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

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'-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 are 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 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 in 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 built a multiple linear regression model based on 3'-UTR *k*-mer (sequence motif length of *k*) compositions. Because the number and position of CPE and PAS elements can strongly influence their activities[18], we used two variables for each *k*-mer of length 5 or 6 nucleotides (nt): 1) a count metric, specifying the total number of a *k*-mer in 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]. The prediction values from each of the ten iterations were concatenated and compared to the measured values. The linear model performed moderately well and explained ~30% of the variance of measured tail-length changes of mRNAs in frog oocytes (Supplementary Fig. 1a; Spearman correlation coefficient $R_s = 0.58$; Pearson correlation coefficient $R_p = 0.55$). In this model, the CPE and the PAS, two known motifs mediating cytoplasmic polyadenylation, were among the top-ranked *k*-mers 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 built an integrated neural network model, called poly(A)-tail length AI (PAL-AI) (Supplementary Fig. 1d, e). PAL-AI integrated a convolution neural network with a recurrent neural network, using an architecture similar to but simpler than that of Saluki[34]. PAL-AI was trained and tested using the sequences of endogenous mRNA 3' UTRs (using the last 1053 nt for UTRs >1053 nt) as the only input, and the same 10-fold cross-validation strategy as used for the linear model. The trained model predicted tail-length changes with high accuracy, significantly out-performing the linear model and explaining more than 60% of the variance of measured values (Fig. 1a; $R_s = 0.79$, $R_p = 0.8$).

Fig. 1

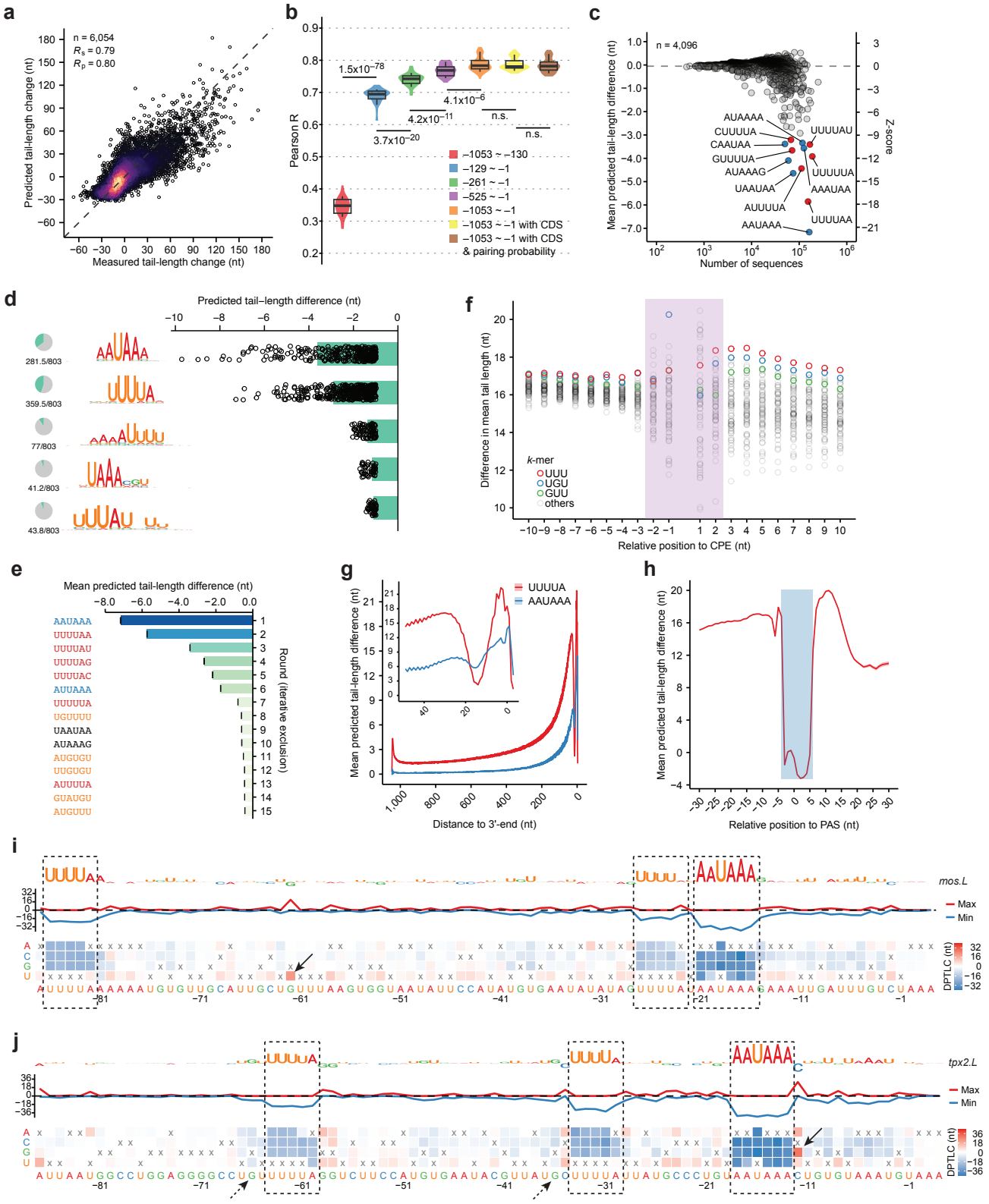

a
n = 6,054
R_s = 0.79
R_p = 0.80
Predicted tail-length change (nt)
Measured tail-length change (nt)
b
Pearson R
1.5x10−78
3.7x10−20
4.2x10−11
4.1x10−6
n.s.
n.s.
−1053 ~ −130
−129 ~ −1
−261 ~ −1
−525 ~ −1
−1053 ~ −1
−1053 ~ −1 with CDS
−1053 ~ −1 with CDS & pairing probability
c
n = 4,096
Mean predicted tail-length difference (nt)
Z-score
Number of sequences
AUAAAA
CUUUUA
CAAUAA
GUUUUA
AUAAAG
UAAUAA
AUUUUA
AAUAAA
UUUUAU
UUUUUA
AAAUAA
UUUUAA
d
Predicted tail–length difference (nt)
281.5/803
359.5/803
77/803
41.2/803
43.8/803
e
Mean predicted tail-length difference (nt)
AAUAAA
UUUUAA
UUUUAU
UUUUAG
UUUUAC
AUUAAA
UUUUUA
UGUUUU
UAAUAA
AUAAAG
AUGUGU
UUGUGU
AUUUUA
GUAUGU
AUGUUU
Round (iterative exclusion)
f
Difference in mean tail length (nt)
k-mer
UUU
UGU
GUU
others
Relative position to CPE (nt)
g
UUUUA
AAUAAA
Mean predicted tail-length difference (nt)
Distance to 3'-end (nt)
h
Mean predicted tail-length difference (nt)
Relative position to PAS (nt)
i
mos.L
Max
Min
DPTLC (nt)
j
tpx2.L
Max
Min
DPTLC (nt)

Fig. 1: PAL-AI predicts tail-length changes during frog oocyte maturation.

a) Performance of PAL-AI. Plotted is the relationship between tail-length changes predicted by the model and those observed in frog oocytes between 0 and 7 h post-progesterone treatment. Each point represents a unique poly(A) site of an endogenous mRNA. Colors indicate the density of points.

b) Performance observed when training PAL-AI using either different regions of the mRNAs as input sequences or an additional feature designed to score structural accessibility (key). Plotted are 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. *P* values were calculated from one-sided, paired *t*-tests; CDS, coding sequence; n.s., not significant.

c) PAL-AI-predicted consequences of motif loss. Shown for each 6-mer is the mean difference between predicted tail-length changes of mutants compared to those of the wild-type. These mean differences are plotted as a function of the mean number of analyzed mutants. Red points represent motifs that contained a CPE and blue points represent those that contained a complete or partial PAS.

d) Sequence motifs associated with PAL-AI-predicted changes in poly(A)-tail length. Shown are sequence logos generated from 8-mers most associated with the largest differences in predicted tail lengths 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.

e) 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.

f) 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].

g) 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 the CPE near position 17 was attributed to disruption of the PAS (Supplementary Fig. 2e). Shaded areas along the lines indicate standard error. The inset shows results for the last 50 nt of the 3' UTR.

h) 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.

i) Impact of single-nucleotide substitutions on predicted tail-length changes. Shown are results for the last 87 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). Letter x's indicate non-substitutions (wild type). 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 motifs (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.

j) Impact of single-nucleotide substitutions in the *tpx2.L* mRNA 3' UTR on predicted tail-length changes, plotted as in (**i**). 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].

We tested different model inputs and evaluated the Pearson R value between the measured and predicted tail-length changes. Reducing the length of input 3'-UTR sequences by omitting 5' regions of long 3' UTRs reduced performance, but only mildly, as the model trained on only the last 129 nt of 3' UTRs retained the power to predict nearly 50% of the variance (Fig. 1b; average R_p = 0.69). In contrast, the model lost most of its predictive power when the last 129 nt of 3' UTRs were omitted from training (Fig. 1b; average R_p = 0.35), supporting the idea that the 3'-most regions of 3' UTRs impart the strongest effect on tail-length changes[18]. Furthermore, incorporating additional features, such as sequences from coding regions and base-pairing probabilities predicted by RNAfold[35], did not improve the overall model performance (Fig. 1b), implying that these features, at least as parameterized in our model, contribute minimally to tail-length changes.

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. 1c). Sequence logos generated from the most impacted 8-mers also matched the CPE and the PAS (Fig. 1d). 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. 2a, 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. 1e). Likewise, we used an analogous procedure to reexamine tail-length changes associated with the gain of 6-mers (Supplementary Fig. 2c). As expected, the top 6-mers identified in these analyses contained a PAS or CPE.

In addition to the PAS and CPE, our analyses also showed that PAL-AI recognized a different class of 6-mers, which were associated with modest but statistically significant tail lengthening (Fig. 1e, Supplementary Fig. 2c). Interestingly, these motifs each contained a UGU or GUU trimer, which might be recognized by other RNA-binding proteins such as CELF or DAZL in frog oocytes[36–38]. Another possibility is that some of these motifs, such as UGUUUU and GUUUUU), which resembled the CPE, might be weakly bound by CPEB1. A third possibility is that the presence of these motifs correlates with tail lengthening but is not causal. To explore this latter possibility, we examined the efficacy of these motifs in arbitrary sequences used in a massively parallel reporter 3' UTR library[18]. Because these arbitrary sequences were not biological, they were not susceptible to noncausal correlative associations, and thus any association of a motif with a tail-length change was presumably causal. Analysis of these reporters revealed an association between these UGU/GUU-containing 6 mers and tail lengthening, but it was small — only a 0.64 nt increase on average for the top UGU/GUU-containing 6 mer UGUUUU in the absence of a CPE, which was substantially less than that of a CPE (17.7 nt on average)[18]. In the same library, a somewhat larger increase in tail length was observed for UGU and GUU (1.8 and 1.4 nt above an average effect of a 3-mer, respectively) when a CPE was also present in the 3' UTR (Supplementary Fig. 2d). This added tail-lengthening effect was observed in all CPE-neighboring positions more than 2 nt from the CPE (Fig. 1f), suggesting that these motifs might play a small auxiliary role in tail lengthening beyond merely modulating the context of the CPE. Moreover, for some 3' UTRs, the co-presence of multiple copies of these UGU/GUU-containing motifs might still contribute to detectable tail lengthening[18], perhaps through binding to CPEB1-interacting proteins, such as DAZL[39,40], which in turn could facilitate recruitment of CPEB1 to adjacent CPE-like motifs.

We carried out additional in silico mutagenesis by inserting either the CPE or the PAS at each position along the 3' UTRs. As expected[18], PAL-AI predicted the strongest

impact on polyadenylation when these elements were introduced near the 3' ends of the 3' UTRs (Fig. 1g). An added CPE was most favorable when located near the PAS on either its 5' or 3' side (Fig. 1h), but not in the position most likely to disrupt the PAS, which is typically ~17 nt from the end of the 3' UTR (Supplementary Fig. 2e)[41], consistent with the positional effects determined from analyses of millions of reporters[18].

To gain more insights into properties PAL-AI learned at the transcript level, we examined predicted consequences of all possible point substitutions at the last 100 nt of the *mos.L* mRNA. In silico mutations that perturbed either the PAS at position –21 or the CPE at position –86 caused considerable decreases in predicted tail lengthening, whereas a G-to-U mutation at position –62 created a new CPE, thereby causing an increase in predicted tail lengthening (Fig. 1i). *mos.L* also contains a second CPE falling near the PAS, at position –27. This PAS-proximal CPE was expected to be suboptimal for cytoplasmic polyadenylation because its separation from the PAS of only 1 nt was too short[18]. Notably, PAL-AI recognized this proximity feature and correctly predicted only mild mutational outcomes when this motif was disrupted (Fig. 1i).

In another example, the *tpx2.L* mRNA, PAL-AI identified two CPEs predicted to be important for tail lengthening. Furthermore, PAL-AI predicted decreases in tail lengthening when Gs at positions –68 and –36 (both at –2 positions relative to CPEs) were mutated and an increase in tail-length change for a C-to-G substitution at position –12 (+1 relative to the PAS) (Fig. 1j)—all observations consistent with the contextual features of the CPE and the PAS identified from analyses of millions of reporters[18]. Similar results were observed for other examples of known substrates of cytoplasmic polyadenylation (Supplementary Fig. 2f, g).

Together, these analyses indicated that PAL-AI successfully captured the sequence elements and their contextual features important for promoting cytoplasmic polyadenylation, thereby enabling it to accurately predict poly(A) tail-length changes of frog mRNAs during oocyte maturation.

PAL-AI predicts poly(A) tail-length change of injected mRNAs

Next, we asked if PAL-AI was able to predict tail-length changes of injected reporter mRNAs. To this end, we amplified a small subset of a 3'-UTR reporter library (N60-PAS^{mos})[18] to generate a library with only ~35,000 variants (Fig. 2a), which we called N60(LC)-PAS^{mos}. The reduced sequence complexity of this library enabled acquisition of tail-length distributions of individual mRNA sequences. We injected this mRNA library into prophase I-arrested frog oocytes, collected RNA samples at 0 and 7 h post-progesterone treatment, and used high-throughput sequencing[18] to identify the 3' UTR variant and determine its tail-length distribution at each of the two stages of maturation (Fig. 2b).

When using tail-length changes measured for these injected mRNAs to test our model trained on endogenous mRNAs, our model performed well (Fig. 2c). Indeed, correlation coefficients observed when testing on the injected mRNAs (R_s= 0.76, R_p = 0.82) resembled those observed when testing on endogenous mRNAs (Fig. 1A, R_s= 0.79, R_p = 0.80). However, the model appeared to under-predict the magnitude of the tail-lengthening effects (Fig. 2c). A couple of possibilities could explain the systematic difference in the magnitude of tail-lengthening observed for injected mRNAs compared to that predicted by

PAL-AI. First, the sequence lengths of the training and the testing datasets were markedly different. Endogenous mRNAs used for training PAL-AI had few 3' UTRs shorter than 100 nt (420 out of 6054), whereas 3' UTRs of the injected mRNAs were all relatively short (81 nt). Second, due to experimental variability, particularly the 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. Regardless of the reasons for this difference, when we re-trained the model with measurements from injected mRNAs, PAL-AI performed well when tested on held-out data (Fig. 2d, R_s= 0.82, R_p = 0.89). 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. 3a–c). This approach of training on synthetic, nonbiological reporter sequences also ensures that these learned elements and features are causative for tail-length changes, rather than correlative with causative features.

Fig. 2

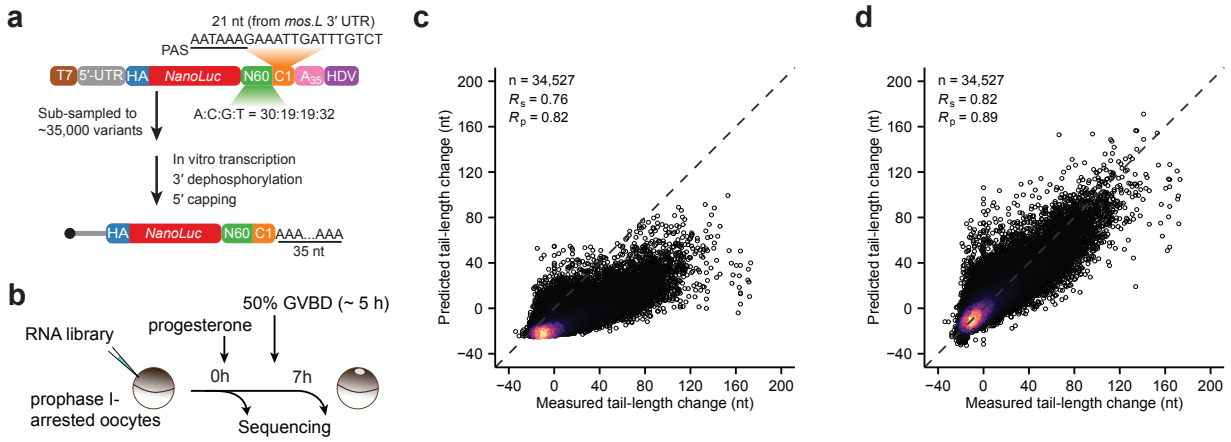


Fig. 2: PAL-AI predicts tail-length changes of synthetic mRNAs in frog oocytes.

a) Schematic of the N60(LC)-PAS^{mos} mRNA library used for injection.

b) Experimental scheme for mRNA library injection and sample collection.

c) Performance of 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. 1a.

d) Performance of a model trained from data generated from the injected N60(LC)-PAS^{mos} mRNA library. 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. 1a.

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

The ability of the model 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 and analyzed the effect of each substitution on tail-length changes in maturing oocytes. Of these 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]. For each mRNA, each nucleotide of the 3' UTR was substituted to one of the three alternative nucleotides, resulting in 3000 variants (Fig. 3a). These variants, together with the non-substituted wild-type molecules, were injected into oocytes and analyzed for tail-length changes.

Fig. 3

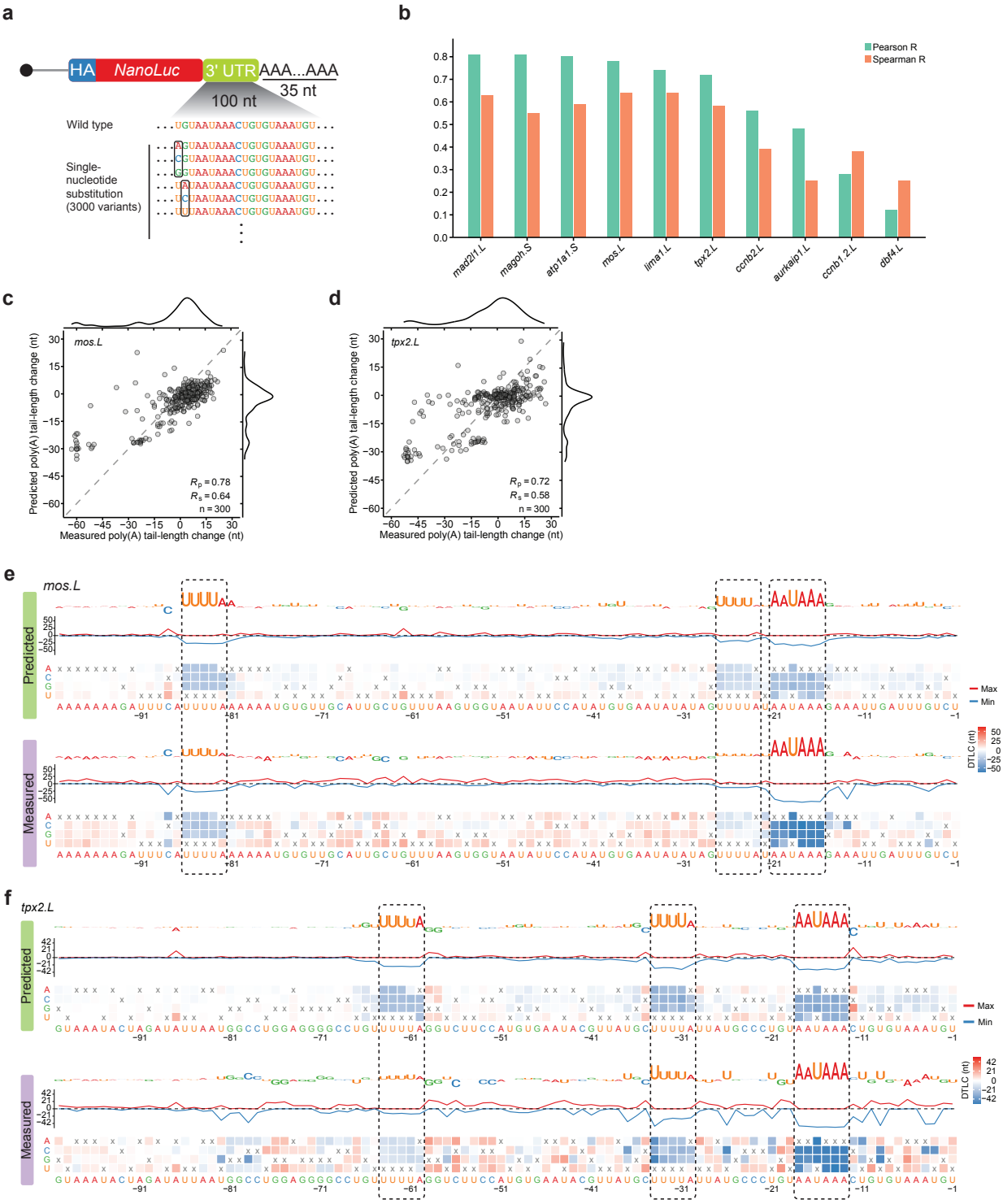


Fig. 3: Experiments validate predicted effects of single-nucleotide substitutions.

a) Schematic of the single-nucleotide mutagenesis mRNA library 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 observed between predicted and measured effects of single-nucleotide substitutions. Pearson and Spearman *R* values for each mRNA in the single-nucleotide mutagenesis library were calculated based on the difference in tail-length changes predicted by PAL-AI and those measured for each single-nucleotide substitution, comparing between 0 h and 7 h post progesterone treatment in frog oocytes.

c) PAL-AI prediction of the effects of single-nucleotide substitutions of the *mos.L* mRNA. Plotted for each single-nucleotide substitution is the relationship between the difference in tail-length changes predicted by PAL-AI and that observed between each single-nucleotide substitution and the wild type, comparing between 0 h and 7 h post progesterone treatment in frog oocytes. On the sides are density distributions for the predicted and measured differences in tail-length changes.

d) PAL-AI prediction of the effects of single-nucleotide substitutions of the *tpx2.L* mRNA; otherwise, as in (**c**).

e) 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). Letter x's indicate non-substitutions (wild type). 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 motifs (the CPE and PAS) that are important for poly(A)-tail lengthening.

f) Impact of single-nucleotide substitutions on predicted and measured tail-length changes for the *tpx2.L* mRNA; otherwise as in (**e**).

When comparing the predicted difference in tail-length change to that measured between each mutant and the wild-type, PAL-AI achieved high accuracy ($R_p > 0.7$) for six mRNAs (*atp1a1.S*, *lima1.L*, *mad2l1.L*, *magoh.S*, *mos.L*, and *tpx2.L*; Fig. 3b–f, Fig. 4a, b, Supplementary Fig. 4a–f) and moderate accuracy ($R_p \sim 0.5$) for two mRNAs (*aurkaip1.L* and *ccnb2.L*; Fig. 3b, Supplementary Fig. 5a, b, d, e). Concordance between predicted and experimental outcomes was particularly clear for substitutions that disrupted either a CPE or PAS (Fig. 3e, f, Fig. 4a, Supplementary Fig. 4d–f). For example, ablation of any of the three motifs (two CPEs and one PAS) within *mos.L* mRNA led to substantial decreases in tail lengthening, as predicted by PAL-AI (Fig. 3e). The predicted weaker contribution to cytoplasmic polyadenylation by the CPE proximal to the PAS compared to that by the CPE further upstream was also validated (Fig. 3e). Similar results illustrating PAL-AI's ability to predict the importance of these motifs and their contextual features were also observed for a few other mRNAs (Fig. 3f, Supplementary Fig. 4). Notably, the predicted differences in tail-length changes observed between many variants and the wild-type were smaller than those measured, perhaps due to the same reasons previously put forward for explaining the under-prediction of changes for the injected N60(LC)-PAS^{mos} library (Fig. 2c). These included longer UTR lengths for the mRNAs used for training and possible differences in maturation times for different oocyte batches.

In another example, *mad2l1.L* mRNA had no CPE within the last 100 nt of its 3' UTR but contained seven UGU/UGUUU motifs (Fig. 4a). Mutations within these motifs caused modest but consistent reductions in tail lengthening (Fig. 4a, b), agreeing with their detectable but weak roles in cytoplasmic polyadenylation. In contrast, substitutions that introduced a single CPE led to large increases in tail lengthening, almost all of which were correctly predicted by PAL-AI (Fig. 4a, b).

Fig. 4

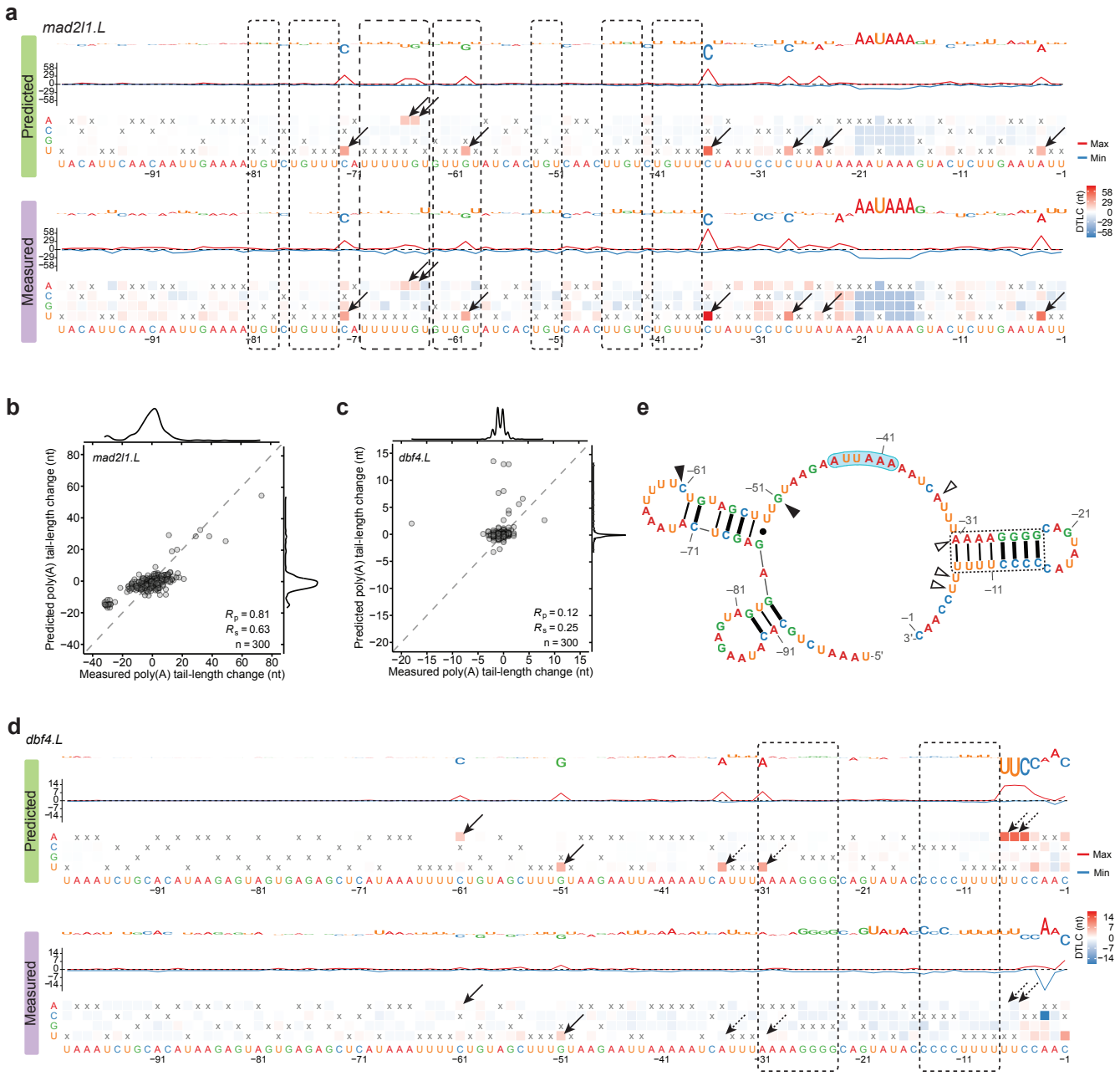


Fig. 4: Auxiliary motifs and structural accessibility influence tail-length changes.

a) Impact of single-nucleotide substitutions on predicted and measured tail-length changes for the *mad2l1.L* mRNA. Dashed rectangles indicate UGU/UGUUU motifs that modestly promote poly(A)-tail lengthening. The arrows point to instances of new CPEs generated by substitutions and their associated increase in tail-length changes. Otherwise, this panel is as in **Fig. 3e**.

b) PAL-AI prediction of the effects of single-nucleotide substitutions of the *mad2l1.L* mRNA. Otherwise, this panel is as in **Fig. 3c**.

c) PAL-AI prediction of the effects of single-nucleotide substitutions of the *dbf4.L* mRNA. Otherwise, this panel is as in **Fig. 3c**.

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.

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, 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.

Despite achieving good predictions for the six aforementioned mRNAs (R_p = 0.72–0.81), PAL-AI performed only moderately well on two mRNAs (*aurkaip1.L* and *ccnb2.L*, R_p = 0.48 and 0.56, Supplementary Fig. 5b, e) and did not perform well at all on the other two mRNAs (*dbf4.L* and *ccnb1.2.L*, R_p = 0.12 and 0.28, Fig.4c and Supplementary Fig. 6a). The wide-type *dbf4.L* mRNA, which had no CPE and a relatively weak PAS (AUUAAA), had little tail lengthening, and thus most single-nucleotide substitutions had negligible opportunity to reduce tail lengthening (Fig. 4d), which largely explained their relatively weak effects for this mRNA. PAL-AI did predict moderate increases in tail lengthening for a few substitutions that would create CPEs predicted to lengthen tails, but by no more than 14 nt. Although some of these predicted increases were validated (positions –61 and –51), others, particularly those downstream the PAS (positions –35, –31, –7 and –6) were not (Fig. 4d). To understand these discrepancies, we predicted the secondary structure of the last 100nt of the wild-type *dbf4.L* 3' UTR using EternaFold[42]. The substitutions at positions –61 and –51 both created CPEs in loop regions, while the substitutions at positions –35, –31, –7 and –6 all created CPEs with nucleotides paired in a stem loop (Fig. 4d, e), which likely prevented these CPEs from being recognized by the CPEB1 protein.

For *aurkaip1.L* and *ccnb2.L* mRNAs, PAL-AI had difficulty predicting increased tail lengthening caused by substitutions in consecutive positions within a few UTR segments (*aurkaip1.L*: positions –34 to –30 and –19 to –15, Supplementary Fig. 5a; *ccnb2.L* : positions –60 to –40 and –15 to –10, Supplementary Fig. 5d). These substitutions appeared to destabilize stem loops that were sequestering PAS elements in the wild-type UTRs (Supplementary Fig. 5c, f), thus making these elements more accessible for CPSF binding. In support of this hypothesis, substitutions that further stabilized these stem loops reduced tail lengthening (*aurkaip1.L*: A-to-G at position –35 and C-to-U at position –14, Supplementary Fig. 5a, c; *ccnb2.L*: C-to-G at position –54 and C-to-G at position –16, Supplementary Fig. 5d, f). Together, these analyses suggest that sequences within the 3' UTR can indirectly affect poly(A) tail-length changes by modulating the structural accessibility of the CPE and the PAS, a mechanism that likely applies to select mRNAs[33] and thus was not effectively learned by PAL-AI.

As for the *ccnb1.2.L* mRNA, PAL-AI incorrectly predicted the effects of a CPE that overlapped with the PAS (Supplementary Fig. 6a, b). Binding of CPEB1 to this overlapping CPE presumably prevented CPSF from recognizing the PAS, thus inhibiting cytoplasmic polyadenylation as previously suggested[43]. Supporting this hypothesis, the C-to-U mutation at position –18 of *magoh.S* mRNA created an overlapping CPE, which reduced tail lengthening (Supplementary Fig. 4f).

Despite some discrepancies between the predictions and experimental results, these analyses demonstrated that, overall, PAL-AI can predict tail-length changes caused by single-nucleotide mutations within mRNA 3' UTRs.

PAL-AI trained from frog oocyte data predicts poly(A) tail-length change in mammalian oocytes

Because principles of poly(A) tail-length control during oocyte maturation are highly conserved among frogs, mice, and humans[18], we asked if PAL-AI trained from frog oocyte data could predict tail-length changes during mammalian oocyte maturation. For mouse

oocytes, PAL-AI achieved moderate accuracy (Fig. 5a; R_s= 0.55, R_p = 0.57). The performance of the model was likely affected by unmatched developmental stages (metaphase I [MI] in frogs and metaphase II [MII] in mice) and also global deadenylation that was independent of 3'-UTR sequences[9,18]. Similar predictive power for PAL-AI was also observed for human oocytes when tail-length changes were examined in MII using published datasets (Fig. 5b; R_s= 0.6, R_p = 0.58), even though the human measurements were acquired using a different sequencing protocol and platform (PacBio)[8]. Indeed, when we examined tail-length changes in stage-matched human oocytes, comparing metaphase I (MI) and GV stages, the performance for PAL-AI improved (Supplementary Fig. 7a; R_s= 0.68, R_p = 0.6).

Fig. 5

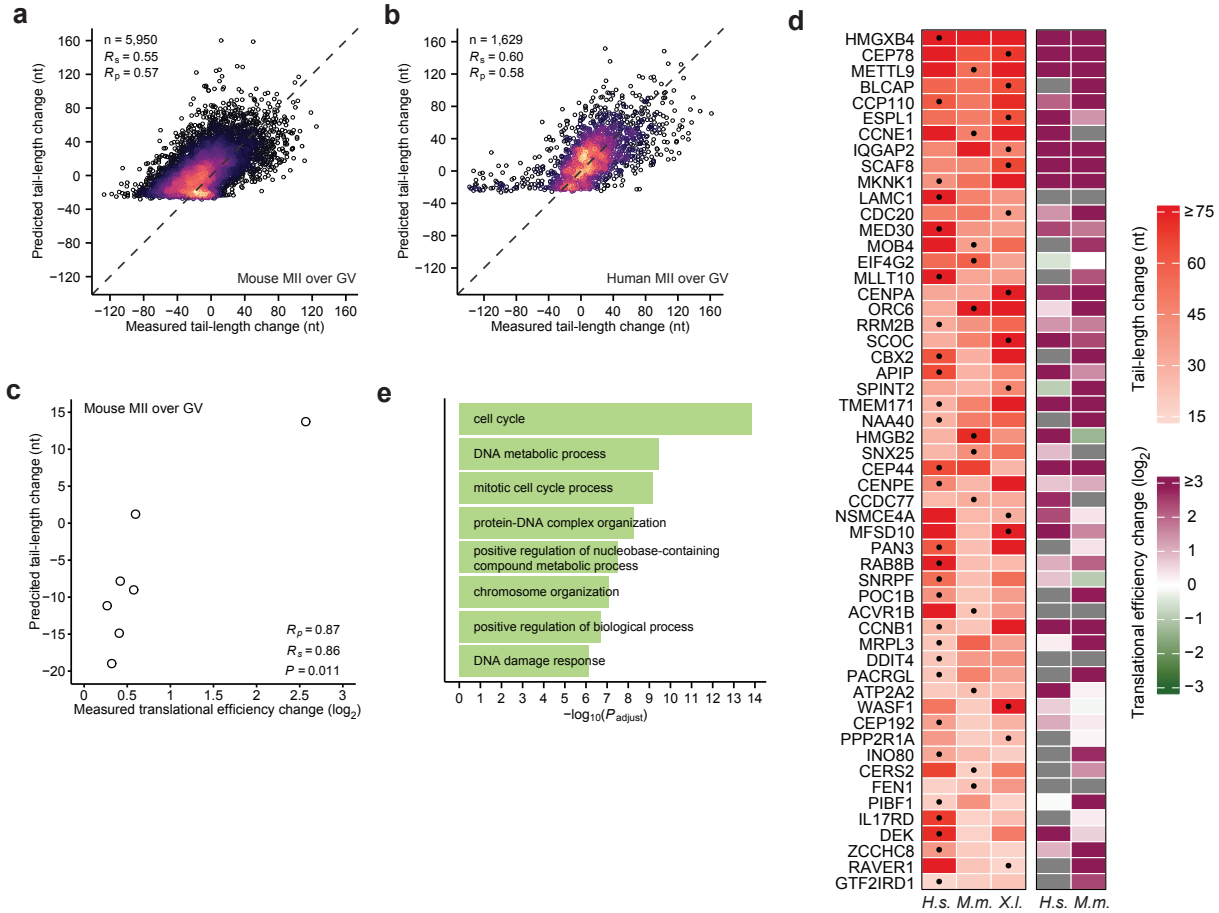


Fig. 5: PAL-AI predicts tail-length changes of endogenous mRNAs of mouse and human oocytes.

a) Performance of frog-trained PAL-AI on mouse mRNAs. Plotted is the relationship between tail-length changes predicted by PAL-AI to occur upon mouse GV-to-MII oocyte maturation, and those that were measured. Each point represents a unique poly(A) site of an endogenous mRNA. Otherwise, this panel is as in Fig. 1a.

b) Performance of frog-trained PAL-AI on human mRNAs. Otherwise, this panel is as in Fig. (**a**).

c) Ability of predicted tail-length changes to explain translational efficiency changes. Plotted is the relationship between PAL-AI-predicted tail-length changes and translational efficiency changes measured by seven mRNA reporters upon mouse GV-to-MII oocyte maturation, as in Xiong et al. (2022)[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 PAL-AI-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,

In frog, mouse, and human oocytes, tail-length changes strongly correlate with translational efficiency changes[3,9,18]. As a result, our model-predicted tail-length changes were expected to be informative regarding up- and down-regulation of mRNA translation not only for endogenous mRNAs but also for any mRNAs with known 3'-UTR sequences. To confirm this expectation, we predicted tail-length changes for a series of reporter mRNAs with different 3'-UTR sequences used for assessing translational regulation during mouse oocyte maturation[6]. Our model-predicted tail-length changes corresponded well with the reported translational efficiency changes (Fig. 5c; R_s= 0.87, R_p = 0.86).

Although tail-length control during oocyte maturation is highly conserved among vertebrates, tail lengthening and associated translational upregulation are known to be required for mRNAs from only a few genes. Conserved tail lengthening of mRNAs from orthologous genes may shed light on the functional significance of tail lengthening during oocyte maturation and early development, but mRNAs from only 19 genes with conserved tail lengthening have been identified[18]. Part of the reason for this small number is that most mRNAs could not be evaluated because they lacked tail-length measurements in at least one of the three species[18]. To overcome this limitation, we used our model-predicted tail-length changes for cases in which tail-length measurements were unavailable, resulting in the identification of another 276 genes (54, 117, and 105 with one, two, and three predicted values respectively) whose mRNA poly(A) tails were either extended or predicted to be extended ≥ 15 nt in all three species (Fig. 5d, Supplementary Fig. 7b, c). Most of these mRNAs had increased translational efficiency, and for those for which translational efficiency changes had not been measured, PAL-AI predicted tail-length increases could be informative. A Gene Ontology analysis of these genes revealed enrichment of molecular pathways such as cell cycle and chromosome organization, which are crucial for meiosis, and DNA metabolic processes and protein-DNA complex organization, which presumably prepare the zygote for cell divisions of early embryonic development (Fig. 5e).

Variants predicted to disrupt poly(A)-tail lengthening are under negative selection in humans and throughout vertebrates

Next, we asked if genetic variants predicted to influence tail length during oocyte maturation have been under evolutionary selection. To this end, we used our model to predict tail-length changes of mRNAs for each 3'-UTR-mapping human genomic variant (~6.5 million) reported in the All of Us Research Program[10]. Then we predicted the effect size of each variant, as the difference between its predicted tail-length change and that of the reference allele (Fig. 6a). Because the 3' ends of mRNAs had the largest influence on tail-length changes (Fig. 1g), we focused our analysis on variants that occurred within the last 100 nt of each 3' UTR. To examine the evolutionary constraints, we calculated the average effect size of all variants with respect to each position of the reference allele and compared it to the conservation at that position of the reference allele, as indicated by the phyloP score, a metric based on the alignment of 100 vertebrate genomic sequences[44,45]. Variants predicted to have the most disruptive outcomes (reduce net tail length of ≥ 25 nt) had significantly higher phyloP scores than those predicted to have moderate outcomes (reduce net tail length ≥ 5 and < 25 nt), and those predicted to have moderate outcomes had significantly higher phyloP scores than those predicted to have minimal outcomes

(reduce net tail length < 5 nt) (Fig. 6b; Mann–Whitney U test, $P < 10^{-63}$ and $P < 10^{-307}$ respectively). Similar results were obtained when PAL-AI was applied to datasets reported in the Genome Aggregation Database (gnomAD v4.1)[11] (Supplementary Fig. 8a, b). These analyses indicated that variants predicted to disrupt tail lengthening during human oocyte maturation have been evolutionarily constrained in the vertebrate lineage.

Fig. 6

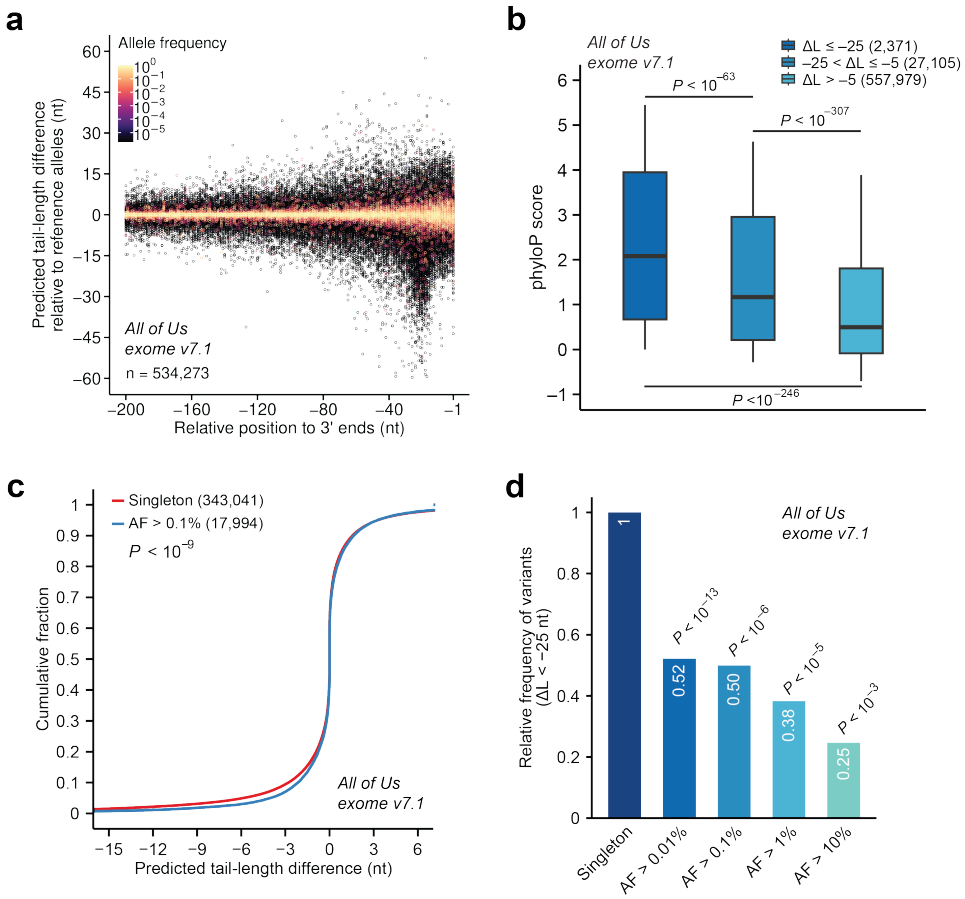


Fig. 6. Variants predicted to disrupt poly(A)-tail lengthening in the human population are selected against.

a) Predicted effects on the tail-length change for human variants. Shown is the PAL-AI-predicted difference between the tail-length change of each variant reported in the All of Us Research Program (exome region v7.1) and that of the reference allele, plotted as a function of the variant position in its 3' UTR. Colors indicate allele frequencies.

b) Higher phyloP scores associated with variants predicted to be more disruptive of tail lengthening. Shown are distributions of phyloP scores for single-nucleotide reference alleles, grouped by average PAL-AI-predicted differences in tail-length changes (Δ L) of each variant with respect to each reference allele, as calculated in (**a**). Only variants within the last 100 nt of 3' UTRs were included in this analysis; *P* value, Mann–Whitney U test. Indicated in the brackets (key) are numbers of reference alleles in each group. Each box-whisker shows the 10th, 25th, 50th, 75th, and 90th percentile.

c) Decreased predicted tail-length changes of singletons compared to common variants. Shown are cumulative distributions of PAL-AI-predicted differences between tail-length changes of variants reported in the All of Us Research Program (exome region v7.1) and those of their reference alleles for singletons (red line) or common variants (allele frequency > 0.1%). Only variants within the last 100 nt of 3' UTRs were included in this analysis; *P* value, Mann–Whitney U test.

d) Enrichment among singletons of variants predicted to severely disrupt poly(A)-tail lengthening. Shown are relative frequencies of variants predicted to decrease tail-length change by more than 25 nt (ΔL < −25 nt) among cohorts of variants with different allele frequencies reported in the All of Us Research Program (exome region v7.1). Only variants within the last 100 nt of 3' UTRs were included in this analysis. Values of relative frequency are indicated within the bars; *P* values were from Fisher's exact tests performed for each cohort against singletons.

We also asked if genetic variants predicted to influence tail-length control during oocyte maturation have functional consequences for human health. Again, we applied PAL-AI to the All of Us Research dataset and focused on variants that occurred within the last 100 nt of each 3' UTR. On average, singleton variants were predicted to be more disruptive of poly(A)-tail lengthening (negative difference of tail-length changes) than were common variants (allele frequency > 0.1%) (Fig. 6c; Mann–Whitney U test, $P < 10^{-9}$). This difference persisted when excluding all variants that would introduce or ablate PAS motifs (Supplementary Fig. 8c), which were presumably also under negative selection due to their nuclear role in specifying mRNA cleavage and polyadenylation[46]. Moreover, severe disruptive outcomes (those predicted to reduce net tail length by > 25 nt) were 2.6-fold more enriched in singleton variants than in variants with allele frequency > 1% (Fisher's Exact test, $P < 10^{-5}$) and 4-fold more than they were in variants with allele frequency > 10% (Fisher's Exact test, $P < 10^{-3}$) (Fig. 6d). In contrast, large-effect-sized gain-of-function variants (those predicted to increase net tail length by > 25 nt) were not significantly more enriched in singleton variants than in common alleles (Supplementary Fig. 8d). Similar results were observed when PAL-AI was applied to the gnomAD v4.1 datasets[11] (Supplementary Fig. 8e–h). Together, these analyses indicated that variants that disrupt poly(A) tail-length control, particularly those that disrupt tail lengthening during oocyte maturation, are under negative selection in humans.

Discussion

We developed PAL-AI, a neural network model that predicted tail-length changes in frog, mouse, and human oocytes using only 3'-UTR sequence information. PAL-AI captured consequential sequence motifs and contextual effects without any feature generation, thereby allowing us to predict tail-length changes of mRNAs with any 3'-UTR sequence. Moreover, PAL-AI helped us identify UGU/GUU-containing motifs contributing to poly(A)-tail lengthening. Although previously implicated in regulating translation[36,39,40,47–49], these motifs had not been associated with tail-length control. Our analyses of the injected mRNA reporter libraries further demonstrated that their tail-lengthening effects were causal rather than correlative with other features of endogenous mRNAs. Individually, these motifs contributed weakly to polyadenylation, but multiple copies did result in considerable tail lengthening, as illustrated by the *mad2l1.L* mRNA. Their precise roles in affecting tail lengths might be context-dependent, as some might serve as CPE-like motifs promiscuously bound by CPEB1, whereas others might be recognized by RNA-binding proteins such as DAZL, which also interacts with CPEB1 and thus might help facilitate recruiting CPEB1 to the 3' UTR.

The ability of PAL-AI to predict tail-length changes of both endogenous mRNAs and injected mRNAs enabled us to capture the consequences of most single-nucleotide substitutions. For a few mRNAs, PAL-AI under-performed in predicting the outcomes of substitutions—often those substitutions that appeared to alter the structural accessibility of CPE or PAS elements. However, when we combined base-pairing probability, as predicted by RNAfold[35], with the 3'-UTR sequences as the input for our model, the performance did not significantly improve, which suggested that more accurate

parameterization of mRNA structure would be needed for the model to account for structural accessibility. Perhaps incorporating data from in-oocyte structural probing of mRNAs will further improve the model predictions[50–53].

Because of the strong influence of tail length on translational efficiency in oocytes, we envision that our model can be used to engineer mRNA 3'-UTR sequences for desired protein expression in these systems. Furthermore, by predicting tail-length changes that had not yet been measured in frog, mouse, or human oocytes, our model helped identify a group of genes that appear to undergo conserved mRNA tail-lengthening during oocyte maturation. Although for mRNAs of most of these genes the importance of translational up-regulation during oocyte maturation and early embryogenesis still needs to be examined, knowledge of these genes can inform efforts to identify mutations that affect human female fertility.

Indeed, genetic variants predicted by PAL-AI to disrupt tail-length control during oocyte maturation have been under negative selection in vertebrates including humans. Presumably, some of these variants perturb sequence elements or contextual features important for cytoplasmic polyadenylation, thus leading to failed translational activation of mRNAs encoding proteins important for oocyte maturation or early embryonic development. Presumably with larger data sets and more sensitive analyses, global negative selection will also be detected for polyadenylation-promoting gain-of-function mutations of mRNAs that should be repressed during these developmental stages.

Past efforts to identify genetic causes of female infertility have primarily focused on variants in coding regions, perhaps in part due to the use of exome sequencing[23]. However, 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[54]. Given the importance of tail-length and translational control in human oocytes and early embryos, as well as our analyses of variant abundancies, 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, and the Whitehead Institute Genome Technology Core for sequencing. D.P.B. is an investigator of the Howard Hughes Medical Institute.

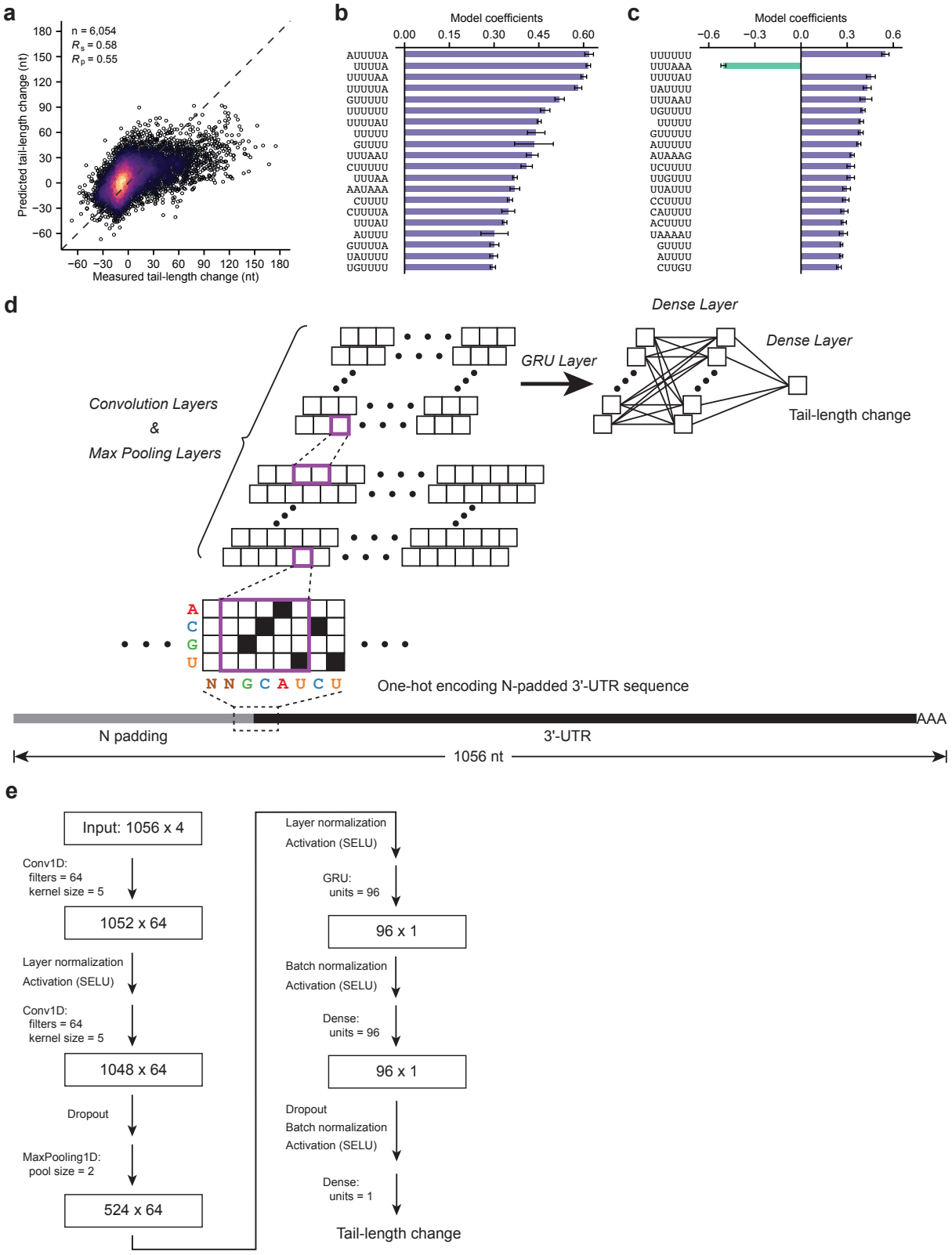
Author contributions

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

Competing interests

The authors declare no competing interests.

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) Performance of the multiple linear regression model; otherwise, as in Fig. 1a.

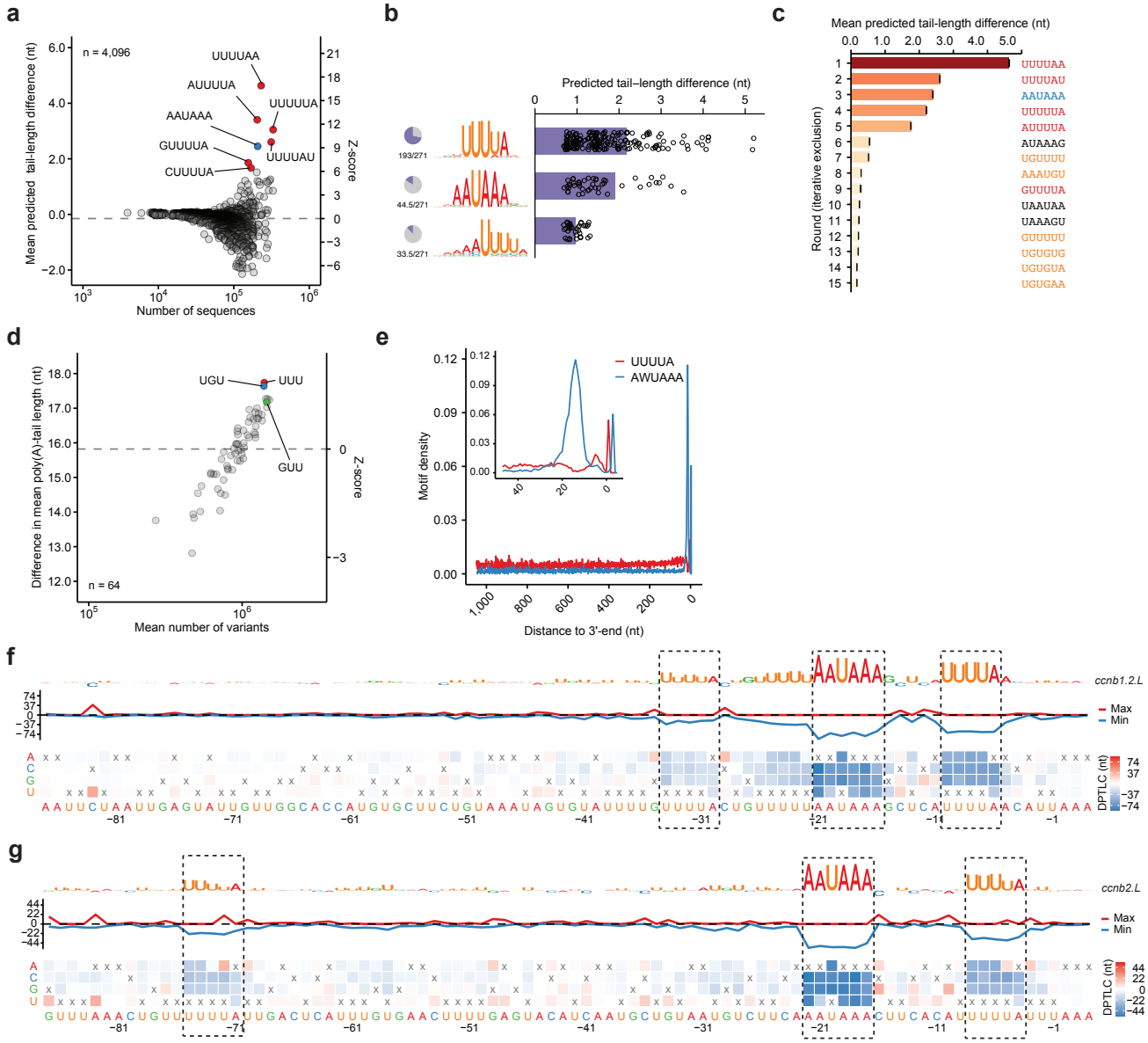
b) Top-ranked features as the number of *k*-mers from the multiple linear regression model. Plotted are mean coefficients of the 20 top-ranked (by absolute value) features as the number of *k*-mers from the model. Error bars indicate standard deviation of coefficients of the 10 models from 10-fold cross validation.

c) Top-ranked features as the position metric of *k*-mers from the multiple linear regression model. 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.

Supplementary Fig. 2



Supplementary Fig. 2: Additional data supporting sequence elements and contextual features learned by PAL-AI.

a) As in **Fig. 1c**, except for consequences of gain of sequence motifs.

b) As in **Fig. 1d**, except for 8-mers most associated with the largest differences in predicted tail length between mutants in which 8-mers were gained and the wild-type.

c) As in **Fig. 1e**, except for top 6-mers associated with increased predicted tail-length change when gained in the in silico mutagenesis analysis.

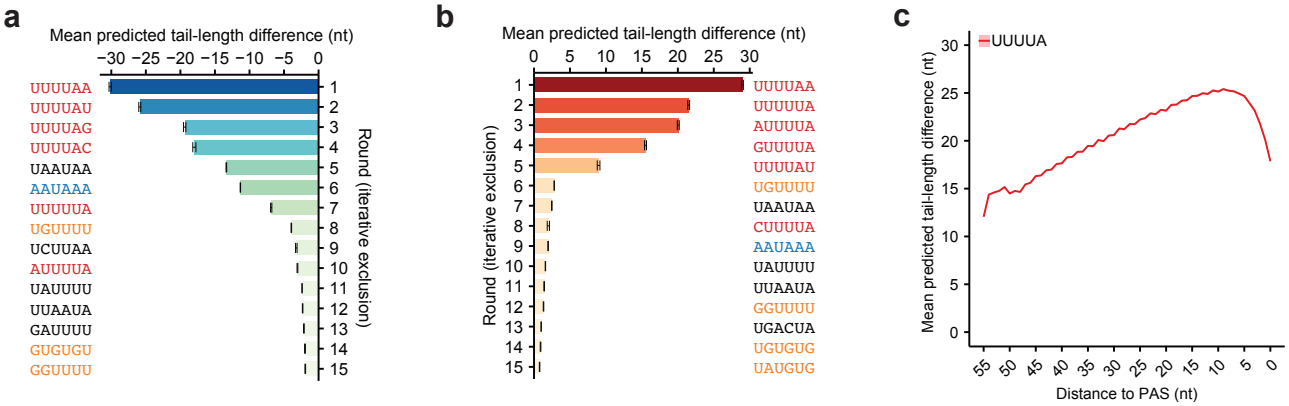
d) Tail-length changes associated with each 3-mer in the 3' UTRs of the N60-PAS^{mos} library[18], comparing between 0 and 7 h post-progesterone treatment. Plotted for each 3-mer are differences in mean tail lengths observed for mRNAs with that 3-mer. Only variants containing one CPE (UUUUA) were included in this analysis.

e) Densities of the PAS and the CPE along the last 1053 nt of frog mRNA 3' UTRs. The inset shows the results for the last 50 nt of the 3' UTR.

f) Impact of single-nucleotide substitutions on predicted tail-length changes for the *ccnb1.2.L* mRNA 3' UTR; otherwise, as in Fig. 1i.

g) Impact of single-nucleotide substitutions on predicted tail-length changes for the *ccnb2.L* mRNA 3' UTR; otherwise, as in Fig. 1i.

Supplementary Fig. 3



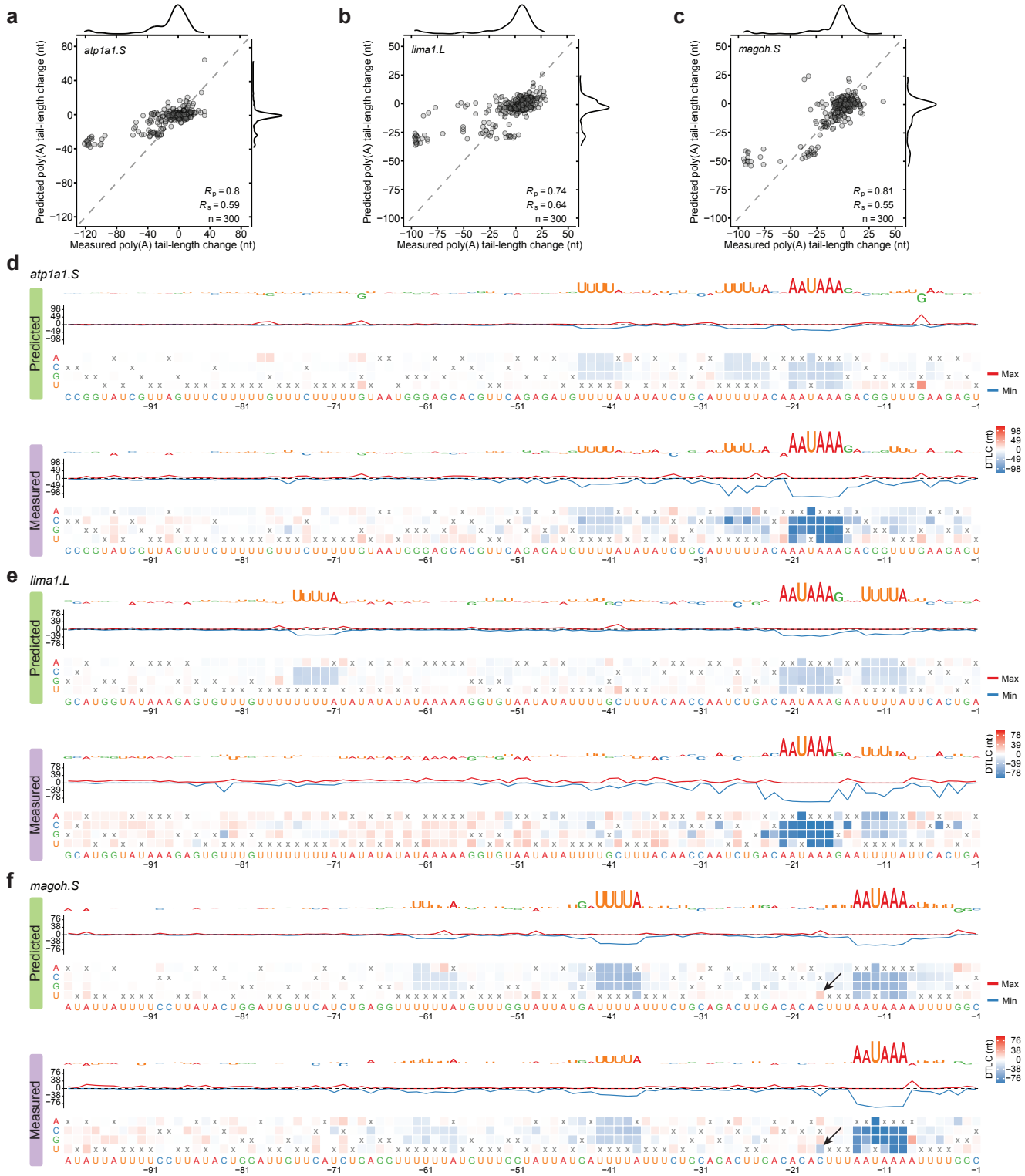
Supplementary Fig. 3. Additional data supporting the prediction of tail-length changes for synthetic mRNAs in frog oocytes.

a) 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. 1e.

b) As in (**a**), except for top 6-mers associated with increased predicted tail-length change when gained in the in silico mutagenesis analysis.

c) 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. 1g.

Supplementary Fig. 4



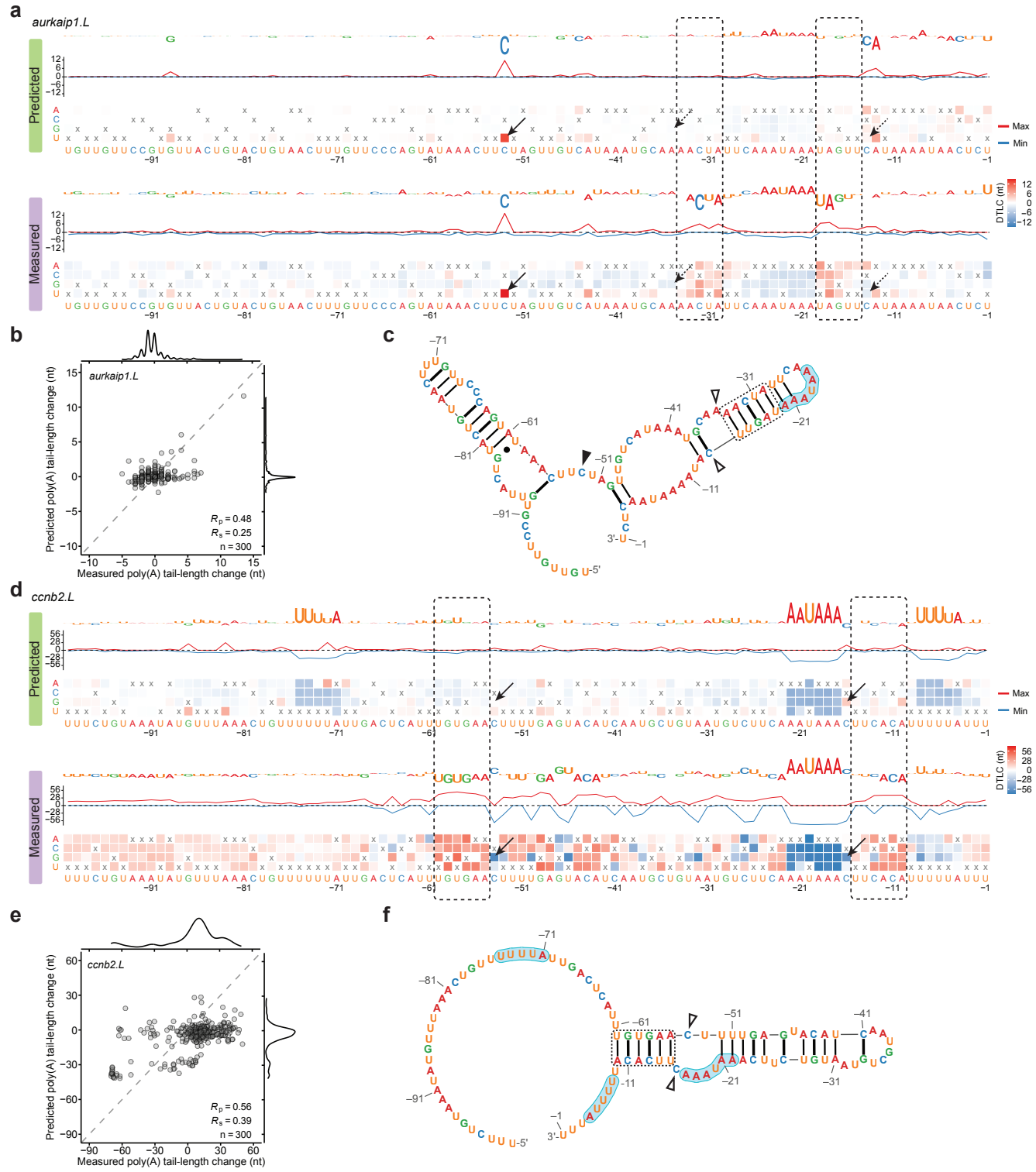
Supplementary Fig. 4: Additional experimental validation of predicted effects of single-nucleotide substitutions.

a) PAL-AI prediction of the effects of single-nucleotide substitutions of the *atp1a1.S* mRNA. Otherwise, this panel is as in Fig. 3c,

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. 3e**.
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 difference in tail-length changes of a C-to-U substitution that creates an overlapping CPE with the PAS.

Supplementary Fig. 5



Supplementary Fig. 5: 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) where the PAS is located. Otherwise, this panel is as in **Fig. 4d**.

b) PAL-AI prediction of the effects of single-nucleotide substitutions of the *aurkaip1.L* mRNA. Otherwise, this panel is as in **Fig. 3c**.

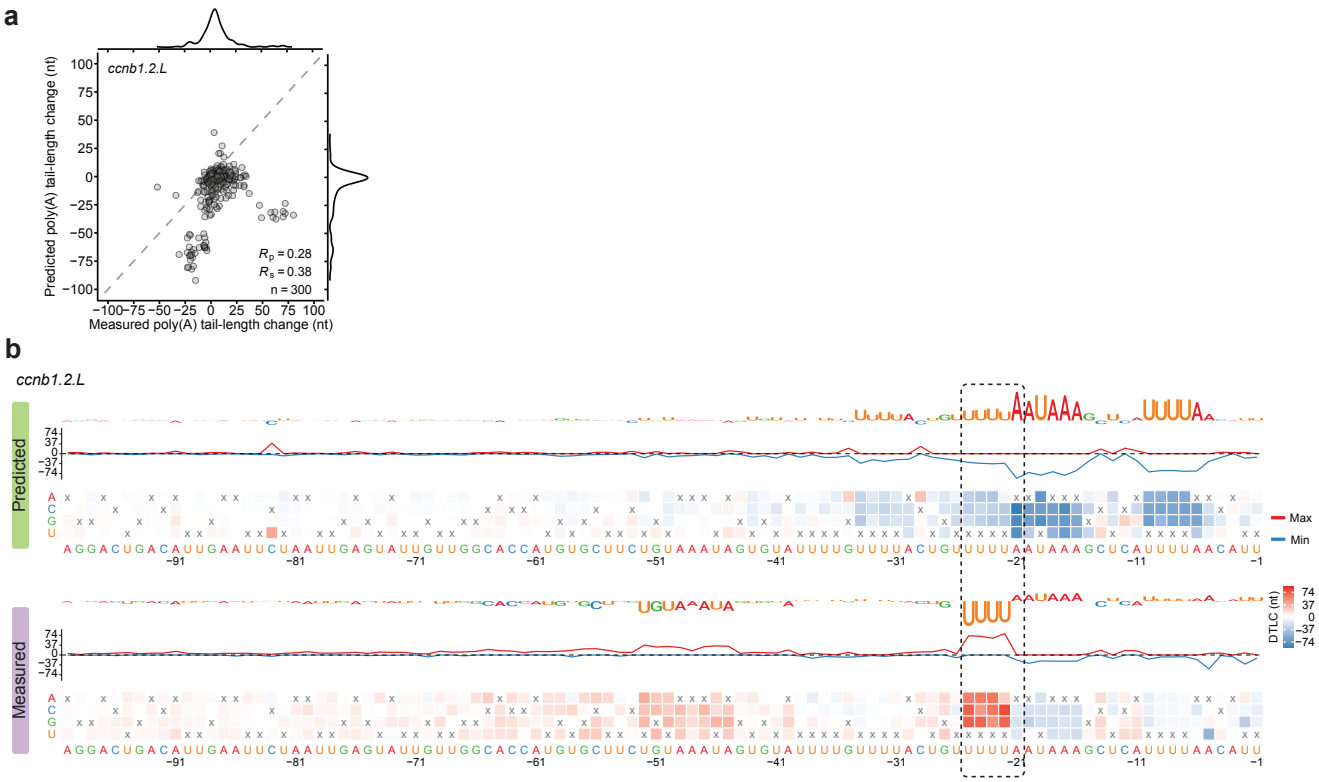
c) EternaFold-predicted maximum expected accuracy secondary structure of the last 100 nt of the *aurkaip1.L* 3' UTR. The dashed rectangle indicates the stem of a hairpin, as highlighted in (**a**). The triangles indicate positions of either a C-to-U substitution (solid) that creates a CPE in a predicted 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 where the PAS is located.

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 a C-to-G substitution would extend the predicted stem indicated with the dashed rectangle.

Supplementary Fig. 6

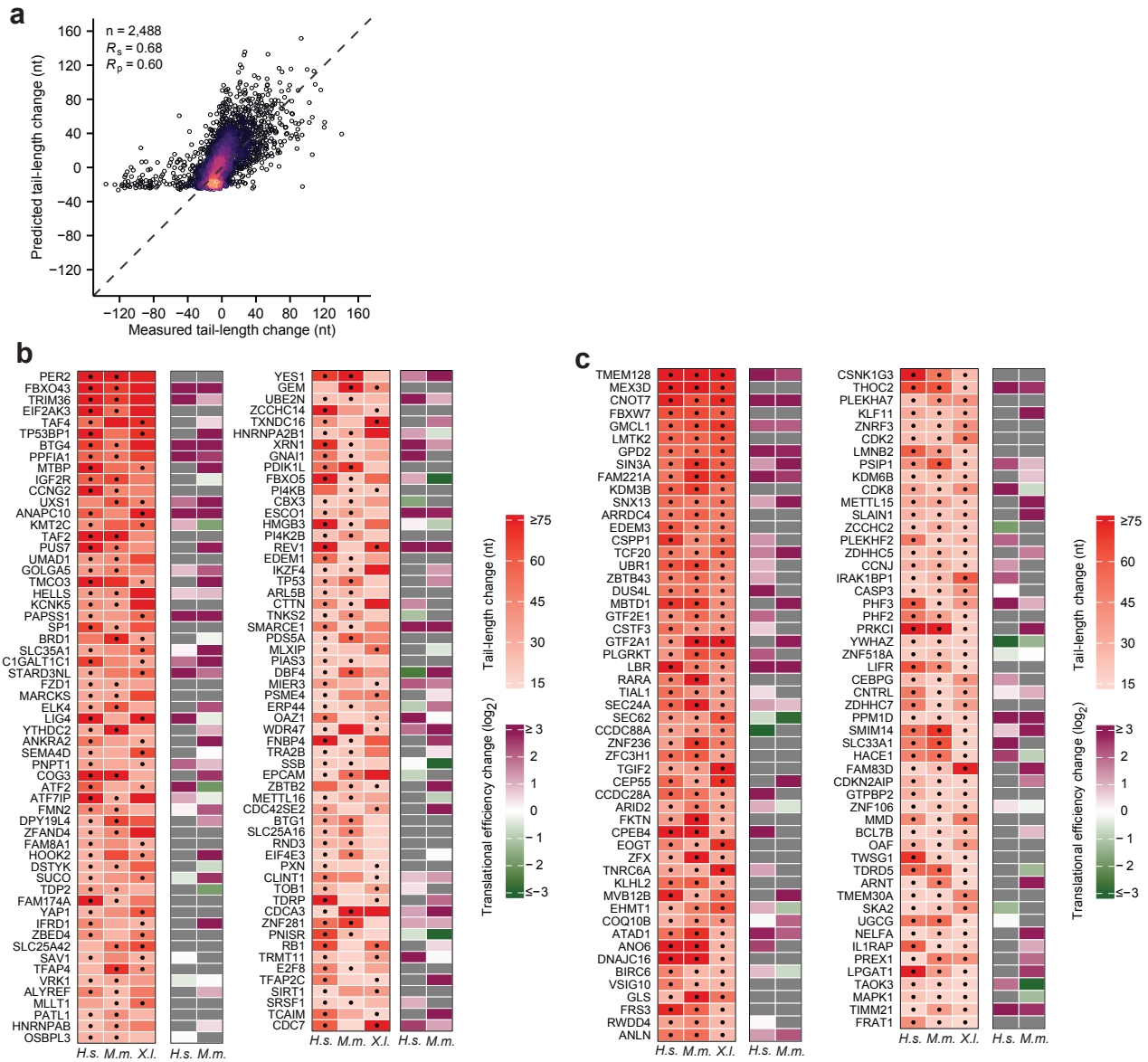


Supplementary Fig. 6: Additional experimental validation of predicted effects of single-nucleotide substitutions exemplifying consequences of CPE-PAS overlap.

a) PAL-AI prediction of the effects of single-nucleotide substitutions of the *ccnb1.2.L* mRNA. Otherwise, this panel is as in **Fig. 3c**.

b) Impact of single-nucleotide substitutions on predicted and measured tail-length changes for the *ccnb1.2.L* mRNA. The dashed rectangle indicates the CPE that overlaps with a PAS. Otherwise, this panel is as in **Fig. 3e**.

Supplementary Fig. 7



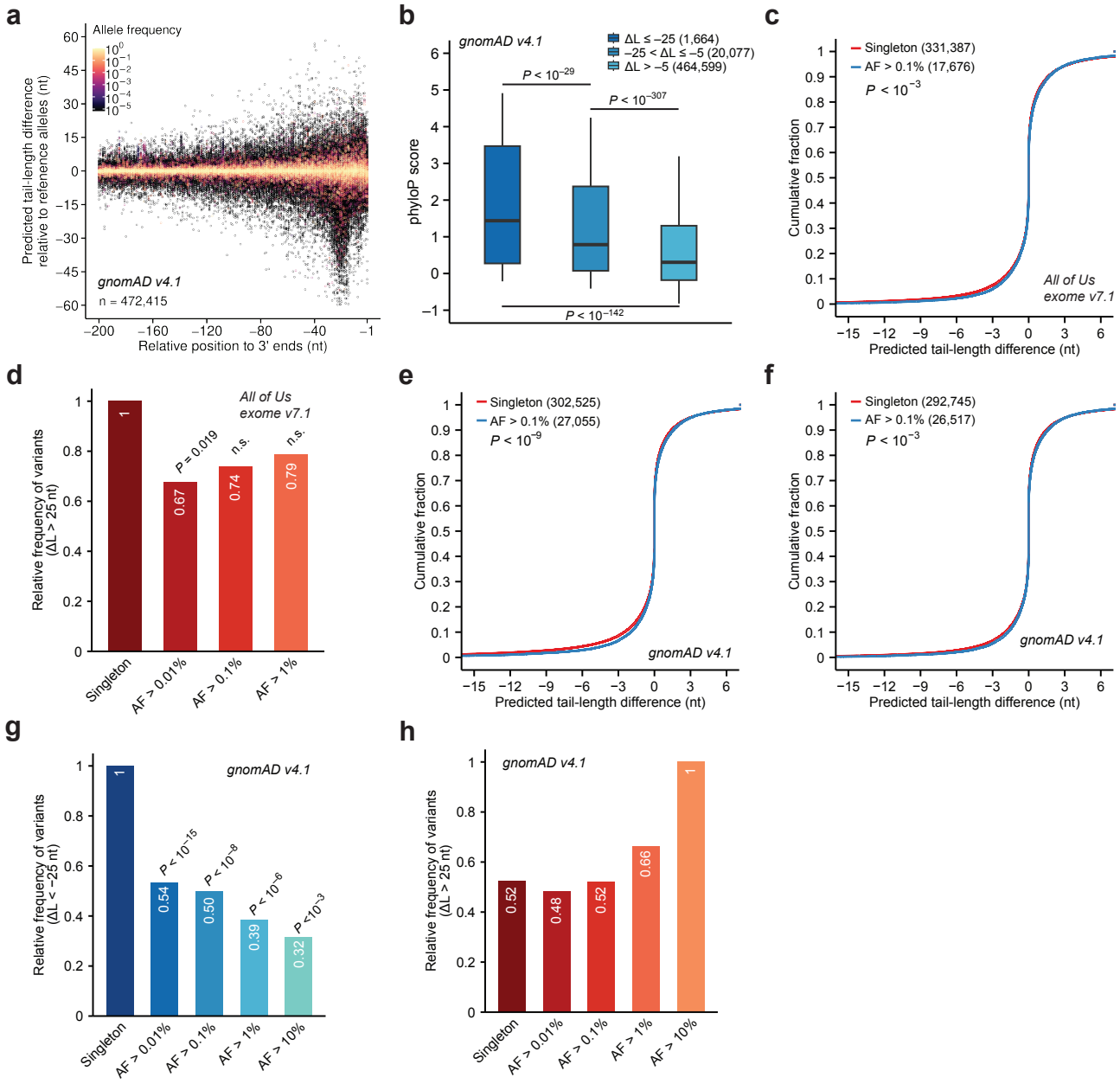
Supplementary Fig. 7: Additional data supporting the prediction of tail-length change for endogenous mRNAs in mouse and human oocytes.

a) Performance of frog-trained PAL-AI on human mRNAs, as in Fig. 5b, except that measured tail-length changes were compared between human GV and MI oocytes.

b) Genes predicted or shown to have substantial mRNA tail-lengthening. This panel is as in Fig. 5d, except genes with PAL-AI-predicted values in two of the three species are shown.

c) Genes predicted to have substantial mRNA tail-lengthening. This panel is in (**b**), except genes with PAL-AI-predicted values in all three species are shown.

Supplementary Fig. 8



Supplementary Fig. 8. Additional data supporting negative selection of human variants predicted to disrupt poly(A)-tail lengthening.

a) Analysis of gnomAD data, showing predicted effects on tail-length change for human variants. This panel is as in Fig. 6a, but for human variants in the gnomAD v4.1 dataset.

b) Higher phyloP scores associated with variants predicted to be more disruptive of tail lengthening. This panel is as in Fig. 6b, but for human variants in the gnomAD v4.1 dataset.

c) Decreased tail-length changes of singletons compared to common variants, when excluding effects on PAS elements. This panel is as in Fig. 6c, except variants that introduced or eliminated PAS elements were excluded.

d) No significant enrichment among singletons of variants predicted to strongly increase poly(A)-tail lengthening. This panel is as in Fig. 6d, except it examines variants predicted to increase tail-length change by more than 25 nt (ΔL > 25 nt).

e) Analysis of gnomAD data, showing decreased predicted tail-length changes of singletons compared to common variants. This panel is as in Fig. 6c, but for human variants in the gnomAD v4.1 dataset.

f) Analysis of gnomAD data, showing decreased predicted tail-length changes of singletons compared to common variants, when excluding effects on PAS elements. This panel is as in (**c**), but for human variants in the gnomAD v4.1 dataset.

g) Analysis of gnomAD data, showing enrichment among singletons of variants predicted to severely disrupt poly(A)-tail lengthening. This panel is as in Fig. 6d, but for human variants in the gnomAD v4.1 dataset.

h) Analysis of gnomAD data, showing no significant enrichment among singletons of variants predicted to strongly increase poly(A)-tail lengthening. This panel is as in (**d**), but for human variants in the gnomAD v4.1 dataset.

Methods

Linear regression models for poly(A) tail-length changes. The poly(A) tail-length change for each isoform of frog oocyte endogenous mRNAs was calculated 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 that were not annotated as "uncertain" in previously curated annotations[18] were included. Positions and numbers of all *k*-mers in the length of 5–6 nt within the last 1000 nt of the 3' UTR of each mRNA isoform were used as features for the linear regression model. For mRNA isoforms with 3' UTRs shorter than 1000 nt, the entire 3' UTR was examined. MRNA isoforms with 3' UTRs shorter than 10 nt were excluded. Each 3' UTR was appended with a tri-nucleotide sequence "AAA" before extracting *k*-mer positions and numbers. The number of each *k*-mer N^k was counted with the *oligonucleotideFrequency* function from the R package *Biostrings*[55]. A composite position feature value P^k was calculated for each *k*-mer to integrate information on all positions of the *k*-mer for an mRNA isoform, defined as

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

where d_i represents the distance between the *i*-th *k*-mer and the 3' end. The linear regression model for the tail-length change of the *i*-th mRNA isoform L_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 a set of coefficients to be optimized. All features were fit and scaled using *RobustScaler* from the Python *scikit-learn* package[56]. Data were stratified into 10 folds by preserving the proportion of mRNA isoforms that contained a CPE (UUUUA) within 3' UTRs in each fold using *StratifiedKFold* from Python *scikit-learn* package[56]. The model was trained using *SGDRegressor* from the Python *scikit-learn* package[56]. Hyperparameters for the model were optimized with *RandomizedSearchCV* from the Python *scikit-learn* package[56]. The final parameters used for regression were "penalty='elasticnet', l1_ratio = 0.555, alpha = 0.0509, power_t = 0.255, eta0 = 0.00000689, max_iter = 100, n_iter_no_change = 10". The model performance was evaluated in a 10-fold cross-validation procedure. In this procedure, the model was trained with 9-fold data and tested on the remaining fold. This step was repeated 5 times and the best model, evaluated by the R squared between the measured tail-length changes and the predicted tail-length changes, was kept to predict the remaining fold. The training set and the testing set were rotated among the 10 folds of data. Finally, the predicted tail-length changes from each fold as the testing set were concatenated and compared with the measured tail-length changes.

PAL-AI for poly(A) tail-length changes. The tail-length change data used for neural network models were the same as those used for linear regression models. For the input, RNA sequences from the last 1053 nt of the 3' UTR of each mRNA isoform were used. Each 3' UTR was appended with a tri-nucleotide sequence "AAA" at the 3' end. The resulting 3'

UTRs shorter than 10 nt were removed. For the retained 3' UTRs shorter than 1056 nt, “N”s were padded to the 5' ends to make the total length of each sequence 1056nt. The sequences were one-hot encoded into four tracks, with “N” represented by four zeros.

The model integrated a convolution neural network and a recurrent neural network in an architecture similar to that adopted by Saluki[34]. A total of four blocks constituted the model. The first block contained an input layer and a 1-dimensional convolution layer, followed by layer normalization and SELU activation. The second block contained repeats of the same mini-block, which comprised a 1-dimensional convolution layer, a dropout layer, and a 1-dimensional max-pooling layer followed by layer normalization and SELU activation. The third block contained a gated recurrent unit (GRU) layer (single direction going forward) followed by batch normalization and SELU activation. The fourth block contained a dense layer, a dropout layer followed by batch normalization and SELU activation, and a final dense layer with one output value for the tail-length change. All convolution layers had the same number of filters and kernel size. All convolution layers, as well as the GRU layer and the dense layer had the same kernel regularizer. All dropout layers had the same dropout rate. As in linear regression models, input data were stratified into 10 folds, and the proportion of mRNA isoforms that contained a CPE (UUUUA) within 3' UTRs in each fold was kept roughly the same.

The model was constructed using the Python *tensorflow* package[57] and was trained using the Adam optimizer on batches of 50 input mRNA isoforms with the mean squared error as the loss and the negative R squared as the metric. The training was performed for 30 epochs or stopped early if no improvement was observed for the loss of the validation set after 10 epochs. Extensive optimizations were performed with *gp_minimize* from the Python *scikit-optimize* package[58] to find the best model for the following hyperparameters (values in the final model indicated in the parentheses): number of filters (64) and kernel size (5) of the convolution layers, the number of units of the dense layer (96), regularization factors of the kernel regularizer (l1 = 0.01, l2 = 0.000001), the dropout rate (0.1), the pool size of the max-pooling layer (2), the number of units of the GRU layer (96), the number of repeated mini-blocks in the second block (1), the learning rate of the optimizer (0.001). In total, the model had 78,593 trainable parameters. Evaluation of the model performance was performed in a 10-fold cross-validation procedure similar to that in linear regression models. The final trained model contained an ensemble of 10 models, each of which was the best model (highest R squared) among the five models repeated on training and testing of the data split in one of the ten ways. All downstream predictions made with this model were performed with all 10 models, and the average of the predicted values was used as the final prediction result.

When comparing the influence of different regions of 3' UTR on the model performance, sequences corresponding to the indicated regions were extracted as the input and the shape of the input layer was changed accordingly. When the coding regions were included in the input, coding sequences were padded to the 5' ends of triple-A-appended 3' UTRs that were shorter than 1056 nt. “N”s were padded to the resulting sequences if they were still shorter than 1056 nt. A fifth track was then added with 1 representing nucleotides within the 3' UTR or 0 otherwise. When secondary structure information was included, triple-A-appended 3'-UTR sequences (last 1053 nt plus 3 As)

were used for prediction with *RNAfold*[35]. The pairing probability was calculated for each nucleotide and added as a sixth track in addition to the fifth track that labeled coding regions. For nucleotides in the coding regions and padded "N"s, 0 was assigned to the sixth track.

The neural network model built for the N60(LC)-PAS^{mos} library had the same architecture as that used for endogenous mRNAs. Training and testing were performed for the N60(LC)-PAS^{mos} library in the same way as that for endogenous mRNAs.

In silico mutagenesis (ISM). For endogenous mRNAs, all 3' UTRs that were not annotated as "uncertain" in previously curated annotations[18] were used. For the N60(LC)-PAS^{mos} library, all 3' UTRs in the curated reference (see **Variant and tail-length analysis for the injected mRNA libraries**) were used. Sequences were truncated to 1053 nt (if longer than 1053 nt) and appended with "AAA" at the 3' end. The resulting 3' UTRs shorter than 10 nt were discarded. The retained 3' UTRs shorter than 1056 nt were padded with "N"s at the 5' end to make the total length of each sequence 1056nt. The sequences were one-hot encoded into four tracks and the neural network model was used to predict tail-length changes of wild-type sequences. For ISM, each nucleotide of each 3' UTR was mutated to three possible alternative nucleotides, and PAL-AI was used to predict the tail-length change of each single-nucleotide substitution mutant. The difference was calculated between the predicted tail-length change of the wild type and that of the mutant.

When examining the tail-length change associated with the loss or the gain of a *k*-mer, the differences between the predicted tail-length change of the wild type and those of all the single-nucleotide substitution mutants that disrupted the original *k*-mer or introduced a new *k*-mer were averaged for all appearances of this *k*-mer within 3' UTRs of all mRNA isoforms.

In some cases, *k*-mer-associated differences in predicted tail-length changes were examined in an iterative exclusion process. In each round of the iteration, if the mutation of a mutant disrupted (when examining *k*-mer loss) or introduced a *k*-mer (when examining *k*-mer gain) in the *k*-mer exclusion list (starting with an empty list in the first round), it would be excluded from the analysis. 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 result of the *k*-mer with the largest tail-length increase (when examining *k*-mer gain) or decrease (when examining *k*-mer loss) was recorded, and this *k*-mer was added to the exclusion list before starting the next round. The iteration ended when either there was no *k*-mer associated with statistically significant increased (or decreased) predicted tail-length change (Bonferroni correction-adjusted $P < 0.05$ in a Welsh's t-test) or no *k*-mer had a Z-score higher than 2 (when examining *k*-mer gain) or lower than –2 (when examining *k*-mer loss).

Motif insertion analysis. Filtering and encoding of the input sequences were performed in the same way as that in ISM. A motif was inserted into a position of a 3' UTR of an mRNA by replacing the sequence of the same length at the same position with the motif. PAL-AI was used to predict the tail-length change of the wild-type 3' UTR and that of the motif-inserted mutant. The difference was calculated between the tail-length changes as the insertion

outcome. When examining the relative positions between the PAS and the CPE, only mRNAs with 3' UTRs containing at least a PAS within the last 100 nt were included in the analysis. The relative distance was calculated between a CPE and the PAS closest to the 3' end.

Preparation of the N60(LC)-PAS^{mos} library. The DNA template for the N60(LC)-PAS^{mos} library was generated by PCR with oligos KXU024 and KXU068 from 2 amol of the DNA template for the N60-PAS^{mos} library used previously[18] in a 100-µl reaction with the KAPA HotStart HiFi Kits (Roche, KK2502). The PCR products were purified with agarose gels (Lonza, 50004) and the GeneJet Gel Extraction Kit (Thermo Fisher, K0692). The mRNA library used for oocyte injection was in vitro transcribed and capped, followed by removal of the 3'-end cyclic phosphate as described previously.[18]

Preparation of the single-nucleotide mutagenesis library. The oligo pool for the single-nucleotide mutagenesis library was synthesized as SurePrint HiFi Oligo from Agilent Technologies (Supplementary Table 2). Sequences of the pool contain a variable region derived from the last 100 nt of the 3' UTR of one of the ten frog oocyte mRNAs (*atp1a1.S*, *ccnb1.2.L*, *ccnb2.L*, *lima1.L*, *mad2l1.L*, *magoh.S*, *mos.L*, *tpx2.L*, *aurkaip1.L* and *dbf4.L*). Each nucleotide at each position along the 3' UTR was mutated to the alternative three possible nucleotides, generating 3010 variants (10 wild-type sequences and 3000 mutants). A 35-nt poly(A) fragment was appended at the 3' end of the variable region. Two constant sequences were then added to the two ends for PCR amplification (24 nt at the 5' end and 20 nt at the 3' end). In addition to the forward strand, the reverse strand of each library sequence was also included in the oligo pool.

The DNA template for in vitro transcription was assembled from three DNA fragments by overlapping PCR using the KAPA HiFi HotStart Kit. Fragment 1 (F1) was amplified from the N60-PAS^{mos} DNA template[18] with primer sets KXU024 and KXU236. Fragment 2 (F2) was amplified from the oligo pool for the single-nucleotide mutagenesis library with primer sets KXS067 and KXU237. Fragment 3 (F3) was amplified from the C071 plasmid[4] with primer sets KXU110 and KXU068. F1 and F2 were joined before the product was joined with F3 by overlapping PCR with a standard 100-µl reaction containing 10 pmol of each fragment, 2 ml KAPA HiFi enzyme, 3 ml 10 mM dNTP, and 20 ml 5x KAPA HiFi buffer for 15 cycles, with an annealing temperature of 66°C. All PCR products were purified with agarose gels and the GeneJet Gel Extraction Kit. The mRNA library was in vitro transcribed and capped, followed by removal of the 3'-end cyclic phosphate as described previously.[18]

Preparation of the reporter mRNA poly(A) tail-length standard mix. Each of the four reporter mRNA poly(A) tail-length standards was assembled from two fragments by overlapping PCR. Both the first (F1) and the second (F2) fragments were amplified using the KAPA HiFi HostStart Kits following the manufacturer's suggested protocol from plasmid C071 constructed in the previous study[4] with the following primer sets: KXU024 and one of KXU290, KXU291, KXU292, and KXU293 [for F1 of poly(A) standards with a length of 10, 50, 90, or 130 nt]; KXU068 and one of KXUm020, KXUm021, KXUm022 and KXUm023 [for F2 of poly(A) standards with a length of 10, 50, 90, or 130 nt]. Overlapping PCR reactions were

performed to join F1 and F2 in a 50-µl reaction containing 100 fmol of each fragment, 1 µl KAPA HiFi enzyme, 1.5 µl 10 mM dNTP, and 10 µl 5x KAPA HiFi buffer. Five PCR cycles were performed following the manufacturer's suggested protocol with an annealing temperature of 63°C. When finished, oligos KXU024 and KXU068 were added to a final concentration of 300 nM with 1 µl KAPA HiFi enzyme, 1.5 µl 10 mM dNTP, and 10 µl 5x KAPA HiFi buffer to a total volume of 100 µl. Twenty-four cycles of PCR were performed with the same settings. PCR products were purified with agarose gels and the GeneJet Gel Extraction Kit. Each RNA was in vitro transcribed and capped, followed by removal of the 3'-end cyclic phosphate as described previously[18]. Four mRNA poly(A) tail-length standards were mixed in equal molar and used for downstream applications. Each standard contained a unique 4-nt barcode in the 3' UTR introduced by one of the primers KXU290, KXU291, KXU292, and KXU293.

Reporter mRNA library injection and sample collection. Frog oocytes used for the injection were obtained from Xenopus1 (12005). Healthy Stage V and VI oocytes were hand-picked and transferred to 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 mM Na_2HPO4) supplemented with 100 µg/ml Gentamicin (Thermo Fisher, 15750060) at 18°C overnight (> 16 hr) for recovery before injections. Injections were performed at 23°C with PLI-100 Plus Pico-Injector. For the N60(LC)-PAS^{mos} library, 4 nl mRNA (0.1 pmol/µl) was injected per oocyte.

Frog oocytes were matured in vitro in OR-2 buffer supplemented with 10 µg/ml progesterone (Millipore Sigma, P0130) diluted from a 10 mg/ml stock in ethanol. The time it took for frog oocytes to mature varied between different batches. In most cases, 50% of oocytes reached GVBD (displaying a white spot in the animal pole) between 3 hr and 5 hr and 100% of oocytes reached GVBD by 7 hr after incubation with progesterone.

Frog oocytes (100 as a group) were collected at the indicated time after progesterone treatment. OR-2 buffer was removed and oocytes were washed three times with ice-cold buffer RL (20 mM HEPES pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 1% (v/v) Triton X-100, 100 µg/ml cycloheximide, cOmplete protease inhibitor cocktail [MilliporeSigma, 11836170001, 1 tablet per 10 ml buffer], 10 µl/oocyte). After removing all wash buffer, oocytes were lysed in buffer RL (10 µl per oocyte) supplemented with 200 units/ml SUPERase•In (10 µl/oocyte) by vigorous shaking and pipetting. Lysates were cleared by centrifugation at 5000 g at 4°C for 10 min. The supernatant was mixed with 3x volume TRIzol LS (Thermo Fisher, 10296010). Note that the "0 h post progesterone treatment" sample was prepared by collecting the supernatant of the lysate of untreated oocytes and combining it with the uninjected mRNA library before the Tri-Reagent was added to the mix. All RNAs isolated with the TRIzol-LS were performed with Phasemaker tubes (Thermo Fisher, A33248) according to the manufacturer's protocol. Total RNAs were resuspended in 40 µl water (0.4 µl/oocyte).

Library preparation and sequencing of the injected mRNA libraries. Total RNAs (10 µl, ~30 µg) isolated from injected oocytes were combined with 10 amol with reporter mRNA poly(A) tail-length standard mix and then mixed with 8 pmol KXSH009, 8 pmol KXSH010, and 2x SSC (0.3 M NaCl, 30 mM sodium citrate pH 7.0) in a total of 50 µl. The RNA and oligos were annealed by incubation at 70°C for 5 min and then slowly cooling to 23°C at

0.1°C/sec. The annealed mixture was combined with 40 µl MyOne Streptavidin C1 beads (Thermo Fisher, 65002) and incubated for 20 min at 23°C on a thermal mixer, shaking with 15 sec on and 1 min 45 sec off. The supernatant was separated from the beads with a magnetic rack and removed. The beads were washed twice with 300 µl 1xB&W buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl) and once with 300 µl 2x SSC. The RNA was eluted from the beads first with 100 µl 10 mM HEPES pH 7.5 at 65°C for 3 min and then second with 100 µl water at 65°C for 3 min. The eluates were combined, precipitated with ethanol, and resuspended in 6.5 µl water.

The anti-sense oligo-enriched RNA was ligated to a pre-adenylated 3' adapter in a 10 µl reaction containing 5 µM 3' adapter KXS347a, 50 mM HEPES pH 7.5, 10 mM $MgCl_2$, 10 mM dithiothreitol, 1 unit/µl T4 RNA ligase 1 (New England Biolabs, M0204S). The ligation reaction was incubated at 23°C for 150 min. After ligation, RNA was extracted with phenol/chloroform, precipitated with ethanol, and resuspended in 11.4 µl water. The ligated RNA was mixed with 0.6 µl 100 µM reverse transcription primer KXS348 and 1 µl 10 mM dNTPs in a total volume of 13 µl, incubated at 65°C for 5 min, and cooled on ice for 1 min. The annealed RNA was reverse transcribed in a 20 µl reaction containing 1x SSIV Buffer, 5 mM dithiothreitol, 1 unit/µl SUPERase•In, and 200 units SuperScript IV Reverse Transcriptase (ThermoFisher, 18090050) at 55°C for 15 min. After reverse transcription, RNA was hydrolyzed with 3.3 µl 1 M NaOH at 90°C for 10 min followed by neutralization with 24 µl 1 M HEPES pH 7.5. The cDNA library was ethanol-precipitated, resuspended in water, and amplified in a 50 µl PCR reaction with KXS037 and a barcoded primer (KXS057) that hybridized to a constant region of the library using the KAPA HiFi HotStart Kits following the manufacturer's suggested protocol for 11~14 cycles. The PCR-amplified library was cleaned up twice with AMPure XP beads (Beckman Coulter, A63881) with a beads-to-sample ratio of 1.2.

Sequencing was performed on an Illumina HiSeq 2500, with a custom run of two reads. The first read was sequenced with the primer KXS067 for 281 cycles. The second read was sequenced in the same direction with a standard TruSeq sequencing primer for 10 cycles. Additionally, to build a reference sequence, the N60(LC)-PAS^{mos} library made from the 0 h post progesterone sample (uninjected) was also sequenced on AVITI, with a custom run of two reads in the same direction for 299 cycles and 10 cycles respectively, using the same primers as those for the HiSeq.

Variant and tail-length analysis for the injected mRNA libraries. To build a reference sequence for the N60(LC)-PAS^{mos} library, sequencing reads (read 1) of the 0 h post progesterone sample (uninjected) from AVITI were trimmed down to 81 nt from 5' ends. Trimmed reads were examined for the presence of the constant region in the 3' UTR of mRNA poly(A) tail-length standards (ACCAGCCTCAAGAACACCCGAATGG) with 7 allowed mismatches in the first 25 nt, allowing positional offsets between –3 and +3 nt. The remaining reads were examined for the expected constant sequence AATAAAGAAATTGATTTGTCT at position 61–81, allowing positional offsets between –3 and +3 nt. Reads with a 21-nt sequence that had no more than six mismatches to the constant sequence at any allowed positions were retained. Retained reads that had a non-zero positional offset were modified to make the offset value zero: "N" bases were added to the

beginning of the read if the offset value was negative, while the first few bases at the start of the read were removed if the offset value is positive. Then the modified reads were collapsed using *UMICollapse*[59]. Clusters with more than 100 reads were kept. The primary sequence of each cluster was used as the reference sequence for the N60(LC)-PAS^{mos} library.

The sequencing run performed on HiSeq 2500 for the N60(LC)-PAS^{mos} and the single-nucleotide mutagenesis libraries stuttered at cycle 26 in read 1 and reported the same bases for cycle 26 and cycle 27. As a result, the base at cycle 27 was removed from all sequencing reads before downstream analysis.

For the N60(LC)-PAS^{mos} libraries, sequencing reads (read 1) from the HiSeq run were mapped to the reference sequence using *STAR* (v2.7.1)[60] with the parameters '--outFilterMultimapNmax 1 --alignEndsType Local –clip3pNbases 199 --outSAMattributes All --outSAMtype BAM SortedByCoordinate'. Poly(A)-tail lengths of the mapped reads were determined the same way as those of the N60-PAS^{mos} libraries described previously[18].

For the single-nucleotide mutagenesis libraries, sequencing reads (read 1) from the HiSeq run were mapped to the reference sequence (the last 100 nt wild-type sequence of the 3' UTR of ten selected frog oocyte mRNAs) using *STAR* (v2.7.1)[60] with the parameters '--outFilterMultimapNmax 1 --outFilterMismatchNmax 1 --alignEndsType EndToEnd –clip3pNbases 180 --outSAMattributes All --outSAMtype BAM SortedByCoordinate'. After mapping, the CIGAR label in the bam file was used to determine the type of mutation for each variant. Poly(A)-tail lengths of the mapped reads were determined the same way as those of the N60(LC)-PAS^{mos} libraries, except that the starting position of the poly(A) tail was set at 101 of the read 1. Only the wild-type and single-nucleotide substitution mutants were kept. For each retained variant, poly(A) tags (read clusters) that contained the same read 2 (serving as a unique molecular identifier) sequences were grouped and collapsed, and the median tail length of the grouped poly(A) tags was used as the poly(A)-tail length of the collapsed tag.

Prediction of poly(A) tail-length changes of the frog oocyte-injected mRNA libraries. For both N60(LC)-PAS^{mos} and the single-nucleotide mutagenesis libraries, the tail-length change for each mRNA variant was calculated between median tail lengths measured at 0 and 7 h post progesterone treatment, requiring at least 50 poly(A) tags at both time points. The whole 3' UTR of each mRNA variant was used and encoded in the same way as those of frog mRNAs used for training PAL-AI. Predictions of tail-length changes were made with PAL-AI trained with frog oocyte endogenous mRNAs.

RNA secondary structure prediction. *EternaFold* (v1.3.1)[42] was used to predict secondary structures of the 3' UTR sequences of the single-nucleotide mutagenesis library with parameters "--params parameters/EternaFoldParams.v1". The dot-bracket output was used to make figures with *RNAcanvas*[61].

Prediction of poly(A) tail-length changes in mouse and human oocytes. Mouse oocyte 3' UTR annotations were obtained from a previous study[18]. Poly(A) tail-length measurements in mouse GV oocytes and MII oocytes were obtained from a previous

study[18] and merged for each pair of replicates. The tail-length change for each mRNA 3' UTR isoform was calculated between the median tail length in GV oocytes and that in MII oocytes, requiring at least 50 poly(A) tags in both merged datasets. PAL-AI trained with frog oocyte endogenous mRNAs was used to predict tail-length changes in mouse oocytes. The input sequences of mouse mRNA 3' UTRs were filtered and encoded in the same way as those of frog mRNAs used for training the model.

For human oocytes, 3' UTR annotations were obtained from a previous study[18]. PAIso-seq[8] bam files were downloaded from Genome Sequence Archive for Human (GSA-Human: https://ngdc.cncb.ac.cn/gsa-human/) with the accession number HRA001911. The reverse-complement sequence of the soft-clipped region at the 5' end of each read was extracted from the bam file. The poly(A)-tail length was determined by the length of the longest consecutive As (allowing individual non-A nucleotides within the stretch) that started before the third nucleotide of the extracted sequence. The tail-length change for each mRNA 3' UTR isoform was calculated between the median tail length in GV oocytes and that in MI or MII oocytes, requiring at least 50 poly(A) tags in both datasets. Prediction of tail-length change for human oocytes was performed the same way as that for mouse oocytes.

For assessing the relationship between predicted tail-length changes and measured translational efficiency changes of injected reporter mRNAs in mouse oocytes, source data were obtained from a previous study (Fig. 4g in the original publication).[6] The translational efficiency change of each reporter was determined by the difference between the mean of normalized fluorescence signals among GV oocytes and that among MI oocytes. The sequences of reporter mRNA 3' UTRs were deduced from the cloning and mutagenesis primers used in this study. Tail-length changes were predicted for these sequences using PAL-AI trained with frog endogenous mRNAs.

Poly(A) tail-length and translational efficiency changes of homologous genes in frogs, mice, and humans. The frog-to-human gene ID conversion table was downloaded from Xenbase. The mouse-to-human gene ID conversion table was downloaded from the Ensembl website. Measured tail-length changes were compared between GV and MII oocytes in humans, GV and MII oocytes in mice, and frog oocytes 0 and 7 h post progesterone treatment. For each mRNA isoform with a unique 3' end in one of the three species, the tail-length change predicted by PAL-AI trained in frog oocyte endogenous mRNAs was used if the measured data was not available. If a gene had multiple mRNA isoforms, the dominant isoform (> 50% by tag count in GV oocytes of humans and mice or frog oocytes 0 h post progesterone treatment) was chosen to represent that gene. For each human gene, if there were more than one homologous gene in another species (mice or frogs), the mean tail-length changes of all the homologous genes from that species were used.

Translational efficiencies measured by ribosome profiling and mRNA-seq were obtained for mouse oocytes from a published study[6], and for human oocytes from another published study[7]. Data were reprocessed as described previously[18]. Translational efficiency changes were compared between GV and MII oocytes in both humans and mice.

For gene ontology (GO) analysis, genes that had substantial tail lengthening (≥ 15 nt, by measurement or prediction) in human, mouse, and frog oocytes during maturation processes were selected and ranked by the minimum of tail-length changes of the three species. The ranked gene name list was used as the query input for *gost* from the R package *gprofiler2*[62] with the parameters "ordered_query = TRUE, user_threshold = 0.05, correction_method = "g_SCS". The top 8 non-overlapping categories (not belonging to the same parental GO category, ranked by Benjamini-Hochberg-adjusted *P* values) in the GO:Biological Processes from the output were selected.

Prediction of poly(A) tail-length changes for variants in the All of Us Research Program and variants in gnomAD. Human oocyte 3' UTR annotations were obtained from a previous study[18]. MRNA 3' UTR isoforms shorter than 10 nt were removed. To avoid losing variants due to incomplete annotations, the "uncertain" 3' UTRs in the annotations were kept for this analysis if their 3' ends were in the PolyA_DB_v3.2 database[63].

For the data reported in the All of Us Research Program[10], all single-nucleotide polymorphism and indel variants in the short-read whole-genome sequencing callsets encompassing human exome regions (version 7.1) were intersected with human oocyte 3' UTR genomic locations and variants that overlapped were kept. The sequence of the mRNA 3' UTR associated with each variant was derived by replacing the reference sequence with the alternative sequence of the variant. Three As were appended to the 3' end. The tail-length change for each retained variant was predicted by PAL-AI trained with frog endogenous mRNAs. The mutational outcome was measured by the difference between the tail-length change of each variant and that of the reference allele (wild-type). If a variant intersected with more than one 3' UTRs, the one associated with the largest absolute difference in tail-length change was registered.

For the data reported in gnomAD[11], genomic variants (v4.1, from 76,215 human individuals) were downloaded from gnomad.broadinstitute.org. Variants that passed all filters (indicated by "PASS" in the "FILTER" column in the VCF file) and intersected with that 3' UTR at genomic locations were retained. The tail-length change for each retained variant and the mutational outcome were obtained the same as that for variants reported in the All of Us Research Program.

For the conservation analysis, phyloP scores were downloaded from the UCSC Genome Browser database[45]. Only phyloP scores of single-nucleotide reference alleles in either the All of Us Research dataset or the gnomAD dataset were retained. The mutational outcome for each variant was calculated by the difference between the tail-length change of the reference allele and that of the variant. For each reference allele, mutational outcomes of all variants with respect to the reference allele were averaged.

Quantification and statistical analysis

Graphs were generated and statistical analyses were performed using R[64]. 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 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.

Other publicly available data analyzed in this study were indicated in 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.

References

1. Eichhorn, S. W. *et al.* mRNA poly(A)-tail changes specified by deadenylation broadly reshape translation in Drosophila oocytes and early embryos. *eLife* **5**, e16955 (2016).

2. Lim, J., Lee, M., Son, A., Chang, H. & Kim, V. N. mTAIL-seq reveals dynamic poly(A) tail regulation in oocyte-to-embryo development. *Genes & Development* **30**, 1671–1682 (2016).

3. Yang, F., Wang, W., Cetinbas, M., Sadreyev, R. I. & Blower, M. D. Genome-wide analysis identifies cis-acting elements regulating mRNA polyadenylation and translation during vertebrate oocyte maturation. *RNA (New York, N.Y.)* **26**, 324–344 (2020).

4. Xiang, K. & Bartel, D. P. The molecular basis of coupling between poly(A)-tail length and translational efficiency. *eLife* **10**, e66493 (2021).

5. Liu, Y. *et al.* Poly(A) tail length is a major regulator of maternal gene expression during the mammalian oocyte-to-embryo transition. *bioRxiv* 2021.08.29.458052 (2021) doi:10.1101/2021.08.29.458052.

6. Xiong, Z. *et al.* Ultrasensitive Ribo-seq reveals translational landscapes during mammalian oocyte-to-embryo transition and pre-implantation development. *Nat Cell Biol* 1–13 (2022) doi:10.1038/s41556-022-00928-6.

7. Zou, Z. *et al.* Translatome and transcriptome co-profiling reveals a role of TPRXs in human zygotic genome activation. *Science* **378**, abo7923 (2022).

8. Liu, Y. *et al.* Remodeling of maternal mRNA through poly(A) tail orchestrates human oocyte-to-embryo transition. *Nat Struct Mol Biol* 1–16 (2023) doi:10.1038/s41594-022-00908-2.

9. Lee, K., Cho, K., Morey, R. & Cook-Andersen, H. An extended wave of global mRNA deadenylation sets up a switch in translation regulation across the mammalian oocyte-to-embryo transition. *Cell Rep.* **43**, 113710–113710 (2024).

10. Investigators, T. A. of U. R. P. G. *et al.* Genomic data in the All of Us Research Program. *Nature* 1–7 (2024) doi:10.1038/s41586-023-06957-x.

11. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

12. Richter, J. D. & Lasko, P. Translational control in oocyte development. *Cold Spring Harbor Perspectives in Biology* **3**, a002758 (2011).

13. Teixeira, F. K. & Lehmann, R. Translational Control during Developmental Transitions. *Csh Perspect Biol* **11**, a032987 (2019).

14. Conti, M. & Kunitomi, C. A genome-wide perspective of the maternal mRNA translation program during oocyte development. *Semin Cell Dev Biol* (2023) doi:10.1016/j.semcdb.2023.03.003.

15. Sheets, M. D., Fox, C. A., Hunt, T., Woude, G. V. & Wickens, M. The 3'-untranslated regions of c-mos and cyclin mRNAs stimulate translation by regulating cytoplasmic polyadenylation. *Genes & Development* **8**, 926–938 (1994).

16. Moor, C. H. de & Richter, J. D. The Mos pathway regulates cytoplasmic polyadenylation in Xenopus oocytes. *Molecular and Cellular Biology* **17**, 6419–6426 (1997).

17. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**, 66–71 (2014).

18. Xiang, K., Ly, J. & Bartel, D. P. Control of poly(A)-tail length and translation in vertebrate oocytes and early embryos. *Dev. Cell* **59**, 1058-1074.e11 (2024).

19. McGrew, L. L., Dworkin-Rastl, E., Dworkin, M. B. & Richter, J. D. Poly(A) elongation during Xenopus oocyte maturation is required for translational recruitment and is mediated by a short sequence element. *Gene Dev* **3**, 803–815 (1989).

20. Radford, H. E., Meijer, H. A. & Moor, C. H. de. Translational control by cytoplasmic polyadenylation in Xenopus oocytes. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1779**, 217–229 (2008).

21. Aoki, F., Hara, K. T. & Schultz, R. M. Acquisition of transcriptional competence in the 1-cell mouse embryo: Requirement for recruitment of maternal mRNAs. *Mol. Reprod. Dev.* **64**, 270–274 (2003).

22. Meneau, F., Dupré, A., Jessus, C. & Daldello, E. M. Translational Control of Xenopus Oocyte Meiosis: Toward the Genomic Era. *Cells 2020, Vol. 9, Page 1502* **9**, 1502 (2020).

23. Yatsenko, S. A. & Rajkovic, A. Genetics of human female infertility. *Biol. Reprod.* **101**, 549–566 (2019).

24. Zheng, W. *et al.* Homozygous Mutations in BTG4 Cause Zygotic Cleavage Failure and Female Infertility. *Am. J. Hum. Genet.* **107**, 24–33 (2020).

25. Liu, R. *et al.* A novel homozygous missense variant in BTG4 causes zygotic cleavage failure and female infertility. *J. Assist. Reprod. Genet.* **38**, 3261–3266 (2021).

26. Varnum, S. M. & Wormington, W. M. Deadenylation of maternal mRNAs during Xenopus oocyte maturation does not require specific cis-sequences: a default mechanism for translational control. *Gene Dev* **4**, 2278–2286 (1990).

27. Fox, C. A. & Wickens, M. Poly(A) removal during oocyte maturation: a default reaction selectively prevented by specific sequences in the 3' UTR of certain maternal mRNAs. *Genes & Development* **4**, 2287–2298 (1990).

28. Hake, L. E. & Richter, J. D. CPEB is a specificity factor that mediates cytoplasmic polyadenylation during Xenopus oocyte maturation. *Cell* **79**, 617–627 (1994).

29. Dickson, K. S., Bilger, A., Ballantyne, S. & Wickens, M. P. The Cleavage and Polyadenylation Specificity Factor in Xenopus laevis Oocytes Is a Cytoplasmic Factor Involved in Regulated Polyadenylation. *Mol Cell Biol* **19**, 5707–5717 (1999).

30. Charlesworth, A., Meijer, H. A. & Moor, C. H. de. Specificity factors in cytoplasmic polyadenylation. *Wiley Interdiscip Rev Rna* **4**, 437–461 (2013).

31. Ivshina, M., Lasko, P. & Richter, J. D. Cytoplasmic Polyadenylation Element Binding Proteins in Development, Health, and Disease. *Annual Review of Cell and Developmental Biology* **30**, 393–415 (2014).

32. Reyes, J. M. & Ross, P. J. Cytoplasmic polyadenylation in mammalian oocyte maturation. *Wiley Interdisciplinary Reviews: RNA* **7**, 71–89 (2016).

33. Weill, L., Belloc, E., Castellazzi, C. L. & Méndez, R. Musashi 1 regulates the timing and extent of meiotic mRNA translational activation by promoting the use of specific CPEs. *Nature Structural & Molecular Biology* **24**, 672–681 (2017).

34. Agarwal, V. & Kelley, D. R. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol* **23**, 245 (2022).

35. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithm Mol Biol* **6**, 26 (2011).

36. Collier, B., Gorgoni, B., Loveridge, C., Cooke, H. J. & Gray, N. K. The DAZL family proteins are PABP-binding proteins that regulate translation in germ cells. *EMBO J.* **24**, 2656–2666 (2005).

37. Bauermeister, D., Claußen, M. & Pieler, T. A novel role for Celf1 in vegetal RNA localization during Xenopus oogenesis. *Dev. Biol.* **405**, 214–224 (2015).

38. Dominguez, D. *et al.* Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell* **70**, 854-867.e9 (2018).

39. Martins, J. P. S. *et al.* DAZL and CPEB1 regulate mRNA translation synergistically during oocyte maturation. *J. Cell Sci.* **129**, 1271–1282 (2016).

40. Yang, C.-R. *et al.* The RNA-binding protein DAZL functions as repressor and activator of mRNA translation during oocyte maturation. *Nat. Commun.* **11**, 1399 (2020).

41. Wu, X. & Bartel, D. P. Widespread Influence of 3'-End Structures on Mammalian mRNA Processing and Stability. *Cell* **169**, 905-917.e11 (2017).

42. Wayment-Steele, H. K. *et al.* RNA secondary structure packages evaluated and improved by high-throughput experiments. *Nat Methods* 1–9 (2022) doi:10.1038/s41592-022-01605-0.

43. Piqué, M., López, J. M., Foissac, S., Guigó, R. & Méndez, R. A Combinatorial Code for CPE-Mediated Translational Control. *Cell* **132**, 434–448 (2008).

44. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).

45. Raney, B. J. *et al.* The UCSC Genome Browser database: 2024 update. *Nucleic Acids Res.* **52**, D1082–D1088 (2023).

46. Linder, J., Koplik, S. E., Kundaje, A. & Seelig, G. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol* **23**, 232 (2022).

47. Padmanabhan, K. & Richter, J. D. Regulated Pumilio-2 binding controls RINGO/Spy mRNA translation and CPEB activation. *Genes Dev.* **20**, 199–209 (2006).

48. Chen, J. *et al.* Genome-wide analysis of translation reveals a critical role for deleted in azoospermia-like (Dazl) at the oocyte-to-zygote transition. *Genes & Development* **25**, 755–766 (2011).

49. Mikedis, M. M. *et al.* DAZL mediates a broad translational program regulating expansion and differentiation of spermatogonial progenitors. *eLife* **9**, e56523 (2020).

50. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2013).

51. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).

52. Spitale, R. C. *et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **advance online publication**, 486–490 (2015).

53. 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).

54. Sang, Q., Ray, P. F. & Wang, L. Understanding the genetics of human infertility. *Science* **380**, 158–163 (2023).

55. DebRoy, H. P. and P. A. and R. G. and S. Biostrings: Efficient manipulation of biological strings. https://bioconductor.org/packages/Biostrings (2021).

56. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

57. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Preprint at https://www.tensorflow.org/.

58. Head, T. *et al. Scikit-Optimize/Scikit-Optimize: V0.5.2 (v0.5.2)*. (2018).

59. Liu, D. Algorithms for efficiently collapsing reads with Unique Molecular Identifiers. *PeerJ* **7**, e8275 (2019).

60. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

61. Johnson, P. Z. & Simon, A. E. RNAcanvas: interactive drawing and exploration of nucleic acid structures. *Nucleic acids Res.* **51**, W501–W508 (2023).

62. 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).

63. 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).

64. Team, R. C. R: A Language and Environment for Statistical Computing. (2021).