RNA-GPS predicts high-resolution RNA subcellular localization and highlights the role of splicing

KEVIN E. WU,[1,2,3] KEVIN R. PARKER,[3] FURQAN M. FAZAL,[3] HOWARD Y. CHANG,[3,4] and JAMES ZOU[1,2]

[1]Department of Computer Science, Stanford University, Stanford, California 94305, USA
[2]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, California 94305, USA
[3]Center for Personal and Dynamic Regulomes, Stanford University School of Medicine, Stanford, California 94305, USA
[4]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA

ABSTRACT

Subcellular localization is essential to RNA biogenesis, processing, and function across the gene expression life cycle. However, the specific nucleotide sequence motifs that direct RNA localization are incompletely understood. Fortunately, new sequencing technologies have provided transcriptome-wide atlases of RNA localization, creating an opportunity to leverage computational modeling. Here we present RNA-GPS, a new machine learning model that uses nucleotide-level features to predict RNA localization across eight different subcellular locations—the first to provide such a wide range of predictions. RNA-GPS's design enables high-throughput sequence ablation and feature importance analyses to probe the sequence motifs that drive localization prediction. We find localization informative motifs to be concentrated on 3′-UTRs and scattered along the coding sequence, and motifs related to splicing to be important drivers of predicted localization, even for cytotopic distinctions for membraneless bodies within the nucleus or for organelles within the cytoplasm. Overall, our results suggest transcript splicing is one of many elements influencing RNA subcellular localization.

Keywords: localization mechanism; machine learning model; RNA localization; splicing in localization

INTRODUCTION

Subcellular localization of RNA transcripts is critical to cellular function, development, and organization (Lécuyer et al. 2007; Buxbaum et al. 2015; Chin and Lécuyer 2017). For example, mRNA transcript localization has been found to be an efficient mechanism for controlling spatial gene expression and localization of subsequently translated proteins (Martin and Ephrussi 2009; Jung et al. 2012). Transcript localization is also a widely observed phenomenon; the majority (~80%) of RNA transcripts exhibit asymmetric distribution across the cellular volume in both human and Drosophila cells (Benoit Bouvrette et al. 2018). Errant transcript localization may also play a pathogenic role; errors in transcript localization have been found in patients with various diseases like spinal muscular atrophy, fragile X syndrome, and Alzheimer's disease (Chin and Lécuyer 2017). Such errors have also been implicated in various forms of cancer (Cooper et al. 2009; Smart et al. 2018). Refining our understanding of this key cellular process would have great implications for basic biology, and possibly even for downstream biomedical applications. In this paper, we set out to achieve this by developing and interpreting a machine learning model that predicts RNA localization from nucleotide sequences.

It is currently accepted that transcript sequence plays a large role in driving transcript localization. Experimental studies have shown that RNA binding proteins (RBPs) typically interact with either primary sequence motifs or secondary structures to guide localization (Ryder and Lerit 2018). Since secondary structure is itself largely determined by primary sequence (Capriotti and Marti-Renom 2010), both localization mechanisms ultimately depend on the transcript sequence. Additional studies have identified Alu repeats that drive localization of long RNAs in human cells (Lubelsky and Ulitsky 2018), along with other sequence motifs that specifically drive nuclear localization of long noncoding RNAs (lncRNAs) (Zhang et al. 2014),

Corresponding authors: howchang@stanford.edu, jamesz@stanford.edu
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.074161.119.

further highlighting the role of transcripts' primary sequence in determining their localization.

Consequently, several works use transcript sequence and/or sequence-derived features to computationally model RNA localization patterns. Most of these works focus on predicting broad nuclear versus cytoplasm localization. For example, deepLncRNA predicts nuclear versus cytoplasmic localization of lncRNAs with a neural network that considers *k*-mer features, the presence of RBP sequence motifs, and genomic loci information (Gudenas and Wang 2018). Recent approaches, such as RNATracker, aim to predict localization to multiple sites by applying complex neural networks to the transcript sequence, which may be further annotated with computationally inferred secondary structure (Yan et al. 2019). Measures of splicing activity have been leveraged in models predicting localization as well (Zuckerman and Ulitsky 2019). Other related works like lncLocator focus on predicting localization of a subclass of RNA corresponding to lncRNAs (Cao et al. 2018).

Despite recent progress, relatively little is known about sequence factors that drive localization to more specific cellular compartments—beyond simple nucleus versus cytoplasm—for RNA broadly. Moreover, of the aforementioned works, only RNATracker provides a meaningful attempt at model interpretation in an effort to relate its performance to biological mechanisms. As the promise of large data sets, in biology or otherwise, often lies beyond simply building a model, but in the insights gained through understanding the nuances of that model, this lack of focus on model interpretation represents a missed opportunity to elucidate the biological mechanisms underlying localization.

We aim to address these challenges in this paper. We leveraged the recently developed APEX-seq technology (based on proximity biotinylation of endogenous RNAs) and data (Fazal et al. 2019) to develop a new model, RNA-GPS, that can predict transcript localization for eight different localizations simultaneously. This is the first method that can predict such highly granular RNA localization, to the best of our knowledge. RNA-GPS incorporates biological knowledge in its design and is directly interpretable. We demonstrate that RNA-GPS predicts localization more accurately compared to several neural network-based approaches, which are also more challenging to interpret (Ghorbani et al. 2019; Gilpin et al. 2019). We present evidence that RNA-GPS not only achieves strong performance but does so by learning meaningful colocalization patterns. The results of our interpretation methods consistently implicate splicing, splicing factor proteins, and the effects of splicing as factors in transcript localization.

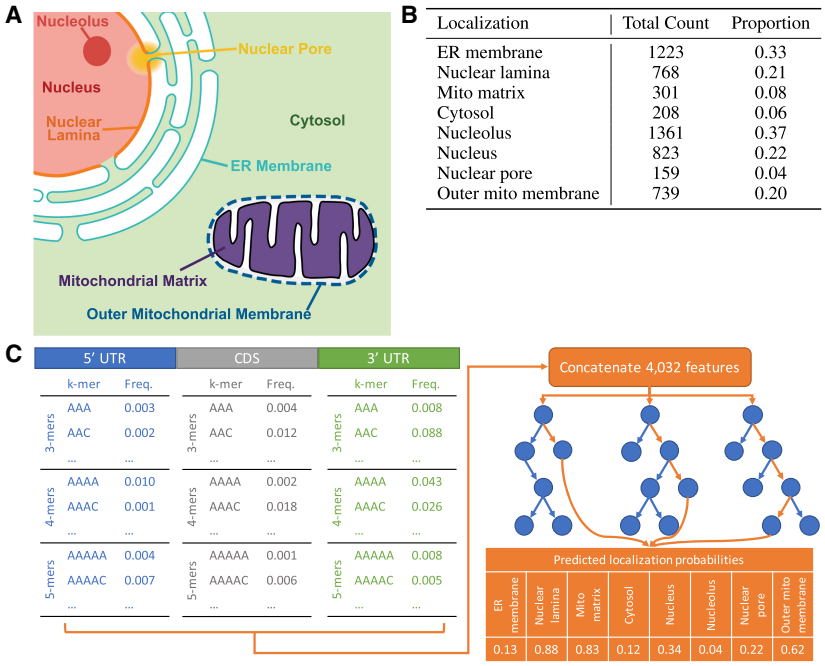
RNA-GPS contributes to multiple types of analysis. Its prediction framework quantifies how much subcellular localization information is contained in different parts of the RNA. This suggests interesting biological insights. For example, we find that the localization signal concentrates around the 3′-UTR and splicing motifs. RNA-GPS also enables researchers to carry out in silico perturbations to estimate how certain RNA sequence modifications alter localization preference. This leads to new biological hypotheses and can even help design synthetic transcripts with prescribed localization tendencies. Similar types of in silico analysis have recently been shown to be powerful applications of machine learning in other areas of genomics (Zou et al. 2019). Finally, RNA-GPS broadens the scope of RNA subcellular analysis by predicting localization in new sequences and in new environments, where experimental data is not currently available.

RESULTS

RNA-GPS predicts the localization of an RNA transcript to eight subcellular compartments—the cytosol, endoplasmic reticulum, mitochondrial matrix, outer mitochondrial membrane, nucleus, nucleolus, nuclear lamina, and nuclear pore (Fig. 1A). For each transcript, RNA-GPS creates a set of features by first segmenting the transcript sequence into the 5′ untranslated region (UTR), coding sequence (CDS), and 3′-UTR, and then quantifies the length-normalized *k*-mer frequencies within each, for *k* equals to 3, 4, and 5 (Fig. 1C). This featurization method critically captures important spatial information regarding where in the transcript a *k*-mer is present. RNA-GPS then predicts the probability that the transcript localizes to each of the eight compartments using a random forest model. Note that transcripts often localize to multiple compartments (Supplemental Fig. S1C,D), and hence we make an independent prediction for each compartment.

RNA-GPS accurately predicts localization to the eight subcellular compartments

We trained RNA-GPS to predict these eight localizations using a data set of $n = 3660$ transcripts derived from APEX-seq results measuring transcript localization in human HEK293T cells (Fazal et al. 2019). Most (but not all) of our transcripts are protein coding (Supplemental Fig. S1E), and each localizes to one or more of eight subcellular compartments: cytosol, endoplasmic reticulum, mitochondrial matrix, outer mitochondrial membrane, nucleus, nucleolus, nuclear lamina, and nuclear pore. The first four of these localizations are cytoplasmic, whereas the latter four are nuclear (Fig. 1A), and the proportion of transcripts localizing to each compartment is shown in Figure 1B. We split this data set into training (80%, $n = 2928$), validation (10%, $n = 366$), and test (10%, $n = 366$) sets. The validation set was used for model architecture and hyperparameter tuning, and the test set was used to perform a final evaluation; all subsequent results and interpretations in the



Localization	Total Count	Proportion
ER membrane	1223	0.33
Nuclear lamina	768	0.21
Mito matrix	301	0.08
Cytosol	208	0.06
Nucleolus	1361	0.37
Nucleus	823	0.22
Nuclear pore	159	0.04
Outer mito membrane	739	0.20

FIGURE 1. Summary of the RNA localization data and RNA-GPS. (*A*) RNA-GPS is trained on APEX-seq data which localizes transcripts to eight subcellular compartments. Localizations considered cytoplasmic lie within the green cytoplasmic region, and nuclear localizations lie within the red nuclear region in the *upper left*. (*B*) The number of positive transcripts for each localization, as well as the corresponding proportion. In total, there are $n = 3660$ transcripts. (*C*) Schematic of RNA-GPS workflow. The algorithm first partitions the sequence into the 5′ untranslated region (UTR), CDS, and 3′-UTR. For each segment, we generate a *k*-mer featurization for $k = 3, 4, 5$ (*left*), which is then used as input to a random forest that outputs a vector of predicted probabilities of localization to each compartment that need not sum to 1 (*right*). The orange arrows trace a possible path through the random forest.

primary text are reported on the test set (and in some cases, with additional cross-validation).

RNA-GPS achieves an overall area under the receiver operating characteristic curve (AUROC) of 0.77 and an area under the precision-recall curve (AUPRC) of 0.49 on the held-out test transcripts (Fig. 2A). RNA-GPS's performance is highly consistent when performing 10-fold cross validation (Supplemental Fig. S4C) and is also consistent across a range of transcript lengths (Supplemental Fig. S4E). To ensure that the prediction performance was not artificially inflated by sequence similarity across train and test sets, we removed all test sequences with significant similarity to training sequences according to BLAST (Altschul et al. 1990) ($n = 110$), and found a very similar overall AUROC of 0.75 and AUPRC of 0.43 on the remaining sequences ($n = 256$). For individual compartments, RNA-GPS also achieves consistently high AUROC and accuracy values (Fig. 2B).

To contextualize RNA-GPS's performance, we performed several additional analyses. First, we compared RNA-GPS to a previous state-of-the-art method, deepLncRNA (Gudenas and Wang 2018), which we adapted, reimplemented, and retrained (see Materials and Methods section). Since deepLncRNA can only predict binary nuclear versus cytoplasmic localization, we "collapsed" and retrained RNA-GPS to predict this binary output as well. On this simplified task, RNA-GPS substantially outperforms deepLncRNA with an AUROC of 0.85 versus 0.74, as evaluated on the test set (Fig. 2C; Supplemental Fig. S3A). RNA-GPS also outperforms deepLncRNA's original reported test set AUROC of 0.787 (Gudenas and Wang 2018). This shows that RNA-GPS's design inherently surpasses that of previous approaches, even on relatively simple localization prediction tasks.

We then sought to contextualize the eight-compartment localization performance of RNA-GPS using a simple baseline ("Baseline" in Fig. 2A). For this, we trained a random forest classifying binary nuclear versus cytoplasmic localizations and combined its predictions with random sublocalizations to nuclear and cytoplasmic compartments. This baseline reflects the performance of a model that is capable of distinguishing nuclear from cytoplasmic sequences, but not much else. The performance of this baseline is substantially worse than RNA-GPS though it does significantly outperform a purely random classifier. This demonstrates that predicting fine-grained transcript localization is a nontrivial extension of nuclear versus cytoplasmic prediction, and one that RNA-GPS successfully learns.

Next, we benchmarked several additional machine learning models—state-of-the-art convolutional and recurrent algorithms, as well as other tree-based methods. We trained and tested each of these methods on the same multilocalization data set as RNA-GPS. We adapted RNA-GPS's segment-wise featurization for the boosted tree and Basset-3 models, in hopes of teasing apart the impact of featurization versus modeling strategies. RNA-GPS significantly outperforms all evaluated deep learning based approaches, likely because its featurization strategy elegantly captures local sequence patterns in a way that is largely agnostic of transcripts' highly variable sequence lengths (Supplemental Fig. S1B)—an intrinsic property of RNA transcripts that deep learning networks often struggle with. The boosted tree model, consisting of eight individual boosted trees each trained to predict localization to one compartment, achieves the most comparable performance. It is also interesting to note that Basset-3, with its

A

Model Type	Model	AUROC	AUPRC
Random	$U(0,1)$	0.49	0.22
Tree	**RNA-GPS**	**0.77**	**0.49**
	Baseline	0.65	0.31
	Boosted tree	0.76	0.48
Convolutional	Basset	0.63	0.26
	Basset-3	0.69	0.34
	Basset-pool	0.50	0.20
Recurrent	LSTM	0.62	0.26
	GRU	0.68	0.33
Conv. & Rec.	RNATracker	0.69	0.32

B

Localization	AUROC	AUPRC	Acc.
ER membrane	0.78	0.68	0.69
Nuclear lamina	0.74	0.48	0.74
Mito matrix	0.71	0.33	0.64
Cytosol	0.86	0.49	0.58
Nucleolus	0.69	0.60	0.65
Nucleus	0.76	0.55	0.68
Nuclear pore	0.80	0.32	0.82
Outer mito mem.	0.79	0.52	0.77
Overall	0.77	0.49	0.70

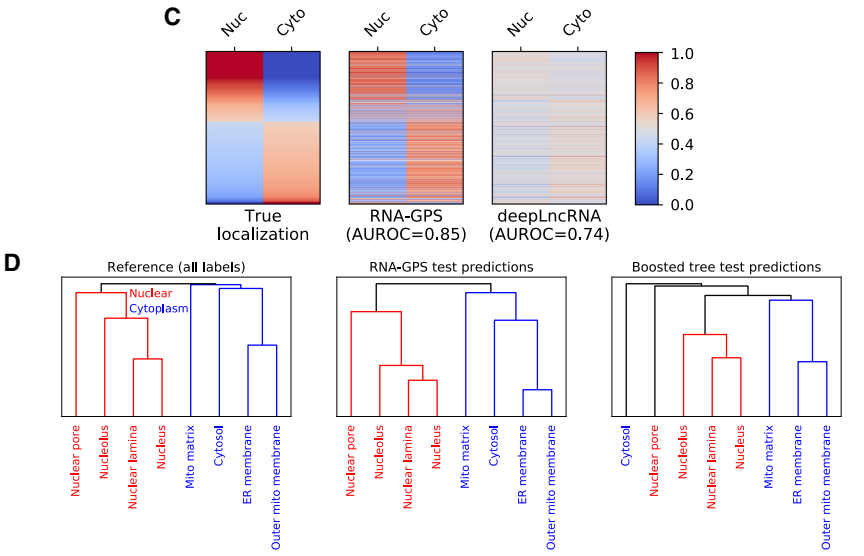


FIGURE 2. Summary of RNA-GPS prediction results. (*A*) Test performance for RNA-GPS and several additional models we developed. RNA-GPS exhibits the best performance for both overall AUROC and AUPRC. See Supplemental Figure S4C for cross-validation results, and Materials and Methods section for detailed description of each model. (*B*) Detailed performance breakdown of RNA-GPS on each of eight different localizations in the test set. AUROC is consistently high, while there is more variance in AUPRC; these curves are shown in Supplemental Figure S4A,B. (*C*) Heatmaps visualizing the output of RNA-GPS (*center*) and deepLncRNA (*right*) on test data for binary nuclear/cytoplasmic localization prediction, where each row represents one of 916 test set genes. The *left* plot shows the ground truth colored by (clipped) $\log_2$ fold changes normalized to [0, 1]. For RNA-GPS, we see clear regions of nuclear/cytoplasmic predictions, whereas the vast majority of deepLncRNA's predictions are "ambivalent" with only small differences separating positive and negative predictions. (*D*) Hierarchical clustering of eight localization compartments using colocalization patterns. Plots are colored to indicate predominantly nuclear (red) or cytoplasmic (blue) subtrees and localizations. Note that although the boosted tree achieves high AUROC and AUPRC, it discordantly separates the cytosol and nuclear pore from the cytoplasmic/nuclear subtrees, respectively. RNA-GPS exactly mirrors the ground-truth clustering, suggesting that its performance is achieved via learning a biologically relevant understanding of localization patterns. This pattern is consistent across cross-validation folds (Supplemental Fig. S4D).

segment-aware featurization, outperforms the default Basset implementation, which lends support to the effectiveness of our featurization strategy. We further evaluated the boosted tree model (highest performing in-house competitor), RNATracker (highest performing existing model), and RNA-GPS across cross-validation folds (Supplemental Fig. S4C). We found that while RNA-GPS and the boosted tree both perform consistently well across folds, RNATracker exhibits large variability in its performance, likely due to overfitting.

We can go beyond summary statistics like AUROC to show that RNA-GPS better captures true biological signal, compared to the aforementioned boosted tree. By clustering cellular compartments based on colocalization patterns, we see that RNA-GPS successfully recapitulates true colocalization patterns (Fig. 2D). Examining the hierarchical clustering of cellular compartments based on observed colocalizations (left), we see that the first split separates nuclear and cytoplasmic localizations, a distinction concordant with biological intuition. Hierarchical clustering of RNA-GPS's predictions (center) mirrors this split, but clustering the boosted tree's predictions incorrectly separates the cytosol and nuclear pore from their cytoplasmic and nuclear relatives (right). Furthermore, RNA-GPS fully reproduces the ground truth relationship between all eight compartments, even past the initial nuclear/cytoplasmic split. Quantifying this difference, RNA-GPS's localization clustering has a Robinson-Foulds distance of 0—which is optimal—compared to the true clustering, whereas the boosted tree clustering has a Robinson-Foulds distance of 4. Across all cross-validation folds, we see that RNA-GPS achieves comparable or better distance on 80% of the folds compared to the boosted tree (Supplemental Fig. S4D). This suggests that RNA-GPS successfully learns localization patterns across multiple cellular localizations, likely as a consequence of having a single, unified internal representation in its model.

Given the apparent importance of the distinction between cytoplasmic and nuclear compartments, we also experimented with tiered models that incorporate this separation as a "biological prior." We did this by training a random forest predicting cytoplasmic and nuclear localization and combined its predictions with two subsequent random forest models, one trained only to predict cytoplasmic sublocalizations and one trained only to predict nuclear sublocalizations, thus mimicking the presumed hierarchical localization process. This is similar to the aforementioned baseline model, except with trained models for sublocalization instead of random predictors. We found

that this "biological prior" approach did not outperform RNA-GPS. This suggests that RNA-GPS is already learning the biological distinction between nuclear and cytoplasmic localizations without explicit encouragement.

Another known biological property of transcript localization is the mechanistic role of RNA binding proteins (RBPs) in regulating this process (Ryder and Lerit 2018)—a property included in the featurization schemes of models like DeepLncRNA (Gudenas and Wang 2018). We attempted to improve RNA-GPS's performance by augmenting our feature space with similar features quantifying enrichment of known RBP binding sites but did not see an improvement (Supplemental Fig. S2). This suggests that RNA-GPS has already learned signals correlated with these motifs, despite having access to only relatively short, unordered *k*-mer features.

Finally, we evaluate how well RNA-GPS generalizes to cell types other than the HEK293T cell line that it was trained on (Supplemental Fig. S3B). As most publicly available transcript localization data sets distinguish only between nuclear versus cytoplasmic localizations, we evaluated the reduced version of RNA-GPS predicting this binary outcome. Given $n = 7641$ localized transcripts measured on the HeLa-S3 cell line (ENCODE Project Consortium 2012), RNA-GPS achieves an AUROC of 0.83. Similarly, on a set of $n = 6359$ localized transcripts measured from the K562 cell line (ENCODE Project Consortium 2012), RNA-GPS achieves an AUROC of 0.82. Both these values are quite similar to the AUROC of 0.85 on the held-out APEX-seq test set, suggesting that RNA-GPS and its predictions can generalize to cell types that it has never before encountered.

Overall, we see consistent evidence of RNA-GPS's strong performance across a variety of RNA localization prediction tasks and contexts. RNA-GPS outperforms prior, often highly complex models, and does so without incorporating prior biological knowledge like known RBP binding sites or assumptions regarding hierarchies in localization. RNA-GPS manages to even recapitulate colocalization patterns, suggesting that it is recognizing true biological signals based on sequence features alone.

Segment-level interpretation of RNA-GPS

RNA-GPS featurizes the 5′ UTR, CDS, and 3′-UTR transcript segments separately (see Materials and Methods section). This scheme naturally lends itself to an ablation study evaluating the relative impact of each segment on localization, where we zero-out each segment's corresponding features and observe, for each transcript, changes to our eight predicted localization probabilities (without retraining), the results of which are shown in Figure 3A. These *P*-values are computed using paired *t*-tests that separately evaluate the impact to positive (exhibiting significant enrichment) and negative (no significant enrichment) localizations (to avoid Simpson's paradox). Overall, ablating the 5′ UTR has no significant effect, while ablating the CDS or 3′-UTR both result in a significant drop in model performance ($\alpha = 0.05$), both in causing the positive localizations to receive lower, less confident scores and the negative localizations to erroneously receive higher scores. For each of the three segment ablations, we also evaluated its impact to RNA-GPS's overall performance in predicting localization to each compartment. These results are shown in Figure 3B and echo the importance of the CDS and 3′-UTR.

For further validation, we also performed the reverse study—training and evaluating variants of RNA-GPS using only features from the 5′ UTR, CDS, or 3′-UTR—and observed similar results. The AUROC and AUPRC for the models trained on each segment is shown in Supplemental Figure S5B. We see that the model trained using features derived from the 5′ UTR had overall performance far poorer than those trained using only the CDS or the 3′-UTR. These results strongly indicate that not only do different segments of the transcript indeed play different roles in localization, but that the 5′ UTR appears to play the least role in driving this process, corroborating prior studies emphasizing the role of the 3′-UTR (Mayr 2018), and supporting our choice to featurize each segment separately.

Motif-level interpretation of RNA-GPS

We take advantage of RNA-GPS's tree-based architecture to identify important *k*-mer features and subsequently assemble them into human-comprehensible sequence motifs (see Materials and Methods section and Fig. 4A). From an original feature space of 4032 features, we identified ∼150–300 *k*-mer features important for predicting each localization (exact counts shown in Fig. 4B). After reassembling these *k*-mers into motifs and annotating them with known RNA binding protein (RBP) binding sites, we found several hits for each localization (Fig. 4B,C). We find that a majority of these localization-driving RBPs have been previously implicated in splicing. Using DAVID (version 6.8) (Huang et al. 2009a,b) to match these RBPs against biological process terms produces, as a top hit, "mRNA splicing, via spliceosome" with an associated $P = 3.6 \times 10^{-15}$, along with several other significant splicing-related terms. These splicing-oriented results are robust and reproducible across different choices of hyperparameters used in the *k*-mer assembly methodology (Supplemental Fig. S6). We see a few RBPs that are particularly highlighted (with more than 10 of their binding sites occurring across our reconstructed motifs): HuR, PTBP1, RBM4, and TIA1, all of which have been experimentally implicated in splicing. HuR, or Hu antigen R, has been found to help stabilize transcripts and directly facilitate their transport (Tran et al. 2003; Doller et al. 2008), but more importantly has also been found to play a role in

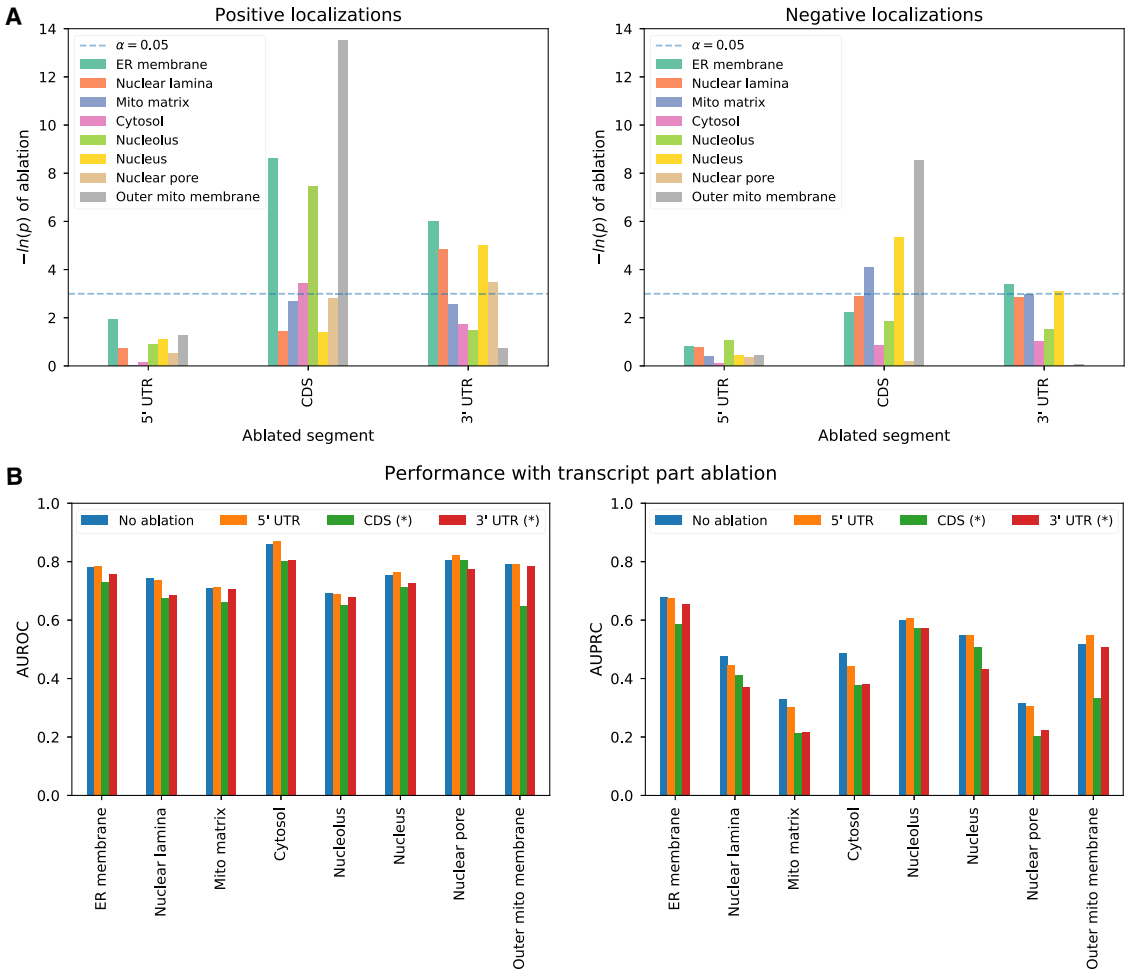
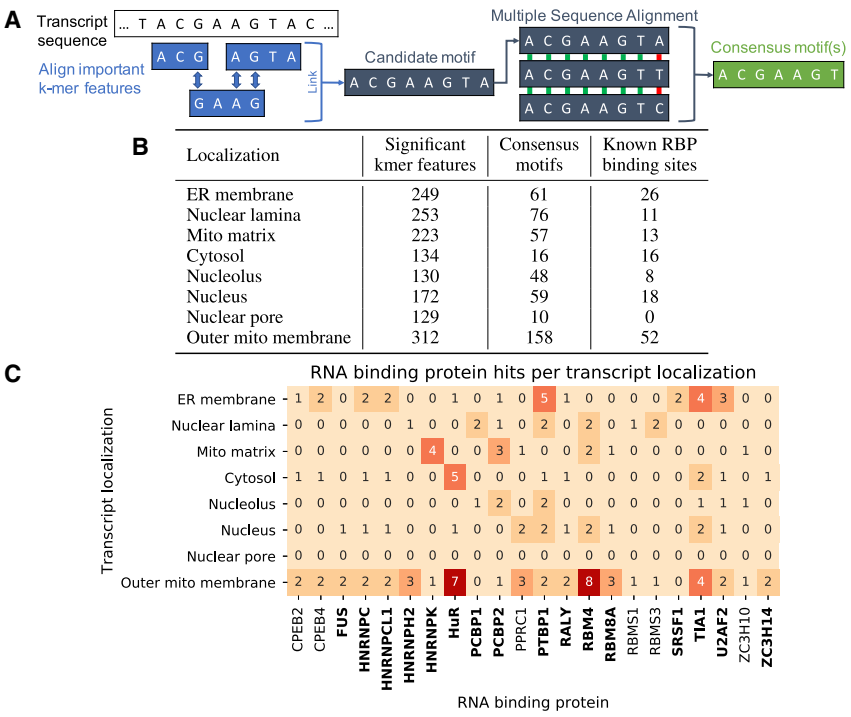


FIGURE 3. Summary of segment ablation results. (*A*) Barplots showing negative-log *P*-values for observed changes in model predictions upon ablating each segment, computed using paired *t*-tests for positive and negative localizations, on the test set. The significance level of $\alpha = 0.05$ is shown by the dotted line. Bars that exceed this indicate that the associated ablation causes a significant impact to RNA-GPS's ability to correctly predict the corresponding (positive or negative) localization. Ablating the 5′ UTR does not result in any significant changes in positive or negative predictions across the board, whereas ablating the CDS and 3′-UTR both result in significant losses in the model's confidence in correct/positive localizations, and significant (erroneous) gains in our model's predictions for incorrect/negative localizations. Exact *P*-values can be found in Supplemental Figure S5A. (*B*) Barplots showing per localization model performance upon ablating each segment of the transcript, compared to no-ablation, full sequence baseline. We see that ablating the CDS and 3′-UTR both result in consistent drops in performance, while this is not the case for the 5′ UTR. Asterisks indicate ablations that cause significant changes in model predictions (i.e., having a significant impact to compartment-wise AUROC or AUPRC, respectively). This shows that the example-level impact shown in *A* also manifests when evaluating our data set holistically.

regulating transcript splicing (Izquierdo 2008; Akaike et al. 2014). PTBP1, or polypyrimidine tract binding protein 1, regulates mRNA splicing during neuronal differentiation (Yap et al. 2012; Vuong et al. 2016). RBM4 (RNA binding motif protein 4) couples with SRSF1, which is also identified by our interpretation method albeit at a lower enrichment, to form an antagonistic splicing regulation mechanism (Chang and Lin 2019). TIA1 is a well-known splicing factor (Del Gatto-Konczak et al. 2000) that might even auto-regulate splicing of its own isoforms (Le Guiner et al. 2001).

In order to determine whether these reconstructed motifs were truly probing the internal logic of RNA-GPS, or simply "lucky" artifacts of the interpretation and/or modeling methods, we ablated occurrences of the motifs we identified by replacing them with "N" bases (Fig. 5A). Making 385 such ablations spanning our test set transcripts, we found a consistent drop in classifier AUROC for all eight localizations. To contextualize the magnitude of this performance drop, we constructed a baseline of 1000 different random sequence ablation experiments, each containing an average of 426 individual sequence ablations with similar properties (e.g., length) to our original motifs. We found that ablating the identified motifs resulted in a more significant performance drop than nearly 96.6% of random ablations based on *t*-statistics (Fig. 5B).



Localization	Significant kmer features	Consensus motifs	Known RBP binding sites
ER membrane	249	61	26
Nuclear lamina	253	76	11
Mito matrix	223	57	13
Cytosol	134	16	16
Nucleolus	130	48	8
Nucleus	172	59	18
Nuclear pore	129	10	0
Outer mito membrane	312	158	52

FIGURE 4. Summary of motif interpretation results. (*A*) Methodology for reconstructing consensus motifs (green) from *k*-mer features that have been flagged as significant (blue). Significant *k*-mer features are first aligned back to a transcript sequence, and neighboring *k*-mers are then computationally "ligated" to create candidate motifs (gray). By doing this for many transcript sequences, we create many candidate motifs (gray), which are collectively used to construct a multiple sequence alignment that lets us isolate conserved consensus sequence motifs (green). (*B*) Table of counts showing the number of significant features and resulting consensus motifs they generate for each localization, as well as the number of known RNA binding protein binding motifs that occur within those consensus motifs. These specific RBPs are then visualized in *C*; RPBs with only one hit are omitted for clarity. Counts in this heatmap represent the number of RBP binding sites found in the motifs driving localization to each compartment. RBPs experimentally implicated in splicing regulation are in bold. Overall, the RBPs we identified significantly enrich the "mRNA splicing, via spliceosome" ontology term with $P = 3.6 \times 10^{-15}$.

Thus, the *P*-value associated with identifying a set of motifs as impactful as our set by chance is $P = 0.034$. This suggests that the specific motifs we isolated significantly influence RNA-GPS's localization predictions, much more so than a randomly chosen set of RBP motifs would.

We used a second, computationally complementary interpretation of RNA-GPS to further validate our findings. Rather than piecing together sequence *k*-mers and annotating them using RBPs' position weighted matrices (PWMs), we directly ablate all known RBP PWMs, observing which RBPs have the largest impact (see Materials and Methods section). The process is conceptually similar to that shown in Figure 5A, except using all RBP PWMs instead of assembled sequence motifs. Doing so, we derived a list of 55 RBPs whose individual ablation resulted in weakened localization signals (Supplemental Fig. S7A). As before, DAVID returns a top hit of "mRNA splicing, via spliceosome" with an associated *P*-value of 5.2×10^{-19}. We see the emphasis on HuR recapitulated here, along with an emphasis on one of aforementioned RBM4's related proteins, RBM5. In addition, there appears to be a strong enrichment for RBPs in the serine and arginine rich splicing factor (SRSF) proteins, especially SRSF1, SRSF10, and SRSF9, as well as for LIN28A, another well-known splicing factor (Yang et al. 2015). These similar splicing-focused results obtained from two different interpretation techniques suggest that RNA-GPS learns to use splicing factor binding sites to inform its predictions.

To verify the generalizability of our observations, we perform the same PWM ablation study on the GRU neural network model. GRU uses the original nucleotide sequence instead of *k*-mers, and thus represents a very different approach to predicting localization. Despite the different features used by RNA-GPS and GRU, ablating the GRU model recapitulates the emphasis on splicing factors, particularly highlighting the HuR and SRSF RBPs, and with DAVID reporting the same top hit of "mRNA splicing, via spliceosome" with $P = 2.1 \times 10^{-31}$ (Supplemental Fig. S7B). While this does not directly evaluate RNA-GPS, this does corroborate the sequence patterns that RNA-GPS appears to be learning, suggesting that this is indeed a signal useful for predicting localization.

We investigated whether the computationally identified splicing motifs from RNA-GPS directly correlate with experimental data on localization. We focused on the intersection of motifs identified by both interpretations of RNA-GPS ($n = 11$, requiring intersected motifs to have been identified with respect to the same transcript part and localization compartment). We used each to stratify the APEX-seq transcripts into two categories: those that contained the motif in the prescribed transcript region (according to the same methodology as in the ablation study), and those that did not. Using a Chi-squared test to compare localization patterns across these groups, we found that in all cases, presence of the motif was significantly associated with increased localization (Supplemental Fig. S8). This further supports the finding that the splicing-based motifs identified by RNA-GPS are correlated with localization, though there are likely to be other determinants.

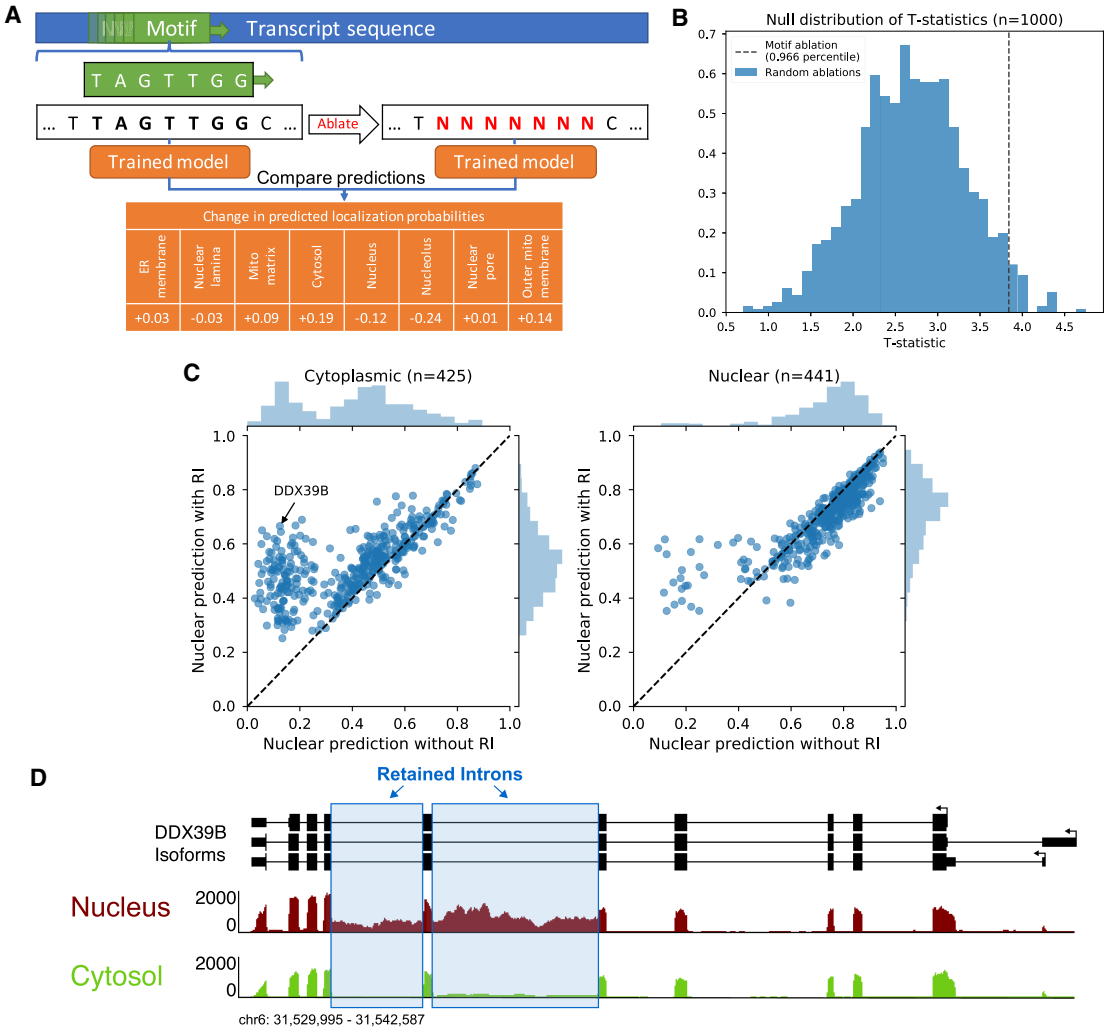


FIGURE 5. Evaluating statistical and biological significance of the motifs identified in Figure 4. (*A*) Shows procedure for taking consensus motifs (from Fig. 4C) and ablating them in silico, which allows us to evaluate how important they are for predicting localization. The motif (green) is scanned across each transcript sequence (blue). Each match is ablated by replacing with "N" bases (red), and we compute the difference in predicted probabilities. We use this methodology to construct *B*, which shows how important the motifs from Figure 4C (black line) are relative to ablating random motifs (blue distribution). We reject the null hypothesis that we are assembling and identifying random motifs with $P = 0.034$; this suggests that the motifs we identified containing splicing-focused RBPs are relevant for predicting localization. To further validate RNA-GPS, we asked whether the nucleus–cytoplasm version of RNA-GPS could distinguish different splice isoforms of the same transcript, specifically those with and without retained intron (RI) sequences. These results are shown in *C*. Each point represents a single transcript's predicted nuclear localization with and without RI; *left* and *right* subplots depict transcripts localizing to the cytoplasm or nucleus according to their RI-free isoform, respectively. Adding RI increases RNA-GPS's predicted nuclear localization, especially for transcripts originally measured to be localized to the cytoplasm (Supplemental Fig. S9, left subplot). (*D*) An example of this behavior. *DDX39B* predominantly localizes to the cytosol, but has an isoform with retained introns (blue boxes) that localizes to the nucleus, which is identified by APEX-seq and correctly predicted by RNA-GPS.

Validating RNA-GPS's predictions with retained intron isoforms

Experimental data on retained intron sequences provides a good opportunity to further test RNA-GPS's predictions. We took $n = 1434$ APEX-seq transcripts with measurable intron retention in the nucleus, using rMATS (Shen et al. 2014) to identify these differential splicing events. We then evaluated the binary nucleus/cytoplasm version of RNA-GPS on the intron-free isoforms of these transcripts, compared to its predictions on the isoforms with retained introns (considered to be part of the CDS for featurization). We found that without retained introns, 69.3% of sequences were predicted to be nuclear, whereas with retained introns, this proportion rose to 80.4%. This increase mirrors experimental observations whereby retained introns correspond to greater nuclear localization (Yap et al. 2012; Braunschweig et al. 2014; Yoshimoto et al. 2017).

Furthermore, this increase in RNA-GPS predicted nuclear localization was most pronounced in transcripts previously predicted and/or measured to be cytoplasmic (Fig. 5C; Supplemental Fig. S9). Transcripts like *DDX39B* (Fig. 5D) show that in the original APEX-seq data, there are substantial differences in splicing patterns for transcripts localizing to different regions of the cell—differences recapitulated by RNA-GPS. Recall that RNA-GPS normalizes its features by transcript length, and thus responds directly to shifts in sequence composition (and not to inflated *k*-mer counts from longer transcripts); this, combined with the fact that RNA-GPS was not exposed to any intronic sequences during its training, makes this analysis a strong experimental validation of the algorithm's inference on the role of splicing for nuclear localization.

DISCUSSION

Here we present RNA-GPS, a computational model capable of predicting, simultaneously, an RNA transcript's localization to eight different subcellular localizations given only sequence-based *k*-mer information. While RNA-GPS provides unprecedented granularity in predicting localization, it also outperforms prior computational models of localization on simpler nuclear/cytoplasm prediction. RNA-GPS's success can largely be attributed to a featurization strategy that leverages natural segmentation of transcripts to encode spatial sequence information without resorting to complex machine learning models that are often relatively difficult to train and interpret. As a result, RNA-GPS sets a new benchmark for both model performance and interpretability. We further show that, within the scope of additional available data sets, RNA-GPS appears to generalize well to other cell lines, and thus may be useful as a tool for predicting localization of yet unquantified transcripts in a variety of cells.

The success of our model also highlights several key aspects regarding the biology of RNA transcript localization within the cell. Notably, our ability to predict localization based solely on *k*-mer features indicates that much of the signal driving localization can be found within transcripts' primary sequence. This echoes findings from prior computational works (Zuckerman and Ulitsky 2019) and indicates some degree of modularity with regard to transcript localization mechanisms in the cell. Furthermore, we showed that within a transcript, localization signals are not evenly distributed across the entire sequence but are rather concentrated about the coding sequence and 3′-UTR segments. Indeed, the relative importance of the 3′-UTR has been documented in numerous studies (Sylvestre et al. 2003; Andreassi and Riccio 2009; Tushev et al. 2018; Ciolli Mattioli et al. 2019). Put more generally, knowing that a motif appears at all in a transcript is not sufficient to predict its localization—where the motif appears is also important.

In addition to predicting localization, we also subsequently applied various interpretation techniques to RNA-GPS as a lens for understanding the precise mechanisms driving localization. Specifically, we find one signal consistently emphasized by RNA-GPS and corroborated by interpretation of auxiliary models: the role of splicing factors in subcellular localization. This phenomenon is further highlighted by the fact that the addition of retained introns to the transcripts substantially increased RNA-GPS's prediction for nuclear localization. This result is consistent with previous experiments, where intron retention has been found to result in nuclear retention of transcripts (Yap et al. 2012; Braunschweig et al. 2014; Yoshimoto et al. 2017). Similarly, Fazal et al. identified cases where different isoforms of the same transcript localized to different regions, which is suggestive of the impact of splicing (Fazal et al. 2019). Transcript splicing has been found to play a direct role in localization of many transcripts, including *ESRP1* (Yang and Carstens 2017) and *DROSHA* (Link et al. 2016), as well. In their computational localization model, Zuckerman and Ulitsky found splicing efficiency (i.e., the relative proportion of transcripts that have been spliced) to be a predictor of nuclear versus cytoplasmic localization (Zuckerman and Ulitsky 2019). Compared to these prior works, our results go further by suggesting that the role of splicing goes well beyond binary nuclear/cytoplasmic localization, influencing transcript targeting to regions as specific as the nuclear lamina for example.

Such a splicing-driven paradigm, if truly responsible for some degree of transcript localization, could have several implications. Firstly, as splicing occurs early on in the process of RNA biogenesis, its importance lends support to the theory that localization is largely determined upon transcript maturation. Furthermore, just as different exons can be alternatively spliced to create different versions of proteins, a similar underlying combinatorial process could allow for highly complex and specific addressing of transcripts using different combinations of address/zip codes. One might even reasonably speculate that increased cellular diversification and increased cytotopic locations in different cell types might necessitate increasingly complex splicing patterns, a correlation that has been noted and reaches its apex in brain tissues (Pan et al. 2008).

In the broader scheme of RNA localization, it is important to keep in mind that the splicing signal we have identified and discussed is likely but one of many pieces of the puzzle. For example, given a localization signal (splicing-related or otherwise), how do transcripts physically transport to the correct localization? A popular notion is that these transcripts are transported along the cytoskeleton by motor protein complexes (Delanoue and Davis 2005; Soundararajan and Bullock 2014). This raises the question of what is actually signaling the recruitment and "addressing" of these complexes to go to the right places; in the presumed context of splicing, might the combination of

splice junctions and retained intron sequences contain such information? Furthermore, since increased RNA transcript length has been observed to affect transcript degradation (and thus their ability to accumulate and apparently localize in a region) (Neymotin et al. 2016), how does varying transcript length induced by differential splicing interact with transcript localization?

While our work makes strides in computational modeling of transcript localization, it is worth pointing out its limitations, as well as additional angles for exploration. Computational works, including this one, cannot comprehensively distill all underlying mechanisms; such model interpretations should be viewed as hypotheses-generation for experimental follow-up and suggestive evidence, not as definitive proof. One of the simplifications we made was to represent each gene with a single, most highly expressed transcript. The primary motivating factor for this was that transcript-level expression quantification was noisy, owing in part to poorly characterized isoforms. This simplification may be hiding valuable sequence information, especially information related to splice variants. Additional featurization strategies, such as the inclusion of secondary structure information, or features regarding chemical modifications to RNA transcripts (Roundtree et al. 2017) may also further improve model performance and shed more light on the mechanisms of localization. Our methodology for assembling *k*-mer features into contiguous motifs is also a point of potential improvement and could further sharpen our findings. An additional layer of complexity is that RNA binding proteins themselves may localize to different areas of the cell, thus influencing their ability to interact with RNA transcripts in pre or post-splicing states. Finally, while our interpretation focuses on identifying short motifs that are correlated with localization, it would also be interesting to investigate larger sequence motifs and interactions across the transcript that could impact localization.

In summary, RNA-GPS is a computational model of transcript localization that, when compared to existing approaches, provides more granular predictions of localization while remaining a relatively simple, intuitive model. We leverage the interpretability of RNA-GPS to probe biological mechanisms underlying transcript localization and implicate splicing as possible mechanism driving localization. While splicing's involvement in transcript localization has been previously studied, we provide new evidence suggesting that its role may be more important than initially understood. We then discuss potential implications of a splicing-driven paradigm of localization. Our work can lead to follow-up studies, both computational and experimental, that further elucidate our understanding of RNA transcript localization. Some of the methods and modeling strategies we presented here may be more generally applicable to other sequence analysis problems.

MATERIALS AND METHODS

Data set

Our primary data set is drawn from the APEX-seq results produced by Fazal et al. (2019). This data set measures localization of 20,852 transcripts at the cytosol, endoplasmic reticulum, mitochondrial matrix, outer mitochondrial membrane, nucleus, nucleolus, nuclear lamina, and nuclear pore. The first four of these localizations are cytoplasmic, whereas the latter four are nuclear (see Fig. 1A for illustration). This data set provides an expression value for each localization/transcript pair, which we then convert to a $\log_2$ fold change score and corresponding adjusted *P*-value quantifying transcript enrichment at that cellular compartment compared to the rest of the cell using DESeq2 (Love et al. 2014) (version 3.9). We define a transcript to be significantly enriched at a cellular compartment if it has a $\log_2$ fold change greater than 0, along with an adjusted *P*-value less than 0.05. Figure 1B lists the number of significantly enriched transcripts at each localization in our full data set; these values are broken down for different data splits in Supplemental Figure S1A. In cases where we have more than one transcript isoform measured for a given gene, we use the most abundant isoform to simplify our modeling. We retain only transcripts with at least one positive localization, which leaves 3660 significantly localized genes/transcripts for our eight-way localization problem. These transcripts are predominantly protein-coding; the exact proportion of transcript types in our data set is shown in Supplemental Figure S1E. Many of these transcripts have more than 1 localization (Supplemental Fig. S1C,D); we thus formulate our machine learning task as a multilabel prediction problem. For our binary cytoplasm versus nucleus classification problem, this same process yields a set of $n = 9155$ APEX-seq transcripts that localize either to the nucleus or cytoplasm, with approximately half the transcripts localizing to each. For processing ENCODE data, we follow an identical process for computing differential expression and identifying significantly localized transcripts, which results in $n = 7641$ transcripts for the HeLa-S3 cell line and $n = 6359$ transcripts for the K562 cell line. For both these cell lines, approximately 25% of the transcripts are also observed in our APEX-seq data set.

We do not include an explicit set of negative examples (i.e., transcripts with no significant localization). This is because it is difficult to definitively say that a transcript does not localize to any compartments due to limitations inherent to the data. Thus, instead of including transcripts with no significant localization to any compartment, we use transcripts with significant localizations to other compartments as negative examples.

Within the APEX-seq data sets, we reserve 10% of the data for testing and 10% for validation, leaving 80% of the examples for training. ENCODE data sets were only used for testing, and thus were not split. As is common practice, we used the validation set to tune modeling approaches and hyperparameters, and the test set to perform a final evaluation and model interpretation. When performing 10-fold cross-validation, we rotated our testing and training folds such that data point was used exactly once as a testing example, and exactly once as a validation example.

Featurization

The basis of our featurization scheme is *k*-mer featurization. Canonically, *k*-mer featurization splits a sequence of length *l*

into subsequences of length k where $k \ll l$, using a stride length of 1 (such that each k-mer subsequence overlaps $k-1$ bases with the previous k-mer). Totaled k-mer counts are then normalized by the total number of k-mer subsequences, such that the sum over the feature vector equals 1. This normalization helps prevent input sequence length from drastically altering k-mer featurization, which is especially important in our case given the large variability in transcript lengths (Supplemental Fig. S1B).

We make several key modifications to the canonical k-mer featurization scheme. Firstly, instead of using a single value of k, we instead featurize using three different values of $k \in [3, 4, 5]$. Second, and perhaps more importantly, instead of featurizing the entire transcript sequence at once, we isolate the 5′ untranslated region (5′ UTR), coding sequence (CDS), and 3′ untranslated region (3′-UTR), and featurize each segment individually. This leverages the biological intuition that the same motifs can have varying functionality based on their location in a transcript. Prior works also suggest that different regions of the transcript play distinct roles in localization, with many citing the 3′-UTR as a particularly key regulator of localization (Sylvestre et al. 2003; Mayr 2018). At the same time, this segment-wise approach helps restore some spatial information regarding where motifs occur in a sequence, information that is usually lost in canonical k-mer featurization approaches. Overall, our featurization strategy produces $4^3 + 4^4 + 4^5 = 1344$ values for each of the aforementioned three transcript regions, for a total of $1344 \times 3 = 4032$ features. A schematic of our featurization strategy is shown in left portion of Figure 1C.

Modeling

We built our model using Python 3.7.3 using the random forest implementation from scikit-learn (version 0.21.3). We use a single random forest model to predict eight probabilities, each corresponding to the predicted likelihood that the transcript localizes to that region (note that these probabilities need not sum to 1, as localization to one compartment does not preclude localization to others). Intuitively, using a single model encourages the model to learn a single "understanding" of the data common to all localizations, hopefully resulting in a more general and biologically meaningful model. Since we have a large feature space, we limited each tree to use only $\sqrt{n}$ features to avoid overfit, where n denotes the number of features. In addition, we performed hyperparameter tuning over the number of estimators, tree criterion, maximum tree depth, and minimum samples per leaf, with the objective of maximizing the area under the receiver operator characteristic (AUROC) on the validation set. For reference, a toy illustration of a random forest is shown in the right portion of Figure 1C. Combined with the aforementioned featurization scheme, this completes RNA-GPS.

Additional reference models

All subsequent models, like RNA-GPS, predict eight probabilities, one corresponding to each localization compartment, without requiring that those probabilities sum to 1. Whenever possible, we include models (or model variants) that also leverage the featurization-by-transcript-segment approach described above, in order to better tease out the impact of modeling decisions versus featurization decisions. The following models are included to contextualize performance of RNA-GPS relative to other approaches.

A biologically intuitive way to understand subcellular localization is to frame it as a hierarchical process: A transcript might first need to "decide" if it is to remain in the nucleus or export to the cytoplasm, before subsequently localizing to compartments within the nucleus or cytoplasm. We use this view of localization to build our naive baseline model. Several works have previously demonstrated the viability of predicting cytoplasmic versus nuclear localization of RNA sequences (Gudenas and Wang 2018; Zuckerman and Ulitsky 2019), so we start by training a random forest to predict nuclear and cytoplasmic localization using the same 4032-dimensional, transcript-segment-aware feature space, and subsequently use a random uniform distribution U(0, 1) to predict further localization within each general localization area. This establishes a "half-guessing" baseline for our models, a performance level that indicates having learned nothing beyond coarse nuclear and cytoplasmic localization.

As an additional tree-based approach, we used boosted trees to model our localization problem. This model utilizes the same featurization as RNA-GPS. Since the implementation of boosted trees provided by XGBoost (Chen and Guestrin 2016) (version 0.82) does not support multilabel predictions, we trained eight independent boosted tree models, one corresponding to each localization. The resulting model provides a reference not only for a different tree methodology, but also for comparing the effectiveness of having eight independent models versus the single-model approach adopted by RNA-GPS.

We also benchmarked several neural network model architectures, all of which were implemented using PyTorch (version 1.0.1). The neural network models we tried can be categorized into fully connected, convolutional, and recurrent approaches. Fully connected networks are, as their name implies, a series of fully connected layers. Convolutional networks learn sets of "receptive fields," which are applied across the input to produce predictions; in the context of sequence analysis, these receptive fields are similar to position weight matrices (PWMs) that are scanned across the sequence. The architecture of convolutional models often requires fixed-size inputs, though strategic usage of pooling layers can enable processing of variable-sized inputs. Recurrent networks are a class of models specifically designed to handle an ordered input of information, whether that is a sentence or a sequence of nucleotide bases. Recurrent networks are designed from the ground up to handle variable sized inputs.

For fully connected models, we adapted the featurization and architecture approaches of deepLncRNA (Gudenas and Wang 2018). This architecture does not make any design choices specific to lncRNA, and thus should generalize well beyond long noncoding RNA transcripts. Specifically, we apply the same k-mer featurization they described with $k \in [2, 3, 4, 5]$ combined with an RBP PWM hit count featurization using the set of RBP PWMs described by Ray et al. (2013), creating a total feature space of 1553 features (versus 1582 for deepLncRNA itself). We exclude deepLncRNA's features describing genomic position, as some of these labels are highly correlated with our transcript localizations. We applied the same training procedure as described in the original paper: using the same dropout rates, using stochastic gradient descent, coupled with L1 and L2 regularization, with hyperparameters tuned to optimize accuracy on the validation set.

For convolutional models, we adopt the Basset architecture, as it has seen much success in understanding sequence motifs in DNA (Kelley et al. 2016) and theoretically is capable of learning nucleotide sequence motifs in general. The vanilla Basset architecture requires a fixed-length input of one-hot encoded nucleotide bases, and we give it the trailing 1000 bp on the 3′ end of each transcript sequence. We also develop several in-house variants of this architecture. Just as we designed RNA-GPS to leverage information across the 5′ UTR, CDS, and 3′ UTR, we likewise developed a derivative of Basset that consumes the trailing 250 bp of those same three transcript segments, for a total of $3 \times 250 = 750$ input bases; we call this approach Basset-3. We also modified Basset to process arbitrary-length input sequences by introducing a pooling layer to aggregate information across the entire sequence, and by increasing the number of kernels learned by the model to compensate for the resulting loss in positional information; we call this approach Basset-pool. We trained all of our convolutional models using the Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.001, using binary cross entropy as our loss function. Vanilla Basset and Basset-3 were trained using minibatches of 64 examples, while Basset-pool was trained using single examples.

For recurrent models, we experimented with long short-term memory (LSTM) and gated recurrent unit (GRU) style architectures, both of which consume raw nucleotide sequences of variable length. Both LSTMs and GRUs feature a "gating" mechanism that helps networks more easily handle long inputs (Chung et al. 2014). For the LSTM, we use an embedding layer of 32 dimensions, followed by a LSTM layer with 64 dimensions, followed by a fully connected layer mapping the 64 dimensions to our eight-dimensional output space. For the GRU, we use an embedding layer of 32 dimensions, followed by two GRU layers with 64 hidden dimensions, followed by a fully connected layer and sigmoid activation to map the 64 hidden dimensions to our eight-dimensional output space. Our recurrent models were trained using stochastic gradient descent with a learning rate of 0.001, using binary cross entropy as our loss function. No batching was used in training for either architecture, as we found that batching resulted in inferior performance on the validation set.

Finally, we reimplemented RNATracker (Yan et al. 2019), a prior work that merges convolutional and recurrent LSTM layers, along with an attention layer. We closely followed the authors' Keras-based implementation to recreate, in PyTorch, the "canonical" RNATracker network architecture with a fixed-size input of 4000 bases from the 3′ end of the input sequence (shorter sequences are 0-padded), but used a sigmoid activation for our final output rather than a softmax activation, as this better matched our multilabel classification problem. Similarly, while the original authors used Kullback–Leibler (KL) divergence for their loss function, we use binary cross-entropy to fit our multilabel context. We also incorporated RNATracker's featurization scheme, which encodes "N" bases as a vector of [0.25, 0.25, 0.25, 0.25] rather than a vector of [0, 0, 0, 0].

Model interpretation via feature importance

We performed feature ablation studies to evaluate feature importance for tree-based models. Using the test set, we take each feature, shuffle its values across examples, and evaluate the difference in AUROC for each localization. Shuffling preserves the distribution of each feature's values but removes the correct correspondence between examples and that feature's value. We repeat this procedure 15 times to calculate a mean, standard deviation, and *z*-score for each feature's impact to each localization's AUROC. We consider *z*-scores $z \leq -2$ to denote significant features, that is, features that carry significant localization signals whose loss has a deleterious effect on the model's ability to correctly predict localization.

Given a set of "significant" *k*-mer features, we then tile these short sequences back to each transcript, linking together consecutive (i.e., being separated by 1 bp or less) *k*-mers together to form longer sequences, which we call candidate motifs. Recall that we featurize the 5′ UTR, CDS, and 3′-UTR separately; tiling and linking together features is done separately as well, where a *k*-mer found to be "significant" for the CDS is only tiled against CDS regions. We then use these candidate motifs to construct a multiple sequence alignment (using ClustalO version 1.2.4 [Sievers et al. 2011]), and retain only subsequences consistently found across ≥3 candidate motifs and are at least 7 bp long. These conceptually represent motifs that are consistently observed across multiple examples. Figure 4A illustrates this process. Finally, we use TomTom (version 5.0.2, [Gupta et al. 2007]) to annotate these motifs with known RBP binding sites, using a database of 102 such RBP PWM matrices (Ray et al. 2013). Note that a single motif may contain multiple RBP binding sites.

We can then evaluate the importance of these sequence motifs by ablating them and observing the difference in localization predictions. This ablation is done by replacing exact matches to the motif with "N" bases and observing the difference in model output, as illustrated in Figure 5A. As we previously discussed, a lack of significant localization does not necessarily indicate lack of localization; thus, we only track differences in true positive localizations when performing these ablations in order to obtain a cleaner signal.

Model interpretation via motif ablation

We also interpret models by ablating RBP PWMs, using the same set of PWMs as we did for annotation (Ray et al. 2013). Note that while the previously discussed motifs are assembled from *k*-mers, and do not account for positional variation, PWMs are precomputed and do account for positional variation. The process for PWM ablation is very similar to that of motif ablation: We scan each PWM across each transcript, ablate high-scoring PWM hits, and observe the impact to model prediction. To find matches between the PWM matrix of probabilities and a given transcript, we adapt the methodology proposed by Gudenas and Wang in their DeepLncRNA model (Gudenas and Wang 2018). Specifically, we calculate a background base distribution for that transcript and normalize the PWM matrix by this distribution. We then take the $\log_2$-transform of the normalized matrix and calculate the maximum possible score *s* of this log-transformed PWM by summing, across columns, the largest value in each column (which correspond to positions). We then "slide" the log-transformed PWM across the query sequence with a stride length of 1, flagging any hits of greater than $0.9 \times s$ (Gudenas and Wang use a cutoff of 0.8 instead, which we elevate to increase specificity of the hits). We ablate these hits by replacing the corresponding

bases with "N" bases and run the ablated transcript sequence through the model (refeaturizing if necessary), observing the difference in model output. After performing these PWM ablations, the question remains: Which ones have a significant effect? To this end, we use cutoffs to identify PWMs that either elicit a large drop in probabilities, or a consistent drop, or both; namely, we flag PWM motifs that either cause at least an average reduction of 0.01 in predicted probabilities, or cause a nonzero drop in predicted probabilities for at least 72/366 (~20%) of test examples.

Metrics and plotting

Metrics were generated using functions available in scikit-learn, and plots were generated using a combination of matplotlib (version 3.0.3) and seaborn (version 0.9.0). *P*-value correction for multiple hypothesis testing was performed using the statsmodels package (version 0.11.1). We primarily evaluated models based on the area under the receiver operator characteristic (AUROC), area under the precision recall curve (AUPRC), and accuracy. When reporting overall (i.e., not per-class) AUROC and AUPRC values, we average across the per-class performance metric for each class, so as not to downplay the impact of less common localizations (that are likely to be more difficult to predict). When reporting test set accuracy, we use the validation set to determine a threshold for positive/negative predictions that maximizes Youden's J-statistic and use those thresholds to determine accuracy on the test set. Conceptually, this method finds the optimal cutoff where the true positive rate is high, and the false positive rate is low.

DATA DEPOSITION

All APEX-Seq data used to train and evaluate RNA-GPS is available through the Gene Expression Omnibus (GEO) under accession GSE116008. Additional data used to validate nuclear versus cytoplasmic localization is available from ENCODE. The RNA-GPS software is available at https://github.com/wukevin/rnagps. Some preprocessed files (such as those pertaining to retained introns) are available on GitHub as well.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

COMPETING INTEREST STATEMENT

H.Y.C. is affiliated with Accent Therapeutics, Boundless Bio, 10× Genomics, Arsenal Bio, and Spring Discovery. K.R.P. is a consultant for Maze Therapeutics.

ACKNOWLEDGMENTS

We thank the members of the Chang and Zou laboratories for helpful discussions. F.M.F. is supported by the Arnold O. Beckman postdoctoral fellowship and by a National Institutes of Health (NIH) K99/R00 award from NHGRI (HG010910). J.Z. is supported by National Science Foundation (NSF) CCF 1763191, NIH R21 MD012867-01, NIH P30AG059307, and grants from the Silicon Valley Foundation and the Chan-Zuckerberg Initiative. H.Y.C. is supported by RM1-HG007735 and R01-HG004361. H.Y.C. is an Investigator of the Howard Hughes Medical Institute.

Author contributions: H.C. and J.Z. conceived and supervised the project. K.P. and F.F. gathered and preprocessed data, including determining significant localization and defining retained intron sequences. K.W. performed data analysis, modeling, and model interpretation with input from all authors. K.W. wrote the manuscript with input from all authors.

Received November 26, 2019; accepted March 19, 2020.

REFERENCES

Akaike Y, Masuda K, Kuwano Y, Nishida K, Kajita K, Kurokawa K, Satake Y, Shoda K, Imoto I, Rokutan K. 2014. HuR regulates alternative splicing of the TRA2β gene in human colon cancer cells under oxidative stress. *Mol Cell Biol* **34:** 2857–2873. doi:10.1128/MCB.00333-14

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410. doi:10.1016/S0022-2836(05)80360-2

Andreassi C, Riccio A. 2009. To localize or not to localize: mRNA fate is in 3'UTR ends. *Trends Cell Biol* **19:** 465–474. doi:10.1016/J.TCB.2009.06.001

Benoit Bouvrette LP, Cody NAL, Bergalet J, Lefebvre FA, Diot C, Wang X, Blanchette M, Lécuyer E. 2018. CeFra- seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in *Drosophila* and human cells. *RNA* **24:** 98–113. doi:10.1261/rna.063172.117

Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. 2014. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* **24:** 1774–1786. doi:10.1101/gr.177790.114

Buxbaum AR, Haimovich G, Singer RH. 2015. In the right place at the right time: visualizing and understanding mRNA localization. *Nat Rev Mol Cell Biol* **16:** 95–109. doi:10.1038/nrm3918

Cao Z, Pan X, Yang Y, Huang Y, Shen HB. 2018. The lncLocator: a subcellular localization predictor for long non- coding RNAs based on a stacked ensemble classifier. *Bioinformatics* **34:** 2185–2194. doi:10.1093/bioinformatics/bty085

Capriotti E, Marti-Renom MA. 2010. Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics* **11:** 322. doi:10.1186/1471-2105-11-322

Chang HL, Lin JC. 2019. SRSF1 and RBM4 differentially modulate the oncogenic effect of HIF-1α in lung cancer cells through alternative splicing mechanism. *Biochim Biophys Acta* **1866:** 118550. doi:10.1016/J.BBAMCR.2019.118550

Chen T, Guestrin C. 2016. *XGBoost: a scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 785–794. ACM, San Francisco.

Chin A, Lécuyer E. 2017. RNA localization: making its way to the center stage. *Biochim Biophys Acta* **1861:** 2956–2970. doi:10.1016/J.BBAGEN.2017.06.011

Chung J, Gulcehre C, Cho K, Bengio Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv 1412.3555.

Ciolli Mattioli C, Rom A, Franke V, Imami K, Arrey G, Terne M, Woehler A, Akalin A, Ulitsky I, Chekulaeva M. 2019. Alternative 3' UTRs direct localization of functionally diverse protein isoforms

in neuronal compartments. *Nucleic Acids Res* **47:** 2560. doi:10.1093/NAR/GKY1270

Cooper TA, Wan L, Dreyfuss G. 2009. RNA and disease. *Cell* **136:** 777–793. doi:10.1016/J.CELL.2009.02.011

Del Gatto-Konczak F, Bourgeois CF, Le Guiner C, Kister L, Gesnel MC, Stévenin J, Breathnach R. 2000. The RNA-binding protein TIA-1 is a novel mammalian splicing regulator acting through intron sequences adjacent to a 5′ splice site. *Mol Cell Biol* **20:** 6287–6299. doi:10.1128/mcb.20.17.6287-6299.2000

Delanoue R, Davis I. 2005. Dynein anchors its mRNA cargo after apical transport in the *Drosophila* blastoderm embryo. *Cell* **122:** 97–106. doi:10.1016/J.CELL.2005.04.033

Doller A, Akool ES, Huwiler A, Müller R, Radeke HH, Pfeilschifter J, Eberhardt W. 2008. Posttranslational modification of the AU-rich element binding protein HuR by protein kinase Cδ elicits angiotensin II-induced stabilization and nuclear export of cyclooxygenase 2 mRNA. *Mol Cell Biol* **28:** 2608–2625. doi:10.1128/MCB.01530-07

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74. doi:10.1038/nature11247

Fazal FM, Han S, Parker KR, Kaewsapsak P, Xu J, Boettiger AN, Chang HY, Ting AY. 2019. Atlas of subcellular RNA localization revealed by APEX-seq. *Cell* **178:** 473–490.e26. doi:10.1016/J.CELL.2019.05.027

Ghorbani A, Abid A, Zou J. 2019. Interpretation of neural networks is fragile. *Proc AAAI Conf Artif Intell* **33:** 3681–3688. doi:10.1609/aaai.v33i01.33013681

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. 2019. *Explaining explanations: an overview of interpretability of machine learning*. Proceedings, 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018, pp, 80–89. IEEE, NY.

Gudenas BL, Wang L. 2018. Prediction of lncRNA subcellular localization with deep learning from sequence features. *Sci Rep* **8:** 16385. doi:10.1038/s41598-018-34708-w

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble W. 2007. Quantifying similarity between motifs. *Genome Biol* **8:** R24. doi:10.1186/gb-2007-8-2-r24

Huang DW, Sherman BT, Lempicki RA. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37:** 1–13. doi:10.1093/nar/gkn923

Huang DW, Sherman BT, Lempicki RA. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4:** 44–57. doi:10.1038/nprot.2008.211

Izquierdo JM. 2008. Hu antigen R (HuR) functions as an alternative pre-mRNA splicing regulator of Fas apoptosis- promoting receptor on exon definition. *J Biol Chem* **283:** 19077–19084. doi:10.1074/jbc.M800017200

Jung H, Yoon BC, Holt CE. 2012. Axonal mRNA localization and local protein synthesis in nervous system assembly, maintenance and repair. *Nat Rev Neurosci* **13:** 308–324. doi:10.1038/nrn3210

Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26:** 990–999. doi:10.1101/gr.200535.115

Kingma DP, Ba J. 2014. Adam: a method for stochastic optimization. arXiv 1412.6980.

Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, Hughes TR, Tomancak P, Krause HM. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* **131:** 174–187. doi:10.1016/J.CELL.2007.08.003

Le Guiner C, Lejeune F, Galiana D, Kister L, Breathnach R, Stévenin J, Del Gatto-Konczak F. 2001. TIA-1 and TIAR activate splicing of alternative exons with weak 5′ splice sites followed by a U-rich stretch on their own pre-mRNAs. *J Biol Chem* **276:** 40638–40646. doi:10.1074/jbc.M105642200

Link S, Grund SE, Diederichs S. 2016. Alternative splicing affects the subcellular localization of Drosha. *Nucleic Acids Res* **44:** 5330–5343. doi:10.1093/nar/gkw400

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/s13059-014-0550-8

Lubelsky Y, Ulitsky I. 2018. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555:** 107–111. doi:10.1038/nature25757

Martin KC, Ephrussi A. 2009. mRNA localization: gene expression in the spatial dimension. *Cell* **136:** 719–730. doi:10.1016/J.CELL.2009.01.044

Mayr C. 2018. What are 3′ UTRs doing? *Cold Spring Harb Perspect Biol* **11:** a034728. doi:10.1101/cshperspect.a034728

Neymotin B, Ettorre V, Gresham D. 2016. Multiple transcript properties related to translation affect mRNA degradation rates in *Saccharomyces cerevisiae*. *G3 (Bethesda)* **6:** 3475–3483. doi:10.1534/g3.116.032276

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40:** 1413–1415. doi:10.1038/ng.259

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499:** 172–177. doi:10.1038/nature12311

Roundtree IA, Evans ME, Pan T, He C. 2017. Dynamic RNA modifications in gene expression regulation. *Cell* **169:** 1187–1200. doi:10.1016/j.cell.2017.05.045

Ryder PV, Lerit DA. 2018. RNA localization regulates diverse and dynamic cellular processes. *Traffic* **19:** 496–502. doi:10.1111/tra.12571

Shen S, Park JW, Zx L, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. 2014. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-seq data. *Proc Natl Acad Sci* **111:** E5593–E5601. doi:10.1073/pnas.1419161111

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7:** 539. doi:10.1038/msb.2011.75

Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, Fugmann T, Wong KK, Van Allen EM. 2018. Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol* **36:** 1056–1058. doi:10.1038/nbt.4239

Soundararajan HC, Bullock SL. 2014. The influence of dynein processivity control, MAPs, and microtubule ends on directional movement of a localising mRNA. *Elife* **3:** e01596. doi:10.7554/eLife.01596

Sylvestre J, Margeot A, Jacq C, Dujardin G, Corral-Debrinski M. 2003. The role of the 3′ untranslated region in mRNA sorting to the vicinity of mitochondria is conserved from yeast to human cells. *Mol Biol Cell* **14:** 3848–3856. doi:10.1091/mbc.e03-02-0074

Tran H, Maurer F, Nagamine Y. 2003. Stabilization of urokinase and urokinase receptor mRNAs by HuR is linked to its cytoplasmic accumulation induced by activated mitogen-activated protein kinase-activated protein kinase 2. *Mol Cell Biol* **23:** 7177–7188. doi:10.1128/mcb.23.20.7177-7188.2003

Tushev G, Glock C, Heumüller M, Biever A, Jovanovic M, Schuman EM. 2018. Alternative 3′ UTRs modify the localization, regulatory potential, stability, and plasticity of mRNAs in neuronal

compartments. *Neuron* **98:** 495–511.e6. doi:10.1016/J.NEURON.2018.03.030

Vuong JK, Lin CH, Zhang M, Chen L, Black DL, Zheng S. 2016. PTBP1 and PTBP2 serve both specific and redundant functions in neuronal pre-mRNA splicing. *Cell Rep* **17:** 2766–2775. doi:10.1016/J.CELREP.2016.11.034

Yan Z, Lécuyer E, Blanchette M. 2019. Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics* **35:** i333–i342. doi:10.1093/bioinformatics/btz337

Yang Y, Carstens RP. 2017. Alternative splicing regulates distinct subcellular localization of Epithelial splicing regulatory protein 1 (Esrp1) isoforms. *Sci Rep* **7:** 3848. doi:10.1038/s41598-017-03180-3

Yang J, Bennett BD, Luo S, Inoue K, Grimm SA, Schroth GP, Bushel PR, Kinyamu HK, Archer TK. 2015. LIN28A modulates splicing and gene expression programs in breast cancer cells. *Mol Cell Biol* **35:** 3225–3243. doi:10.1128/MCB.00426-15

Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev EV. 2012. Coordinated regulation of neuronal mRNA steady- state levels through developmentally controlled intron retention. *Genes Dev* **26:** 1209–1223. doi:10.1101/gad.188037.112

Yoshimoto R, Kaida D, Furuno M, Burroughs AM, Noma S, Suzuki H, Kawamura Y, Hayashizaki Y, Mayeda A, Yoshida M. 2017. Global analysis of pre-mRNA subcellular localization following splicing inhibition by spliceostatin A. *RNA* **23:** 47–57. doi:10.1261/rna.058065.116

Zhang B, Gunawardane L, Niazi F, Jahanbani F, Chen X, Valadkhan S. 2014. A novel RNA motif mediates the strict nuclear localization of a long noncoding RNA. *Mol Cell Biol* **34:** 2318–2329. doi:10.1128/MCB.01673-13

Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. 2019. A primer on deep learning in genomics. *Nat Genet* **51:** 12–18. doi:10.1038/s41588-018-0295-5

Zuckerman B, Ulitsky I. 2019. Predictive models of subcellular localization of long RNAs. *RNA* **25:** 557–572. doi:10.1261/rna.068288.118