was evaluated in 10-fold cross-validation. In this procedure, the 9-fold data were used for model training and validation (further split into 90% for training and 10% for validation, stratified by CPE count) while the remaining fold was used for testing the model performance. After each epoch, the model was evaluated on the validation set. Training stopped early if validation loss did not improve over 10 consecutive epochs, and the best model (based on the lowest

917 validation loss) was saved. This training/validation procedure was repeated 5 times,
918 accounting for differences in weight initializations, and the model with the highest R^2
919 (between measured and predicted tail-length changes) for the test data was selected to
920 predict the remaining fold. The procedure was rotated across all 10 folds. Final model
921 performance was assessed by concatenating predictions from all test folds and comparing
922 them with the measured tail-length changes.
923
924 **PAL-AI for poly(A) tail-length changes**. The following datasets were used to train or
925 validate different PAL-AI models: 1) the frog endogenous mRNA dataset: same as that used
926 in the linear regression model; 2) the N60(LC)-PAS^{mos} library dataset: tail-length changes of
927 the N60(LC)-PAS^{mos} library mRNAs injected into frog oocytes, comparing median tail
928 lengths measured at 0 and 7 h post-progesterone treatment during oocyte maturation,
929 requiring at least 50 poly(A) tags at both time points; 3) the single-nucleotide mutagenesis
930 library dataset: tail-length changes of the single-nucleotide mutagenesis library mRNAs
931 injected into frog oocytes, comparing median tail lengths measured at 0 and 7 h post-
932 progesterone treatment during oocyte maturation, requiring at least 50 poly(A) tags at both
933 time points; 4) the mouse endogenous mRNA dataset: tail-length changes of mouse
934 endogenous mRNAs, comparing between median tail lengths measured in GV and MII
935 oocytes, requiring at least 50 poly(A) tags in both replicate-merged datasets as described
936 previously[18]; 5) the human endogenous mRNA dataset: tail-length changes of human
937 endogenous mRNAs, comparing between median tail lengths measured in GV and MII
938 oocytes, requiring at least 50 poly(A) tags in both datasets as described previously[18].
939 To encode sequences as input for PAL-AI, RNA sequences from the specified
940 regions of each mRNA isoform's 3' UTR were appended at the 3' end with the tri-nucleotide
941 sequence "AAA". For isoforms with 3' UTRs shorter than the specified region, the entire 3'
942 UTR was used, and "N"s were padded at the 5' end. Isoforms whose AAA-appended 3' UTRs
943 were shorter than 10 nt were excluded. Sequences were one-hot encoded into four
944 channels, with "N" represented by a zero vector. When coding regions were included, they
945 were padded to the 5' end of the AAA-appended 3' UTRs. If the total length was still shorter
946 than the required input length, "N"s were padded at the 5' end. A fifth channel was added,
947 where a value of 1 marked 3'-UTR positions and 0 marked coding or padded regions. When
948 secondary structure information was included, *RNAfold*[35] was used to predict structure
949 from the AAA-appended 3' UTR. The unpairing probability of each nucleotide was added as
950 a sixth input channel. Nucleotides from the coding region and padded "N"s were assigned
951 a value of 0 in this sixth channel. The inputs for different final PAL-AI models were
952 summarized in Table 1.
953 The base PAL-AI model PAL-AI-s integrated a convolutional neural network with a
954 recurrent neural network, in an architecture similar to Saluki[34]. It consisted of four
955 modules: 1) Input block, which included an input layer, a 1D convolution, a normalization
956 layer, and an activation layer. 2) Convolutional block, which included repeated structure,
957 with each repeat comprising a 1D convolution layer, a normalization layer, an activation
958 layer, a 1D maxpooling layer, and a dropout layer. 3) Recurrent block, which included a
959 recurrent layer (GRU, BiGRU, LSTM, or BiLSTM), a normalization layer, and an activation
960 layer. 4) Dense block, which contains a dense layer, a dropout layer, a normalization layer,

961 an activation layer, and a final dense layer with a single output for the tail-length change. All
962 convolution layers used the same number of filters and kernel size. The convolutional
963 layers, the recurrent layer, and the first dense layer used the same kernel initializer and
964 regularizer. All dropout layers had the same dropout rate. Layer normalization was used in
965 all blocks unless Leaky ReLU was the activation function, in which case batch
966 normalization was used in the dense block. All activation layers used the same function—
967 either SELU, SiLU (Swish), or Leaky ReLU. If SELU was used, the Lecun normal initializer
968 and Alpha Dropout were applied. Otherwise, the HE normal initializer was used.
969 The PAL-AI-s model was implemented using the Python *TensorFlow* package[64] and
970 trained with the Adam optimizer. Loss functions used were either the mean squared error
971 (MSE) or the mean absolute error (MAE). The hyperparameter tuning, cross-validation
972 strategy, and model performance evaluation followed the same procedures as the linear
973 regression model.
974 For the N60(LC)-PAS^{mos} library, PAL-AI-s used the same architecture. Sequences in
975 the 60-nt random region and constant region were appended with "AAA" and one-hot
976 encoded. Training and evaluation were performed as described for endogenous mRNAs.
977 The PAL-AI-m and PAL-AI-ms models shared the same base architecture as PAL-AI-
978 s, with three key modifications. First, an additional input feature was included to indicate
979 the dataset group for each mRNA sequence. Second, rather than a single output, the final
980 dense layer outputs multiple values, with one per dataset group. Third, a custom gathering
981 layer was added after the dense layer to select the appropriate group-specific output
982 based on the input dataset label. For the PAL-AI-m model, the frog endogenous mRNA
983 dataset (output head 1) and the N60(LC)-PAS^{mos} library dataset (output head 2) were used.
984 For the PAL-AI-ms model, the frog endogenous mRNA dataset (output head 1), the human
985 endogenous mRNA dataset (output head 2), and the mouse endogenous mRNA dataset
986 (output head 3) were used. All input sequences were encoded as described for PAL-AI-s.
987 Training and evaluation followed the same procedure as PAL-AI-s, with the following
988 differences to accommodate multiple groups. First, when splitting the training and
989 validation set, the validation set contributed 10% of its data for the group with the smallest
990 dataset, and validation sets from all other groups were downsampled to match this
991 number. Second, for training, data from larger groups were similarly downsampled per
992 epoch to match the smallest group, and batches were alternated across groups. Third, the
993 average of negative R^2 values (between measured and predicted tail-length changes)
994 across validation groups was used as the objective to minimize during hyperparameter
995 optimization.
996 A ResNet-based variant of PAL-AI was also tested, replacing the second module of
997 PAL-AI-s with a dilated convolutional ResNet architecture[36] comprising seven groups with
998 dilation rates of 1, 2, 4, 8, 4, 2, and 1. Each group contained four blocks, each consisting of
999 a dilated convolution layer, layer normalization, an activation layer, a second dilated
1000 convolution layer, another layer normalization, a residual skip connection, and a final
1001 activation layer.
1002 For all PAL-AI models, an optional input representing the initial tail length could be
1003 provided. When included, this value was concatenated to the input of the final dense layer.

Although incorporating this feature improved performance for some datasets, it reduced generalizability, as initial tail-length measurements were not available for all mRNAs.

For all predictions made with PAL-AI, input sequences were encoded using the same scheme as during model training. To compare model performance on the single-nucleotide mutagenesis library, predictions were generated using either PAL-AI-s (endogenous mRNA-trained) or PAL-AI-m (endogenous mRNA-specific output head 1). Unless otherwise specified, final predictions were computed as the average output across the best-performing models from each fold of the 10-fold cross-validation.

Hyperparameters for each PAL-AI model were optimized using *Optuna*[65], with final values summarized in Table 1.

Table 1: Parameters used in each PAL-AI model.

	PAL-AI-s (frog endogenous mRNAs)	**PAL-AI-s [N60(LC)-PAS^{mos} library]**	**PAL-AI-m**	**PAL-AI-ms**	**PAL-AI (ResNet)**
Length of input sequences (nt)	2000	84	2000	2000	2000
Use of coding sequences	Yes	No	Yes	Yes	No
Use of group label input	No	No	Yes	Yes	No
Number of filters for 1D convolution layers	64	96	64	64	96
Kernel size for 1D convolution layers	5	4	4	5	4
Number of repeats of the convolutional block	1	1	1	2	Not applicable
Number of units of the dense layer	64	96	128	128	32
Size of the maxpooling layer	2	2	2	2	Not applicable
Recurrent layer	GRU	GRU	GRU	GRU	GRU
Number of units of the recurrent layer	96	128	128	128	64
Activation function	Leaky ReLU	Leaky ReLU	Leaky ReLU	Leaky ReLU	Leaky ReLU

Kernel regularizer L1	0.001	0.000001798	0.0000628	0.0000043	0.006823
Kernel regularizer L2	0.0045	0.001257	0.00019765	0.00344	0.000006507
Dropout rate	0.3	0.2	0.2	0.2	0.2
Learning rate	0.003	0.00030457	0.001934	0.00068	0.0005537
Loss function	MSE	MSE	MSE	MAE	MAE
Batch size	128	64	128	32	64
Epochs	40	100	100	100	100
Trainable parameters	75,713	138,689	109,826	135,043	2,116,129
Non-trainable parameters	128	192	256	256	10,816

In silico mutagenesis (ISM). For frog and human endogenous mRNAs, all 3' UTRs were analyzed except those previously annotated as "uncertain"[18]. For the N60(LC)-PAS^{mos} library, all 3' UTRs in the curated reference were used. Input sequences were constructed and encoded as described for PAL-AI. Tail-length changes of the wild-type sequences were predicted using PAL-AI. For ISM, each nucleotide within 3' UTRs was individually substituted with each of the three possible alternative nucleotides (the fifth channel annotating the 3' UTR was not altered), and the tail-length change of each mutant was predicted with corresponding PAL-AI models: PAL-AI-s (frog endogenous mRNA-trained) for frog endogenous mRNAs, PAL-AI-s (N60(LC)-PAS^{mos} library-trained) for N60(LC)-PAS^{mos} library mRNAs, and PAL-AI-ms (human-specific output head 2) for human endogenous mRNAs. The difference between the predicted tail-length change of the wild-type and each mutant was then calculated.

To examine the tail-length effect of losing or gaining a *k*-mer, differences between the predicted tail-length change of the wild-type and those of all single-nucleotide substitution mutants that disrupted the original *k*-mer or introduced a new k-mer were averaged across all instances of that *k*-mer within the last 300 nt of 3' UTRs from all mRNA isoforms.

In some cases, *k*-mer-associated differences in predicted tail-length changes were examined in an iterative exclusion process. Only mutations within the last 300 nt of 3' UTRs were considered. In each round of the iteration, mutations that disrupted (for *k*-mer loss) or introduced (for *k*-mer gain) any *k*-mer already present in an exclusion list (initially empty) were ignored. The average difference in predicted tail-length change between the wild-type and the mutant was then calculated for all *k*-mers that were not in the exclusion list. The *k*-mer with the largest tail-length increase (for *k*-mer gain) or decrease (for *k*-mer loss) was recorded and added to the exclusion list before starting the next round. The iteration continued until no *k*-mer showed a statistically significant increase (or decrease) in predicted tail-length change (Welsh's t-test; Bonferroni-adjusted $P > 0.05$).

Motif logos associated with predicted tail-length changes were generated as previously described[18].

1047 **Motif insertion analysis.** Filtering and encoding of input sequences were performed as in
1048 the ISM analysis. To insert a motif, the sequence of the same length at a specified position
1049 within the 3' UTR of an mRNA was replaced with the motif. The fifth channel annotating the
1050 3' UTR was left unchanged. PAL-AI was then used to predict the tail-length changes of both
1051 the wild-type and motif-inserted mutant. The difference between these predictions was
1052 used as the insertion outcome. When analyzing the relative positioning of the PAS and CPE
1053 elements, only mRNAs whose 3' UTRs contained at least a PAS within the last 100 nt were
1054 included. For these mRNAs, the relative distance was calculated between each CPE and
1055 the PAS closest to the 3' end.
1056
1057 **Reanalysis of reporter libraries for *k*-mer-associated tail-length changes**. Poly(A) tail-
1058 length measurements of the frog oocyte-injected N60-PAS^{mos} library at 0 and 7 h post-
1059 progesterone treatment and CPE^{mos}-N60 library at 0 and 5 h post-progesterone treatment
1060 were obtained from a previous study[18]. For the N60-PAS^{mos} library, the analysis was
1061 restricted to reporter variants containing exactly one instance of the UUUUA motif within
1062 the 3' UTR. For the CPE^{mos}-N60 library, the analysis was restricted to reporter variants not
1063 containing any AAUAAA or AUUAAA motifs within the 3' UTR. For each *k*-mer, all variants
1064 containing that *k*-mer were identified, and the mean tail length across those variants was
1065 computed. To assess the positional effect of specific *k*-mers, only variants in which the *k*-
1066 mer appeared at a defined position were included in the calculation.
1067
1068 **Preparation of the N60(LC)-PAS^{mos} library.** The DNA template for the N60(LC)-PAS^{mos}
1069 library was generated by PCR using primers KXU024 and KXU068, and 2 amol of the N60-
1070 PAS^{mos} DNA template used previously[18]. The reaction (100-µl total volume) was performed
1071 with the KAPA HiFi HotStart Kit (Roche, KK2502). PCR products were gel-purified using
1072 agarose gels (Lonza, 50004) and the GeneJet Gel Extraction Kit (Thermo Fisher, K0692). The
1073 resulting mRNA library was in vitro transcribed and capped, followed by removal of the 3'-
1074 end cyclic phosphate as previously described[18].
1075
1076 **Preparation of the single-nucleotide mutagenesis library.** The oligo pool was
1077 synthesized as SurePrint HiFi Oligo (Agilent Technologies; Supplementary Table 2). Each
1078 oligo contained a variable region derived from the last 100 nt of the 3' UTR of one of ten frog
1079 oocyte mRNAs (*atp1a1.S*, *ccnb1.2.L*, *ccnb2.L*, *lima1.L*, *mad2l1.L*, *magoh.S*, *mos.L*, *tpx2.L*,
1080 *aurkaip1.L*, and *dbf4.L*). At each position, every nucleotide was mutated to the three
1081 alternative nucleotides, yielding 3010 sequences (10 wild-type and 3000 mutants). A 35-nt
1082 poly(A) fragment was appended to the 3' end and constant sequences (24 nt at the 5' end
1083 and 20 nt at the 3' end) were added for PCR amplification. Both forward and reverse strands
1084 were included in the pool.
1085 The DNA template for in vitro transcription was assembled by overlapping PCR using
1086 the KAPA HiFi HotStart Kit. Three fragnments were joined sequentially. Fragment 1 (F1) was
1087 amplified from the N60-PAS^{mos} template[18] with primers KXU024 and KXU236. Fragment 2
1088 (F2) was amplified from the oligo pool with primers KXS067 and KXU237. Fragment 3 (F3)
1089 was amplified from the plasmid C071[4] with primers KXU110 and KXU068. F1 and F2 were
1090 first joined, and the resulting product was joined with F3 in a 100-µl PCR reaction

1091 containing 10 pmol of each fragment, 2 µl KAPA HiFi enzyme, 3 µl 10 mM dNTP, and 20 µl 5×
1092 KAPA HiFi buffer for 15 cycles, with an annealing temperature of 66°C. PCR products were
1093 purified using agarose gels and the GeneJet Gel Extraction Kit. The mRNA library was then
1094 in vitro transcribed and capped, followed by removal of the 3'-end cyclic phosphate as
1095 described previously[18].
1096
1097 **Preparation of the reporter mRNA poly(A) tail-length standard mix.** Each of the four
1098 poly(A) tail-length standards was constructed from two fragments (F1 and F2) joined by
1099 overlapping PCR. Both fragments were amplified from plasmid C071 using the KAPA HiFi
1100 HostStart Kit and the following primer sets: KXU024 and one of KXU290, KXU291, KXU292,
1101 or KXU293 (for F1); KXU068 and one of KXUm020, KXUm021, KXUm022, or KXUm023 (for
1102 F2, yielding tail lengths of 10, 50, 90, or 130 nt). Overlapping PCR was performed in a 50-µl
1103 reaction containing 100 fmol of each fragment, 1 µl KAPA HiFi enzyme, 1.5 µl 10 mM dNTPs,
1104 and 10 µl 5× KAPA HiFi buffer. After 5 cycles (annealing temperature 63°C), primers KXU024
1105 and KXU068 were added to a final concentration of 300 nM along with 1 µl KAPA HiFi
1106 enzyme, 1.5 µl 10 mM dNTP, and 10 µl 5× KAPA HiFi buffer to bring the final volume to 100
1107 µl. Twenty-four additional PCR cycles were performed. PCR products were purified using
1108 agarose gels and the GeneJet Gel Extraction Kit. Each RNA was in vitro transcribed and
1109 capped, followed by removal of the 3'-end cyclic phosphate as described previously[18]. The
1110 four mRNA standards were mixed at equimolar concentrations. Each contained a unique 4-
1111 nt barcode in the 3' UTR, introduced by KXU290, KXU291, KXU292, or KXU293. Sequences
1112 of the four tail-length standards are provided in Supplementary Table 3.
1113
1114 **Reporter mRNA library injection and sample collection**. Frog oocytes were obtained
1115 from Xenopus1 (12005). Healthy Stage V–VI oocytes were hand-picked and transferred to
1116 OR-2 buffer (5 mM HEPES pH 7.6, 82.5 mM NaCl, 2.5 mM KCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 1
1117 mM Na_2HPO_4) supplemented with 100 µg/ml Gentamicin (Thermo Fisher, 15750060), and
1118 incubated at 18°C overnight (> 16 hr) for recovery. Injections were performed at 23°C using
1119 a PLI-100 Plus Pico-Injector. For both the N60(LC)-PAS^{mos} library and the single-nucleotide
1120 mutagenesis library, 4 nl of mRNA (0.1 pmol/µl) was injected per oocyte.
1121 Frog oocytes were matured in vitro in OR-2 buffer supplemented with 10 µg/ml
1122 progesterone (Millipore Sigma, P0130), diluted from a 10 mg/ml ethanol-disolved stock.
1123 Maturation timing varied across batches: in most cases, 50% of oocytes underwent
1124 germinal vesicle breakdown (GVBD; indicated by a white spot on the animal pole) within 3–
1125 5 hr, and 100% reached GVBD by 7 h post-progesterone addition.
1126 At indicated time points post-progesterone treatment, groups of 100 frog oocytes
1127 were collected. After removing OR-2 buffer, oocytes were washed three times with ice-cold
1128 buffer RL (20 mM HEPES pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 1% Triton X-100, 100 µg/ml
1129 cycloheximide, cOmplete protease inhibitor cocktail [MilliporeSigma, 11836170001, 1
1130 tablet/10 ml], 10 µl/oocyte). Following the washes, buffer was completely removed and
1131 oocytes were lysed in buffer RL (10 µl per oocyte) supplemented with 200 U/ml
1132 SUPERase•In (10 µl/oocyte) by vigorous shaking and pipetting. Lysates were clarified by
1133 centrifugation at 5,000 g for 10 min at 4°C. Supernatants were mixed with 3 volumes of
1134 TRIzol LS (Thermo Fisher, 10296010).

1135 The 0 hr post-progesterone sample was prepared by collecting the supernatant from
1136 lysed, untreated oocytes and mixing it with the uninjected mRNA library prior to TRIzol LS
1137 addition. All TRIzol-LS RNA extractions were performed using Phasemaker tubes (Thermo
1138 Fisher, A33248) following the manufacturer's instructions. Total RNA was resuspended in
1139 40 µl water (0.4 µl/oocyte).
1140
1141 **Library preparation and sequencing of the injected mRNA libraries.** Total RNA (10 µl,
1142 ~30 µg) from injected oocytes was combined with 10 amol of reporter mRNA poly(A)-tail-
1143 length standard mix. To this, 8 pmol each of oligos KXSH009 and KXSH010, and 2× SSC (0.3
1144 M NaCl, 30 mM sodium citrate pH 7.0), were added in a final volume of 50 µl. The RNA-oligo
1145 mixture was incubated at 70°C for 5 min and then cooled to 23°C at a rate of 0.1°C/sec for
1146 annealing. The annealed RNA was incubated with 40 µl MyOne Streptavidin C1 beads
1147 (Thermo Fisher, 65002) for 20 min at 23°C on a thermal mixer (15 sec on and 1 min 45 sec
1148 off). The supernatant was separated from the beads with a magnetic rack and removed.
1149 Beads were washed twice with 300 µl 1× B&W buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA,
1150 1 M NaCl) and once with 300 µl 2× SSC. RNA was eluted from beads with two sequential
1151 incubations: first in 100 µl 10 mM HEPES pH 7.5 at 65°C for 3 min, then in 100 µl water at
1152 65°C for 3 min. Eluates were combined, ethanol-precipitated, and resuspended in 6.5 µl
1153 water.
1154 Anti-sense oligo-enriched RNA was ligated to a pre-adenylated 3' adapter (KXS347a)
1155 in a 10 µl reaction containing 5 µM adapter, 50 mM HEPES pH 7.5, 10 mM $MgCl_2$, 10 mM
1156 dithiothreitol, and 1 U/µl T4 RNA ligase 1 (NEB, M0204S). Ligation was carried out at 23°C
1157 for 150 min. Ligated RNA was extracted with phenol/chloroform, ethanol-precipitated, and
1158 resuspended in 11.4 µl water. The ligated RNA was mixed with 0.6 µl 100 µM primer KXS348
1159 and 1 µl 10 mM dNTPs in a total volume of 13 µl, incubated at 65°C for 5 min, and cooled on
1160 ice for 1 min. Reverse transcription was performed in a 20 µl reaction using SuperScript IV
1161 (ThermoFisher, 18090050), with 1× SSIV Buffer, 5 mM dithiothreitol, 1 U/µl SUPERase•In,
1162 and 200 U enzyme, incubated at 55°C for 15 min. After reverse transcription, RNA was then
1163 hydrolyzed with 3.3 µl 1 M NaOH at 90°C for 10 min, followed by neutralization with 24 µl 1
1164 M HEPES pH 7.5. The resulting cDNA was ethanol-precipitated, resuspended in water, and
1165 amplified in a 50 µl PCR reaction using primer KXS037 and a barcoded primer (KXS057)
1166 targeting a constant region of the library. Amplification was performed using the KAPA HiFi
1167 HotStart Kit following the manufacturer's suggested protocol for 11~14 cycles. The final
1168 PCR-amplified product was cleaned up twice with AMPure XP beads (Beckman Coulter,
1169 A63881) at a 1.2× beads-to-sample ratio.
1170 Sequencing was performed on an Illumina HiSeq 2500, with a custom run of two
1171 reads: a 281-cycle first read using primer KXS067, and a 10-cycle second read using a
1172 standard TruSeq sequencing primer. To generate a reference, the uninjected 0 hr N60(LC)-
1173 PAS^{mos} library sample was also sequenced on AVITI using the same primers, with custom
1174 read lengths of 299 and 10 cycles.
1175
1176 **Variant and tail-length analysis for the injected mRNA libraries**. To build a reference
1177 sequence for the N60(LC)-PAS^{mos} library, read 1 from the 0 hr post-progesterone
1178 (uninjected) sample sequenced on AVITI was trimmed to 81 nt from the 5' end. Trimmed

1179 reads were screened for the presence of a known constant region from the 3' UTR of poly(A)
1180 tail-length standards (ACCAGCCTCAAGAACACCCGAATGG). A maximum of seven
1181 mismatches was allowed within the first 25 nt of this sequence, permitting positional
1182 offsets from –3 to +3 nt. Next, the remaining reads were examined for the expected
1183 constant sequence AATAAAGAAATTGATTTGTCT at positions 61–81, again allowing offsets
1184 from –3 and +3 nt. Reads containing a 21-nt segment with no more than six mismatches to
1185 the constant sequence at any allowed position were retained. For reads with a non-zero
1186 offset, alignment was corrected to zero: if the offset was negative, "N" bases were
1187 prepended to the read; if positive, the corresponding number of bases was trimmed from
1188 the start of the read. These modified reads were then clustered using *UMICollapse*[66], and
1189 clusters with more than 100 supporting reads were retained. The consensus (primary)
1190 sequence from each cluster was used to construct the reference sequence for the
1191 N60(LC)-PAS^{mos} library.
1192 During the HiSeq 2500 run for both the N60(LC)-PAS^{mos} and single-nucleotide
1193 mutagenesis libraries, a sequencing artifact occurred at cycle 26 of read 1, resulting in a
1194 repeated base call at cycle 26 and 27. Consequently, base 27 was removed from all reads
1195 prior to downstream analysis.
1196 For the N60(LC)-PAS^{mos} libraries, read 1 sequences from the HiSeq run were aligned
1197 to the reference using *STAR* (v2.7.1)[67] with the following parameters '--
1198 outFilterMultimapNmax 1 --alignEndsType Local –clip3pNbases 199 --outSAMattributes All
1199 --outSAMtype BAM SortedByCoordinate'. Poly(A)-tail lengths for mapped reads were
1200 determined as previously described for the N60-PAS^{mos} libraries[18].
1201 For the single-nucleotide mutagenesis libraries, read 1 sequences from the HiSeq
1202 run were aligned to the reference consisting of the last 100 nt of the 3' UTR from ten
1203 selected frog oocyte mRNAs (wild-type sequences) using *STAR* (v2.7.1)[67] with these
1204 parameters '--outFilterMultimapNmax 1 --outFilterMismatchNmax 1 --alignEndsType
1205 EndToEnd –clip3pNbases 180 --outSAMattributes All --outSAMtype BAM
1206 SortedByCoordinate'. Mutations were classified using the CIGAR string in the resulting BAM
1207 file. Poly(A)-tail lengths for these mapped reads were determined as for the N60(LC)-PAS^{mos}
1208 libraries, with the tail start position defined as position 101 of read 1. Only wild-type and
1209 single-nucleotide substitution variants were retained. For each variant, reads sharing the
1210 same read 2 sequence (serving as a unique molecular identifier, UMI) were grouped and
1211 collapsed. The median tail length of each collapsed was used as the representative
1212 poly(A)-tail length.
1213 Reference sequences of both the N60(LC)-PAS^{mos} and single-nucleotide
1214 mutagenesis libraries are provided in Supplementary Table 3.
1215
1216 **RNA secondary structure prediction.** RNA secondary structures for the 3' UTR sequences
1217 in the single-nucleotide mutagenesis library were predicted using *EternaFold* (v1.3.1)[46] with
1218 the parameters "--params parameters/EternaFoldParams.v1". The resulting dot-bracket
1219 notations were visualized using *RNAcanvas*[68].
1220
1221 **Analysis of mouse oocyte data sequenced by Nanopore**. Poly(A)-tail lengths of mouse
1222 oocyte mRNAs obtained via Oxford Nanopore Technologies (ONT) sequencing were

retrieved from a previous study[9]. Tail-length changes were calculated by comparing the median tail lengths between GV and MII oocytes, requiring at least 50 poly(A) reads in datasets from both stages. As ONT-reported tail-length measurements were reported per gene rather than per mRNA isoform, comparisons were made with the primary isoform for each gene, defined based on Hi-Seq data[18] or predicted using PAL-AI-ms (mouse-specific output head 3). Analyses were restricted to genes whose primary isoform represented more than 90% of expression in GV oocytes.

Prediction of poly(A) tail-length changes of reporter mRNAs during mouse oocyte maturation. To evaluate the relationship between predicted tail-length changes and experimentally measured translational efficiency changes of reporter mRNAs injected into mouse oocytes, source data were obtained from a previous study (see Fig. 4g of the original publication)[6]. Translational efficiency change was calculated as the difference between the mean normalized fluorescence signals in GV versus MII oocytes (24 hr). Reporter mRNA 3' UTR sequences were inferred from the cloning and mutagenesis primers used in this study. Tail-length changes for these sequences were predicted using PAL-AI-ms (mouse-specific output head 3).

Cross-species analysis of poly(A) tail-length and translational efficiency changes. Gene orthology tables (frog-to-human from Xenbase, mouse-to-human from Ensembl) were used to map homologous genes. Measured tail-length changes were compared between GV and MII oocytes in humans and mice, and between 0 and 7 h post-progesterone treatment in frog oocytes. For mRNA isoforms with unique 3' ends in one of the three species, predicted tail-length changes from PAL-AI-ms were used if experimental data were unavailable. For genes with multiple mRNA isoforms, the dominant isoform (> 50% by tag count in GV oocytes of humans or mice, or in 0 hr frog oocytes) was selected. If a human gene had multiple homologs in frog or mouse, the average tail-length change across all the homologs was used.

Translational efficiency data derived from ribosome profiling and mRNA-seq were reprocessed[18] and compared between GV and MII oocytes in humans[7] and mice[6].

For gene ontology (GO) analysis, genes showing substantial tail lengthening (≥ 15 nt, by measurement or prediction) during oocyte maturation in all three species were selected. These genes were ranked by the smallest tail-length change across three species, and the ranked list was submitted to the *gost* function in the R package *gprofiler2*[69] with the parameters “ordered_query = TRUE, user_threshold = 0.05, correction_method = "g_SCS". The top eight non-overlapping GO:Biological Process categories (not from the same parent term and ranked by Benjamini-Hochberg-adjusted *P* value) were reported.

Evolutionary constraint associated with tail-length changes. PhyloP scores from multiple sequence alignments of 99 vertebrate genomes to the human genome were obtained from the UCSC Genome Browser[49]. For each nucleotide within the last 100 nt of human 3' UTRs with an available phyloP score, the mean PAL-AI-predicted differences in tail-length changes when this nucleotide was substituted with each of the three alternative nucleotides (from ISM analysis) were calculated.

Prediction of poly(A) tail-length changes for human genetic variants from the All of Us Research Program and gnomAD. MRNA 3' UTR isoforms shorter than 10 nt were excluded from analysis. To maximize variant coverage despite incomplete transcript annotations, the "uncertain" 3' UTRs in the annotations were retained if their 3' ends were supported by entries in the PolyA_DB_v3.2 database[70].

For the data reported in the All of Us Research Program[10], all single-nucleotide polymorphism and indel variants from short-read whole-genome sequencing callsets encompassing human exome regions (version 8) were intersected with human oocyte 3' UTR genomic coordinates. Variants overlapping these regions were selected. For each variant, the associated 3' UTR sequence was reconstructed by substituting the reference allele with the alternative allele, followed by appending three adenines (AAA) at the 3' end. Tail-length changes were predicted using PAL-AI-ms (human-specific output head 2). The mutational outcome was defined as the difference in predicted tail-length change between the variant and corresponding wild-type (reference) sequence. In cases where a variant overlapped multiple 3' UTRs, the one with the largest absolute predicted difference in tail-length change was used for downstream analysis.

For the data reported in gnomAD[11], genomic variants from v4.1 (covering 76,215 individuals) were downloaded from gnomad.broadinstitute.org. Variants were included if they passed all quality filters (indicated by "PASS" in the "FILTER" column in the VCF file) and intersected annotated 3' UTR genomic locations. Tail-length changes and mutational outcomes were computed using the same approach as for the All of Us Research Program.

Quantification and statistical analysis

Graphs were generated and statistical analyses were performed using R[71]. Statistical parameters, including the value of n, statistical test, and statistical significance (*P* value) are reported in the figures or their legends. No statistical methods were used to pre-determine sample size.

Data availability

All standard sequencing data are available in the Gene Expression Omnibus under the accession number GSE280422. Raw intensity data for reporter mRNA tail-length sequencing cannot be deposited in public databases due to their large sizes and are available upon request.

Oligo sequences used in this study are listed in Supplementary Table 1. Sequences of the oligo library used for the single-nucleotide mutagenesis library are listed in Supplementary Table 2. Sequences of the N60(LC)-PAS^{mos} library, the single-nucleotide mutagenesis library, and the tail-length standards are listed in Supplementary Table 3.

Other publicly available data analyzed in this study are indicated in the relevant sections of **Methods**.

Code availability

The code for PAL-AI is written in Python 3.8 and available at https://github.com/coffeebond/PAL-AI. Reporter mRNA tail-length sequencing data analyses were performed using a custom script written in Python 2.7 and available at https://github.com/coffeebond/MPRA_tail_seq. The codes used to generate the figures are available at https://github.com/coffeebond/PAL-AI_paper.

1316 **References**
1317
1318 1. Eichhorn, S. W. *et al.* mRNA poly(A)-tail changes specified by deadenylation broadly
1319 reshape translation in Drosophila oocytes and early embryos. *eLife* **5**, e16955 (2016).
1320 2. Lim, J., Lee, M., Son, A., Chang, H. & Kim, V. N. mTAIL-seq reveals dynamic poly(A) tail
1321 regulation in oocyte-to-embryo development. *Genes & Development* **30**, 1671–1682
1322 (2016).
1323 3. Yang, F., Wang, W., Cetinbas, M., Sadreyev, R. I. & Blower, M. D. Genome-wide analysis
1324 identifies cis-acting elements regulating mRNA polyadenylation and translation during
1325 vertebrate oocyte maturation. *RNA (New York, N.Y.)* **26**, 324–344 (2020).
1326 4. Xiang, K. & Bartel, D. P. The molecular basis of coupling between poly(A)-tail length and
1327 translational efficiency. *eLife* **10**, e66493 (2021).
1328 5. Liu, Y. *et al.* Poly(A) tail length is a major regulator of maternal gene expression during
1329 the mammalian oocyte-to-embryo transition. *bioRxiv* 2021.08.29.458052 (2021)
1330 doi:10.1101/2021.08.29.458052.
1331 6. Xiong, Z. *et al.* Ultrasensitive Ribo-seq reveals translational landscapes during
1332 mammalian oocyte-to-embryo transition and pre-implantation development. *Nat Cell Biol*
1333 1–13 (2022) doi:10.1038/s41556-022-00928-6.
1334 7. Zou, Z. *et al.* Translatome and transcriptome co-profiling reveals a role of TPRXs in
1335 human zygotic genome activation. *Science* **378**, abo7923 (2022).
1336 8. Liu, Y. *et al.* Remodeling of maternal mRNA through poly(A) tail orchestrates human
1337 oocyte-to-embryo transition. *Nat Struct Mol Biol* 1–16 (2023) doi:10.1038/s41594-022-
1338 00908-2.
1339 9. Lee, K., Cho, K., Morey, R. & Cook-Andersen, H. An extended wave of global mRNA
1340 deadenylation sets up a switch in translation regulation across the mammalian oocyte-to-
1341 embryo transition. *Cell Rep.* **43**, 113710–113710 (2024).
1342 10. Investigators, T. A. of U. R. P. G. *et al.* Genomic data in the All of Us Research Program.
1343 *Nature* 1–7 (2024) doi:10.1038/s41586-023-06957-x.
1344 11. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human
1345 genomes. *Nature* **625**, 92–100 (2024).
1346 12. Richter, J. D. & Lasko, P. Translational control in oocyte development. *Cold Spring*
1347 *Harbor Perspectives in Biology* **3**, a002758 (2011).
1348 13. Teixeira, F. K. & Lehmann, R. Translational Control during Developmental Transitions.
1349 *Csh Perspect Biol* **11**, a032987 (2019).
1350 14. Conti, M. & Kunitomi, C. A genome-wide perspective of the maternal mRNA translation
1351 program during oocyte development. *Semin Cell Dev Biol* (2023)
1352 doi:10.1016/j.semcdb.2023.03.003.
1353 15. Sheets, M. D., Fox, C. A., Hunt, T., Woude, G. V. & Wickens, M. The 3'-untranslated
1354 regions of c-mos and cyclin mRNAs stimulate translation by regulating cytoplasmic
1355 polyadenylation. *Genes & Development* **8**, 926–938 (1994).
1356 16. Moor, C. H. de & Richter, J. D. The Mos pathway regulates cytoplasmic polyadenylation
1357 in Xenopus oocytes. *Molecular and Cellular Biology* **17**, 6419–6426 (1997).
1358 17. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling
1359 reveals an embryonic switch in translational control. *Nature* **508**, 66–71 (2014).

1360 18. Xiang, K., Ly, J. & Bartel, D. P. Control of poly(A)-tail length and translation in vertebrate
1361 oocytes and early embryos. *Dev. Cell* **59**, 1058-1074.e11 (2024).
1362 19. McGrew, L. L., Dworkin-Rastl, E., Dworkin, M. B. & Richter, J. D. Poly(A) elongation
1363 during Xenopus oocyte maturation is required for translational recruitment and is
1364 mediated by a short sequence element. *Gene Dev* **3**, 803–815 (1989).
1365 20. Radford, H. E., Meijer, H. A. & Moor, C. H. de. Translational control by cytoplasmic
1366 polyadenylation in Xenopus oocytes. *Biochimica et Biophysica Acta (BBA) - Gene*
1367 *Regulatory Mechanisms* **1779**, 217–229 (2008).
1368 21. Aoki, F., Hara, K. T. & Schultz, R. M. Acquisition of transcriptional competence in the 1-
1369 cell mouse embryo: Requirement for recruitment of maternal mRNAs. *Mol. Reprod. Dev.*
1370 **64**, 270–274 (2003).
1371 22. Meneau, F., Dupré, A., Jessus, C. & Daldello, E. M. Translational Control of Xenopus
1372 Oocyte Meiosis: Toward the Genomic Era. *Cells 2020, Vol. 9, Page 1502* **9**, 1502 (2020).
1373 23. Yatsenko, S. A. & Rajkovic, A. Genetics of human female infertility. *Biol. Reprod.* **101**,
1374 549–566 (2019).
1375 24. Zheng, W. *et al.* Homozygous Mutations in BTG4 Cause Zygotic Cleavage Failure and
1376 Female Infertility. *Am. J. Hum. Genet.* **107**, 24–33 (2020).
1377 25. Liu, R. *et al.* A novel homozygous missense variant in BTG4 causes zygotic cleavage
1378 failure and female infertility. *J. Assist. Reprod. Genet.* **38**, 3261–3266 (2021).
1379 26. Varnum, S. M. & Wormington, W. M. Deadenylation of maternal mRNAs during Xenopus
1380 oocyte maturation does not require specific cis-sequences: a default mechanism for
1381 translational control. *Gene Dev* **4**, 2278–2286 (1990).
1382 27. Fox, C. A. & Wickens, M. Poly(A) removal during oocyte maturation: a default reaction
1383 selectively prevented by specific sequences in the 3' UTR of certain maternal mRNAs.
1384 *Genes & Development* **4**, 2287–2298 (1990).
1385 28. Hake, L. E. & Richter, J. D. CPEB is a specificity factor that mediates cytoplasmic
1386 polyadenylation during Xenopus oocyte maturation. *Cell* **79**, 617–627 (1994).
1387 29. Dickson, K. S., Bilger, A., Ballantyne, S. & Wickens, M. P. The Cleavage and
1388 Polyadenylation Specificity Factor in Xenopus laevis Oocytes Is a Cytoplasmic Factor
1389 Involved in Regulated Polyadenylation. *Mol Cell Biol* **19**, 5707–5717 (1999).
1390 30. Charlesworth, A., Meijer, H. A. & Moor, C. H. de. Specificity factors in cytoplasmic
1391 polyadenylation. *Wiley Interdiscip Rev Rna* **4**, 437–461 (2013).
1392 31. Ivshina, M., Lasko, P. & Richter, J. D. Cytoplasmic Polyadenylation Element Binding
1393 Proteins in Development, Health, and Disease. *Annual Review of Cell and Developmental*
1394 *Biology* **30**, 393–415 (2014).
1395 32. Reyes, J. M. & Ross, P. J. Cytoplasmic polyadenylation in mammalian oocyte
1396 maturation. *Wiley Interdisciplinary Reviews: RNA* **7**, 71–89 (2016).
1397 33. Weill, L., Belloc, E., Castellazzi, C. L. & Méndez, R. Musashi 1 regulates the timing and
1398 extent of meiotic mRNA translational activation by promoting the use of specific CPEs.
1399 *Nature Structural & Molecular Biology* **24**, 672–681 (2017).
1400 34. Agarwal, V. & Kelley, D. R. The genetic and biochemical determinants of mRNA
1401 degradation rates in mammals. *Genome Biol* **23**, 245 (2022).
1402 35. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithm Mol Biol* **6**, 26 (2011).

1403 36. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv*
1404 (2015) doi:10.48550/arxiv.1512.03385.
1405 37. Hwang, H., Jeon, H., Yeo, N. & Baek, D. Big data and deep learning for RNA biology. *Exp.*
1406 *Mol. Med.* **56**, 1293–1321 (2024).
1407 38. Collier, B., Gorgoni, B., Loveridge, C., Cooke, H. J. & Gray, N. K. The DAZL family
1408 proteins are PABP-binding proteins that regulate translation in germ cells. *EMBO J.* **24**,
1409 2656–2666 (2005).
1410 39. Bauermeister, D., Claußen, M. & Pieler, T. A novel role for Celf1 in vegetal RNA
1411 localization during Xenopus oogenesis. *Dev. Biol.* **405**, 214–224 (2015).
1412 40. Dominguez, D. *et al.* Sequence, Structure, and Context Preferences of Human RNA
1413 Binding Proteins. *Mol Cell* **70**, 854-867.e9 (2018).
1414 41. Martins, J. P. S. *et al.* DAZL and CPEB1 regulate mRNA translation synergistically during
1415 oocyte maturation. *J. Cell Sci.* **129**, 1271–1282 (2016).
1416 42. Yang, C.-R. *et al.* The RNA-binding protein DAZL functions as repressor and activator of
1417 mRNA translation during oocyte maturation. *Nat. Commun.* **11**, 1399 (2020).
1418 43. Wu, X. & Bartel, D. P. Widespread Influence of 3'-End Structures on Mammalian mRNA
1419 Processing and Stability. *Cell* **169**, 905-917.e11 (2017).
1420 44. Hoque, M. *et al.* Analysis of alternative cleavage and polyadenylation by 3' region
1421 extraction and deep sequencing. *Nat. Methods* **10**, 133–139 (2013).
1422 45. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant
1423 polyadenylation signal usage in human genes. *Genome Research* **10**, 1001–1010 (2000).
1424 46. Wayment-Steele, H. K. *et al.* RNA secondary structure packages evaluated and
1425 improved by high-throughput experiments. *Nat Methods* 1–9 (2022) doi:10.1038/s41592-
1426 022-01605-0.
1427 47. Piqué, M., López, J. M., Foissac, S., Guigó, R. & Méndez, R. A Combinatorial Code for
1428 CPE-Mediated Translational Control. *Cell* **132**, 434–448 (2008).
1429 48. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral
1430 substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
1431 49. Raney, B. J. *et al.* The UCSC Genome Browser database: 2024 update. *Nucleic Acids*
1432 *Res.* **52**, D1082–D1088 (2023).
1433 50. Linder, J., Koplik, S. E., Kundaje, A. & Seelig, G. Deciphering the impact of genetic
1434 variation on human polyadenylation using APARENT2. *Genome Biol* **23**, 232 (2022).
1435 51. Padmanabhan, K. & Richter, J. D. Regulated Pumilio-2 binding controls RINGO/Spy
1436 mRNA translation and CPEB activation. *Genes Dev.* **20**, 199–209 (2006).
1437 52. Chen, J. *et al.* Genome-wide analysis of translation reveals a critical role for deleted in
1438 azoospermia-like (Dazl) at the oocyte-to-zygote transition. *Genes & Development* **25**, 755–
1439 766 (2011).
1440 53. Mikedis, M. M. *et al.* DAZL mediates a broad translational program regulating expansion
1441 and differentiation of spermatogonial progenitors. *eLife* **9**, e56523 (2020).
1442 54. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide
1443 probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**,
1444 701–705 (2013).
1445 55. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel
1446 regulatory features. *Nature* **505**, 696–700 (2014).

56. Spitale, R. C. *et al*. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **advance online publication**, 486–490 (2015).
57. Zhang, J., Fei, Y., Sun, L. & Zhang, Q. C. Advances and opportunities in RNA structure experimental determination and computational modeling. *Nat Methods* **19**, 1193–1207 (2022).
58. Kaufmann, I., Martin, G., Friedlein, A., Langen, H. & Keller, W. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *The EMBO Journal* **23**, 616–626 (2004).
59. Lackford, B. *et al*. Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO J*. **33**, 878–889 (2014).
60. Stroup, E. K. & Ji, Z. Deep learning of human polyadenylation sites at nucleotide resolution reveals molecular determinants of site usage and relevance in disease. *Nat. Commun*. **14**, 7378 (2023).
61. Sang, Q., Ray, P. F. & Wang, L. Understanding the genetics of human infertility. *Science* **380**, 158–163 (2023).
62. DebRoy, H. P. and P. A. and R. G. and S. Biostrings: Efficient manipulation of biological strings. https://bioconductor.org/packages/Biostrings (2021).
63. Pedregosa, F. *et al*. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
64. Abadi, M. *et al*. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Preprint at https://www.tensorflow.org/.
65. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. *arXiv* (2019) doi:10.48550/arxiv.1907.10902.
66. Liu, D. Algorithms for efficiently collapsing reads with Unique Molecular Identifiers. *PeerJ* **7**, e8275 (2019).
67. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
68. Johnson, P. Z. & Simon, A. E. RNAcanvas: interactive drawing and exploration of nucleic acid structures. *Nucleic acids Res*. **51**, W501–W508 (2023).
69. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2-- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research* **9 (ELIXIR)**, (2020).
70. Wang, R., Nambiar, R., Zheng, D. & Tian, B. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res*. **46**, gkx1000- (2017).
71. Team, R. C. R: A Language and Environment for Statistical Computing. (2021).