nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-019-0164-5

Human 5′ UTR design and variant effect prediction from a massively parallel translation assay

Paul J. Sample[1,5], Ban Wang[1,5], David W. Reid[2], Vlad Presnyak[2], Iain J. McFadyen[2], David R. Morris[3] and Georg Seelig[1,4*]

The ability to predict the impact of *cis*-regulatory sequences on gene expression would facilitate discovery in fundamental and applied biology. Here we combine polysome profiling of a library of 280,000 randomized 5′ untranslated regions (UTRs) with deep learning to build a predictive model that relates human 5′ UTR sequence to translation. Together with a genetic algorithm, we use the model to engineer new 5′ UTRs that accurately direct specified levels of ribosome loading, providing the ability to tune sequences for optimal protein expression. We show that the same approach can be extended to chemically modified RNA, an important feature for applications in mRNA therapeutics and synthetic biology. We test 35,212 truncated human 5′ UTRs and 3,577 naturally occurring variants and show that the model predicts ribosome loading of these sequences. Finally, we provide evidence of 45 single-nucleotide variants (SNVs) associated with human diseases that substantially change ribosome loading and thus may represent a molecular basis for disease.

The sequence of the 5′ UTR is a primary determinant of translation efficiency[1,2]. While many *cis*-regulatory elements within human 5′ UTRs have been characterized individually, the field still lacks a means to accurately predict protein expression from 5′ UTR sequence alone, limiting the ability to estimate the effects of genome-encoded variants and the ability to engineer 5′ UTRs for precise translation control. Massively parallel reporter assays (MPRAs; methods that assess thousands to millions of sequence variants in a single experiment) coupled with machine learning have proven useful in addressing similar challenges by producing quantitative biological insight that would be difficult to obtain through traditional approaches[3–9].

Earlier MPRAs designed to learn aspects of 5′ UTR *cis* regulation relied on fluorescence-activated cell sorting[10,11] or growth selection[12] to stratify libraries by activity. These techniques require the expression of a single library variant per cell, which must be transcribed within the cell from a DNA template, making it difficult to distinguish between the effects of transcriptional and translational control. Polysome profiling[13] overcomes this limitation by enabling single cells to translate tens to hundreds of in vitro-transcribed (IVT) and transfected mRNA variants. Polysome profiling has been used extensively to measure translation of native RNA isoforms[14,15], but isolating the role of 5′ UTR regulation has been difficult owing to differences in the size and sequence of the concomitant coding sequences (CDSs) and 3′ UTRs.

Here we report the development of an MPRA that measures the translation of hundreds of thousands of randomized 5′ UTRs via polysome profiling and RNA sequencing. We then use the data to train a convolutional neural network (CNN) that can predict ribosome loading from sequence alone.

Results

MPRA design and validation. To build a model capable of predicting ribosome loading of human 5′ UTR variants and designing new 5′ UTRs for targeted expression (Fig. 1a), we first created a library with 280,000 gene sequences consisting of a random 5′ UTR and a constant region containing the CDS for enhanced green fluorescent protein (eGFP) and a 3′ UTR (Fig. 1b). Specifically, the 5′ UTR of each construct began with 25 nucleotides of defined sequence used for PCR amplification, followed by 50 nucleotides of fully random sequence before the eGFP CDS. HEK293T cells were transfected with IVT library mRNA, cells were collected after 12 h and polysome fractions were then collected and sequenced (Supplementary Fig. 1). For a given UTR, the relative counts per fraction were multiplied by the number of ribosomes associated with each fraction and the resulting values were summed to obtain a measured mean ribosome load (MRL; Supplementary Note 1). Below, we refer to the entire workflow required to measure the MRL of all 5′ UTRs in a library, that is, library transfection, polysome profiling, high-throughput sequencing and MRL analysis, as a ‘polysome profiling experiment’. We initially focused on the first 50 bases upstream of the CDS to specifically investigate the regulatory signals that mediate the initiation of translation beyond ribosomal recruitment to the 5′ cap. Intriguingly, variants within the 50-nucleotide window directly adjacent to the start codon are under stronger negative selection than those further upstream[16], providing another motivation to focus on this window.

To validate our approach, we asked whether it captured known aspects of translation regulation. Translation initiation is largely dependent on start codons and their context and position relative to a CDS[12,17]. Our data clearly showed the expected decrease in ribosome loading for sequences with either out-of-frame upstream start codons (uAUGs) (Fig. 1c) or upstream open reading frames (uORFs)[18,19] (Supplementary Fig. 2b). On average, we observed considerably lower use of CUG and GUG as alternative start codons as compared to AUG (Fig. 1c and Supplementary Figs. 3 and 4), in contrast to other reports that have shown widespread usage of non-AUG start sites[15,20,21]. This difference is possibly due to these alternative start codons being used more often under stress conditions[22]. However, we found that CUG and GUG start codons could

[1]Department of Electrical Engineering, University of Washington, Seattle, WA, USA. [2]Moderna, Cambridge, MA, USA. [3]Department of Biochemistry, University of Washington, Seattle, WA, USA. [4]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA. [5]These authors contributed equally: Paul J. Sample, Ban Wang. *e-mail: gseelig@uw.edu

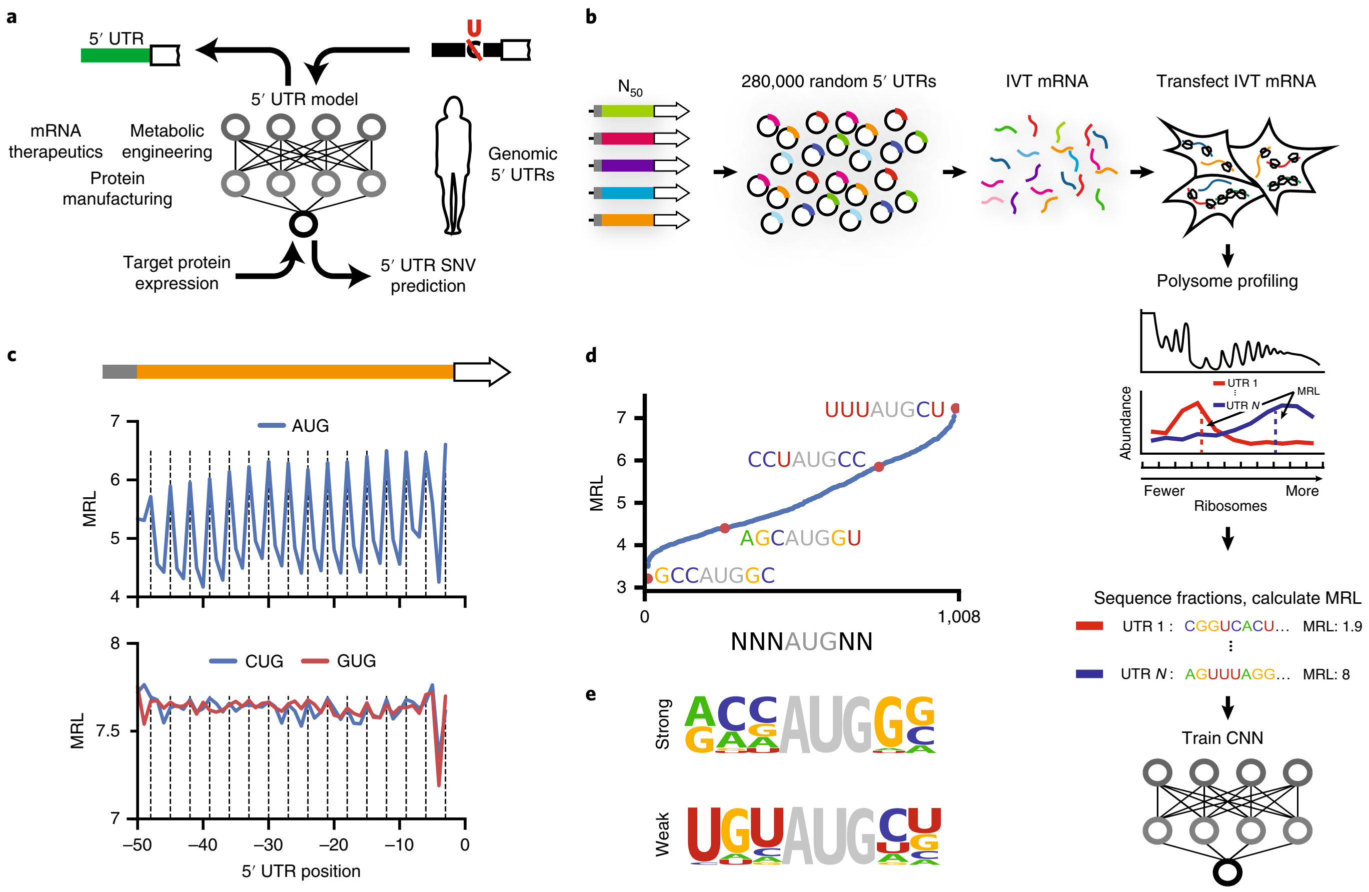

Fig. 1 | A library of 280,000 random 50-nucleotide oligomers as 5′ UTRs for eGFP. a, A 5′ UTR model capable of predicting translation from sequence information was used to evaluate the effect of 5′ UTR SNVs, and to engineer new sequences for optimal protein expression. **b**, A library of 280,000 members was built by inserting a T7 promoter followed by 25 nucleotides of defined 5′ UTR sequence, a random 50-nucleotide sequence and the eGFP CDS into a plasmid backbone. IVT library mRNA was produced by in vitro transcription from a linearized DNA template obtained by PCR from the plasmid library. Cells transfected with IVT library mRNA were grown for 12 h before polysome profiling. Read counts per fraction were used to calculate MRL for each UTR, and the resulting data were used to train a CNN. **c**, Out-of-frame uAUGs reduce ribosome loading (vertical lines indicate positions that are in frame with the eGFP CDS). A similar but much weaker periodicity is observed for CUGs and GUGs. **d**, The repressive strength of all out-of-frame variations of NNNAUGNN. **e**, Nucleotide frequencies calculated for the 20 most repressive (strong) and least repressive (weak) TIS sequences.

impact ribosome loading, especially when surrounded by strong sequence context as detailed in Supplementary Figs. 3 and 4. The region surrounding the start codon, known as the translation initiation site (TIS) or the Kozak sequence, is a primary determinant of whether a ribosome will begin translation. We scored the repressive strength of all out-of-frame TISs by finding the mean MRL of sequences with all permutations of NNNAUGNN (except where NNN was AUG) (Fig. 1d). Using the 20 most repressive and 20 least repressive sequences, we calculated nucleotide frequencies for the strongest and weakest TISs. This analysis recapitulated the importance of a purine (A or G) at position −3 relative to the AUG and a G at position +4 (refs. [10,23,24]) (Fig. 1e). Ultimately, these data suggest that each TIS sequence can uniquely tune translation initiation to a fine degree. Translation initiation and elongation are also affected by RNA secondary structure that forms within 5′ UTRs and CDSs, with the strongest structures (i.e., lowest free energy) showing the most negative effect on translation[17,25]. By calculating UTR minimum free energy (MFE)[26] and comparing it to UTR MRL, we captured and quantified this repressive effect of secondary structure on ribosome load[17,25] (Supplementary Fig. 2c).

Modeling 5′ UTRs and ribosome loading. We set out to develop a model, Optimus 5-Prime (Supplementary Code), that could quantitatively capture the relationship between 5′ UTR sequences and their associated MRLs. To this end, we trained a CNN with 260,000 sequences from the 280,000-member eGFP library. The remaining 20,000 sequences were withheld for testing. After an exhaustive grid search to find optimal hyperparameters (Fig. 2a; Methods), Optimus 5-Prime could explain 93% of MRL variation in the test set (Fig. 2b). A model trained on data from another polysome profiling experiment performed similarly (Supplementary Fig. 5a). By comparison, position-specific *k*-mer (*k* between 1 and 6) linear models could at best explain 66% of the variation in the test set (Supplementary Figs. 6 and 7, and Supplementary Table 1).

Up to this point, we used MRL as a simple measure for translation, but the raw data also captured how often a given sequence occurred in each polysome fraction. We thus set out to build a model capable of predicting the full polysome distribution for a given sequence. Using a similar network architecture but with 14 linear outputs representing the polysome fractions (Supplementary Fig. 8), the model captured the relationship between 5′ UTR sequence and the distribution of ribosome occupancy on test data remarkably well (Fig. 2c), explaining on average 83% of variation across all fractions (Fig. 2d). To test whether MRL predictions corresponded to actual protein expression, we selected and synthesized mRNAs containing ten different UTRs from the library with a wide range of observed

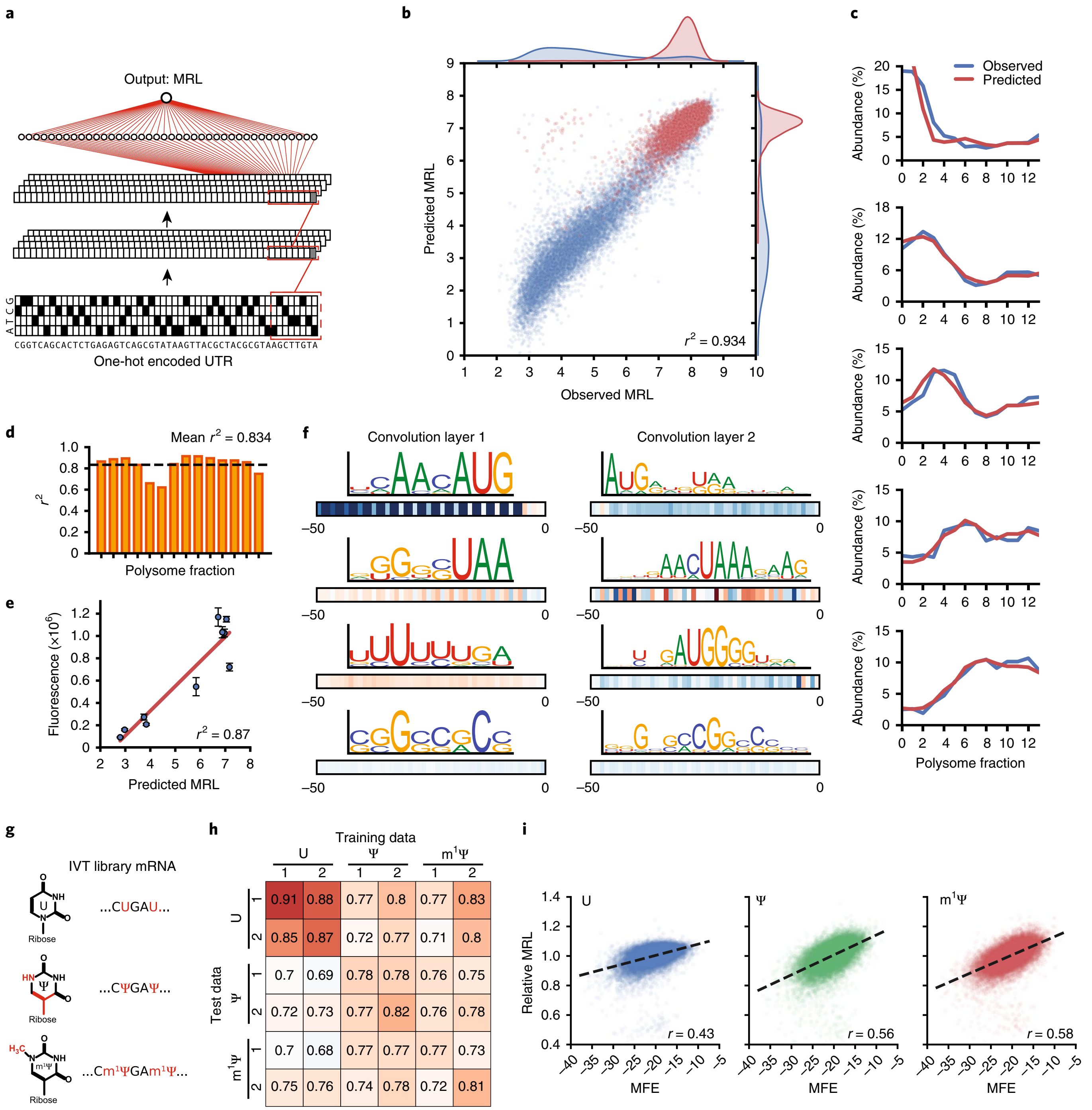

Fig. 2 | Modeling 5′ UTR sequences and ribosome loading. a, Schematic of Optimus 5-Prime: a one-hot encoded 5′ UTR sequence is fed into a CNN composed of three convolution layers and a fully connected layer to produce a linear output predicting MRL. **b**, Optimus 5-Prime trained on 260,000 UTRs and tested on 20,000 held-out sequences could explain 93% of the variability in observed MRLs. Blue dots represent sequences with a uAUG, and red dots represent sequences without a uAUG (*n* = 20,000). **c**, A similar model was trained to predict the distribution of polysome profiles for an individual 5′ UTR. The observed (blue) and predicted (red) polysome distribution of 5 randomly selected example UTRs from the 20,000-member test set spanning MRLs from four to eight (top to bottom) are shown. **d**, The performance of the polysome profile model per fraction ranged from $r^2 = 0.621$ to $r^2 = 0.915$ with an average of $r^2 = 0.834$ across all fractions (*n* = 20,000). **e**, eGFP expression for ten UTRs selected from the library was evaluated via eGFP fluorescence using IncuCyte live-cell imaging (*n* = 3, mean ± s.e.m.). Predicted MRL and fluorescence are highly correlated ($r^2 = 0.87$, *n* = 10). For details, see Supplementary Table 2. **f**, Visualization of four out of 120 filters from the first convolution layer (left) and four out of 120 filters from the second convolution layer (right). Boxes below show the correlation (Pearson's *r*, color scales in Supplementary Fig. 10) between filter activation and MRL at each UTR position. Filters learned important regulatory motifs such as start and stop codons, uORFs, and (G+C)-rich regions that are likely involved in the formation of secondary structure. **g**, IVT mRNA from the eGFP library was generated with Ψ or m^{1}Ψ in place of uridine and evaluated by polysome profiling and modeling. **h**, Model performance (r^2) with training and testing on different datasets. The unmodified RNA (U) models perform best with uridine data, whereas the Ψ and m^{1}Ψ models perform equally well with Ψ and m^{1}Ψ test data (*n* = 20,000). **i**, Ribosome loading as a function of MFE. The correlation between MFE and MRL is lower for unmodified RNA than for RNA containing Ψ and m^1 Ψ (Pearson's *r* of 0.43, 0.56 and 0.58, respectively; *n* = 19,976).

MRLs. We then transfected HEK293T cells with these mRNAs and measured eGFP fluorescence using IncuCyte live-cell imaging. Fluorescence and predicted MRL were highly correlated ($r^2=0.87$), and the most poorly translated sequence showed 15-fold-less fluorescence than the most strongly translated sequence (Fig. 2e). We also tested Optimus 5-Prime on 77 5′ UTRs that were previously designed by Ferreira et al.[27] and characterized using a fluorescent reporter system in six different cell lines. UTRs were designed to result in a range of expression levels by inserting one or multiple uORFs. The MRL predictions from our model correlated well with the independently reported fluorescence levels (r^2 between 0.73 and 0.85; Supplementary Fig. 9).

To determine whether Optimus 5-Prime would generalize to other coding sequences, we built a separate degenerate 5′ UTR mRNA library with an mCherry CDS replacing the CDS for eGFP. Following the polysome profiling and modeling procedure described above, we found that the model, although only trained on the eGFP library, still performed well, explaining 77% and 78% of the variation in MRL from two independent polysome profiling experiments performed with this new reporter library (Supplementary Fig. 5). The decrease in accuracy is explained in part by differences between the protocols for eGFP and mCherry polysome profiling (Methods).

Finally, to aid interpretation of the model we applied visualization techniques developed in computer vision and recently popularized in computational biology[4,8,28]. Visualization of the filters in the first and second convolution layer revealed recognizable motifs including strong TIS sequences (for example, ACCAUG), stop codons (UAA, UGA and UAG), uORFs, non-canonical start codons (CUG and GUG) and sequences composed of repeated CG or AU elements that are likely involved in the formation of secondary structure (Fig. 2f and Supplementary Fig. 10). Of note, several filters did not fall into either of these categories and also did not match previously described position–weight matrices (PWMs) for RNA-binding proteins (Tomtom[29] and the *Homo sapiens* RBP database[30]), suggesting the possibility for previously undescribed regulatory interactions.

Evaluation of mRNA containing pseudouridine and 1-methyl pseudouridine. The two uridine analogs pseudouridine (Ψ) and 1-methylpseudouridine ($m^1Ψ$) are widely used for mRNA therapeutics because they can increase mRNA stability and help modulate the host immune response[31,32]. We applied our method to transcripts containing either Ψ or $m^1Ψ$ instead of uridine (Fig. 2g) and found that the model trained on the unmodified uridine library could explain between 69% and 73% of the measured variability in the Ψ and between 68% and 76% of the measure variability in the $m^1Ψ$ polysome profiling data (Fig. 2h). Prediction accuracy could be further improved by training the models directly on data from the modified RNAs (the same held-out library sequences were used in all test sets to ensure consistency). This is likely due to the model learning the impact of Ψ and $m^1Ψ$ on the formation of secondary structure[33]. In line with this, MRL was more positively correlated with the predicted MFE of a UTR for Ψ ($r=0.56$) and $m^1Ψ$ ($r=0.58$) than it was for uridine ($r=0.43$) (Fig. 2i).

5′ UTR design for targeted ribosome loading. As a further test of our model's capabilities, we asked whether it could be used to engineer completely new functional 5′ UTRs. A tool capable of designing 5′ UTRs to obtain specific levels of protein expression would be a valuable asset for mRNA therapeutics and metabolic engineering. While there has been some success in this effort in prokaryotes, yeast and even mammalian cells[27,34–36], a fully rational approach to designing functional 5′ UTRs has not yet been implemented. We developed a genetic algorithm that iteratively edits an initial random 50-nucleotide sequence (not contained in the library of 280,000 sequences) until it is predicted by the model to load a target number of ribosomes and thus achieve an intended level of translation activity (Fig. 3a). The model used for this process was developed before Optimus 5-Prime in Fig. 2 and differs slightly in terms of network architecture (Methods) and performance ($r^2=0.92$) (Supplementary Fig. 11; Methods). We designed two sets of UTRs for testing. The sequences in the first set were designed to target specific MRLs from three to nine and a no-limit maximum MRL (Fig. 3b). The second set was designed to follow the stepwise evolution of a UTR. For this second set, we first used the algorithm to select for sequences with low ribosome loading and then, after 800 iterations, to select for high ribosome loading. Each unique sequence generated by the algorithm as the UTR evolved was synthesized and tested (Fig. 3c and Supplementary Fig. 12a–d). We did this for 20 sequences in which uAUGs were allowed and another 20 in which uAUGs were not allowed. Sequences containing uAUGs and those without uAUGs could both span the full MRL range.

Of the 12,000 total UTRs evolved for targeted expression in the first set, the median MRL for MRL targets of three through eight followed the expected trend from low to high, with low variability within each group. For the UTRs in the second set that were evolved stepwise, predicted MRLs closely matched the trend of the observed MRLs along the trajectory. While we created sequences with high ribosome loading (Supplementary Fig. 12e), in both sets, the prediction from the model and the observed MRL eventually diverged as the model produced UTRs with very high predicted MRLs. We suspected that the divergence between predictions and measurements at very high MRL values might reflect the unusual sequence composition of the maximally evolved UTRs, which often contained multiple long poly(U) sequences that were rarely seen in the random library. We corrected the model by training it (Fig. 3d) for four additional iterations with 6,082 UTRs from the target MRL sublibrary, which had a much higher frequency of homopolymers, and 2,695 previously unseen random UTRs. Re-evaluation of held-out sequences from the 'target MRL' library showed a dramatic improvement as compared to the original model (r^2 improved from 0.386 to 0.772) (Fig. 3e and Supplementary Fig. 13a), as did the predicted loading of the sequences that evolved stepwise (Fig. 3c and Supplementary Fig. 12a–d). Using this expanded dataset, we retrained the Optimus 5-Prime model from Fig. 2, which showed increased accuracy with all sublibraries and unchanged performance with random library sequences (Supplementary Fig. 13b). This improvement led us to use the retrained version of Optimus 5-Prime from this point on.

Predicting the effect of human 5′ UTR variants on ribosome loading. We next set out to establish whether a model trained only on synthetic sequences can predict how human 5′ UTR sequences control translation. Assessing model performance on endogenous transcripts is challenging owing to the confounding contributions of 3′ UTRs and CDSs. As an alternative approach, we synthesized and tested via polysome profiling a 5′ UTR library consisting of the first 50 nucleotides preceding the start codon of 35,212 common human transcripts as well as 5′ UTR fragments carrying 3,577 variant sequences from the ClinVar database[37] that occur within these regions. The same eGFP context as the randomized library was used in this alternative approach. The top 25,000 sequences by read coverage, which includes 22,747 common and 2,253 variant sequences, were used for downstream analysis. Using the retrained model, we were able to explain 82% of the observed variation in MRL for the common and SNV-containing 5′ UTR sequences (Fig. 4a). Despite being trained on random sequences, the model was able to learn the *cis*-regulatory rules of human 5′ UTR sequences that lie directly upstream of a CDS.

Genetic variants play a major role in phenotypic differences between individuals[38], and how these sequences affect translation is only beginning to be understood[39,40]. However, existing

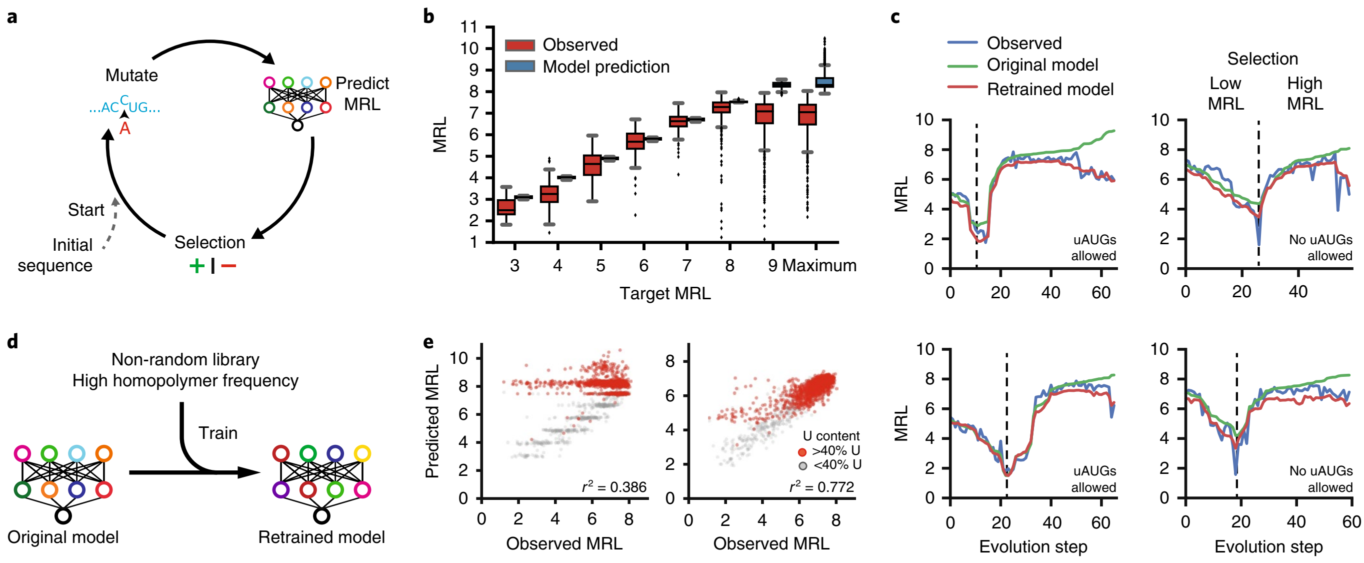

Fig. 3 | Design of new 5′ UTRs. a, Diagram of a genetic algorithm that was used in conjunction with Optimus 5-Prime to evolve sequences to target specific levels of ribosome loading. **b**, Comparison of the predicted and observed MRLs for evolved 5′ UTRs for targeted ribosome loading. All 16 box plots are defined in terms of the sample size, minima, median, maxima and percentiles (Supplementary Table 3). **c**, Stepwise evolution analysis. Randomly initialized UTRs were first evolved for low ribosome loading and then for high ribosome loading (selection change at dashed line). Four of 80 (Supplementary Fig. 11a–d) examples are shown. Examples on the left were permitted to have uAUGs, while those on the right were not. Each unique sequence that was generated during the evolution process was synthesized and tested by polysome profiling. The original Optimus 5-Prime prediction (green) and the observed MRL (blue) eventually diverged, but the predictions from the retrained Optimus 5-Prime (red) more accurately reflected the data. **d**, The original Optimus 5-Prime was retrained using sequences from the designed library with a high frequency of poly(U), poly(C), poly(A) and poly(G) stretches, which occurred rarely in the random library. **e**, The accuracy of the retrained Optimus 5-Prime (right) was greater than that of the original model (left) when predicting MRL for the sequences with a high frequency of poly(U) stretches (red) generated by the genetic algorithm (r^2 of 0.386 (original) and 0.772 (retrained), $n = 2{,}146$).

approaches to this problem, such as quantitative trait locus analysis and genome-wide association studies are limited to common variants and cannot scale to the enormous number of rare 5′ UTR variants occurring in the human population. In contrast, a model-based approach can in principle be used to score the impact of any 5′ UTR variant on translation. With this in mind, we investigated the ability of the Optimus 5-Prime model to predict the effect of disease-relevant variants by testing its performance when predicting the difference in MRL for pairs of wild-type ('common') and SNV-containing 5′ UTR sequences (measured as the $\log_2$-transformed difference in MRL). The majority of SNVs had little to no effect, but 45 had $\log_2$-transformed differences greater than 0.5 or less than −0.5 (Supplementary Table 4; ClinVar SUB4797518). Overall, Optimus 5-Prime could explain 56% of the observed difference in MRL (Fig. 4b) and accurately predicted the direction of effect for 64% of the variants. The relatively lower predictive accuracy as compared to direct prediction of variant effects is a consequence of the increased noise that results from comparing two measurements. Moreover, a majority of variants did not affect translation, resulting in a large cluster of variants for which the difference in MRL change was close to zero where measurements were dominated by noise. Importantly, the model could explain 77% of the variance for variants with measured $\log_2$-transformed differences of greater than 0.5 or less than −0.5 in comparison to the common sequence (Supplementary Fig. 14a). As an example, one of the ClinVar variants with sizeable differences in MRL, rs867711777, is found in the 5′ UTR of the *CPOX* gene and shows a $\log_2$ difference of −0.89. Depletion of *CPOX* reduces heme biosynthesis and is the cause of hereditary coproporphyria[41]. The large MRL difference suggests that this SNV, labeled as uncertain in the ClinVar database, could be pathogenic. The rs376208311 variant lies in the 5′ UTR of the ribosomal subunit gene *RPL5* and showed a $\log_2$-transformed difference in MRL of −0.87. This variant is associated with Diamond–Blackfan anemia, which can be caused by disruption or downregulation of *RPL5* (ref. [42]). Another SNV in the 5′ UTR of *TMEM127*, rs121908813, is implicated in familial pheochromocytoma, a condition characterized by tumors found in the neuroendocrine system that secrete high levels of catecholamines[43]. In our assay, the variant 5′ UTR showed a $\log_2$-transformed difference in MRL of −1.5 as compared to the wild-type 5′ UTR sequence. *TMEM127* acts as a tumor suppressor, and decreased expression of it could explain the observed pathogenicity of this variant. For the three examples, the model predicted that these specific SNVs, all of which introduce an upstream start codon, would most dramatically affect ribosome loading (Fig. 4c). We also identified 2,308 additional SNVs resulting from errors in oligonucleotide synthesis, and found that 103 of them showed $\log_2$-transformed MRL changes of greater than 0.5 or less than −0.5 (Supplementary Fig. 14b).

Modeling human 5′ UTRs of varying length. Human 5′ UTR sequences vary in length from tens to thousands of nucleotides with a median length of 218 nucleotides[44,45]. Because only 13% of human 5′ UTRs are shorter than 50 nucleotides and can be covered by Optimus 5-Prime, we next asked whether the approach introduced here could be extended to longer 5′ UTRs. To this end, we first created a 5′ UTR library where the length of the random sequence upstream of the start codon ranged from 25 to 100 nucleotides, which would increase the coverage of human 5′ UTRs to 29%. After polysome profiling and RNA sequencing, we retained 83,919 distinct 5′ UTRs spanning the entire length distribution from 25 to 100 nucleotides. As observed with the 50-nucleotide library, sequences containing uAUGs had a lower median MRL than sequences of similar length not containing them. Moreover, for sequences not containing uAUGs, the MRL slightly increased with length, likely because longer transcripts can accommodate more ribosomes (Supplementary Fig. 15a). We then retrained our model to capture and predict the

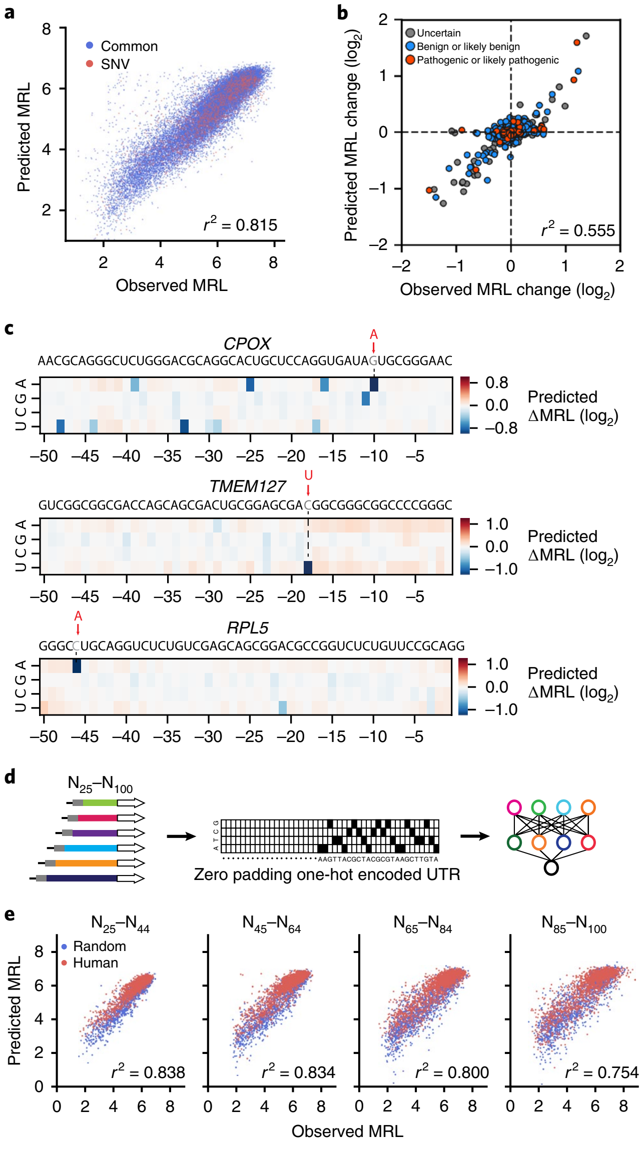

Fig. 4 | Model performance with human 5′ UTRs and generalization to 5′ UTRs of varying length. a, The retrained Optimus 5-Prime model could explain 82% of the observed variation in MRL for 22,747 common and 2,253 SNV-containing human 5' UTR sequences ($n = 25{,}000$). **b**, The $\log_2$-transformed difference in MRL between an SNV-containing UTR and the correspnding common sequence was compared to the predicted difference between the two ($r^2 = 0.555$, $n = 1{,}597$). SNV classification labels are from the ClinVar database. **c**, In silico saturation mutagenesis and model prediction of the difference in MRL for all 5′ UTR variants of *CPOX*, *TMEM127* and *RPL5*. The three annotated ClinVar variants rs867711777 (*CPOX*, G>A), rs121908813 (*TMEM127*, C>U) and rs376208311 (*RPL5*, C>A) are predicted to have the most dramatic effect on ribosome loading. **d**, A library of 76,319 random 5′ UTRs with varying lengths from 25 to 100 nucleotides (N25–N100) was used to train the generalized Optimus 5-Prime model. Sequences were one-hot encoded and zero padded to 100 nucleotides if shorter than 100 nucleotides. **e**, Random ($n = 7{,}600$; blue dots) and human ($n = 7{,}600$; red dots) sequences were tested using the generalized Optimus 5-Prime model. One hundred sequences of each length (25–100 nucleotides) are represented. Model accuracy (r^2 between 0.754 and 0.838) is shown for the prediction of MRLs for 5′ UTRs of different lengths (from left to right, $n = 4{,}000$, $n = 4{,}000$, $n = 4{,}000$ and $n = 3{,}200$).

impact of both sequence and length on MRL. To accommodate sequences up to 100 nucleotides in length, we increased the width of the input layer but otherwise retained the same network architecture as before (Fig. 4d). To ensure that 5′ UTRs of all lengths would be represented equally, we took the 100 5′ UTRs with the deepest read coverage at each length (~10% of the library) as the test set, rather than using the top 10% of the entire population. The remaining 90% of UTRs were used for training. In fact, we found that the average number of sequencing reads per library member rapidly decreased with increasing UTR length, likely because of the decreasing yield of full-length sequences for longer 5′ UTRs (Supplementary Fig. 15b). We also created a second test set consisting of 7,600 human 5′ UTRs, corresponding to 100 UTRs for each length from 25 to 100 nucleotides. The generalized Optimus 5-Prime model performed well on both the human ($r^2 = 0.78$) and random ($r^2 = 0.84$) sequences (Supplementary Fig. 15c,d) and on 5′ UTRs of any length (Fig. 4e; r^2 between 0.75 and 0.84). The slight decrease in performance observed for longer 5′ UTRs is due to lower read coverage for longer sequences and the concomitant decrease in the quality of the test set. These results suggest that the approach we developed here is not limited to fixed-length UTRs and could be extended even beyond a 100-nucleotide window by synthesizing correspondingly longer 5′ UTRs for model training.

Discussion

The method developed here, which combines polysome profiling of a randomized 5′ UTR library with deep learning, has provided a wealth of information detailing the relationship between the 5′ UTR sequence preceding a CDS and regulation of translation. The data and model enabled quantitative assessment of secondary structure, uAUGs and uORFs, Kozak sequences and other *cis*-regulatory sequence elements in the context of unmodified mRNA and Ψ- and m^{1}Ψ-modified mRNA. Optimus 5-Prime, the CNN trained on the data, has excellent performance, explaining up to 93% of MRL variation in the test set and up to 82% of variation for truncated human UTRs. In future work, this approach could be further generalized to include the impact of the mRNA 5′ terminus including the 5′ cap structure, and even 3′ UTR sequence on ribosome loading. Our model also proved capable of predicting the effect of disease-relevant 5′ UTR variants on translation, even suggesting mechanisms of action. Of note, predictions are not limited to common variants or even to those that have been previously described; instead, the model can be used to screen every possible SNV, insertion or deletion in the 100 bases upstream of a start codon, of which there are millions in the human genome, and select for further study those that have the strongest impact on ribosome loading and thus have the highest likelihood of being pathogenic. Finally, using Optimus 5-Prime and a genetic algorithm, we were able to engineer new 5′ UTR sequences for targeted ribosome loading, enabling applications in synthetic biology and precision medicine that are even more forward looking.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41587-019-0164-5.

Received: 20 May 2018; Accepted: 21 May 2019;
Published online: 1 July 2019

References

1. Araujo, P. R. et al. Before it gets started: regulating translation at the 5′ UTR. *Comp. Funct. Genom.* **2012**, 475731 (2012).
2. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* **11**, 113–127 (2010).
3. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878 (2016).

4. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
5. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
6. Kleftogiannis, D., Kalnis, P. & Bajic, V. B. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.* **43**, e6 (2015).
7. Liu, F., Li, H., Ren, C., Bo, X. & Shu, W. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *Sci. Rep.* **6**, 28517 (2016).
8. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
9. Zhao, W. et al. Massively parallel functional annotation of 3′ untranslated regions. *Nat. Biotechnol.* **32**, 387–391 (2014).
10. Noderer, W. L. et al. Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.* **10**, 748 (2014).
11. Kosuri, S. et al. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **110**, 14024–14029 (2013).
12. Cuperus, J. T. et al. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res.* **27**, 2015–2024 (2017).
13. Zuccotti, P. & Modelska, A. in *Post-Transcriptional Gene Regulation* (ed. Dassi, E.) 59–69 (Humana Press, 2016).
14. Floor, S. N. & Doudna, J. A. Tunable protein synthesis by transcript isoforms in human cells. *elife* **5**, e10921 (2016).
15. Wang, X., Hou, J., Quedenau, C. & Chen, W. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. *Mol. Syst. Biol.* **12**, 875 (2016).
16. Whiffin, N. et al. Characterising the loss-of-function impact of 5′ untranslated region variants in whole genome sequence data from 15,708 individuals. Preprint at https://www.biorxiv.org/content/10.1101/543504v1 (2019).
17. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416 (2016).
18. Morris, D. R. & Geballe, A. P. Upstream open reading frames as regulators of mRNA translation. *Mol. Cell. Biol.* **20**, 8635–8642 (2000).
19. Johnstone, T. G., Bazzini, A. A. & Giraldez, A. J. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* **35**, 706–723 (2016).
20. Lee, S. et al. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl Acad. Sci. USA* **109**, E2424–E2432 (2012).
21. Reuter, K., Biehl, A., Koch, L. & Helms, V. PreTIS: a tool to predict non-canonical 5′ UTR translational initiation sites in human and mouse. *PLoS Comput. Biol.* **12**, e1005170 (2016).
22. Starck, S. R. et al. Translation from the 5′ untranslated region shapes the integrated stress response. *Science* **351**, aad3867 (2016).
23. Hinnebusch, A. G. The scanning mechanism of eukaryotic translation initiation. *Annu. Rev. Biochem.* **83**, 779–812 (2014).
24. Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**, 283–292 (1986).
25. Kozak, M. Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. *Proc. Natl Acad. Sci. USA* **83**, 2850–2854 (1986).
26. Zadeh, J. N. et al. NUPACK: analysis and design of nucleic acid systems. *J. Comput. Chem.* **32**, 170–173 (2011).
27. Ferreira, J. P., Overton, K. W. & Wang, C. L. Tuning gene expression with synthetic upstream open reading frames. *Proc. Natl Acad. Sci. USA* **110**, 11284–11289 (2013).
28. Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A deep neural network for predicting and engineering alternative polyadenylation. *Cell* in press https://doi.org/10.1016/j.cell.2019.04.046 (2019).
29. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
30. Ray, D. et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
31. Karikó, K. et al. Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Mol. Ther.* **16**, 1833–1840 (2008).
32. Anderson, B. R. et al. Incorporation of pseudouridine into mRNA enhances translation by diminishing PKR activation. *Nucleic Acids Res.* **38**, 5884–5892 (2010).
33. Kierzek, E. et al. The contribution of pseudouridine to stabilities and structure of RNAs. *Nucleic Acids Res.* **42**, 3492–3501 (2014).
34. Seo, S. W. et al. Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metab. Eng.* **15**, 67–74 (2013).
35. Jensen, M. K. & Keasling, J. D. Recent applications of synthetic biology tools for yeast metabolic engineering. *FEMS Yeast Res.* **15**, 1–10 (2015).
36. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27**, 946–950 (2009).
37. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
38. Hernandez, R. D. et al. Singleton variants dominate the genetic architecture of human gene expression. Preprint https://doi.org/10.2139/ssrn.3151998 (2018).
39. Battle, A. et al. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
40. Cenik, C. et al. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res.* **25**, 1610–1621 (2015).
41. Wang, B. & Bissell, D. M. *Hereditary Coproporphyria* (University of Washington, 2012). .
42. Boria, I. et al. The ribosomal basis of Diamond–Blackfan anemia: mutation and database update. *Hum. Mutat.* **31**, 1269–1279 (2010).
43. Qin, Y. et al. Germline mutations in *TMEM127* confer susceptibility to pheochromocytoma. *Nat. Genet.* **42**, 229–233 (2010).
44. Mignone, F. et al. Untranslated regions of mRNAs. *Genome Biol.* **3**, reviews0004.1 (2002).
45. Leppek, K., Das, R. & Barna, M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.* **19**, 158–174 (2018).

Acknowledgements

We would like to thank A. Rosenberg and J. Linder for helpful discussions on data analysis and modeling. We would also like to thank M. Moore, A. Hsieh and Y. Lim for constructive comments on the manuscript. We are grateful to C. Wang for providing fluorescence data[27]. This work was supported by a sponsored research agreement by Moderna and National Institutes of Health grant R01HG009892 to G.S.

Author contributions

P.J.S. and B.W. designed and performed experiments, performed data analysis and modeling, and wrote the manuscript. D.W.R. performed fluorescence validation experiments. V.P. and I.M. wrote the manuscript. D.R.M. helped design polysome profiling. G.S. designed experiments and wrote the manuscript.

Competing interests

P.J.S., B.W., G.S. and DRM declare no competing interests. D.R., V.P. and I.M. are employees and shareholders of Moderna.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0164-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to G.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Randomized 50-nucleotide oligomer 5′ UTR library. A vector (pET 28) encoding a T7 promoter followed by 25 nucleotides of a defined 5′ UTR (GGGACATCGTAGAGAGTCGTACTTA) and the eGFP CDS was linearized with AgeI to allow for insertion of the 5′ UTR library between the defined sequence and the CDS. The defined 25-nucleotide sequence allows for PCR amplification after reverse transcription. Two nucleotides were changed at positions +11 (C to A) and +14 (C to T) in the eGFP CDS to introduce stop codons (TAA) in frame −1 and −2 relative to the ATG. The oligonucleotide (Supplementary Table 5, primer 282) that was used for library insertion contains the defined 5′ UTR, followed by 50 nucleotides of randomized bases and 21 nucleotides that overlap the eGFP CDS (including the ATG start site) (IDT). A reverse primer (primer 283) complementary to the 21-nucleotide eGFP overlap was used to produce a double-stranded product via Klenow extension with Klenow polymerase I (NEB). The vector and insert were assembled by Gibson reaction (NEB) and the product was electroporated into 5-alpha electrocompetent *Escherichia coli* (NEB). A small portion of the electroporated bacteria was plated and resulted in ~750,000 colony-forming units, and the rest was grown in liquid culture overnight (all bacteria were grown under kanamycin selection). The isolated plasmid is the eGFP library.

To produce the mCherry library the same process as above was performed, with some modifications. The same defined 5′ UTR that lies upstream of the randomized 50-nucleotide UTR in the eGFP construct was used (primer 252). Klenow extension with primer 253 created the double-stranded insert that was assembled with the AgeI-linearized backbone by Gibson reaction. The mCherry CDS, however, did not have intentionally-placed stop codons.

eGFP library sequence. Bold indicates the defined 5′ end of the 5′ UTR. The 50-nucleotide oligomer random UTR (non-random UTR in the case of the designed library) immediately follows. The underlined sequence corresponds to a truncated BGH poly(A) signal. During in vitro transcription, a 70-nucleotide-long poly(A) tail is added at the 3′ end.

GGGACATCGTAGAGAGTCGTACTTA(N50)atgggcgaattaagtaagggcgagga
gctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccacaagttcagcgtgtcc
ggcgagggcgagggcgatgccacctacggcaagctgaccctgaagttcatctgcaccaccggcaagctgcccgtg
ccctggcccaccctcgtgaccaccctgacctacggcgtgcagtgcttcagccgctaccccgaccacatgaagcagcac
gacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaactacaaga
cccgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaagggcatcgacttcaaggaggac
ggcaacatcctggggcacaagctggagtacaactacaacagccacaacgtctatatcatggccgacaagcagaagaa
cggcatcaaggtgaacttcaagatccgccacaacatcgaggacggcagcgtgcagctcgccgaccactaccagcag
aacacccccatcggcgacggccccgtgctgctgcccgacaaccactacctgagcacccagtccaagctgagcaaag
accccaacgagaagcgcgatcacatggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacga
gctgtacaagttcgaataaagctagcgcctcgactgtgccttctagttgccagccatctgttgtttg

mCherry library sequence. The sequence has the same defined 5′ end and truncated BGH poly(A) signal sequences as the eGFP library.

GGGACATCGTAGAGAGTCGTACTTA(N50)atgcctcccgagaagaagatcaagagcg
tgagcaagggcgaggaggataacatggccatcatcaaggagttcatgcgcttcaaggtgcacatggagggctccgt
gaacggccacgagttcgagatcgagggcgagggcgagggccgcccctacgagggcacccagaccgccaagctgaa
ggtgaccaagggtggcccctgcccttcgcctgggacatcctgtcccctcagttcatgtacggctccaaggcctacgtg
aagcaccccgccgacatccccgactacttgaagctgtccttccccgagggcttcaagtgggagcgcgtgatgaacttc
gaggacggcggcgtggtgaccgtgacccaggactcctccctgcaggacggcgagttcatctacaaggtgaagctgc
gcggcaccaacttcccctccgacggccccgtaatgcagaagaagaccatgggctgggaggcctcctccgagcgg
atgtaccccgaggacggcgccctgaagggcgagatcaagcagaggctgaagctgaaggacggcggccactacg
acgctgaggtcaagaccacctacaaggccaagaagcccgtgcagctgcccggcgcctacaacgtcaa
catcaagttggacatcacctcccacaacgaggactacaccatcgtggaacagtacgaacgcgcc
gagggccgccactccaccggcggcatggacgagctgtacaagtcttaacgcctcgactgtgccttctag
ttgccagccatctgttgtttg

In vitro transcription. A template for in vitro transcription was produced via PCR of the library plasmid with primers 254 and 255 and KAPA Hi-Fi polymerase (Kapa Biosystems). The double-stranded DNA product has a T7 promoter at the 5′ end and a truncated BGH poly(A) signal sequence followed by a 70-nucleotide poly(A) sequence (introduced with primer 254) at the 3′ end. The IVT reaction used the HiScribe T7 high-yield RNA synthesis kit (NEB) and 3′-*O*-Me-m^{7}G(5′)ppp(5′)G RNA cap (NEB) was used as the cap structure analog. The DNA template was digested with DNase I (NEB) and the IVT mRNA was purified using RNA Clean & Concentrator (Zymo Research). This protocol was used to produce the unmodified eGFP IVT mRNA and mCherry IVT mRNA for transfection. For synthesis of individual mRNAs for assessment of expression, linear DNA templates were assembled containing a T7 polymerase promoter, 5′ UTR, coding sequence, 3′ UTR and template-encoded poly(A) tail. mRNA transcription and purification were carried out as described previously[46]. For mRNA libraries containing alternatives to uridine, UTP was replaced with pseudouridine-5′-triphosphate or N^1-methylpseudouridine-5′-triphosphate during transcription. The final mRNAs utilized Cap1 to increase mRNA translation efficiency. After purification, the mRNA was diluted in citrate buffer to the desired concentration.

Transfection of IVT mRNAs. HEK293T cells were plated on 10-cm cell culture dishes 24 h before transfection (between 1 and 2 million cells per plate). At 60% to 80% confluency, cells were transfected with 14.5 μg of library mRNA using Lipofectamine MessengerMAX (Thermo Fisher Scientific) following the manufacturer's protocol. Plates were washed with 10 ml 1× Dulbecco's PBS and 10 ml medium (DMEM with 10% FBS and 1% penicillin–streptomycin) after incubation for 1 h. Cells were lysed 12 h after transfection.

Cell lysis and RNA isolation. For cell lysis and RNA isolation the following solutions were prepared: salt solution (10×; 100 mM NaCl, 100 mM $MgCl_2$, 100 mM Tris-HCl pH 7.5 and RNase-free water[13]), wash buffer (100 μg ml^{-1} cycloheximide (NEB) in RNase-free Dulbecco's PBS (10 ml per plate)) and lysis buffer (1× salt solution, 1% of 20% Triton X-100, 1 mM dithiothreitol, 0.2 U μl^{-1} SUPERase-In (Thermo Fisher Scientific) and 100 μg ml^{-1} cycloheximide). Wash buffer and lysis buffer were chilled throughout the protocol. After 12 h of growth at 37 °C, cells were placed on ice and medium was aspirated. Translating ribosomes were halted by adding 5 ml of wash buffer and cells were then kept at 37 °C for 5 min followed by aspiration on ice. Cells were washed by adding 5 ml of wash buffer and aspirating thoroughly. Cells were then lysed with 300 μl of ice-cold lysis buffer and scraped from the plates, cell clumps were disrupted by pipetting approximately five times and cells were placed into a prechilled microcentrifuge tube. Cells in lysis solution were incubated for 10 min on ice and then triturated by passing through a 25-gauge needle ten times[14]. Debris was cleared by centrifugation at 16,000*g* for 5 min. Supernatant was supplemented with 1.5 μl of 1 U μl^{-1} DNase I (final concentration of 0.005 U μl^{-1}) and placed on ice for 30 min. Lysate was then stored at −80 °C or used directly for polysome profiling.

Polysome profiling. Sucrose gradient buffers contained either 20% or 55% (wt/vol) sucrose as well as 100 mM KCl, 20 mM HEPES pH 7.2 and 10 mM $MgCl_2$. Sucrose (20%, 5.4 ml) was gently layered over 5.4 ml of 55% sucrose in an ultracentrifuge tube. The tube was then sealed with parafilm, placed on its side and left overnight at 4 °C. Approximately 2 h before use, the gradient was returned to an upright position. Once prepared, cell lysate was layered over the gradient and centrifuged for 3 h at 151,000*g* using a Beckman SW-41 Ti rotor. Only for mCherry library, a slightly different protocol was applied. Specifically, sucrose gradient buffers contained either 7% or 47% (wt/vol) sucrose as well as 150 mM NaCl, 20 mM Tris-HCl pH 7.2, 5 mM MgCl2 and 1 mM dithiothreitol. Sucrose (7%, 5.4 ml) was gently layered over 5.4 ml of 47% sucrose in an ultracentrifuge tube. The tube was centrifuged for 1 h and 45 min at 39,000 r.p.m.

Polysome fraction processing and next-generation sequencing. Fractions of 500 μl corresponding to ribosome peaks including the 40S and 60S peaks were individually collected and processed. Five-hundred microliters of TRIzol (Thermo Fisher Scientific) was added to each fraction and the fractions were vortexed. After incubating at room temperature for 5 min, 100 μl of chloroform was added and the mixture was vortexed and then incubated for 5 min at room temperature. Fractions were spun at 13,000 r.p.m. for 10 min and the RNA from the supernatant was purified following the protocol for RNA Clean & Concentrator (Zymo Research). Elution was performed with 15 μl of RNase-free water. The purified RNA was reverse transcribed using SuperScript IV (Thermo Fisher Scientific) and gene-specific primers (primer 289 for eGFP libraries and primer 220 for mCherry libraries). Both reverse transcription primers have unique molecular indices (UMIs). The products were then amplified with overhangs for Illumina-based sequencing; the reverse transcription primers contain barcodes that indicate the polysome fraction from which the reverse transcription product was derived. A custom forward primer for read 1 anneals to the defined 5′ end of the 5′ UTR; the mCherry library and the eGFP library have the same 5′ end sequence. Products were sequenced with the Illumina NextSeq platform using NextSeq 500/550 v2 High Output 75 cycle kits.

Sequence processing. Raw sequence files, separated by their fraction-associated barcodes, were processed with Cutadapt[47], outputting the 50-nucleotide UTR and 9–15 nucleotides corrsponding to the N-terminal region of the CDS. UTRs were clustered and UMIs were counted using Bartender[48]. The eGFP library contained approximately 750,000 unique sequences and the mCherry library contained approximately 500,000 sequences. UTRs were removed if the CDS sequence did not match the intended sequence. Because many of the remaining sequences had very few reads, we took the top 280,000 sequences for the eGFP library and the top 200,000 sequences for the mCherry library. No sequences in the eGFP and mCherry libraries matched. To normalize differences in total read counts between fractions, relative reads were calculated within each fraction. Using these values, the relative distribution of reads for each UTR across the fractions was determined. MRL was calculated by multiplying each fraction's relative distribution of reads by the number of ribosomes associated with each fraction and these values were summed (Supplementary Note 1).

Translation validation. Ten 5′ UTR sequences with a wide range of MRLs were selected from the eGFP library and individually cloned into the same vector as the randomized library. IVT mRNA was synthesized and HEK293T cells were transfected with Lipofectamine 2000 and then monitored for eGFP fluorescence using an IncuCyte S3 live-cell analysis system. Expression was reported as the maximum eGFP fluorescence over a 20.5-h time window.

Convolution neural network for MRL prediction. All code was written in Python 2.7 and all neural network development was done using the Keras (https://keras.io) and TensorFlow backends[49]. For hyperparameter selection, the top 50,000 sequences, in terms of total read counts per UTR, were used. We performed a tenfold cross-validation grid search to exhaustively test hyperparameter combinations of convolution layers (2, 3), convolution filter lengths (8, 10, 12), number of convolution filters (40, 80, 120), number of nodes in the dense layer (40, 80, 120) and dropout probability between all layers (0, 0.2, 0.4). The best hyperparameter combinations were as follows.

First convolution layer: 120 filters (8 × 4), rectified linear unit (ReLU) activation and 0% dropout.

Second convolution layer: 120 filters (8 × 1), ReLU activation and 0% dropout.

Third convolution layer: 120 filters (8 × 1), ReLU activation and 0% dropout.

Dense layer: 40 nodes and 20% dropout.

Output layer: one linear output.

For the unmodified uridine eGFP model, 260,000 UTRs were used for training while 20,000 were used to evaluate the model. The model was trained over three epochs before overfitting occurred. Before training, we first sorted the UTRs on the basis of the number of total reads; those with the highest read counts were used for the test set. UTRs with more reads had higher resolution and so more accurately reflected their MRL as compared to UTRs with fewer reads that were noisier (Supplementary Fig. 15a–c). However, the model performed nearly as well after randomly splitting the training and test sets (Supplementary Fig. 15d).

Polysome profile model. After performing the same grid search as for the model trained to predict the MRL of a sequence, the best hyperparameters for the polysome profile CNN were as follows.

First convolution layer: 120 filters (8 × 4), ReLU activation and 0% dropout.

Second convolution layer: 120 filters (8 × 1), ReLU activation and 0% dropout.

Third convolution layer: 120 filters (8 × 1), ReLU activation and 0% dropout.

Dense layer: 80 nodes and 10% dropout.

Output layer: 14 linear outputs.

Splitting of the training and test sets was performed as in the MRL model.

Model used for evolving new UTRs. First convolution layer: 40 filters (8 × 4), ReLU activation and 0% dropout.

Second convolution layer: 40 filters (8 × 1), ReLU activation and 0% dropout.

Dense layer: 40 nodes, 20% dropout.

Output layer: one linear output.

***k*-mer linear model.** UTR sequences were represented as *k*-mers at each position of the UTR. These position-specific *k*-mers were used as features for training a model via linear regression. 1-mers to 6-mers were tested. Training involved regularization to limit overfitting and fivefold cross-validation. The same training and test sets used in building the CNN were used.

Filter visualization. For each filter, 2,000 8-mers from the eGFP 5′ UTR library that showed the highest activation were selected. From these, PWMs were calculated and used to visualize the sequence compositions that strongly activated each filter. Visualization of the second convolution layer involved a wider sequence window (15 bases) and PWMs were calculated with fewer *k*-mers (maximum 200).

Filter activation by UTR position. For a given filter, the filter's activation at each UTR position was assessed (only the top 100,000 UTRs in terms of total read counts were analyzed). These activations, position by position, were compared to UTR MRLs and a Pearson's *r* value was calculated. Negative values indicate a negative correlation between filter activation and MRL. Positive values indicate that filter activation and MRLs are positively correlated.

Relationship between UTR structure and MRL for uridine, Ψ and $m^1\Psi$. The MFE values for 20,000 UTRs from the eGFP library were calculated using Nupack[26] and compared to the MRLs from the uridine, Ψ and $m^1\Psi$ datasets.

Genetic algorithm for designing new 5′ UTR sequences. The 5′ UTR model used for evolving new sequences was trained with a different architecture than the main model that is described above and used throughout the manuscript. This was because sometime after training this first model, we determined that adding a third convolution layer and additional filters to each layer resulted in improved performance.

All sequence evolutions began with randomized sequences. Over a set number of iterations, a single randomly selected base or two bases with a 50% probability, were introduced and the fitness was evaluated using the model. If the new sequence scored higher or closer to the target MRL, then it was accepted; otherwise, the unchanged sequence was selected.

Evolution for targeted MRLs. We evolved three distinct sets of sequences for targeted expression: sequences without uAUGs and upstream stop codons, sequences where uAUGs and upstream stop codons were allowed and sequences where uAUGs were not allowed but upstream stop codons were. Each set evolved initially random sequences to hit MRLs of 3, 4, 5, 6, 7, 8 and 9 as well as a maximum value. From these sequences, 200 were selected for MRLs 3–7 and 1,000 sequences were selected for MRLs 8 and 9 and the maximum MRL value. In total, including the three sequence conditions, 12,000 sequences were synthesized and tested via polysome profiling. In Fig. 2b, the predicted values are scaled to the observed values within the data. This creates a discrepancy between the categorical names (*x*-axis markers) and the predicted MRLs, which are the values that should be used for the comparison between observed and predicted data.

Stepwise evolution of sequences. As a sequence evolved using our algorithm, a new sequence was created if its score is improved relative to its previous state. We recorded the sequences for these steps and tested their performance relative to the model prediction. Four distinct conditions were used and 20 UTRs for each were evolved, totaling 80 examples of UTR stepwise evolution. UTRs in the first two conditions were evolved to the highest MRL over 800 iterations; one condition allowed for uAUGs and the other did not. The third condition evolved sequences to the lowest MRL over 800 iterations and then changed to select for the highest MRL over 800 iterations while allowing uAUGs. The fourth condition was the same as the third except that uAUGs were not permitted. In total, beginning with 20 sequences for each condition, 7,526 UTRs were generated for analysis.

Selection of human UTR sequences. All human 5′ UTR transcripts from the human genome, as annotated by Ensembl, were retrieved using Biomart[50]. The first 50 nucleotides upstream of the annotated TISs were selected for synthesis, totaling 35,212 sequences.

Selection of 5′ UTR SNVs. All sequence variants in the ClinVar database[37] occurring in the selected UTR regions were synthesized, totaling 3,577 sequences.

Reanalysis of library UTRs. Five thousand eGFP library sequences were selected over a range of MRLs. These were synthesized and tested via polysome profiling with the rest of the designed sequences.

Synthesis of designed 50-nucleotide sequences. All designed and human 5′ UTR sequences were synthesized by CustomArray. Fragments were PCR amplified and cloned into the pET 28 eGFP vector as described above.

Design and synthesis of 5′ UTR sequences of varying length. Random and human 5′ UTR sequences of varying length (25 to 100 nucleotides) were synthesized by Agilent Technologies. All human 5′ UTR transcripts from the human genome, as annotated by Ensembl, were retrieved using Biomart[50]. The first 100 nucleotides upstream of the annotated TISs were selected for synthesis, totaling 17,586 sequences in the length range from 25 to 100 nucleotides. Fragments were PCR amplified and cloned into the pET 28 eGFP vector as described above.

Generalized CNN for 5′ UTRs up to 100 nucleotides in length. The generalized model used the same CNN architecture as the model trained on 50-nucleotide UTRs. Random sequences ($n = 76,319$) with lengths ranging from 25 to 100 nucleotides were used for training. The input space was expanded to 100 nucleotides using one-hot encoding (for sequences shorter than 100 nucleotides, zero padding was used). The top 100 sequences at each length, as measured by total read counts per UTR, were used to test the model's accuracy, resulting in a test set of 7,600 random 5′ UTRs. A test set consisting of 7,600 human 5′ UTRs was created in a similar way: of 15,555 human UTRs that were detected in the sequencing data, the top 100 UTRs by read count at each length were used as part of the test set.

Statistics. All r^2 values are the square of the correlation coefficient of linear least-squares regression. All box plots and violin plots have the median as the center and the first and third quartiles as the upper and lower edges of the box. The upper line is the third quartile plus 1.5 × interquartile range and the lower line is the first quartile minus 1.5 × interquartile range. Maxima and minima are identified. All violin plots include standard deviation. All *P* values were obtained from two-tailed *t* tests and were calculated using the Python package scipy.stats.ttest_ind. Sample sizes and *P* values are indicated in relevant figures. Additional information can be found in the Nature Research Reporting Summary.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The authors declare that all data supporting the findings of this study are available from Gene Expression Omnibus under accession GSE114002.

Code availability

The code for the Optimus 5-Prime model is provided in the Supplementary Code file. All code is also available at https://github.com/pjsample/human_5utr_modeling.

References

46. Richner, J. M. et al. Vaccine mediated protection against Zika virus-induced congenital disease. *Cell* **170**, 273–283 (2017).
47. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
48. Zhao, L., Liu, Z., Levy, S. F. & Wu, S. Bartender: a fast and accurate clustering algorithm to count barcode reads. *Bioinformatics* **34**, 739–747 (2017).
49. Abadi, M. et al. TensorFlow: Large-scale machine laerning on heterogeneous systems. Software available from tensorflow.org (2015).
50. Smedley, D. et al. BioMart—biological queries made easy. *BMC Genomics* **10**, 22 (2009).

nature research

Corresponding author(s): Georg Seelig

Last updated by author(s): May 16, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ □ A description of all covariates tested

☒ □ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☒ □ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☒ □ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

□ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Nextseq 550 system, BCL2FASTQ v2.19, NUPACK v3.0.6, Cutadapt v1.9, Bartender v1.0
Data analysis	https://github.com/pjsample/human_5utr_modeling , Python 2.7, Pandas 0.18, sklearn 0.19, SciPy v1.0.0, Numpy 1.14.0, Keras v2.1, TensorFlow v1.8, BioMart

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

GSE114002

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical test were used to calculate sample sizes. For library diversity, the goal was to get as many unique sequences as possible, which is limited by cloning efficiency. Ultimately, we had 280,000 and 200,000 unique sequences in the EGFP and mCherry libraries, respectively. These sample sizes proved to be adequate based on the model's ability to accurately predict held-out UTR sequences.
Data exclusions	For the eGFP library, the top 280,000 in terms of total sequencing reads were analyzed. 200,000 sequences based on total reads were kept for the mCherry library. All other sequences were not included for analysis or modeling. Sequences with few total reads introduce technical noise and are less likely to reflect true ribosome loading. Exclusion criteria were not established beforehand.
Replication	Two replicates were performed for the eGFP unmodified, eGFP pseudouridine, eGFP m1pseudouridine, and mCherry unmodified polysome profiling experiments. Three replicates were performed for the 10 5' UTRs testing fluorescence. All attempts at replication were successful.
Randomization	For model cross validation sklearn's train_test_split was used. UTRs designed via genetic algorithm were initialized with random sequences. Other types of randomization, such as treatment groups, were not applicable for this study.
Blinding	Blinding was not relevant for this study. All sequence variants were collected en masse without bias and randomization was used when modeling.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
☒	□	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293T, HEK293, PD31, K562, HCT116, CHO-K1, MPC11
Authentication	None of the cell lines have been authenticated.
Mycoplasma contamination	Not tested.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used.