Predicting dynamic cellular protein-RNA interactions using deep learning and *in vivo* RNA structure

Lei Sun^{1,4}, Kui Xu^{1,4}, Wenzhe Huang^{1,4}, Yucheng T. Yang^{2,3,4}, Lei Tang¹, Tuanlin Xiong¹, Qiangfeng Cliff Zhang^{1,5,*}

¹ MOE Key Laboratory of Bioinformatics, Beijing Advanced Innovation Center for Structural Biology, Center for Synthetic and Systems Biology, Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, 100084, Beijing, China

² Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

³ Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

⁴ Co-first author

⁵ Lead Contact

* Correspondence: qc Zhang@tsinghua.edu.cn (Q.C.Z.)

Abstract

Interactions with RNA-binding proteins (RBPs) are crucial for RNA regulation and function. While both RNA sequence and structure are critical determinants, RNA structure is dependent on cellular environment and especially important in regulating dynamic RBP bindings across various conditions. However, how distinct it contributes to RBP binding *in vivo* remains poorly understood. To address this issue, we obtained transcriptome-wide RNA secondary structure profiles in multiple cell-types, and established a deep neural network, PrismNet, that uses *in vivo* RNA structures to accurately predict cellular protein-RNA interactions. With a deep learning “attention” strategy, PrismNet discovers the exact binding nucleotides and their mutational effect. The predicted binding sites are highly conserved and enriched for rare, deleterious genetic variants. Remarkably, dynamic RBP binding sites are enriched for structure-changing variants (riboSNitches), which are often associated with disease, reflecting dysregulated RBP bindings. Our resource enables the analysis of cell-type-specific RNA regulation, with applications in human disease.

Highlights

- 1, A big data resource of transcriptome-wide RNA secondary structure profiles in multiple cell types
- 2, PrismNet, a deep neural network, accurately models the sequence and structural combined patterns of protein-RNA interactions *in vivo*
- 3, RNA structural information *in vivo* is critical for the accurate prediction of dynamic RBP binding in various cellular conditions
- 4, PrismNet can dissect and predict how mutations affect RBP binding via RNA sequence or structure changes
- 5, RNA structure-changing RiboSNitches are enriched in dynamic RBP binding sites and often associated with disease, likely disrupting RBP-based regulation

Introduction

RNA binding proteins (RBPs) play essential roles in regulating the transcription, metabolism, and translation of cellular RNAs (Baltz et al., 2012; Castello et al., 2012; Gerstberger et al., 2014; Licatalosi and Darnell, 2010). Determining RBP binding profiles in different conditions and elucidating their detailed regulatory mechanisms are critical for understanding their functions. However, given the sheer number of RBPs that accounts for close to 10% of the human proteome (Brannan et al., 2016; Hentze et al., 2018), establishing links between RBPs and their targets has been an enormous challenge. To address this question, many high-throughput technologies have been developed to profile and predict RBP binding. Assays such as systematic evolution of ligands by exponential selection (SELEX), RNAcompete, and RNA Bind-n-Seq can characterize the sequence preferences of RBPs *in vitro* (Ellington and Szostak, 1990; Lambert et al., 2014; Ray et al., 2009), and methods like RNA immunoprecipitation (RIP) and UV crosslinking followed by immunoprecipitation (CLIP) and sequencing can identify RBP binding sites *in vivo* (Gilbert and Svejstrup, 2006; Hafner et al., 2010; Licatalosi et al., 2008; Van Nostrand et al., 2016). Nevertheless, despite extensive mapping of RBP binding, the underlying characteristics remains largely unclear, due to the complexities of high non-specific background, interactions with other RBPs and/or cofactors, and existence of alternative RNA secondary structures of the RNA targets (Darnell et al., 2005; Dominguez et al., 2018; Sun et al., 2019).

In addition to the methods of direct measuring, other approaches have been developed to model and predict RBP binding. Traditionally, position-weight-matrices have been used to describe RBP binding determinants and to predict RBP binding targets from RNA sequences (Bailey et al., 2009). Machine learning methods that integrate different types of information also have been developed to more accurately characterizes the binding pattern of RBPs (Li et al., 2010; Maticzka et al., 2014;

Orenstein et al., 2016b). More recently, deep learning (LeCun et al., 2015) approaches have been successfully applied to model protein-RNA interactions and predict RBP binding sites (Alipanahi et al., 2015; Ben-Bassat et al., 2018; Gandhi et al., 2018; Ghanbari and Ohler, 2020; Koo et al., 2018; Zhang et al., 2016). For example, DeepBind was developed to learn RBP binding preferences from RNAcompete data using a deep neural network (Alipanahi et al., 2015).

Although these learning methods successfully capture RBP binding preferences of primary sequence, their prediction accuracy under different physiological states is limited because the RNA sequence is not dependent on *in vivo* conditions. Over the years, several methods have been developed to include RNA structural features of RBP targets in their modeling, but these structures were based on computational prediction rather than *in vivo* analysis (Li et al., 2010; Maticzka et al., 2014; Orenstein et al., 2016b; Zhang et al., 2016). Although RNA structure can be predicted from sequence with some accuracy (Eddy, 2014; Seetin and Mathews, 2012), the predictions do not reflect the dynamic regulations by cellular trans-factors and usually show substantial differences to *in vivo* structures (Rouskin et al., 2014; Spitale et al., 2015). Thus, *in vivo* RNA structure data are essential for accurate modeling and predictions of protein-RNA interactions in physiologically relevant contexts.

Here, we bridge this knowledge gap by determining transcriptome-wide RNA secondary structures in multiple cell types. We then integrate this experimentally-derived information in the construction of a deep discriminative neural network (PrismNet, Protein-RNA Interaction by Structure-informed Modeling using deep neural NETwork) that accurately models and predicts RBP targets *in vivo*. We apply PrismNet to predict how genomic variants in affect RBP binding, especially in the context of human diseases. Specifically, we focused on single nucleotide variants that disrupt RNA structure (riboSNitches) and often associated with human disease, and

discovered that riboSNitches are enriched in dynamic, cell-type-specific RBP binding sites.

Results

RNA structuromes in different cell types reveal prevalence of structurally variable sites and their association with dynamic RBP binding

RNA structure is flexible, and this feature plays an instrumental role in determining the varying protein-RNA interactions in different cellular conditions (Gerstberger et al., 2014; Jankowsky and Harris, 2015; Lewis et al., 2017). To elucidate the global relationship between RNA structure and RBP binding, we used *in vivo* click selective 2'-hydroxyl acylation and profiling experiment (icSHAPE) technology (Spitale et al., 2015) to generated a comprehensive resource of RNA secondary structures in seven cell types: K562, HepG2, HEK293, HEK 293T, HeLa, H9 and mES cells (Figures 1A, S1A). These cell lines were selected based on their availability of rich RBP binding data by CLIP-seq (Figure S1A), which allowed for integrative structural-and-interaction modeling of RBPs in a matched cellular context.

On average, we obtained at least 200 million usable reads for each library replicate after quality control, totaling 4.4 billion reads. We determined RNA secondary structures of the transcripts using icSHAPE-pipe (Li et al., 2019). Our data achieved high coverage of the global transcriptomes (>50,000 transcripts in human; >30,000 transcripts in mouse) as well as high quality (RPKM Pearson correlation coefficient >0.97 between replicates) (Figures S1A-B). For example, our icSHAPE profiling data on 18S rRNA from different human cell lines were highly consistent (Figure S1C) and agreed well with known 18S secondary structures from crystal structures (Figures S1D-E).

Previously, we found that although RNA structure is relatively stable across different subcellular locations, it also contains many structurally variable sites (Sun et al., 2019). These structurally variable sites are hotspots for post-transcriptional regulation, including RBP binding and RNA modification. We found that this is also true when comparing RNA structures across different cell lines, i.e., most of the RNA structures are stable across all cell lines tested, but they also contain a substantial fraction of regions that display structural variability (Figure 1B and S2A-B).

RBP binding could be affected by the diverse cellular environments and thus expected to be dynamic across cell types. We re-analyzed available enhanced CLIP (eCLIP) data and indeed observed very different binding profiles for the same RBPs in different cell lines. For example, on average, only anywhere between ~15% and ~60% of the binding sites are shared in K562 and HepG2 cells (Figure S2C). Importantly, we found a high level of association between these dynamic RBP binding sites and the RNA structurally variable sites between the two cell types (Figure 1B, Figure S2D). As an example, HNRNPM is known to bind poly-U sites with single-stranded structure (Van Nostrand et al., 2016), exemplified by the binding sites in the *LSS* and *FAH* transcripts in K562 cells (Figure 1C and S2E). Accordingly, these sites lose HNRNPM binding when switching from the single-stranded to a more double-stranded conformation in HepG2 cells. Overall, these data suggest that RNA structure could play an essential role in determining the dynamic RBP binding in diverse cellular conditions. It is thus critical to incorporate the *in vivo* RNA structural information into platforms that model and predict RBP binding - and their changes - across diverse cellular conditions.

PrismNet accurately predicts RBP binding *in vivo* using deep learning and RNA structural data

We constructed PrismNet, a deep neural network to accurately model and predict RBP binding, by integrating the *in vivo* RNA secondary structure profiles that we generated with aggregated big data of RBP binding sites. In total, we included 60 RBPs from POSTAR (Hu et al., 2017), 22 RBPs from starBase (Yang et al., 2011), as well as 59 RBPs from ENCODE (Van Nostrand et al., 2016) (Figure S1). Importantly, in contrast to previous methods that have only considered RNA sequences, or relied on computationally predicted RNA structures, PrismNet was constructed to learn protein-RNA interaction determinants from both RNA sequence and *in vivo* structure simultaneously. This approach proved to be vital for capturing the complex interplay between cell type-specific changes in structures and interactions *in vivo*.

For each RBP and an available CLIP experiment, PrismNet trained a model that evaluates every position within the binding sites and calculated a score that reflects the binding contribution of that nucleotide, using the 5,000 most confident peaks from CLIP. As part of the input data, the icSHAPE structure scores of each nucleotide in the same cell type of the CLIP experiment were encoded as a one-dimensional vector, together with the four-dimensional one-hot-encoded sequence (Figure 2A). The PrismNet architecture uses a convolutional layer, a two-dimensional residual block (He et al., 2016) and a one-dimensional residual block connected by max pooling to capture sequence and structural determinants spanning large distances in transcripts. A squeeze-and-excitation (SE) module (Hu et al., 2019) is applied to adaptively recalibrate convolutional channels for learning channel-wise attention (Figure S3A). To mitigate potential overfitting of the PrismNet, we added dropout (Srivastava et al., 2014) layers after every residual block, and other regularizers including weight decay (Hanson and Pratt, 1989) and early stopping in the training stage. Importantly, we also applied SmoothGrad (Wattenberg, 2017) to enable the enhanced saliency maps (Zisserman, 2014) for the visualization and identification of high attention regions

(HARs), which are predicted to be the exact locations of RBP binding nucleotides (Figure 2A).

To demonstrate that PrismNet can accurately model the sequence and structural basis of RBP binding, we performed a proof-of-principle analysis with the RBP IGF2BP1, which plays important roles in regulating RNA stability and localized translation (Huang et al., 2018). Recently, eCLIP uncovered 46,226 IGF2BP1 binding sites in K562 cells (Van Nostrand et al., 2016). We trained PrismNet with this dataset and our icSHAPE data also in K562 cells. We then applied this model to predict the IGF2BP1 binding sites in HepG2 cells, using their corresponding HepG2 icSHAPE data. The predictions were then compared to the HepG2 eCLIP results from the same group, as ground truth. According to eCLIP, the EIF3F transcript contains 14 IGF2BP1 binding sites in K562 cells, and 10 binding sites in HepG2 cells. We found that PrismNet correctly predicted all 10 binding sites within the EIF3F transcript in HepG2 cells with no false positives, by using the model trained in K562 cells. In contrast, DeepBind (Alipanahi et al., 2015), a deep learning model taking sequence only as input, correctly predicted only 2 of the 10 sites (Figure 2B). Interestingly, if we did not include RNA secondary structure data in the training of PrismNet, the sequence-only version correctly predicted only 7 of the 10 sites, indicating that RNA structure information is important for accurate binding site prediction by PrismNet.

We then systematically evaluated the prediction performance of PrismNet by using transcriptome-wide binding sites of all 99 RBPs, and comparing with other state-of-the-art computational methods, including RCK (Orenstein et al., 2016b), GraphProt (Maticzka et al., 2014) and DeepBind (Alipanahi et al., 2015). As mentioned above, DeepBind is the first deep learning model to predict RBP binding sites utilizing only RNA sequences; GraphProt models RBP binding sites based on sequence and predicted structure with graph-kernel features; RCK infers protein-RNA binding

preferences using a k -mer based model with RNA sequences and predicted structure. For every CLIP-seq dataset of an RBP, we split the binding sites into a training and a test set and use the same sets to benchmark all prediction methods. We observed that, overall, PrismNet achieved the highest prediction performance in terms of Area Under the receiver operating characteristic Curve (AUC) (Figure 2C). For most RBPs, PrismNet constantly showed superior prediction performance (Figures S3B-C). Some RBPs show dramatic performance improvement by PrismNet (e.g., SLTM: PrismNet = 0.82 versus GraphPort = 0.61; XRN2: PrismNet = 0.87 versus GraphPort = 0.62; DGCR8: PrismNet = 0.89 versus GraphPort = 0.63).

RNA secondary structure *in vivo* is critical input for accurate prediction of RBP binding

To dissect how RNA sequence and structural information contribute to the accurate predictions of PrismNet, we trained the PrismNet model using different combinations of input data: (i) sequence and experimentally-measured structure, (ii) sequence only, (iii) experimentally-measured structure only, (iv) sequence and structure predicted by RNAshape (Steffen et al., 2006), and (v) sequence and randomly generated structure. As expected, the model with sequence and experimentally-measured structure (i) as input significantly outperformed other models (Figures 2D and S3D). Notably, PrismNet achieved better performance on almost all RBPs over predictions based on sequence and computationally predicted structures (iv) (Figure 2E). It is also interesting to note that PrismNet could fairly accurately predict protein-RNA binding sites using RNA secondary structure data only (iii), although the prediction accuracy was inferior to that only using sequence data only (ii).

Unexpectedly, we observed comparable prediction performance for the three models that use sequence data only (ii), or sequence data with predicted structures (iv) and sequence data with randomly generated structures (v) (Figures 2D and S3D).

A recent study showed that predicted structures did not improve the prediction performance of protein-RNA binding if the deep learning models were appropriately designed (Koo et al., 2018). This finding is consistent with our results: a good deep neural network can implicitly retrieve and use sequence-embedded RNA structure information, just like independent predictions. Fledging separately predicted RNA structures thus cannot further improve the prediction (Figure S3E). However, it was surprising that training PrismNet with randomly generated RNA structures did not lead to a visible deterioration of prediction performance, implicating that PrismNet is robust to the noise in the input RNA structural data (Figure 2D).

Interestingly, the prediction improvement brought by experimentally-measured *in vivo* structural data was more significant for those RBPs with lower performance in the sequence-only model (Figure 2F). The level of improvement by the provided RNA structural information was associated with the type of RNA binding domain in the RBPs (Figure 2F and S3F). For instance, RBPs containing single-stranded RNA (ssRNA)-binding domains, such as the RNA recognition motif (RRM) and the K homology (KH) domains, were more dependent on RNA sequence for target recognition, and the prediction improvement from RNA structure was less when comparing to other domains. On the other hand, RBPs containing a double-stranded RNA (dsRNA)-binding motif (dsRBM, e.g., SND1 and DGCR8) and helicase domains (e.g., DDX42 and DDX3X) were the least accurate in RNA sequence-only predictions, and the improvement from RNA structural information was the most dramatic. These findings confirm that, in general, RNA structure influences the binding of dsRNA-binding domains more than that of ssRNA-binding domains.

Finally, the predicted binding probability from PrismNet correlated well with the binding affinity determined from CLIP experiments, as shown for different RBPs in K562 cells (Spearman correlation coefficient = 0.46, $P=2.2e-16$) (Figures 2G and S3G-

H). Although we only used “1/0” labels to denote the binding and unbinding events in the training dataset, PrismNet apparently learned a quantitative model for RBP binding from the big data of sequence, structure, and protein-RNA interaction. Unexpectedly, cell type-specific binding sites (Figure S2C) generally had lower predicted binding scores compared to common binding sites (Figure 2H, $p = 0$ for unpaired t -test). PrismNet is thus able to predict dynamic RBP bindings with intermediate affinity, which likely play important regulatory roles in response to changes in cellular environments.

RBP binding sites predicted by PrismNet correlates with post-transcriptional regulation

Given that PrismNet could accurately and quantitatively predict RBP binding, we next asked whether the predicted binding sites and binding affinity correlate with RBP-mediated post-transcriptional regulation.

Many of the surveyed RBPs are splicing factors. To investigate the potential concordance between predicted binding sites and alternative splicing, we examined SRSF1, an important splicing factor that functions in both constitutive and alternative pre-mRNA splicing (Pandit et al., 2013). There are over 142,507 SRSF1 binding sites in HepG2 cells, detected by eCLIP (Van Nostrand et al., 2016). We trained PrismNet with this dataset and the HepG2 RNA structural data. PrismNet was then able to predict about 193,584 and 182,886 binding sites in HEK293T and K562 cells, respectively. Of these binding sites, only 133,865 (~70%) were shared in both HEK293T and K562 cell lines, suggesting a cell-type-specific binding pattern for SRSF1 (Figure S4A).

To determine if dynamic SRSF1 binding sites correlate with alternative splicing, we examined the splicing levels of 12 exons that contain cell-type-specific binding

sites in their 5' splice sites. Indeed, we observed a positive correlation between the differential affinity of the binding sites and differential inclusion levels of these exons in K562 vs HEK293T cells (Spearman correlation coefficient = 0.59, $P=0.04$; [Figure 3A, top](#)), suggesting that SRSF1 binding contributes to exon inclusion (Anczukow et al., 2015). We further used RNA-seq data to test all the exons with cell-type-specific SRSF1 binding sites in their 5' splice sites, and found that the differential affinity of the binding sites was modestly correlated with the differential splicing scores of the exons (Spearman correlation coefficient = 0.183, $P<4.895e-6$; [Figure 3A, bottom](#)).

In addition to alternative splicing, RBPs are also essential regulators of translation and degradation (Gerstberger et al., 2014; Hasan et al., 2014). We explored potential associations between the predicted binding sites and translation efficiency as well as RNA half-life for all surveyed proteins. Our analysis recovered many RBPs known to regulate translation efficiency, including the well-studied IGF2BP proteins (Huang et al., 2018), FXR1 (Vasudevan and Steitz, 2007) and LIN28A (Jin et al., 2011) ([Figure 3B](#)). Focusing on LIN28A, a well-known RBP that promotes translation (Jin et al., 2011), we found a positive correlation between the binding score predicted by PrismNet and the translation efficiency of the target transcripts ([Figure 3B](#)). Moreover, we also confirmed many proteins that stabilize RNA, such as LIN28B (Hafner et al., 2013), FMR1 (Zhang et al., 2018) and SND1 (Paukku et al., 2008). We looked into transcripts with predicted binding sites of SND1, a protein that is known to regulate RNA half-life (Paukku et al., 2008), and found that they showed increased stability ([Figures 3C](#)).

This analysis also identified many putative novel regulators of RNA translation and half-life. For example, we found that ALKBH5 (an m⁶A eraser) may promote RNA translation, consistent with that the elimination of excessively deposited m⁶A could reduce translation efficiency (Slobodin et al., 2017) ([Figure S4B](#)). SMNDC1, an

alternative splicing regulator, may promote RNA stability, consistent with a previous finding showing that alternative splicing could be coupled with nonsense-mediated mRNA decay (Saltzman et al., 2008) (Figure S4C). Overall, these data indicate that PrismNet, especially its quantitative prediction of the binding affinity, can be used to discover post-transcriptional regulators and targets.

Verification of the RNA structure code of RBP binding preference

To characterize the binding preferences of RBPs, we developed a computational framework to capture the sequence and structural signatures of each binding site using the saliency map, a technology that has been successfully used in computer vision for visual attention retrieval (Wattenberg, 2017). The important regions in each RBP binding site manifested as high attention regions (HARs) in the saliency map, with the quantification of the contribution of every position to the binding. As a proof-of-concept, we visualized the HARs of a splicing regulator, RBFOX2. The highest attention regions (red, Figure S5A) indicate that a change at these positions - in sequence or structure - would most dramatically change RBFOX2 binding probability. Our maps clearly revealed that HARs were enriched for “GCAUG” within single-stranded structures, consistent with previous studies of RBFOX2 (Jin et al., 2003).

The HAR is a theoretical model that quantifies the sequence and structural contributions of a nucleotide to RBP binding. It thus is important to separately assess their roles by experiments. Previously, experimental validations of a binding motif mainly consider how a sequence mutation affect RBP binding (Alipanahi et al., 2015; Grønning et al., 2019), so here we focused on how a structural change will also influence RBP binding on top of sequence.

We first used different melting-and-folding treatments to perturb RNA structure without altering sequence. Briefly, for a given RBP, we selected PrismNet-predicted

RNA binding sites that form hairpin structures (both *in vivo* and *in vitro*), with the HAR residing on the stem. We heat-denatured the RNA fragment of the binding site, and then either slow-cooled to allow refolding into the hairpin structure or fast-cooled to retain single-stranded conformation (Li et al., 2008; Liu et al., 2015b). For example, PrismNet predicted a double-stranded binding site for SND1 in the transcript encoding eukaryotic translation initiation factor 1 (EIF1) in human K562 cells (Figure 4A). We found that SND1 showed stronger affinity to the slow-cooled vs. fast-cooled RNA fragment, consistent with the higher affinity of SND1 for the double-stranded conformation (Figure 4A). Using this approach, we also validated the structural preferences of TIA1 for single-stranded conformation (Figure 4B). These data validated that PrismNet accurately predicted the structural preference of SND1 and TIA1. Given that the structural context of the HARs predicted by PrismNet generally depend on the conditions, it again confirms the essential importance to integrate the *in vivo* structural information in RBP binding site prediction.

As an orthogonal approach to investigate the effect of structural context of HAR, we perturbed the structure by mutating a distal sequence outside of the binding nucleotides. TIA1 is known to bind a poly-U sequence motif (Meyer et al., 2018), and PrismNet discovered that the motif must be in a single-stranded region for high affinity binding. We synthesized a host sequence fragment in the *CANX* mRNA containing a TIA1 HAR. Mutating the poly-U motif disrupted TIA1 binding, even when the whole sequence was kept single-stranded by snap-cooling, confirming the importance of the poly-U sequence for TIA1 binding. In contrast, a mutation at a distal site resulted in diminished TIA1 binding only under conditions that promote base-pairing with the poly-U sequence, further confirming that TIA1 preferentially binds single-stranded poly-U sequences (Figure 4C). Overall, these results verify that PrismNet has learned the regulatory code encoded in RNA secondary structure, thus allowing one to predict how structural alteration affects RBP binding.

Integrative motifs from PrismNet reveal mechanisms of RBP-RNA recognition

We aggregated and aligned all high attention regions (HARs) to obtain the binding patterns for each RBP, defined as sequence-and-structure integrative motifs (“integrative motifs”). We calculated the integrative motifs for all the RBPs we had surveyed, and used a combined logo to denote a motif by adding a structural component to the nucleotide code (“P” for paired, “U” for unpaired) (Hofacker et al., 1994). The sequence component of the integrative motifs that we discovered were highly consistent with the known sequence preference of these RBPs as documented in the ATtRACT database (Giudice et al., 2016) (e.g. GGA motif for SRSF1, poly-U motif for U2AF2), and with those derived from sequencing experiments by using *k*-mers enrichment (Van Nostrand et al., 2018), or a recent method mCross (Feng et al., 2019) (Figures 5A and S5B-C).

In addition to sequence preferences, all RBPs that we surveyed exhibited preferences for specific RNA structures (Figure 5A). Most RBPs preferred unpaired regions, consistent with previous studies (Dominguez et al., 2018; Li et al., 2010). Importantly, with the structural constraints, integrative motifs had a much lower false positive rate in discovering true RBP binding sites compared to sequence-only motifs (Figures 5A and S5D). For example, respectively 52% (29644/57282) and 25% (20056/79749) of the sites matched by the integrative motif and the sequence-only motif were covered by the experimental CLIP-seq data.

To systematically visualize and compare the sequence and structural specificities for each RBP, we performed hierarchical clustering of all associated integrative motifs. Interestingly, RBPs involved in the same RNA regulatory pathway were generally clustered together (Figure 5B), which can be partially explained by their similar integrative motifs. For example, RBPs preferring the GA (AG/GAA) motifs in single-

stranded RNAs were enriched for regulators of alternative splicing and were clustered together, including SRSF family proteins (e.g. SRSF1, SRSF7 and SRSF9) as well as some hnRNPs (e.g. hnRNPA1, hnRNPM). Many RBPs that bind to U-rich, A-rich or AU-rich motifs in single-stranded RNAs were associated with RNA stability and also clustered together (Figure 5B), consistent with previous studies showing that the AU-rich elements (AREs) in 3' UTRs target host mRNAs for rapid degradation (Vasudevan and Steitz, 2007).

Some RBPs have multiple binding patterns, i.e., integrative motifs. For example, SRSF9 can bind to GUGGA in single-stranded structures and CCGGGA in double-stranded structures, and RBM27 can bind to polyU sequence in single-stranded structures and UCCUC in double-stranded structures. Generally these proteins contain multiple RBDs with different binding preferences and plays diverse regulatory roles (Dominguez et al., 2018; Ray et al., 2013). Overall, these observations suggest that integrative motifs could capture the determinants of RBP binding more accurately and agree with the domain composition and biological functions of the associated RBPs.

High attention regions predicted by PrismNet represent evolutionarily conserved, functionally important sites

An integrative motif of an RBP displays certain levels of flexibility in sequence and/or structural contents at each position (Figure 5B). Variations beyond the tolerance level may disrupt RBP binding and consequently lead to dysregulation. Indeed, using both the PhyloP and the PhastCons scores from UCSC (Pollard et al., 2010; Siepel et al., 2005), we found that the HARs predicted by PrismNet are more evolutionarily conserved than the overall transcripts (Figures 6A and S6A).

SNP density is also an indicator of evolutionary conservation, and has been used to identify constrained regions across the human genome (Sabeti et al., 2007). We analyzed the enrichment of SNPs at HARs versus the whole transcript. First of all, we separated the HARs into three groups according to their attention in primary sequence and/or secondary structure. Briefly, for each binding site, the regions with both high attention sequence and structure components in PrismNet belong to the "sequence & structure" group; the regions with only high attention sequence or structure component are in the "sequence only" or "structure only" group. We then calculated an odds ratio for each group to represent the enrichment of SNPs in HARs over background random positions for every RBP.

We first considered the common SNPs in the dbSNP database (Sherry et al., 2001) and found that all three groups showed a depletion of common SNPs for most RBPs (Figure 6B). Furthermore, while the groups of "sequence only" and "structure only" showed comparable odds ratios, the group of "sequence & structure" were generally more depleted of SNPs (odd ratio: 0.60 (sequence & structure) vs 0.78 (sequence) or 0.75 (structure), $P < 0.001$, two-sides t -test). We confirmed these results using the common SNPs from the 1000 Genomes (Altshuler et al., 2015) (Figures S6B).

Next, we investigated the enrichment of rare variants, which, on the contrary, are considered to be deleterious in human populations (Lek et al., 2016). We found that the HARs of most RBPs are enriched for rare SNVs (Figures 6C and S6C), suggesting that the variants within HARs tend to disrupt the functionality of the predicted binding sites. And again, "sequence & structure" group shows higher enrichment than the other two groups. Together, these data suggest that the HARs, particularly with both sequence and structure signatures, are evolutionary conserved, depleted of common SNPs, with harbored variants tending to be deleterious rare mutations.

We thus further explored the relationship of variants in HARs with human disease. Indeed, genetic mutations deposited in the ClinVar database are enriched in HARs, especially in the “sequence & structure” group (Figure 6D). Recently, large-scale neurogenomic studies have identified a large number of eQTL SNPs and *de novo* mutations associated with psychiatric disorders including autism spectrum disorder (ASD), schizophrenia (SCZ), and bipolar disorder (BD) (An et al., 2018; Gandal et al., 2018). In these studies, the effects of the identified variants were mostly explained by their ability to alter chromatin structure and TF binding. Interestingly, we found that these variants were also more enriched within the HARs (Figures 6E). The odds ratios varied for RBPs within each disorder, with some specific RBPs exhibiting stronger enrichment (Figure 6F). For example, RBFOX2, SND1 and FMR1 binding sites were frequently disrupted by ASD-associated variants, consistent with the roles of these proteins as revealed by previous studies (Hagerman et al., 2011; Martin et al., 2007; Weyn-Vanhentenryck et al., 2014). Notably, again, the “sequence & structure” HAR group shows higher levels of enrichment for these psychiatric disorder-associated variants. Thus, analyzing mutations that are within the PrismNet HARs could nominate putative RBP regulators and targets in complex human disorders.

RNA-structure-disruptive variants (RiboSNitches) are associated with dynamic RBP binding and disease

To further investigate the relationship between mutations in the predicted HARs and human disease, we focused on riboSNitches, a special class of SNPs or SNVs in which different alleles exhibit distinct foci RNA structures (Halvorsen et al., 2010; Wan et al., 2014b) (Figure S7A). RiboSNitches are typically identified by allele-specific structural analysis in the same cells or cells from closely-related individuals (Wan et al., 2014b). Here we developed a novel pipeline to uncover potential riboSNitches by comparing RNA structures at SNP sites with different alleles, and identified thousands of putative riboSNitches for each pair of cell lines (Figure 7A). We intersected this

dataset with the riboSNitches identified in a previous study in human lymphoblastoid cell lines (Wan et al., 2014b) and found a significant overlap (Figure S7B), confirming the validity of the dataset.

Notably, these riboSNitches are strongly associated with dynamic RBP binding (Figures 7A-B). For every RBP in every pair of cell lines, we intersected the riboSNitches with the PrismNet-predicted dynamic binding sites. We found that most of these riboSNitches are located in a dynamic binding site of at least one RBP (Figure 7A), and for most RBPs, their dynamic binding sites are enriched with riboSNitches, particularly for those RBPs with double-stranded structural preference (e.g. DGCR8, SND1) (Figure 7B).

Many of the riboSNitches are also disease-associated variants from the ClinVar database (Landrum et al., 2014). Strikingly, most of these disease-associated riboSNitches are located in dynamic RBP binding sites (Figure S7C). Compared to variants that do not affect RNA structure, riboSNitches were more enriched with disease-associated RBP-binding-disruptive mutations, both at the global level and at the level of individual RBPs as well (Figure 7C). Note that to avoid the possibility of protein truncation or frame shift, here we only considered synonymous variants.

To clarify the regulatory relevance of riboSNitches in the context of disease association, we focused on a putative riboSNitch in the pyridoxamine 5'-phosphate oxidase (PNPO) gene for further validation. This riboSNitch corresponds to a synonymous G-to-A mutation (c.552 G>A; p.184 Leu>Leu) that has been associated with epilepsy, a severe neurological disorder (Richards et al., 2015). We found that HEK293 contains both alleles, whereas K562 contains the G allele (Figure S7D). The locus was indeed a riboSNitch, with very different RNA structures in the two cell lines (Figures 7D-E). We then used PrismNet to scan each RBP for its binding probability

at this local region and found that TARDBP had a much stronger binding affinity in K562 cells than in HEK293 cells (Figure 7F binding probability: 0.658 in K562 versus 0.028 in HEK293), consistent with the experimental CLIP-seq data (Figure S7E, K562: GSM2423707, GSM2423708; HEK293: DRX012638, DRX012639). TARDBP has been implicated in pre-mRNA splicing (Tollervey et al., 2011) and we noticed that the *PNPO* exon containing this riboSNitch exhibited a dramatically reduced inclusion rate in HEK293 cells (Figure 7G, $P=0.013$, t-test). Interestingly, knockdown of TARDBP in K562 cells resulted in substantially reduced *PNPO* exon inclusion, in line with the phenotype of disrupted TARDBP binding in HEK293 cells (Figures 7H and S7F). Collectively, these data indicate that riboSNitches, by disrupting the local RNA structure and regulating RBP binding simultaneously, may function to bridge the disease-associated synonymous mutations to RBP-based post-transcriptional regulation.

Discussion

A substantial body of work has attempted to map RBP binding profiles to gain a more complete and mechanistic understanding of RNA regulation (Baltz et al., 2012; Castello et al., 2012; Gerstberger et al., 2014; Licatalosi and Darnell, 2010). Target recognition by RBPs has been found to be remarkably precise *in vivo* yet usually involves a limited primary RNA sequence space that is rich in low-complexity motifs (Jankowsky and Harris, 2015; Lunde et al., 2007; Singh and Valcarcel, 2005). Binding site predictions with these low-complexity motifs inevitably contain many false positives. Local RNA structure of RBP binding sites is known to contribute critically to the specificity of RBP recognition, however, previous predictive models have not included information on *in vivo* RNA structure due to limited data availability. Here, we resolved these issues by generating a large dataset of RNA structural profiles across diverse mammalian cell lines. Our RNA structurome data shows that RNA structure displays a significant amount of cell line-dependent variability, strongly suggesting its

critical importance in the modeling cell line-specific RBP binding. We implemented these insights into a deep neural network, PrismNet, constructed upon the large amount of cell line-specific RBP binding and RNA structurome data *in vivo* from matched cell lines. PrismNet learns protein-RNA interaction models and thus was able to very accurately predict RBP binding in cellular conditions.

The PrismNet model also allows us to dissect the relative contributions of RNA primary sequence and secondary structure. We verified that the change of RNA structure alone could affect RBP binding, once again highlighting its critical role in the dynamic regulation of RBP binding *in vivo*. Applying PrismNet models will accurately predict how RBP bindings vary together with RNA structural changes in new physiological contexts. Our finding also shows the promise of deep learning models for integrative data analysis and providing biological insights, rather than merely serving as black box classifiers.

By extracting features from a big dataset of binding sites with sequence and structural information, PrismNet learns a quantitative model of specificity determinants using the saliency map, in which the contribution of every position can be visualized. More importantly, with the model, PrismNet can predict mutations with severe consequences on RBP binding, and automatically dissect the perturbation effects resulting from the disruption of normal sequence and/or structure patterning. We identified and validated mutations in the PrismNet high attention regions that are strongly deleterious in the human population, and are implicated with potentially important roles in complex human disorders.

Much work remains to be done to understand post-transcriptional regulation in different physiological contexts, particularly in disease-relevant conditions where mutations affect RNA structure and RBP binding. SNPs with allele-specific RNA

structure, i.e. riboSNitches, are an emerging topic of great interest with ample opportunities for discovery. Although some candidates were found associated with diverse human disorders and phenotypes (Wan et al., 2014b), we still do not know why such correlations exist and the underlying mechanisms. Here we comprehensively surveyed their association with dynamic RBP binding and also human disease. It is thus a fascinating hypothesis that these correlations could be mediated by the dysregulation of RBP binding. To this end, more RNA secondary structure profiles from disease-related tissue types will be valuable for dissecting the mechanisms of riboSNitches in a specific disease or phenotype context. Another important and further step would be to discover small-molecule drugs that could potentially target RNA in a structure-specific manner (Warner et al., 2018). Together, the enhanced understanding of the regulatory mechanisms that govern the formation of RNA structure and its interaction with other factors may pave the way to the discovery of novel therapeutic candidates.

Acknowledgments

We thank members of the Zhang labs for discussion. We thank Meiling Piao for experimental advice, Pan Li for computational advice, and Xiaohua Shen, Yafei Yin, Kehkooi Kee and Nan Wang for the gifts of cell lines. This work is supported by the National Natural Science Foundation of China (Grants No. 31671355, 91740204, and 31761163007). L.S. was supported by the Tsinghua-Peking Center for Life Sciences Postdoctoral Fellowship.

Author Contributions

Q.C.Z. conceived and supervised the project. L.S. performed icSHAPE experiments and other validation experiments. K.X. and W.H designed the deep learning models. W.H., L.S., and Y.T.Y. analyzed all the results. L.T. assisted with experiments, and

T.X. assisted with analysis. Q.C.Z., Y.T.Y., and L.S. wrote the manuscript with inputs from all authors.

Declaration of Interests

The authors declare no competing interests.

Reference

- Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 33, 831-838.
- Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., *et al.* (2015). A global reference for human genetic variation. *Nature* 526, 68-74.
- Amiri, A., Coppola, G., Scuderi, S., Wu, F., Roychowdhury, T., Liu, F., Pochareddy, S., Shin, Y., Safi, A., Song, L., *et al.* (2018). Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* 362.
- An, J.Y., Lin, K., Zhu, L., Werling, D.M., Dong, S., Brand, H., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L., *et al.* (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* 362.
- Anczukow, O., Akerman, M., Clery, A., Wu, J., Shen, C., Shirole, N.H., Raimer, A., Sun, S., Jensen, M.A., Hua, Y., *et al.* (2015). SRSF1-Regulated Alternative Splicing in Breast Cancer. *Molecular cell* 60, 105-117.
- Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J.Y., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* 37, W202-W208.
- Baltz, A.G., Munschauer, M., Schwanhauser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., *et al.* (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Molecular cell* 46, 674-690.
- Ben-Bassat, I., Chor, B., and Orenstein, Y. (2018). A deep neural network approach for learning intrinsic protein-RNA binding preferences. *Bioinformatics* 34, i638-i646.
- Brannan, K.W., Jin, W., Huelga, S.C., Banks, C.A., Gilmore, J.M., Florens, L., Washburn, M.P., Van Nostrand, E.L., Pratt, G.A., Schwinn, M.K., *et al.* (2016). SONAR Discovers RNA-Binding Proteins from Analysis of Large-Scale Protein-Protein Interactomes. *Molecular cell* 64, 282-293.
- Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., *et al.* (2012). Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* 149, 1393-1406.

657 Darnell, J.C., Fraser, C.E., Mostovetsky, O., Stefani, G., Jones, T.A., Eddy, S.R., and
658 Darnell, R.B. (2005). Kissing complex RNAs mediate interaction between the Fragile-X
659 mental retardation protein KH2 domain and brain polyribosomes. *Gene Dev* 19, 903-918.
660 Diederik P. Kingma, J.B. (2014). Adam: A Method for Stochastic Optimization. arxiv
661 *abs/1412.6980*.
662 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson,
663 M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*
664 29, 15-21.
665 Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C.,
666 Lambert, N.J., Van Nostrand, E.L., Pratt, G.A., *et al.* (2018). Sequence, Structure, and
667 Context Preferences of Human RNA Binding Proteins. *Molecular cell* 70, 854-867 e859.
668 Eddy, S.R. (2014). Computational analysis of conserved RNA secondary structure in
669 transcriptomes and genomes. *Annu Rev Biophys* 43, 433-456.
670 Ellington, A.D., and Szostak, J.W. (1990). In vitro Selection of Rna Molecules That Bind
671 Specific Ligands. *Nature* 346, 818-822.
672 Feng, H., Bao, S., Rahman, M.A., Weyn-Vanhentenryck, S.M., Khan, A., Wong, J., Shah, A.,
673 Flynn, E.D., Krainer, A.R., and Zhang, C. (2019). Modeling RNA-Binding Protein
674 Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites. *Molecular cell* 74,
675 1189-1204 e1186.
676 Flynn, R.A., Zhang, Q.C., Spitale, R.C., Lee, B., Mumbach, M.R., and Chang, H.Y. (2016).
677 Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE.
678 *Nature protocols* 11, 273-290.
679 Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van
680 Bakel, H., Varghese, M., Wang, Y., *et al.* (2018). Transcriptome-wide isoform-level
681 dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* 362.
682 Gandhi, S., Lee, L.J., Delong, A., Duvenaud, D., and Frey, B.J. (2018). cDeepbind: A context
683 sensitive deep learning model of RNA-protein binding. *bioRxiv*, 345140.
684 Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding
685 proteins. *Nat Rev Genet* 15, 829-845.
686 Ghanbari, M., and Ohler, U. (2020). Deep neural networks for interpreting RNA-binding
687 protein target preferences. *Genome research* 30, 214-226.
688 Gilbert, C., and Svejstrup, J.Q. (2006). RNA immunoprecipitation for determining RNA-
689 protein associations in vivo. *Curr Protoc Mol Biol Chapter* 27, Unit 27 24.
690 Giudice, G., Sanchez-Cabo, F., Torroja, C., and Lara-Pezzi, E. (2016). ATtRACT-a database
691 of RNA-binding proteins and associated motifs. *Database (Oxford)* 2016.
692 Grønning, A.G.B., Doktor, T.K., Larsen, S.J., Petersen, U.S.S., Holm, L.L., Bruun, G.H.,
693 Hansen, M.B., Hartung, A.-M., Baumbach, J., and Andresen, B.S. (2019). DeepCLIP:
694 Predicting the effect of mutations on protein-RNA binding with Deep Learning. *bioRxiv*,
695 757062.
696 Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying
697 similarity between motifs. *Genome Biol* 8.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M., *et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* *141*, 129-141.

Hafner, M., Max, K.E., Bandaru, P., Morozov, P., Gerstberger, S., Brown, M., Molina, H., and Tuschl, T. (2013). Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition. *RNA* *19*, 613-626.

Hagerman, R., Au, J., and Hagerman, P. (2011). FMR1 premutation and full mutation molecular mechanisms related to autism. *J Neurodev Disord* *3*, 211-224.

Halvorsen, M., Martin, J.S., Broadaway, S., and Laederach, A. (2010). Disease-Associated Mutations That Alter the RNA Structural Ensemble. *Plos Genetics* *6*.

Hanson, S.J., and Pratt, L.Y. (1989). Comparing Biases for Minimal Network Construction with Back-Propagation. Paper presented at: Advances in Neural Information Processing Systems

Hasan, A., Cotobal, C., Duncan, C.D., and Mata, J. (2014). Systematic analysis of the role of RNA-binding proteins in the regulation of RNA stability. *PLoS Genet* *10*, e1004684.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. Paper presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

He, K.M., Zhang, X.Y., Ren, S.Q., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. Paper presented at: 2015 International Conference on Computer Vision (ICCV).

Hentze, M.W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nat Rev Mol Cell Biol* *19*, 327-341.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., *et al.* (2006). The UCSC Genome Browser Database: update 2006. *Nucleic acids research* *34*, D590-D598.

Hinton, A.K.I.S.G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Paper presented at: Advances in Neural Information Processing Systems.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., and Schuster, P. (1994). Fast Folding and Comparison of Rna Secondary Structures. *Monatsh Chem* *125*, 167-188.

Hrdlickova, R., Toloue, M., and Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA* *8*.

Hu, B., Yang, Y.T., Huang, Y., Zhu, Y., and Lu, Z.J. (2017). POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic acids research* *45*, D104-D114.

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2019). Squeeze-and-Excitation Networks. *IEEE Trans Pattern Anal Mach Intell*.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-Excitation Networks. Paper presented at: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

739 Huang, H., Weng, H., Sun, W., Qin, X., Shi, H., Wu, H., Zhao, B.S., Mesquita, A., Liu, C.,
740 Yuan, C.L., *et al.* (2018). Recognition of RNA N(6)-methyladenosine by IGF2BP proteins
741 enhances mRNA stability and translation. *Nat Cell Biol* 20, 285-295.

742 Jankowsky, E., and Harris, M.E. (2015). Specificity and nonspecificity in RNA-protein
743 interactions. *Nat Rev Mol Cell Biol* 16, 533-544.

744 Jin, J., Jing, W., Lei, X.X., Feng, C., Peng, S., Boris-Lawrie, K., and Huang, Y. (2011).
745 Evidence that Lin28 stimulates translation by recruiting RNA helicase A to polysomes.
746 *Nucleic acids research* 39, 3724-3734.

747 Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., and
748 Inoue, K. (2003). A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing
749 via the pentanucleotide GCAUG. *EMBO J* 22, 905-912.

750 Jost Tobias Springenberg, A.D., Thomas Brox, Martin Riedmiller (2014). Striving for
751 Simplicity: The All Convolutional Net. *arxiv abs/1412.6806*.

752 Jung, D., Xiong, J., Ye, M., Qin, X., Li, L., Cheng, S., Luo, M., Peng, J., Dong, J., Tang, F.,
753 *et al.* (2017). In vitro differentiation of human embryonic stem cells into ovarian follicle-like
754 cells. *Nat Commun* 8, 15680.

755 Koo, P.K., Anand, P., Paul, S.B., and Eddy, S.R. (2018). Inferring Sequence-Structure
756 Preferences of RNA-Binding Proteins with Convolutional Residual Networks. *bioRxiv*,
757 418459.

758 Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., and Burge, C.B. (2014).
759 RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity
760 of RNA binding proteins. *Molecular cell* 54, 887-900.

761 Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and
762 Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and
763 human phenotype. *Nucleic acids research* 42, D980-D985.

764 Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat*
765 *Methods* 9, 357-U354.

766 LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436-444.

767 Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-
768 Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.* (2016). Analysis of protein-coding
769 genetic variation in 60,706 humans. *Nature* 536, 285-291.

770 Lewis, C.J., Pan, T., and Kalsotra, A. (2017). RNA modifications and structures cooperate to
771 guide RNA-protein interactions. *Nature reviews Molecular cell biology* 18, 202-210.

772 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis,
773 G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence
774 Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079.

775 Li, P., Shi, R., and Zhang, Q.C. (2019). icSHAPE-pipe: A comprehensive toolkit for
776 icSHAPE data analysis and evaluation. *Methods*.

777 Li, P.T.X., Viereggs, J., and Tinoco, I. (2008). How RNA unfolds and refolds. *Annual Review*
778 *of Biochemistry* 77, 77-100.

779 Li, X., Quon, G., Lipshitz, H.D., and Morris, Q. (2010). Predicting in vivo binding sites of
780 RNA-binding proteins using mRNA secondary structure. *RNA* 16, 1096-1107.

781 Licatalosi, D.D., and Darnell, R.B. (2010). RNA processing and its regulation: global insights
782 into biological networks. *Nat Rev Genet* 11, 75-87.

783 Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer,
784 A.C., Blume, J.E., Wang, X.N., *et al.* (2008). HITS-CLIP yields genome-wide insights into
785 brain alternative RNA processing. *Nature* 456, 464-U422.

786 Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015a). N(6)-methyladenosine-
787 dependent RNA structural switches regulate RNA-protein interactions. *Nature* 518, 560-564.

788 Liu, N., Dai, Q., Zheng, G.Q., He, C., Parisien, M., and Pan, T. (2015b). N-6-
789 methyladenosine-dependent RNA structural switches regulate RNA-protein interactions.
790 *Nature* 518, 560-564.

791 Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for
792 efficient function. *Nat Rev Mol Cell Biol* 8, 479-490.

793 Martin, C.L., Duvall, J.A., Ilkin, Y., Simon, J.S., Arreaza, M.G., WilkeS, K., Alvarez-
794 Retuerto, A., Whichello, A., Powell, C.M., Rao, K., *et al.* (2007). Cytogenetic and molecular
795 characterization of A2BP1/FOX1 as a candidate gene for autism. *Am J Med Genet B* 144b,
796 869-876.

797 Maticzka, D., Lange, S.J., Costa, F., and Backofen, R. (2014). GraphProt: modeling binding
798 preferences of RNA-binding proteins. *Genome Biol* 15, R17.

799 Meyer, C., Garzia, A., Mazzola, M., Gerstberger, S., Molina, H., and Tuschl, T. (2018). The
800 TIA1 RNA-Binding Protein Family Regulates EIF2AK2-Mediated Stress Response and Cell
801 Cycle Progression. *Molecular cell* 69, 622-635 e626.

802 Orenstein, Y., Wang, Y., and Berger, B. (2016a). RCK: accurate and efficient inference of
803 sequence- and structure-based protein-RNA binding models from RNAcompete data.
804 *Bioinformatics* 32, i351-i359.

805 Orenstein, Y., Wang, Y., and Berger, B. (2016b). RCK: accurate and efficient inference of
806 sequence- and structure-based protein-RNA binding models from RNAcompete data.
807 *Bioinformatics* 32, i351-i359.

808 Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G., Li, H., Qiu, J., Huang, J., Yeo, G.W.,
809 Ares, M., Jr., and Fu, X.D. (2013). Genome-wide analysis reveals SR protein cooperation and
810 competition in regulated splicing. *Molecular cell* 50, 223-235.

811 Paszke, A.a.G., Sam and Massa, Francisco and Lerer, Adam and Bradbury, James and
812 Chanan, Gregory and Killeen, Trevor and Lin, Zeming and Gimelshein, Natalia and Antiga,
813 Luca and Desmaison, Alban and Kopf, Andreas and Yang, Edward and DeVito, Zachary and
814 Raison, Martin and Tejani, Alykhan and Chilamkurthy, Sasank and Steiner, Benoit and Fang,
815 Lu and Bai, Junjie and Chintala, Soumith (2019). PyTorch: An Imperative Style, High-
816 Performance Deep Learning Library. Paper presented at: Advances in Neural Information
817 Processing Systems.

818 Paukku, K., Kalkkinen, N., Silvennoinen, O., Kontula, K.K., and Lehtonen, J.Y.A. (2008).
819 p100 increases AT1R expression through interaction with AT1R 3'-UTR. *Nucleic acids*
820 *research* 36, 4474-4487.

821 Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of
822 nonneutral substitution rates on mammalian phylogenies. *Genome research* 20, 110-121.

Priya Goyal, P.D., Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, Kaiming He (2017). Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour (arXiv:1706.02677v2).

Ray, D., Kazan, H., Chan, E.T., Pena Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol* 27, 667-670.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., 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.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., *et al.* (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 17, 405-424.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. *Nat Biotechnol* 29, 24-26.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* 505, 701-705.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., *et al.* (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913-918.

Saltzman, A.L., Kim, Y.K., Pan, Q., Fagnani, M.M., Maquat, L.E., and Blencowe, B.J. (2008). Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. *Molecular and Cellular Biology* 28, 4320-4330.

Scaffidi, P., and Misteli, T. (2006). Lamin A-dependent nuclear defects in human aging. *Science* 312, 1059-1063.

Schueler, M., Munschauer, M., Gregersen, L.H., Finzel, A., Loewer, A., Chen, W., Landthaler, M., and Dieterich, C. (2014). Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biol* 15, R15.

Seetin, M.G., and Mathews, D.H. (2012). RNA structure prediction: an overview of methods. *Methods in molecular biology* 905, 99-122.

Sergey Ioffe, C.S. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arxiv abs/1502.03167*.

Shen, X., Liu, Y., Hsu, Y.J., Fujiwara, Y., Kim, J., Mao, X., Yuan, G.C., and Orkin, S.H. (2008). EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency. *Molecular cell* 32, 491-502.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research* 29, 308-311.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* 15, 1034-1050.

Singh, R., and Valcarcel, J. (2005). Building specificity with nonspecific RNA-binding proteins. *Nat Struct Mol Biol* 12, 645-653.

Slobodin, B., Han, R.Q., Calderone, V., Vrielink, J.A.F.O., Loayza-Puch, F., Elkon, R., and Agami, R. (2017). Transcription Impacts the Efficiency of mRNA Translation via Co-transcriptional N6-adenosine Methylation. *Cell* 169, 326-337.

Solomon, O., Di Segni, A., Cesarkas, K., Porath, H.T., Marcu-Malina, V., Mizrahi, O., Stern-Ginossar, N., Kol, N., Farage-Barhom, S., Glick-Saar, E., *et al.* (2017). RNA editing by ADAR1 leads to context-dependent transcriptome-wide changes in RNA secondary structure. *Nature Communications* 8.

Spitale, R.C., Flynn, R.A., Zhang, Q.C., Crisalli, P., Lee, B., Jung, J.W., Kuchelmeister, H.Y., Batista, P.J., Torre, E.A., Kool, E.T., *et al.* (2015). Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* 519, 486-490.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res* 15, 1929-1958.

Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., and Giegerich, R. (2006). RNASHAPes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* 22, 500-503.

Sun, L., Fazal, F.M., Li, P., Broughton, J.P., Lee, B., Tang, L., Huang, W., Kool, E.T., Chang, H.Y., and Zhang, Q.C. (2019). RNA structure maps across mammalian cellular compartments. *Nat Struct Mol Biol* 26, 322-330.

Tollervey, J.R., Curk, T., Rogelj, B., Briesse, M., Cereda, M., Kayikci, M., Konig, J., Hortobagyi, T., Nishimura, A.L., Zupunski, V., *et al.* (2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nature Neuroscience* 14, 452-U180.

Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J., and Eyra, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* 19, 40.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., *et al.* (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43, 11 10 11-11 10 33.

Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., *et al.* (2018). A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv*, 179648.

Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., *et al.* (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* 13, 508-514.

Vasudevan, S., and Steitz, J.A. (2007). AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2. *Cell* 128, 1105-1118.

Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., *et al.* (2014a). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 505, 706-709.

Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z.Q., Zhang, J.J., Spitale, R.C., Snyder, M.P., Segal, E., *et al.* (2014b). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 505, 706-+.

Warner, K.D., Hajdin, C.E., and Weeks, K.M. (2018). Principles for targeting RNA with drug-like small molecules. *Nat Rev Drug Discov* 17, 547-558.

Wattenberg, D.S.a.N.T.a.B.K.a.F.B.V.e.g.a.M. (2017). SmoothGrad: removing noise by adding noise. *ArXiv abs/1706.03825*.

Weyn-Vanhentenryck, S.M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Xue, C., Herre, M., Silver, P.A., Zhang, M.Q., *et al.* (2014). HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep* 6, 1139-1152.

Yang, J.H., Li, J.H., Shao, P., Zhou, H., Chen, Y.Q., and Qu, L.H. (2011). starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic acids research* 39, D202-209.

Zhang, F., Kang, Y., Wang, M., Li, Y., Xu, T., Yang, W., Song, H., Wu, H., Shu, Q., and Jin, P. (2018). Fragile X mental retardation protein modulates the stability of its m6A-marked messenger RNA targets. *Hum Mol Genet* 27, 3936-3950.

Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., and Zeng, J. (2016). A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic acids research* 44, e32.

Zhu, Y., Xu, G., Yang, Y.T., Xu, Z., Chen, X., Shi, B., Xie, D., Lu, Z.J., and Wang, P. (2019). POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic acids research* 47, D203-D211.

Zisserman, K.S.A.V.A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Paper presented at: Workshop at International Conference on Learning Representations.

FIGURE LEGENDS

Figure 1. Association between RNA structural variations and dynamic RBP bindings can be used to predict RBP bindings in varying cellular context.

(A) Integrative modeling and prediction of RBP bindings by PrismNet using *in vivo* RNA structure and RBP binding sites from matched types of cells. PrismNet can be used to dissect and predict the perturbation effects of disease-associated genetic variations on RBP binding.

(B) Circos plot showing the relationship between RNA structural variations (Δ Structures) and dynamic RBP binding sites (Δ RBP) in HepG2 and K562 cells. The transcripts in the region chr14:22,000,000-24,000,000 is magnified to illustrate that dynamic RBP binding sites is overlapped with RNA structural variations. Numbers show the fraction of overlapped over all RBP binding sites (e.g., 22/63).

(C) RNA structural and HNRNPM binding profiles in HepG2 and K562 cell lines. Top: icSHAPE scores in the two cell lines for transcripts *LSS*. Middle: Binding site of HNRNPM on *LSS* (eCLIP). Bottom: RNA structural models of the HNRNPM binding sites on *LSS* in the two cell lines. Models are constructed by *RNAshape* with icSHAPE score constraints. Green dashed lines indicate the known HNRNPM poly-U binding motif.

Figure 2. PrismNet more accurately predicts RBP binding in cellular conditions and *in vivo* RNA structure is critical to the prediction improvement.

(A) Model architecture of PrismNet. The input features include the RNA sequence encoded in the 4-dimensional one-hot encoding and icSHAPE structural score as the fifth-dimension. The neural network consists of multiple convolutional layers, squeeze-and-excitation networks and residual blocks to capture the joint sequence-and-structural determinants of RBP binding. SE, squeeze-and-excitation.

(B) Predicted versus observed binding sites of IGF2BP1 on the *EIF3F* transcript. Green/black, observed binding sites in K562/HepG2 cells by eCLIP, used as training

and ground truth reference data respectively. Blue/red, true/false positive predictions in HepG2 cells based on the models trained on K562 data.

(C) Violin plot of the overall AUC scores of PrismNet versus other methods in all human 144 PrismNet models of 99 RBPs. $***P < 0.001$ (two-sided paired t-test).

(D) Violin plot of the overall AUC scores of PrismNet models using different types of input data in all human 144 PrismNet models of 99 RBPs. $***P < 0.001$ (two-sided paired t-test).

(E) Scatter plot of AUC scores of PrismNet models using *in vivo* structures versus computationally predicted structures. Each dot represents an RBP.

(F) Scatter plot of AUC improvements of PrismNet versus AUC scores of PrismNet models using only sequence information. Each dot represents an RBP.

(G) Density map of binding scores of IGF2BP1 predicted using PrismNet versus the observed binding scores from eCLIP experiments in K562 cells.

(H) Violin plot of PrismNet-predicted binding probabilities at the binding sites in K562 cells only, HepG2 cells only, or both. $***P < 0.001$ (unpaired t-test).

Figure 3. The quantitative RBP binding of PrismNet prediction correlates with the regulations of splicing, translation and degradation.

(A) Correlation of PrismNet-predicted differential SRSF1 binding and alternative splicing between HEK293T and K562 cells. Top: on experimentally derived alternative splicing scores of 12 exons. Bottom: on transcriptome-wide RNA-seq data.

(B) Correlation between PrismNet-predicted RBP binding and translational efficiency in HepG2 cells. Top: Distribution of correlation coefficients of all RBPs with significance scores. Marked RBPs are known translation regulators. Bottom: Density plot of the predicted LIN28A binding scores versus translation efficiencies for the target transcripts.

(C) Correlation between PrismNet-predicted RBP binding and RNA degradation in HEK293 cells. Top: Distribution of correlation coefficients of all RBPs with significance

scores. Marked RBPs are known RNA degradation regulators. Bottom: Density plot of the predicted SND1 binding scores versus half-lives of the target transcripts.

Figure 4. Validation of PrismNet-predicted effects of RNA sequence mutations and structure changes on RBP binding

(A) Experimental validation of SND1 binding on a site from the *EIF1* transcript (*EIF1*:521-621). Top: The saliency map showing the site with double-stranded conformation, predicted to be bound by SND1. Grey bar, region of synthesized RNA fragment. Green background, duplex region. Bottom left: Predicted *in vitro* secondary structural model. Bottom right: RNA pull-down for the synthesized RNA fragment in different structural conformation (n=3 replicates).

(B) Experimental validation of TIA1 binding on a site from the *SLC35D2* transcript (*SLC35D2*:225-325). Top: The saliency map showing the site with double-stranded conformation, predicted *not* to be bound by TIA1. Grey bar, region of synthesized RNA fragment. Green background, TIA1-binding sequence motif. Bottom left: Predicted *in vitro* secondary structural model. Bottom right: RNA pull-down for the synthesized RNA fragment in different structural conformation (n=3 replicates).

(C) Experimental validation of TIA1 binding on a site from the *CANX* transcript (*CANX*:55-155). Left: The saliency maps showing the site of the wildtype (WT) and sequence mutants in different structural conformation: (i) WT in single-stranded conformation; (ii) HAR sequence mutation, single-stranded conformation, (iii) distal site sequence mutation, single-stranded conformation, (iv) distal site sequence mutation, double-stranded conformation. Grey, region of synthesized RNA fragments. Green background, TIA1-binding sequence motif. Pink background, a distal site. Middle: Predicted *in vitro* secondary structural models. Right: RNA pull-down for the synthesized RNA fragment in different structural conformations (n=3 replicates).

Figure 5. Integrative motifs derived from PrismNet are more specific in capturing transcriptome-wide binding site and clustered with similar post-transcriptional regulatory functions.

(A) Motifs of different RBPs. Left: the integrative motifs from PrismNet, sequence motifs from mCross and ATtRACT. The integrative motif consists of a sequence component (up) and a structural component (down), where “U” for unpaired and “P” for paired nucleotides. Right: The true positive and all matched RBP binding sites on the transcriptome by motif scanning using the integrative and the ATtRACT motifs. True positives are determined by comparing with eCLIP experiments.

(B) Hierarchical clustering of integrative motifs of all human RBPs. Color represents functions in the inner tree, structural preferences in the integrative motif circle, and RBDs in the outer circles. RBPs with red fonts are discussed in the main text.

Figure 6. Conservation and enrichment of genomic variants in PrismNet-predicted high attentione regions (HARs)

(A) Distribution of PhyloP20way conservation scores of HARs, with those of all CLIP binding sites and all transcribed regions as positive and negative controls.

(B) Enrichment of common SNPs (from dbSNP) in HARs and CLIP binding sites for every RBP comparing to random transcript regions.

(C) Enrichment of rare SNVs (from dbSNP) relative to common SNVs in HARs and CLIP binding sites for every RBP.

(D) Enrichment of ClinVar mutation in HARs comparing to random transcript regions.

(E-F) Enrichment of ASD, BD and SCZ-associated variants in all HARs collectively **(E)** or HARs of every RBP respectively **(F)**, comparing to random transcript regions. The RBPs known to associate with psychiatric disorders were highlighted.

All HARs and CLIP-seq binding sites are in K562 cells. In **B**, **C**, and **F**, each dot represent an RBP. RBPs with HARs significantly depleted/enriched of genomic

variants (including common and rare SNVs and disease-associated variants) are shown in color (*for $P < 0.05$, permutation or fisher exact test).

Figure 7. RiboSNitches are associated with RBP dynamic binding sites and human diseases.

(A) The number of all riboSNitches between different cell lines (bottom-left triangle), and those within dynamic RBP binding sites predicted by PrismNet (up-right triangle).

(B) Enrichment of riboSNitches in dynamic RBP binding sites comparing to common binding sites predicted by PrismNet.

(C) Enrichment of riboSNitches relative to CSVs within HARs in human disease (from ClinVar).

In B and C, each dot represents an RBP, with significantly enrichment shown in color (* for $P < 0.05$, fisher exact test).

(D) Structural profiles of the icSHAPE scores around a riboSNitch in the *PNPO* transcript in HEK293 and K562 cells.

(E) Structural models of riboSNitch and flanking sequences in the *PNPO* transcript in HEK293 and K562 cells, predicted by *RNAshape* with the corresponding icSHAPE scores as constraints. Red stars indicate the mutation sites.

(F) Binding probabilities and saliency maps of TARDBP on the *PNPO* transcript in HEK293 and K562 cells.

(G) Splicing of the *PNPO* transcript in K562 and HEK293 cells (n=3 replicates).

(H) Splicing of the *PNPO* transcript in WT and TARDBP knockdown K562 cells. FUS knockdowns are included as a negative control.

Supplemental Figure 1. Probing of RNA structuromes by icSHAPE in different cell lines.

(A) Left: Heatmap of available RBP CLIP datasets in each cell line. RBPs with datasets in at least three cell lines were labeled. Middel: Cell lines with the number of available

CLIP datasets. Right: The number of transcripts with secondary structure probed by icSHAPE in each cell line.

(B) Correlation of RNA expression (RPKM) between replicates of icSHAPE libraries.

(C) Tracks of icSHAPE scores of the human 18S ribosomal RNA.

(D) Structural model of 18S rRNA (141-340). Model is plotted with the *ViennaRNA* web service based on data from the *Comparative RNA Web* (CRW) Site. Nucleotides are colored with icSHAPE scores from HepG2 cells.

(E) ROC curve plot of the agreement of icSHAPE scores from different cell lines with the reference 18S rRNA structure from the *Comparative RNA Web* (CRW) Site. AUC scores for all ROC curves are included.

Supplemental Figure 2. RNA secondary structurally variable sites, dynamic RBP binding profiles, and their associations.

(A) Tracks of icSHAPE scores at the human genome location chrX:153,688,450-153,688,850. The light Gray background color labels a structurally variable site.

(B) Bar plot of the ratio and the number of RNA structurally variable sites between any two human cell lines. Volcano plot shows the statistical significance versus magnitude of RNA structural changes between HepG2 and K562 cells. The red plots are the identified structurally variable sites.

(C) Stacked bar plots of the percentage of dynamic (cell type-specific) and common RBP binding sites in two cell lines from eCLIP datasets: HepG2-specific (red), K562-specific (blue) and common (grey).

(D) The overlap between RNA secondary structurally variable sites (big central circle) and dynamic RBP binding sites (small surrounding circles) in HepG2 and K562 cells. The total numbers and also the overlapped numbers are shown. *P* values were calculated by Fisher's exact test. **P* < 0.05; ***P* < 1 × 10⁻³; ****P* < 1 × 10⁻⁵. Note that this is not a Venn diagram in that the overlaps among the RBP binding sites are not considered.

(E) RNA structural and HNRNPM binding profiles in HepG2 and K562 cell lines. Top: icSHAPE scores in the two cell lines for transcripts *FAH*. Middle: Binding site of HNRNPM on *FAH* (eCLIP). Bottom: RNA structural models of the HNRNPM binding sites on *FAH* in the two cell lines. Models are constructed by *RNAshape* with icSHAPE score constraints. Green dashed lines indicate the HNRNPM binding motif.

Supplemental Figure 3. The architecture and prediction performance of PrismNet.

(A) The architecture of PrismNet with important hyper-parameters.

(B-C) Scatter plot of AUC scores of the predictions by PrismNet versus GraphPort (B) and DeepBind (C). Each dot represents an RBP.

(D) Scatter plot of AUC scores of the predictions by PrismNet with computationally-predicted structure versus with only sequence. Each dot represents an RBP.

(E) Bar plot of AUC scores of the predictions by PrismNet (red), PrismNet with only sequence (dark grey), and PrismNet with computationally-predicted RNA structure (light grey).

(F) Violin plot of AUC improvements of PrismNet predictions using *in vivo* structure for RBP groups with different RNA-binding domains. * $P < 0.05$; ** $P < 0.01$; *** $P < 1 \times 10^{-3}$ (two-sided unpaired t-test).

(G) Density plot of the binding probability of SAFB2, SND1, and RBM27 predicted by PrismNet versus the observed binding scores from eCLIP in K562 cells.

(H) Violin plot of the correlation between PrismNet-predicted binding probability and CLIP binding scores from eCLIP for different proteins.

Supplemental Figure 4. The PrismNet-predicted differential RBP binding, and the correlation of predicted quantitative RBP binding and target translation and degradation.

(A) Venn diagram of SRSF1 binding sites in HEK293T and K562 cells, predicted by PrismNet model trained on the HepG2 SRSF1 eCLIP dataset.

(B) Density plot of the PrismNet-predicted ALKBH5 binding scores versus translation efficiency of the target transcripts.

(C) Density plot of the PrismNet-predicted SMNDC1 binding scores versus half-lives of the target transcripts.

Supplemental Figure 5. Saliency maps of PrismNet-predicted binding sites and Integrative motifs of different RBPs.

(A) Saliency maps of Rbfox2 binding sites derived from the mES PrismNet model. The heatmap tracks display the reponse of the model at each nucleotide, with red color showing high attention (upper: sequence reponse, bottom: structure reponse). The sequence logos on the top are shown to help to display high attention sequence component. The red line at bottom represents icSHAPE scores.

(B) Integrative motifs derived from PrismNet, compared with those provided in eCLIP and RBNS.

(C) Integrative motifs derived from PrismNet, compared with those obtained by mCross.

(D) True positives and all matched binding sites on the transcriptome by motif scanning using the integrative and the ATtRACT motifs of different RBPs. True positives are determined by comparing with eCLIP experiments. Related to main Figure 5A. Here we used a loose criterion in motif scanning to match more sites.

Supplemental Figure 6. Conservation and enrichment of genomic variants in PrismNet-predicted high attention regions (HARs)

(A) Distribution of PhastCons100way conservation scores of HARs, with those of all binding sites from CLIP-seq experiments and all transcript regions as positive and negative controls.

(B) Enrichment of common SNPs (from 1000 Genomes) in HARs and CLIP binding sites for every RBP comparing to random transcript regions.

(C) Enrichment of rare SNVs (from 1000 Genomes) relative to common SNVs in HARs and CLIP binding sites for every RBP.

All HARs and CLIP-seq binding sites are in K562 cells. In **B** and **C**, each dot represents an RBP. RBPs with HARs significantly depleted/enriched of genomic variants are shown in color (*for $P < 0.05$, permutation or fisher exact test).

Supplemental Figure 7. RiboSNitches between different cell lines and their association in RBP dynamic binding sites and human diseases.

(A) RiboSNitches in the *DDOST* and *ID3* transcripts. Left: Tracks of icSHAPE scores around the riboSNitches in HEK293 and HeLa cell lines. The dash lines indicate the riboSNitch sites. Right: RNA structural models constructed by *RNAshape* with icSHAPE score constraints.

(B) The number (bottom-left triangle) and the enrichment (up-right triangle) of detected riboSNitches intersecting with the riboSNitches identified in human lymphoblastoid cell lines (Wan et. al., 2014).

(C) The number of all riboSNitches corresponding to a SNV in ClinVar (bottom-left triangle) and those that are also in dynamic RBP binding sites predicted by PrismNet (up-right triangle).

(D) Validation of the riboSNitch alleles in the *PNPO* gene in HEK293 and K562 cells by Sanger sequencing.

(E) TARDBP binding tracks in K562 (eCLIP) and HEK293 (PARCLIP) in the *PNPO* transcript. The dash line indicates the riboSNitch site.

(F) The knock down efficiency of FUS and TARDBP in K562 cells.

Figure 1

Figure 2

Figure 3

Figure 5

Figure 6

Figure 7

Figure S2

Figure S3

A

B

C

E

D

F

G

H

Figure S4

Figure S5

Figure S6

Figure S7