A systematic delineation of 3′UTR regulatory elements and their contextual associations

Authors: Nan Liang [1,2†], Jing-Yi Li [1,2†], Yang Ding [1,2], Yu Wang [1,2], Jackson Champer [2,3*], Li-Chen Ren [1,2*], Ge Gao [1,2*]

Affiliations:

[1] State Key Laboratory of Gene Function and Modulation Research, School of Life Sciences, Biomedical Pioneering Innovative Center (BIOPIC) and Beijing Advanced Innovation Center for Genomics (ICG), Peking University, Beijing 100871, China.

[2] Center for Bioinformatics (CBI), Peking University, Beijing 100871, China.

[3] Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.

*Corresponding author. Email: jchamper@pku.edu.cn; renlc@mail.cbi.pku.edu.cn; gaog@mail.cbi.pku.edu.cn.

†These authors contributed equally to this work.

Abstract

The 3′ region of genes plays a crucial role in transcriptional and post-transcriptional regulation, yet the functional elements within it remain largely unknown. Here, we assessed the impact of over 1.6 million synthetic 3′UTR sequences on transcript abundance and subcellular localization in A549 cells, revealing a substantial number of relatively independent elements with distinct

regulatory associations. By training a context-aware deep learning model, we found that these regulatory elements exhibit modest context dependency but possess unique preferences individually. Through relatively weak interactions and appropriate context arrangements, multiple elements can synergistically generate significant regulatory effects. Finally, we demonstrated the possibility for applying our model to identify disease-related 3′UTR variants.

Introduction

In human cells, 3′ untranslated regions (3′UTRs) affect both the abundance and localization of transcripts, adding significantly to the plasticity of gene regulation [1–3]. Inspired by studies on promoters and enhancers [4–11], methods based on massively parallel reporter assays (MPRAs) have been well adopted to investigate 3′UTRs [12–23]. However, these studies usually utilize naturally occurring sequences with limited diversity [22,24–26], and the sophisticated interactions between expression abundance and localization are less well studied [1,3,21,27,28]. These limitations may impede a comprehensive analysis of regulatory elements within 3′UTRs, as well as the effective modeling and visualization of their contextual dependencies and interactions.

Random sequences synthesized *in vitro* have helped us reveal the binding motifs and contextual preferences of various of RNA-binding proteins (RBPs) and microRNAs [29,30]. Methods that combine random sequences, MPRAs, and machine learning have been remarkably successful in helping us understand the *in vivo* regulatory codes for RNA splicing, 5’UTRs, alternative polyadenylations and promoters [31–35]. We hypothesize that a more comprehensive and unbiased understanding of 3′UTR regulatory elements and their contextual associative properties could be achieved by using MPRAs to simultaneously measure the effects of a large number of random

3′UTR sequences on transcript abundance and localization, and then learning the relationship between the sequences and regulatory phenotypes through artificial neural networks.

Results

To efficiently assess the functions of vast numbers of nucleic acid sequences, we developed a technique called SEERS (Selective Enrichment of Episomes with Random Sequences). This method utilizes a self-replicating episomal vector [36] (**Supplemental Fig. S1A**) to selectively amplify target cells as they recover from transfection-associated toxicities, thereby reducing the unintended intracellular disturbances commonly seen in canonical MPRA protocols [37] (**Supplemental Fig. S1B**). By jointly analyzing the proportions of reads in subcellularly fractionated DNA and RNA samples, we could easily infer the regulatory effects of each 3′UTR sequence (**Fig. 1A**).

Using SEERS, we evaluated the regulatory effects of 1,664,033 randomly synthesized 45-nt 3′UTR insertions (N45s) on mCherry transcripts (**Fig. 1B**) in A549 cells. Our analyses suggest that these random N45s have sufficient diversity to encompass most of the sequence patterns found in human native 3′UTRs (**Supplemental Fig. S2**). Consistent with previous reports, we found that microRNA-binding sites were significantly overrepresented among 7-mers negatively correlated with transcript abundance (Mann-Whitney $p = 6.7\times10^{-8}$, see also **Supplemental Fig. S3A**). Additionally, the abundances of N45s carrying Pumilio elements, AU-rich elements (ARE), or m6A sites were significantly lower compared to the background (Mann-Whitney $p < 8.4\times10^{-7}$, see also **Supplemental Fig. S3B**). These findings indicate that SEERS effectively captures the known regulatory elements of human 3′UTRs [22,38].

We found that a simple CNN-based "motif counting" model (**Fig. 1C**) fits the data well, implying that the regulatory code of RNAs can largely be interpreted through a combination of multiple motifs [31,39,40]. Our tests showed that the performance of the "motif counting" model plateaued when the length of its convolution kernel reached 8 (**Fig. 1D**). This suggests that these independent motifs ("regulatory elements") are likely no larger than 8 nucleotides, which aligns with the maximum size of the binding motifs for microRNAs and RNA-binding proteins [29,30,41].

We further examined the overall differences in the abundance and localization of mCherry mRNAs compared to background when each 2-8mer was present in the N45 sequences. We found that over 20% (18,925/87,376) of all 2-8mers exhibited significant regulatory correlations (Mann-Whitney $p < 1\times10^{-6}$), although their effects were relatively weak (**Supplemental Fig. S4A**). If a 3′UTR regulatory element played a decisive role in transcript abundance or localization, we would expect to observe significant differences in the abundance or localization of that regulatory element sequence in the endogenous transcriptome compared to non-functional sequences. However, by comparing the regulatory correlations of all significant 8-mers in SEERS with their frequency differences in endogenous polyA+ transcripts from both nuclear and cytoplasmic fractions of A549 cells, we only detected moderate correlations (r = 0.23~0.32, **Supplemental Fig. S5**). This suggests that 3′UTR regulatory elements have detectable but non-dominant effects on the steady-state distribution of full-length transcripts, underscoring the complexity of transcriptional and post-transcriptional regulation [28,38].

We observed that about 8% (238/2,950) of the 8-mers positively correlated with transcript abundance did not exhibit significant strand specificity (Mann-Whitney $p > 0.1$). To investigate whether these 8-mers might represent transcription factor binding sites, we compared them with

the HOCOMOCO and JASPAR databases. The results showed that only 0.8% (2/238, JASPAR) to 4.2% (10/238, HOCOMOCO) of these 8-mers were similar to reported transcription factor binding sites (q-value < 0.05). This proportion is comparable to random 8-mers (3.5% to 4.0%), suggesting that the SEERS experimental results primarily reflect regulatory signals from 3′UTRs rather than enhancers.

By integrating the properties of each 2-8mer in terms of abundance and localization, we observed that they naturally formed three broad functional clusters: 1) low-expression nuclear-enriched elements, 2) high-expression nuclear-enriched elements, and 3) high-expression cytoplasmic-enriched elements (**Fig. 2**). This tripartite classification suggests the existence of three fundamental processing pathways for human 3′UTRs in Pol II transcript regulation [3,42].

We attempted to cluster these 2-8mers based on functional and sequence similarity but found they exhibited extensive mutual similarities that precluded formation of well-defined clusters (**Supplemental Fig. S6A**). This implies that while certain 2-8mers show significant regulatory associations, they likely do not represent independent binding motifs for single *trans*-factors, but may instead be associated with multiple factors. i.e., these elements likely represent composite regulatory signals involving multiple RBPs and/or other factors (e.g., miRNAs), resulting in hybrid patterns that reflect the averaged binding preferences of the participating regulators. We therefore treated each k-mer as an independent regulatory element rather than initially merging them into motif patterns, as such merging could introduce potential biases (**Supplemental Fig. S6B**). Nevertheless, in the aforementioned k-mer landscape (**Fig. 2**), the peripheral 8-mers demonstrated the strongest regulatory effects. We performed targeted nucleotide substitutions in several prominent 8-mers to systematically evaluate the contribution of individual nucleotides to

their regulatory activity, thereby identifying their functional motifs. Some motifs showed similarity to known RBP binding motifs (e.g., TGAAGAAA to SRSF1 and TATTTATT to ZFP36, etc.). However, given that many RBPs share highly similar motifs [29,41,43], we deliberately refrained from mapping regulatory elements in this study to any single *trans*-factor or speculating about potential roles of *trans*-factors.

To further investigate whether the role of these regulatory elements depends on their context, we trained an LSTM-based network. This context-aware model, "TALE" (Transcript Abundance and Localization Estimator), improved predictive performance by approximately 10% compared to the simple "motif counting" model, achieving accuracy comparable to SEERS experimental measurements ($R^2 = 0.83 \sim 0.88$, **Fig. 3A and Supplemental Fig. S7A**). To comprehensively evaluate TALE's predictive capability, we conducted an independent small-scale SEERS experiment. In this independent experiment, each N45 sequence covered a substantially larger number of cells (increased from an average of 20 to approximately 5,500 per sequence), resulting in more robust measurements. The results demonstrated that TALE accurately predicted the experimental outcomes ($R^2 = 0.88 \sim 0.91$, **Supplemental Fig. S7B-C**). Notably, the minimum edit distance between N45 sequences in this independent test set and those in TALE's training set was around 15 (**Supplemental Fig. S7D**). These findings collectively indicate that TALE exhibits strong generalization capability across the N45 sequence space.

Using TALE, we predicted context-dependent variations in regulatory relevance for each element across random genomic contexts (**Methods**) and observed that most elements maintained relatively stable functions (**Fig. 3B**). These elements exhibited significant suppression or enhancement only in a few extreme contexts, which partially explains why the

"context-aware" model showed only ~10% improvement in predictive performance compared to the "motif-counting" model. In contrast, as a negative control, the regulatory relevance of random 8-mers fluctuated substantially across contexts, indicating that they lack inherent regulatory functions and instead act transiently only when chance connections with random contexts fortuitously form functional elements. Similarly, the AU-rich element (TATTTATT) demonstrated strong context dependency, a finding consistent with previous reports [23].

To explore the characteristics of these extreme contexts, we conducted *in silico* directed evolution of the surrounding sequences of each regulatory element using TALE. This analysis revealed widely varying contextual and structural preferences among different elements (**Fig. 3C**). Notably, we found that the enhancing context predicted by TALE for U1 sites closely resembled a previously reported secondary structure [21], while the suppressing context exhibited a secondary structure "locked" by a stem-loop.

To investigate potential patterns of cooperation between regulatory elements, we compared the joint regulatory effect of any two identical or different 5-mers in the SEERS experimental data with the sum of their respective independent effects (we did not have enough N45 to perform this test on 8-mers). We found that these effects were highly compatible (**Fig. 4A**), suggesting that most 5-mers acted additively rather than cooperatively. In other words, 3′UTR regulatory elements tend to function more independently, with their combinations less likely to produce novel effects. Furthermore, we mapped the relationships between all 5-mers that exhibited significantly stronger joint effects than independent effects, revealing no specific clustering pattern among them (**Supplemental Fig. S8**). These results imply relatively weak but complex interactions between the 3′UTR regulatory elements.

Next, we investigated the organization of regulatory elements in 3′UTRs with different regulatory functions. To do so, we employed *in silico* directed evolution using TALE to generate two N45 sequences predicted to have the lowest RNA abundance. Both N45s were found to be enriched in multiple U1 sites (**Fig. 4B**), which aligns with our expectations. U1 sites in 3′UTRs have been shown to trigger the "telescripting" phenomenon [44], leading to 3′UTR splicing and the subsequent loss of the N45 sequence in RNA, thereby significantly reducing RNA/DNA levels (**Supplemental Fig. S9**).

In further *in silico* mutagenesis experiments, we found that TALE not only accurately predicted the regulatory relevance of individual nucleotides in U1 sites, but also inferred functional interdependencies or competition among them. These relationships helped to elucidate the interaction patterns between elements. For example, TALE predicted that mutations in the central GT of the U1 site significantly reduced the regulatory relevance of neighboring nucleotides, but not vice versa. This is consistent with the established consensus that the GT dinucleotide is the core motif of the U1 site [31]. Additionally, TALE predicted that disrupting one U1 site would significantly increase the regulatory relevance of other U1 sites (**Fig. 4B**), supporting their competitive nature as 5′ splice sites.

Similar analyses for N45s in other regulatory contexts revealed that mutations in most nucleotides primarily affect their immediate neighbors, suggesting the presence of short regulatory elements. However, we also observed regional interactions between nucleotides over distances greater than the size of a single element (8 nt) (**Fig. 4B and Supplemental Table S1**).

This suggests that, although the contextual associations between regulatory elements may be weak, multiple elements can still cooperate effectively when arranged optimally.

Causal inference of human 3′UTR variants remains challenging. Although our model was developed using a human lung cancer cell line, evidence suggests that general regulatory patterns in 3′UTRs may be relatively conserved across multiple cell types originating from different tissues [22]. Based on this, we employed TALE to predict the effects of 3′UTR variants annotated in the ClinVar database. The results revealed a significant difference in the distribution of predicted effect sizes between "Pathogenic" and "Benign" SNPs (Kolmogorov-Smirnov p = 0.000133). Using all 15,225 benign SNPs as the null distribution, 43 out of the 190 pathogenic SNPs exhibited significant (empirical $p < 0.05$) aberrant predicted effects (**Fig. 5A and Supplemental Table S2**). Further analysis of the SNPs with the strongest effects showed that these variants create or abolish 5′ splice sites (**Fig. 5B**), potentially activating or suppressing telescripting and ultimately influencing overall mRNA structure. This explains why these subtle nucleotide changes can lead to significant phenotypic consequences, consistent with previous reports [22,45]. Together, these findings demonstrate that the TALE model can help elucidate part of the causal relationship between 3′UTR variants and diseases.

Discussion

In this study, we developed SEERS, a highly efficient, high-throughput experimental technique. By combining it with massive synthetic random sequences and artificial neural networks, we were able to comprehensively and unbiasedly identify 3′UTR regulatory elements and their contextual dependencies. We found that a large and diverse set of short regulatory elements contribute to the regulatory functions of 3′UTRs, affecting both transcript abundance and

subcellular localization. Although these elements function primarily in an additive manner, they can still be organized to create strong synergistic effects through weak contextual connections. Our work provides a methodological prototype and offers new insights for a comprehensive understanding of the gene regulatory functions of 3′UTRs.

Given the potentially limited impact of long-range genomic context on post-transcriptional regulation, plasmid-based MPRA techniques offer unique advantages in studying 3′UTRs. By incorporating synthetic DNA fragments into the 3′UTRs of reporter genes, these assays allow for the unrestricted evaluation of the effects of any sequence or variant. This approach has successfully unveiled numerous regulatory elements within 3′UTRs in past studies [12–23]. However, accurate characterization of these new elements historically relied on extensive mutagenesis and screening of native 3′UTR sequences—a process that is costly, inefficient, and limited in sensitivity, making it challenging to pinpoint regulatory elements at single-nucleotide resolution. In contrast, vast arrays of random sequences can be synthesized *in vitro* at a negligible cost compared to the synthesis of specific DNA sequences. Initially applied to explore DNA-protein binding *in vitro* [46], this use of random sequences pioneered a high-throughput, systematic method for screening specific binding motifs from a theoretically infinite sequence space. Recently, the application of random sequences has expanded to *in vivo* exploration of regulatory elements [31–35], revealing the full landscape of regulatory elements against an almost unbiased backdrop, thereby assessing the functional relevance of each nucleotide within these elements and their contextual associations. This efficient approach has not only allowed for the rediscovery and confirmation of many previously identified elements, but has also significantly expanded the “encyclopedia” of regulatory elements. It is anticipated that in the near future, the

application of random sequences will play an even greater role in the exploration of gene regulation and function, enhancing our understanding in these areas [26].

The unprecedented high resolution of SEERS for regulatory elements revealed the significant regulatory correlations of 2-mers. Our "motif counting" model suggests that the linear summation of 2-mer effects explains one-third of the regulatory functions of our model 3′UTRs (**Fig. 1D**). Our experimental data also indicated that a single 2-mer is sufficient to produce weak but statistically significant regulatory effects on the mRNAs (**Supplemental Fig. S10A**), and these regulatory effects are enhanced as the number of the 2-mer increases (**Supplemental Fig. S10B**). Notably, these significance levels and trends disappeared entirely when the data were shuffled, suggesting that the regulatory associations of 2-mers observed in our data are not artifacts of random generation (**Supplemental Fig. S11**). Moreover, N45 populations completely lacking specific 2-mers exhibited more pronounced differences in regulatory phenotypes (**Supplemental Fig. S12**). These findings collectively suggest that 2-mers may indeed play a critical regulatory role. Although the correlation between nucleotide content and 3′UTR regulatory functions has been reported [18,22,47], our results showed that almost every different 2-mer had a unique, strand-specific effect. This implies that individual 2-mers can exert regulatory effects in a relatively independent manner, rather than relying on collective forces or stable structures formed through base pairing. This phenomenon aligns to some extent with prior studies describing RBPs with relatively simple core binding motifs [29], but the underlying mechanisms remain to be further explored.

In our results, the U1 binding site emerged as the most prominent regulatory element within 3′UTRs, consistent with previous reports [21,22,48]. However, we further discovered that this

regulatory effect is largely governed by the "telescripting" mechanism, which could result in artefacts in MPRA experiments. This not only re-emphasizes the importance of this mechanism in gene regulation but also suggests that extra caution may be necessary in studies involving U1 binding sites.

Methods

SEERS vector and library construction

The SEERS vector (LN110) was derived from the pCXLE-EGFP vector (Addgene #27082) by modification. Briefly, we extended the open reading frame of EGFP by adding a P2A-PuroR sequence and subsequently included an mCherry expression cassette downstream of the EGFP-P2A-PuroR expression cassette (See **Supplemental Table S3** for sequences of all plasmids and oligonucleotides).

The N45 insertion fragments were generated by a single cycle of PCR (98°C for 30 seconds, 66°C for 2 minutes, 72°C for 2 minutes) using NEBNext Ultra II Q5 Master Mix (NEB) with the oligos 110-HindIII-N45 and bGH_R2 (0.5 μM each). These fragments were inserted into LN110 via a HindIII site located next to the stop codon of mCherry, achieving a 100% positive insertion rate (10/10), as verified by Sanger sequencing. All plasmid assemblies were performed using NEBuilder HiFi DNA Assembly Master Mix (NEB). The assembled plasmids were then transformed into XL10-Gold cells, amplified on LB agar plates containing 100 μg/mL carbenicillin at 37°C, followed by purification using the EndoFree Maxi Plasmid Kit (Tiangen).

Cell culture and transfection

A549 cells (obtained from the Cell Resource Center, Peking Union Medical College, and verified by STR analysis) were maintained in DMEM/F-12 medium (HyClone) supplemented with 10% fetal bovine serum (Gibco) and 1X penicillin-streptomycin-amphotericin B solution (Solarbio). The SEERS libraries were co-transfected with pCXWB-EBNA1 (Addgene #37624) at a molar ratio of 4:1. All transfections were performed using either Lipofectamine 3000 (Invitrogen) or TransIT-X2 (Mirus). Twenty-four hours post-transfection, 1 μg/mL puromycin (Gibco) was added to the culture medium and maintained until all cells were 100% GFP-positive (GFP+) and reached approximately 70% confluency for collection.

For the SEERS experiment involving ~1.5 million N45 variants, we conducted 6 biological replicates, each collecting approximately 1 × 10^7 cells. For the SEERS experiment involving ~1,000 N45 variants, we collected approximately 5 × 10^6 cells and performed 4 technical replicates for downstream NGS.

Nucleic acid extraction

Cells were harvested using TrypLE Express (Gibco) and centrifuged at 400 × g for 1 minute. After removing the supernatant, approximately every 5 million cells were resuspended in 175 μL of ice-cold buffer RLN2 (DPBS, 0.5% Nonidet P-40, 10 μg/mL 2-mercaptoethanol) and incubated for 5 minutes. The cells were then gently pipetted and re-centrifuged at 400 × g for 2 minutes, and 150 μL of the cytoplasmic supernatant was collected. The precipitated nuclei were re-washed with 1 mL of RLN2. Each of the two fractions was subsequently dissolved in 600 μL

of RLT Plus buffer (Qiagen). Nucleic acids were extracted from these samples using the AllPrep DNA/RNA Mini Kit (Qiagen).

DNA was not detectable in the purified nucleic acids from all cytoplasmic fractions when assessed with the Qubit dsDNA HS Assay Kit (Invitrogen). The purities of the subcellular fractions were further evaluated by RNA sequencing, using mitochondrial genes and snoRNAs as reference markers.

Standard RNA sequencing and alignment

All library constructions and RNA sequencing were performed by Novogene. The libraries were strand-specific with fragment sizes ranging from 250 to 400 bp and were sequenced on a NovaSeq 6000 platform using 2 × 150 bp paired-end runs. The resulting reads were aligned to the hg38 reference genome and counted by gene using R (R Core Team. 2025) with Rsubread [49]. Only paired reads that aligned uniquely to a single locus were retained.

GO enrichment analysis

We compared RNA sequencing results from two normal A549 samples and two samples collected at the end of the SEERS experiments using DESeq2 [50]. All upregulated genes with a *p*-value $< 1\times10^{-4}$ were selected for GO enrichment analysis, focusing on Biological Process (BP) subontologies, using ClusterProfiler [51].

SEERS amplicon sequencing

Total RNA was reverse transcribed into cDNA using HiScript III RT SuperMix for qPCR (Vazyme) according to the manufacturer's instructions. Subsequently, the N45 region was amplified from 100 ng of DNA or 3 μL of cDNA in a 30 μL reaction using NEBNext Ultra II Q5 Master Mix (NEB) and primers P5-N8-RD1-110-F and P7-N8-RD2-110-R (1 μM each). The cycling conditions were as follows: 98°C for 30 seconds, followed by 25 to 30 cycles of 98°C for 10 seconds, 68°C for 30 seconds, 65°C for 45 seconds, and a final extension at 65°C for 5 minutes. Amplicons of the expected size (295 bp) were excised from an agarose gel and purified using the HiPure Gel Pure Micro Kit (Magen). The purified products were sent to Novogene for sequencing using 2 × 150 bp paired-end runs on a NovaSeq 6000, aiming for approximately 20 million read pairs per amplicon library. For each biological sample, at least three independent library constructions and sequencing runs were performed as technical replicates.

SEERS data processing

We first merged each pair of reads into a single read using NGmerge [52], then extracted the N45 sequences from each read based on the backbone sequences (15 nucleotides on each side) flanking the Hind III insertion site of the SEERS vector. Although the insertion sequence was designed to be “N45”, we observed that a small proportion (~5%) of sequences were not exactly 45 nucleotides in length. Since size selection was performed on the libraries before sequencing, we excluded all data corresponding to inserts longer than 46 or shorter than 44 nucleotides to ensure fairness. For simplicity, the remaining inserts are collectively referred to as “N45s”.

To minimize the impact of mutations and sequencing errors, we counted the number of reads and occurrences of each N45 in 18 experimental replicates (6 biological samples × 3 technical replicates). We found that these N45s formed two distinct populations based on the distribution of these counts. We removed data corresponding to the population of N45s with a low probability of occurrence in the nuclear fraction DNA samples.

From the remaining data, we calculated the average proportion of reads for each N45 in nuclear DNA (DNA), nuclear RNA (Nuc), and cytoplasmic RNA (Cyt) samples. These proportions were then converted into the normalized abundance of each N45 in the nucleus (Nuc/DNA) and cytoplasm (Cyt/DNA). Additionally, we calculated the "RNA/DNA" values, where RNA = Cyt*(16/17) + Nuc*(1/17), based on the measurement by qPCR indicating that the ratio of mCherry transcript molecules in the cytoplasm to the nucleus of A549 cells is approximately 16 to 1.

SEERS *k*-mer profiling

For each *k*-mer, we first identified all N45 sequences containing that *k*-mer and then compared these N45s with the entire population using the Mann–Whitney U test to assess the regulatory relevance of the *k*-mer. If a statistically significant difference (p-value $< 1\times10^{-6}$) existed between the N45 populations of two reverse complementary *k*-mers and the difference exceeded a fold change of 0.5-2, the two *k*-mers were labeled as "strand-specific".

For clustering, two *k*-mers were defined as "adjacent" if they differed by only one edit distance and had similar regulatory properties (both in abundance and localization) within a fold

change of 0.5-2. Using the resulting adjacency matrix, all 2-8mers were clustered with the igraph R package [53] using the "fast greedy" algorithm.

To highlight k-mers with significant regulatory relevance, we analyzed the regulatory contribution of each nucleotide within these k-mers to generate the logos in Fig. 2. The specific methodology is as follows: Taking TGAAGAAA as an example, assuming the regulatory effect (i.e., $\Delta\log_2$[RNA/DNA] or $\Delta\log_2$[Cyt/Nuc]) of this k-mer is R(TGAAGAAA), the regulatory contribution score of nucleotide T is calculated as: R(TGAAGAAA) - median[R(AGAAGAAA), R(CGAAGAAA), R(GGAAGAAA)]. This evaluation was systematically applied to every nucleotide position of each target k-mer. The regulatory contribution scores for individual nucleotides were subsequently used to construct the corresponding logos.

Validation of known 3′UTR regulatory elements

We downloaded the list of predicted binding sites (7-mers) for all miRNAs from TargetScan, selecting human miRNAs using Species ID = 9606, and removed all duplicated 7-mers. The regulatory properties of these 7-mers were queried using the method described above ("SEERS *k*-mer profiling") in the SEERS experimental results. Delta $\log_2$(Cyt/Nuc) was used as the x-axis and Delta $\log_2$(RNA/DNA) as the y-axis to analyze the 2D kernel densities of miRNA-related 7-mers and all 7-mers, respectively, using the kde2d function from the MASS R package. The difference between these two density matrices was used as the z-value for the final ggplot in R.

Similarly, we compared the distribution of N45s carrying known regulatory elements with the background distribution of all N45s.

Construction and training of neural networks

To avoid ambiguity regarding the lack of expression in certain N45s due to insufficient sequencing depth, these N45s were excluded from the data. For enhanced interpretability, the normalized abundance values (Nuc/DNA and Cyt/DNA) of each N45 were $\log_2$-transformed. We observed that N45s with higher DNA read counts exhibited more consistent regulatory effects across experimental replicates (**Supplemental Fig. S13**), likely because they covered a greater number of cells. To ensure high-quality test data, the top 10,000 N45s with the highest DNA read counts were held out as the test set. From the remaining N45s, 20% were randomly selected as the validation set, with the remainder used as the training set (https://github.com/gao-lab/SEERS/blob/main/data/TALE_training_data.tar.bz2). Additionally, 1,000 N45s from another small-scale SEERS experiment were set aside as an independent test set, with no overlap among these datasets.

We found that the following neural network architecture consistently achieved high predictive performance, measured by the mean absolute error (MAE) between predicted values and the validation set: an embedding layer with an input dimension of 46×6 (representing 46 nucleotides: A, T, C, G, N, and padding) and an output dimension of 46×5; two unidirectional LSTM layers with 64 and 32 neurons, respectively (returning hidden states for each time step), followed by a dropout rate of 0.5; a fully connected layer with 64 neurons (ReLU activation),

another dropout of 0.5; and an output layer with two neurons (linear activation). We determined that using the Adam optimizer with a learning rate of 2^-11 resulted in a robust model, which we named the "context-aware" model, or "TALE".

For the "motif counting" model, we constructed a CNN with the following architecture: an embedding layer identical to that of the "context-aware" model; a 1D convolutional layer with stride size 1 and ReLU activation; a global average pooling layer; and an output layer with two neurons (linear activation). For each kernel size in the convolutional layer, we explored the optimal number of filters and learning rates separately.

All networks were constructed and trained using TensorFlow. Model training was performed with the maximum batch size permitted by the GPU memory (11 GB), typically 32,768.

In silico mutagenesis and directed evolution

To calculate the "regulatory relevance score" of a specific nucleotide in a N45, we first mutate this nucleotide into the other three possible nucleotides (A, T, C, G), generating three mutant N45s. We then use the model to predict the regulatory phenotype (RNA/DNA or Cyt/Nuc) differences between the original N45 and these three mutant N45s. The median of these three differences is taken as the final score.

To study the variation in the regulatory relevance of a target element in random contexts, we fixed the target element in the middle of the N45 sequence and randomly filled in the other positions, generating 512 N45 variants. We then predicted the sum of the "regulatory relevance

scores" for each nucleotide within the target element in each N45, defining this as the "regulatory correlation" of that element in the given context.

For directed evolution, we first generated a list of products containing all possible saturation point mutations of a given initial N45 sequence and predicted the regulatory phenotypes of these products. The mutant closest to the desired evolutionary target was selected as the substrate for the next round of mutations. This iterative process continued until no further improvements in sequence fitness could be achieved.

To visualize the impact of a nucleotide at a "mutated position" (1A) on the regulatory relevance score (original score: 1a2a) of another nucleotide at an "affected position" (2A) within an N45 sequence, we mutated 1A to the other three nucleotides (1T, 1C, 1G) respectively. Using the method described above, we evaluated the regulatory relevance scores of 2A after each mutation (1t2a, 1c2a, 1g2a) and calculated the median of 1t2a/1a2a, 1c2a/1a2a, and 1g2a/1a2a as the fold-change value of 2A's regulatory relevance score upon 1A mutation (i.e., the fold-change values shown in Fig. 4B). For any given N45 sequence, this operation was performed for all possible pairs of mutated and affected positions, ultimately generating an association matrix. This matrix was then visualized as a heatmap to depict inter-nucleotide interactions within the N45. To emphasize subtle changes (and prevent large values from dominating the color scale), negative values in the matrix were set to 0, while values greater than 2 were capped at 2.

k-mer additivity test

We first retrieved each N45 population containing a specific *k*-mer or *k*-mer pair. Similar to the method described in "SEERS *k*-mer profiling," we inferred the regulatory power of each *k*-mer or *k*-mer pair by comparing the median regulatory property of the N45 population containing that *k*-mer or *k*-mer pair with the entire N45 population. We denote the regulatory power of *k*-mer1 as "Delta1," that of *k*-mer2 as "Delta2," and the combined regulatory power of *k*-mer1 and *k*-mer2 as "Delta12." Delta1 + Delta2 represents the linear "expected" value of the additive effect between the *k*-mers, while Delta12 represents the "actual" value.

Native 3′UTR variant effect assessment

We first extracted 44 nucleotides upstream and downstream of the target SNP to obtain 89-nucleotide Ref and Alt sequences. Each sequence was split into 45 overlapping 45-mers with a step size of 1 nucleotide. We used the TALE model to predict the regulatory phenotypes of these 45-mers and then compared the differences (Delta Cyt/DNA) between each pair of 45-mers from Ref and Alt. The median of these differences was taken as the final output value for assessing the effect of the SNP.

Data access

All raw sequencing data generated in this study have been deposited in the Genome Sequence Archive for Human (GSA-Human; https://ngdc.cncb.ac.cn/gsa-human/) under accession number

HRA008408. All processed data and code are provided in Supplemental_Code.zip and are also available at https://github.com/gao-lab/SEERS.

Competing interests

Authors declare that they have no competing interests.

Acknowledgments

We thank Drs. C. Li and L. Kong, as well as their labs for their generous supports in experimental facilities. We thank X.W. Liu and L.M. Xiao for their artistic inspirations. This work was supported by funds from National Key Research and Development Program (grant no. 2016YFC0901603) and State Key Laboratory of Gene Function and Modulation Research (formerly State Key Laboratory of Protein and Plant Gene Research). Part of the analyses was supported by the Computing Platform of the Center for Life Sciences of Peking University and the High-performance Computing Platform of Peking University.

Author contributions: N.L., G.G., and L.C.R. conceived the study. N.L. developed SEERS and conducted the experiments. N.L., J.Y.L., and Y.D. analyzed the experimental data. J.Y.L. and N.L. constructed the neural network models and performed downstream analyses. N.L. wrote the manuscript. G.G., L.C.R., J.C., and Y.W. revised the manuscript.

References

1. Mayr, C. What Are 3′ UTRs Doing? *Cold Spring Harb Perspect Biol* a034728 (2018) doi:10.1101/cshperspect.a034728.

2. Zrimec, J. *et al.* Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat Commun* **11**, 6141 (2020).

3. Fazal, F. M. *et al.* Atlas of Subcellular RNA Localization Revealed by APEX-Seq. *Cell* **178**, 473-490.e26 (2019).

4. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**, 1173–1175 (2009).

5. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**, 265–270 (2012).

6. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**, 271–277 (2012).

7. Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

8. van Arensbergen, J. *et al.* Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol* **35**, 145–153 (2017).

9. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).

10. van Arensbergen, J. *et al.* High-throughput identification of human SNPs affecting regulatory element activity. *Nat Genet* **51**, 1160–1169 (2019).

11. Abell, N. S. *et al.* Multiple causal variants underlie genetic associations in humans. *Science* **375**, 1247–1254 (2022).

12. Zhao, W. *et al.* Massively parallel functional annotation of 3′ untranslated regions. *Nat Biotechnol* **32**,

387–391 (2014).

13. Oikonomou, P., Goodarzi, H. & Tavazoie, S. Systematic Identification of Regulatory Elements in Conserved 3′ UTRs of Human Transcripts. *Cell Reports* **7**, 281–292 (2014).

14. Yartseva, V., Takacs, C. M., Vejnar, C. E., Lee, M. T. & Giraldez, A. J. RESA identifies mRNA-regulatory sequences at high resolution. *Nat Methods* **14**, 201–207 (2017).

15. Rabani, M., Pieper, L., Chew, G.-L. & Schier, A. F. A Massively Parallel Reporter Assay of 3′ UTR Sequences Identifies In Vivo Rules for mRNA Degradation. *Molecular Cell* **68**, 1083-1094.e5 (2017).

16. Shukla, C. J. *et al.* High-throughput identification of RNA nuclear enrichment sequences. *The EMBO Journal* **37**, e98452 (2018).

17. Lubelsky, Y. & Ulitsky, I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555**, 107–111 (2018).

18. Litterman, A. J. *et al.* A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. *Genome Res.* **29**, 896–906 (2019).

19. Vejnar, C. E. *et al.* Genome wide analysis of 3′ UTR sequence elements and proteins regulating mRNA stability during maternal-to-zygotic transition in zebrafish. *Genome Res.* **29**, 1100–1114 (2019).

20. Slutskin, I. V., Weinberger, A. & Segal, E. Sequence determinants of polyadenylation-mediated regulation. *Genome Res.* **29**, 1635–1647 (2019).

21. Yin, Y. *et al.* U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature* **580**, 147–150 (2020).

22. Griesemer, D. *et al.* Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247-5260.e19 (2021).

23. Siegel, D. A., Le Tonqueze, O., Biton, A., Zaitlen, N. & Erle, D. J. Massively parallel analysis of human 3′ UTRs reveals that AU-rich element length and registration predict mRNA destabilization. *G3 Genes|Genomes|Genetics* **12**, jkab404 (2022).

24. Yuan, Y., Guo, L., Shen, L. & Liu, J. S. Predicting Gene Expression from Sequence: A Reexamination. *PLOS Computational Biology* **3**, e243 (2007).

25. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nat Rev Genet* **23**, 169–181 (2022).

26. de Boer, C. G. & Taipale, J. Hold out the genome: a roadmap to solving the cis-regulatory code. *Nature* 1–10 (2023) doi:10.1038/s41586-023-06661-w.

27. Ren, J. *et al.* Spatiotemporally resolved transcriptomics reveals the subcellular RNA kinetic landscape. *Nat Methods* **20**, 695–705 (2023).

28. Ietswaart, R. *et al.* Genome-wide quantification of RNA flow across subcellular compartments reveals determinants of the mammalian transcript life cycle. *Molecular Cell* **84**, 2765-2784.e16 (2024).

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

30. McGeary, S. E. *et al.* The biochemical basis of microRNA targeting efficacy. *Science* **366**, eaav1741 (2019).

31. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell* **163**, 698–711 (2015).

32. Sample, P. J. *et al.* Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat Biotechnol* **37**, 803–809 (2019).

33. Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell* **178**, 91-106.e23 (2019).

34. de Boer, C. G. *et al.* Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol* **38**, 56–65 (2020).

35. Sahu, B. *et al.* Sequence determinants of human gene regulatory elements. *Nat Genet* **54**, 283–294 (2022).

36. Yu, J. *et al.* Human Induced Pluripotent Stem Cells Free of Vector and Transgene Sequences. *Science* **324**, 797–801 (2009).

37. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods* **15**, 141–149 (2018).

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

39. Kirk, J. M. *et al.* Functional classification of long non-coding RNAs by k-mer content. *Nat Genet* **50**, 1474–1482 (2018).

40. Yuan, G.-H., Wang, Y., Wang, G.-Z. & Yang, L. RNAlight: a machine learning model to identify nucleotide features determining RNA subcellular localization. *Briefings in Bioinformatics* bbac509 (2022) doi:10.1093/bib/bbac509.

41. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).

42. Villanueva, E. *et al.* System-wide analysis of RNA and protein subcellular localization dynamics. *Nat Methods* 1–12 (2023) doi:10.1038/s41592-023-02101-9.

43. Sun, L. *et al.* Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures. *Cell Res* **31**, 495–516 (2021).

44. Berg, M. G. *et al.* U1 snRNP Determines mRNA Length and Regulates Isoform Expression. *Cell* **150**, 53–64 (2012).

45. Chan, J. J. *et al.* Pan-cancer pervasive upregulation of 3′ UTR splicing drives tumourigenesis. *Nat Cell Biol* 1–12 (2022) doi:10.1038/s41556-022-00913-z.

46. Tuerk, C. & Gold, L. Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science* **249**, 505–510 (1990).

47. Zuckerman, B., Ron, M., Mikl, M., Segal, E. & Ulitsky, I. Gene Architecture and Sequence Composition Underpin Selective Dependency of Nuclear Export of Long RNAs on NXF1 and the TREX Complex. *Molecular Cell* **79**, 251-267.e6 (2020).

48. Goraczniak, R., Behlke, M. A. & Gunderson, S. I. Gene silencing by synthetic U1 Adaptors. *Nat Biotechnol* **27**, 257–263 (2009).

49. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research* **47**, e47 (2019).

50. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

51. Xu, S. *et al.* Using clusterProfiler to characterize multiomics data. *Nat Protoc* **19**, 3292–3320 (2024).

52. Gaspar, J. M. NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics* **19**, 536 (2018).

53. Antonov, M. *et al.* igraph enables fast and robust network analysis across programming languages.

Preprint at https://doi.org/10.48550/arXiv.2311.10260 (2023).

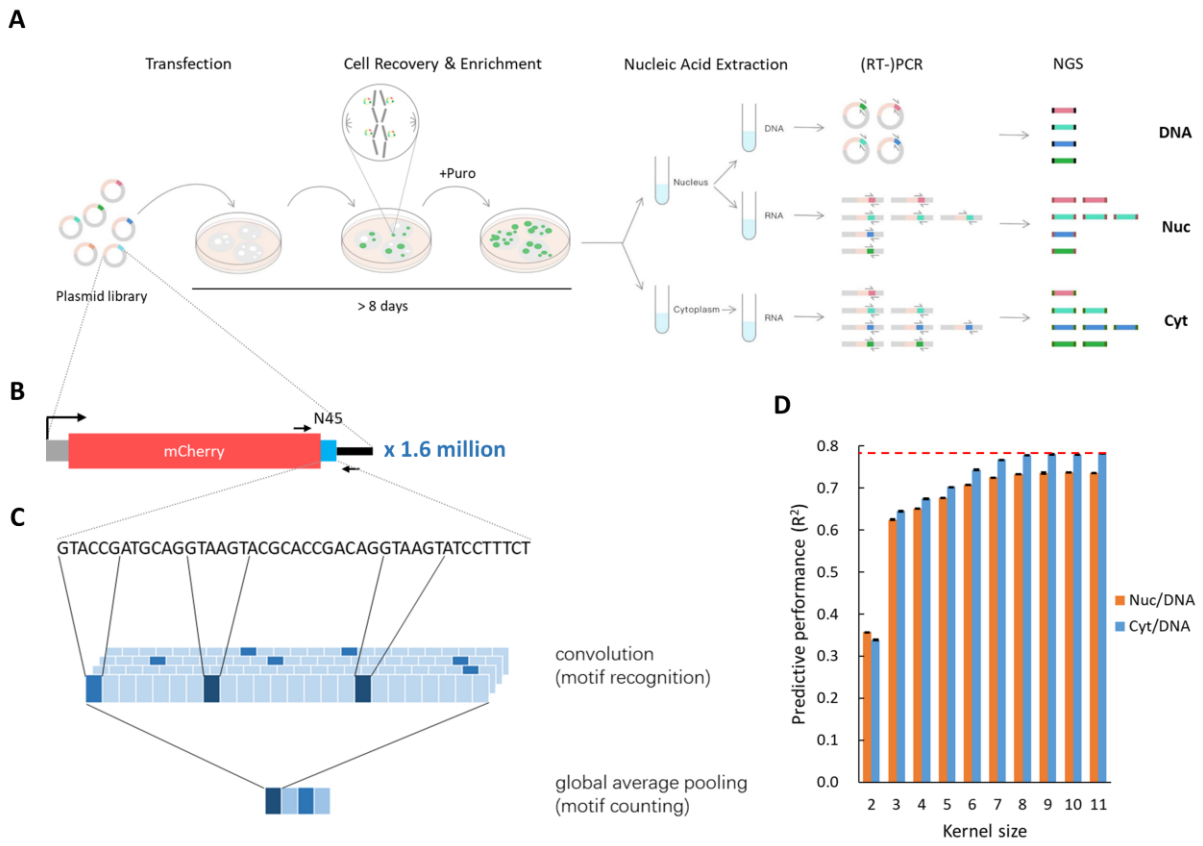

Fig. 1. Exploration of 3′UTR regulatory elements with SEERS. (**A**) The experimental workflow of SEERS. The regulatory effects of each 3′UTR insert could be quantified by its normalized subcellular abundance (Nuc/DNA and Cyt/DNA) as measured by amplicon sequencings (NGS). (**B**) Structure of the mCherry reporter transcript and the N45 insertions. Over 1.6 million N45s were assessed by SEERS for downstream model trainings and evaluations. Small arrows indicate the annealing sites of primers used for amplification and sequencing of the N45 inserts. (**C**) Design of the "motif counting" model (omitting its output layer). (**D**) Predictive performances (coefficient of determination, R^2, with respect to the main

test dataset, n=10,000) of the "motif counting" model for different kernel sizes (n=3 for each size).

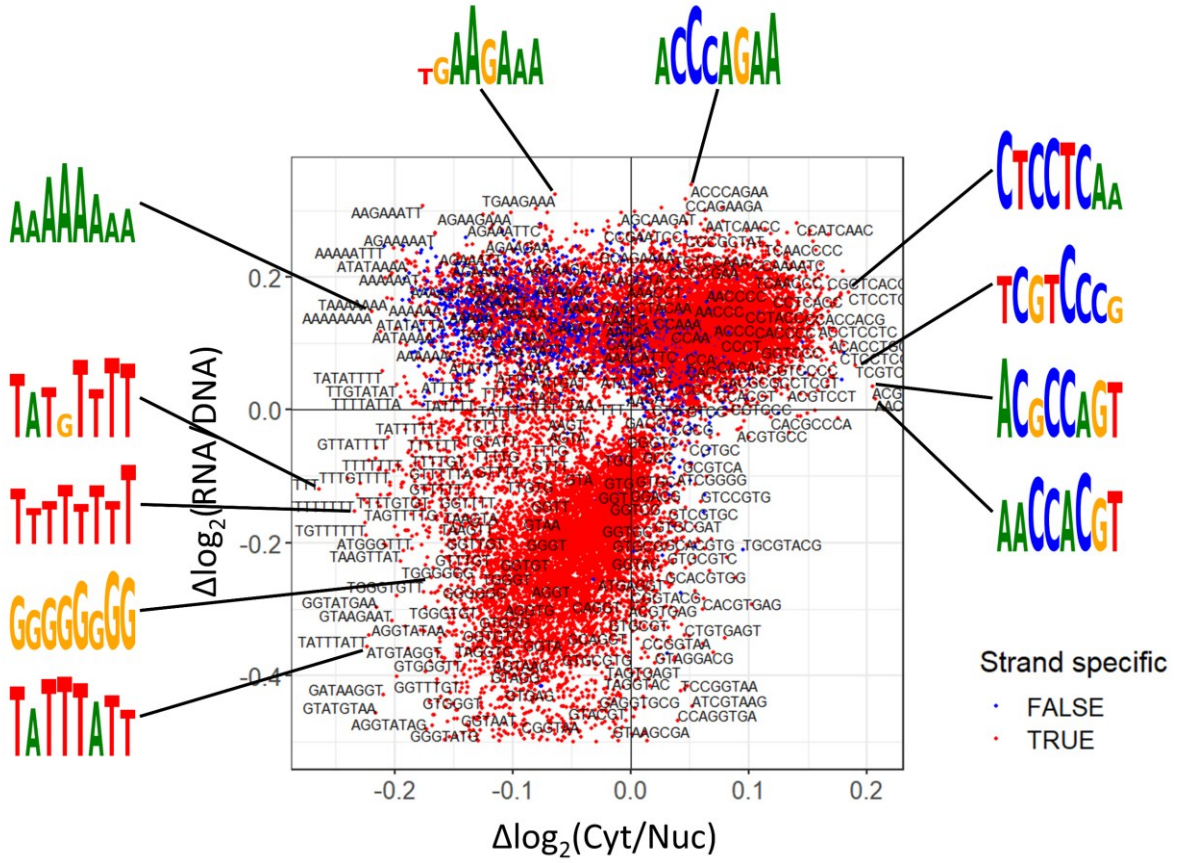

Fig. 2. SEERS identified a large number of elements with diverse regulatory roles in 3′UTRs. The functional landscape of 2-8mers with significant regulatory correlations (n=18,285) is shown, focusing on the core region (see Fig. S4B for the complete landscape). Representative motifs of the top-ranked 8-mers with the strongest regulatory relevance are highlighted. These motifs were determined via nucleotide substitution assays, which assess the contribution of individual nucleotides.

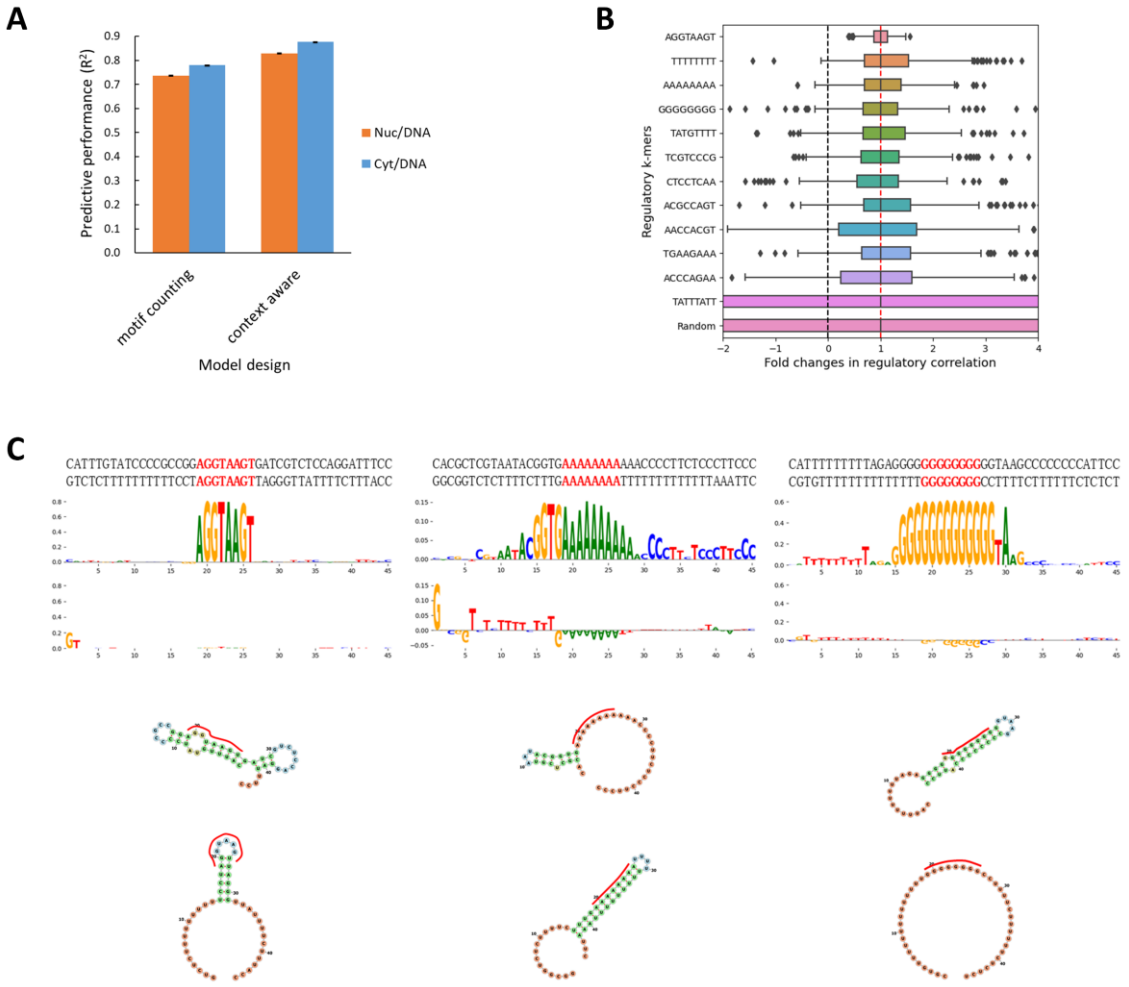

Fig. 3. Context-dependence of the regulatory elements revealed by TALE. (**A**) Comparison of the best predictive performance (coefficient of determination, R^2, with respect to the main test dataset, n=10,000) of the LSTM-based "context-aware" model (TALE) with the CNN-based "motif counting" model (n=3). (**B**) Range of regulatory relevance for each regulatory element (8-mer) predicted by TALE in random context sequences (n=512). Results for random 8-mers are included as negative control in the last row. (**C**) Top: Context sequences optimized by TALE for maximal enhancement (top) or suppression (bottom) for each regulatory element. Middle: the regulatory relevance score (in Nuc/Cyt) of each nucleotide in these sequences. Bottom: their secondary structures as predicted by RNAfold.

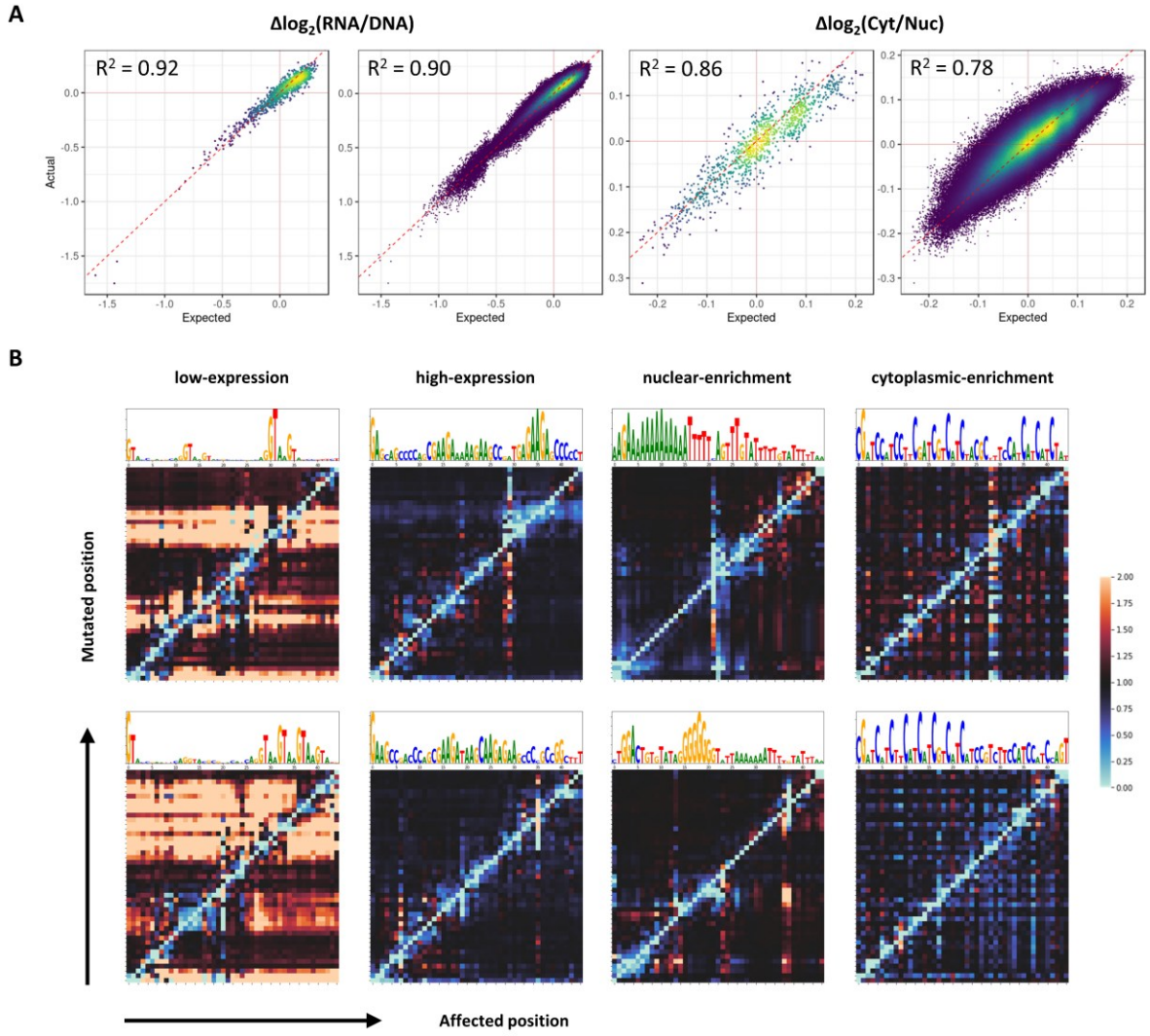

Fig. 4. Weak yet complex interactions exist between the regulatory elements. (**A**) The actual regulatory effects of any two identical (n=1,024, the more sparse plots) or different (n=524,800 pairs, the denser plots) 5-mer combinations in N45, compared with the linear expectation values obtained by simply summing their individual independent effects. (**B**) N45s with different regulatory characteristics predicted by TALE, showing the regulatory relevance score for each nucleotide (logo plot), and the fold changes in the scores of other nucleotides caused by mutations at each nucleotide (heatmap). To highlight subtle changes, all fold-change values beyond the 0-2 range were truncated to 0 and 2.

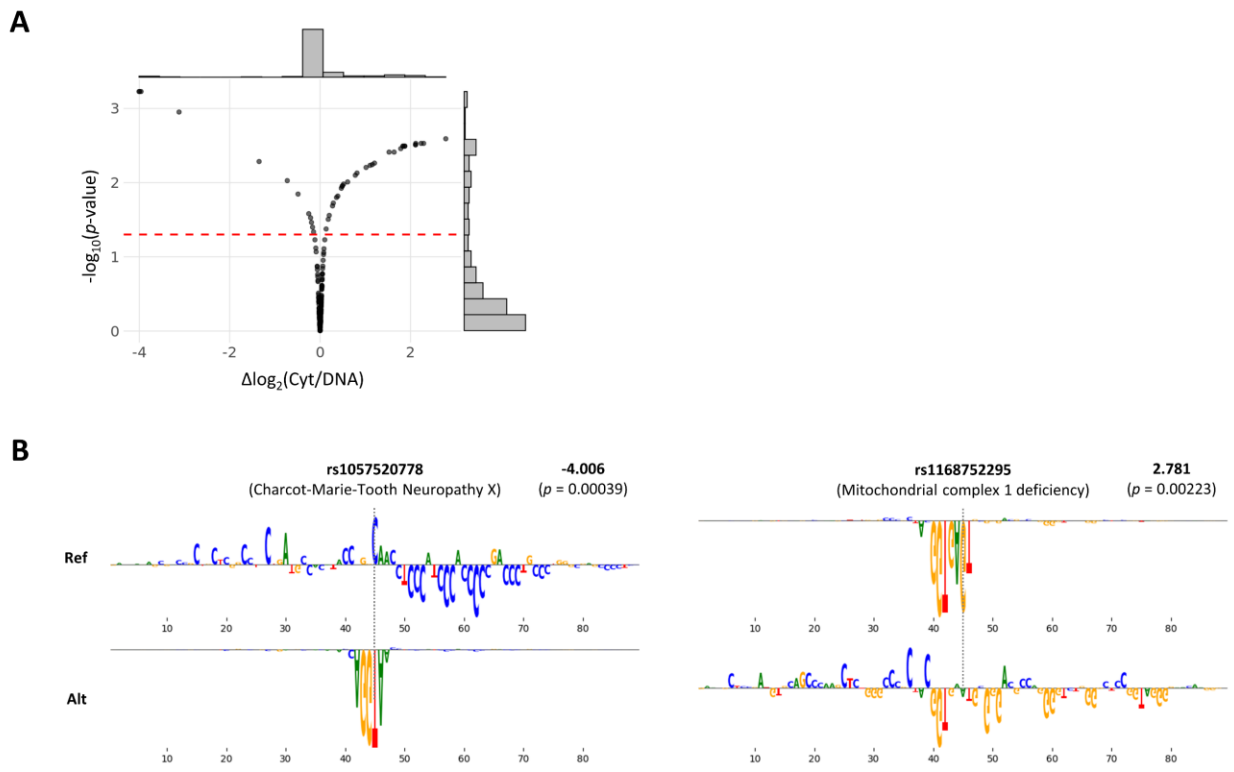

Fig. 5. The TALE model helps explain a subset of disease-associated 3′UTR variants. (**A**) Volcano plot of mutation effects predicted by TALE for all 190 pathogenic 3′UTR SNPs in ClinVar. The vertical axis represents the empirical *p*-values derived using the predicted effects of all 15,225 benign 3′UTR SNPs as the null distribution. The red dashed line indicates the significance cutoff at $p = 0.05$. (**B**) The strongest mutation effects predicted by TALE for pathogenic 3′UTR SNPs in ClinVar result from disruption or creation of 5′ splice sites. Upward nucleotide logos indicate mutations decreasing transcript abundance, while downward logos indicate increased abundance. Bold numerical values represent TALE-predicted mutation effects of the SNP ($\log_2$[Cyt/DNA]), with parenthetical *p*-values derived from empirical comparisons to the null distribution of predicted effects for benign 3′UTR SNPs in ClinVar.