A Deep Boosting Based Approach for Capturing the Sequence Binding Preferences of RNA-Binding Proteins from High-Throughput CLIP-Seq Data

Shuya Li[1,†], Fanghong Dong[2,†], Yuexin Wu[2,3,†], Sai Zhang[2], Chen Zhang[2], Xiao Liu[1], Tao Jiang[4,5], and Jianyang Zeng[2,*]

[1]School of Life Sciences, Tsinghua University, Beijing, China.

[2]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China.

[3]Present address: Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA

[4]Department of Computer Science and Engineering, University of California, Riverside, CA.

[5]MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST / Department of Computer Science and Technology, Tsinghua University, Beijing, China.

[*]To whom correspondence should be addressed. Email: zengjy321@tsinghua.edu.cn.

[†]These authors contributed equally to this work.

Abstract

Characterizing the binding behaviors of RNA-binding proteins (RBPs) is important for understanding their functional roles in gene expression regulation. However, current high-throughput experimental methods for identifying RBP targets, such as CLIP-seq and RNAcompete, usually suffer from the false positive and false negative issues. Here, we develop a deep boosting based machine learning approach, called DeBooster, to accurately model the binding sequence preferences and identify the corresponding binding targets of RBPs from CLIP-seq data. Comprehensive validation tests have shown that DeBooster can outperform other state-of-the-art approaches in predicting RBP targets and recover false negatives that are common in current CLIP-seq data. In addition, we have demonstrated several new potential applications of DeBooster in understanding the regulatory func-

tions of RBPs, including the binding effects of the RNA helicase MOV10 on mRNA degradation, the influence of different binding behaviors of the ADAR proteins on RNA editing, as well as the antagonizing effect of RBP binding on miRNA repression. Moreover, DeBooster may provide an effective index to investigate the effect of pathogenic mutations in RBP binding sites, especially those related to splicing events. We expect that DeBooster will be widely applied to analyze large-scale CLIP-seq experimental data and can provide a practically useful tool for novel biological discoveries in understanding the regulatory mechanisms of RBPs. The scource code of DeBooster can be downloaded from `http://github.com/dongfanghong/deepboost`.

1 Introduction

RNA binding proteins (RBPs) play important roles in multiple aspects of gene expression regulation, such as alternative splicing, RNA modification, mRNA export and localization [1]. Not only does the dysregulation of RBPs induce abnormality, but also the mutations in their binding targets have the potential to cause diseases [2]. So, capturing the intrinsic binding preferences of RBPs and identifying their binding targets in a precise and large-scale manner are essential to understand the regulatory roles of RBPs and reveal their connections to the pathogenesis of complex diseases.

Before the development of high-throughput techniques for characterizing RNA-protein interactions, only a few RBPs were well studied based on the small-scale experiments, such as *in vitro* EMSA [3] and *in vivo* fluorescence methods [4]. Recently, several high-throughput sequencing-based approaches, e.g., CLIP-seq [5–7], SELEX [8, 9] and RNAcompete [10, 11], have been proposed to measure RBP binding sites and binding affinities in a transcriptome-wide manner. However, despite the huge amount of data generated by these techniques, they still suffer from the false positive and false negative issues mainly due to experimental noise and bias [12]. To overcome these drawbacks, various computational models [13–19] have been developed to learn RBP binding preferences and detect putative RBP targets based on abundant experimental data. As many RBPs have been validated to recognize structured regions [20], there is a tendency in recent studies to incorporate the structural features of target RNAs into prediction models, such as MEMERIS [15], GraphProt [17] and our recent deep learning based model [19], where the integration of RNA structural information has been shown to largely boost the prediction performance. Nevertheless, the current transcriptome-wide experimental techniques

for measuring RNA structures are far from maturity. On the other hand, predicting RNA structures using computational models usually requires a substantial amount of additional effort and time, and a predicted RNA structure is generally less accurate compared to that derived from experimental approaches. In addition, systematic integration of both sequence and structural information generally requires a more complex prediction model. So far, it remains largely unknown whether we can derive a sequence based prediction model that only takes RNA seqeuence as input, while still achieving prediction performance comparable to that of the state-of-the-art prediction methods that require both sequence and structural profiles. To fill this gap between modeling accuracy and computational complexity, we develop a deep boosting based model, called DeBooster, that requires only sequence information and can capture RBP binding preferences and predict binding sites from high-throughput CLIP-seq data with high accuracy and efficiency.

Through testing on 24 CLIP-seq datasets, we have shown that even without using RNA structural information, DeBooster can outperform the state-of-the-art methods that take both sequence and structural information as input, including both GraphProt [17] and our previous deep learning based model [19]. In addition, we have performed comprehensive tests to validate the superiority of DeBooster: (i) DeBooster can accurately capture RBP binding preferences and generate RBP binding motifs that are consistent with previous studies in the literature; (ii) DeBooster can successfully carry out the cross-platform prediction task and effectively address the false negative problem that is prevalent in current CLIP-seq data; (iii) In addition to binary classification, DeBooster can be easily extended to solve a regression problem such that the prediction scores closely match the experimentally-measured binding affinities.

In addition to the above extensive validation tests, we have further demonstrated several new possible applications of DeBooster in studying the regulatory roles of RBPs. With an integrative analysis based on other types of data and our prediction results, we not only derive literature-consistent results concerning RBP regulation, but also hope to gain novel insights into the biological rationale of the regulatory roles of RBPs. In particular, we have conformed that the binding targets of the RNA helicase MOV10 predicted by DeBooster are highly associated with the fold changes of mRNA half-lives, providing another evidence on the regulatory functions of RNA helicases on mRNA half-lives. In addition, it has been suggested that a fraction of ADAR binding events might be "non-productive", i.e., these bindings may not trigger any RNA editing [21]. Consistent with this hypothesis, we have also observed a clear discrepancy in the predicted ADAR binding patterns between "non-productive"

and "productive" binding behaviors, which may indicate the existence of different ADAR binding modes to accomplish its diverse regulatory functions. Moreover, we applied DeBooster to study the antagonizing effect of RBP binding on miRNA repression. In particular, it has been known that in the $3'$ UTR of the oncogene *ERBB2*, the RBP ELAVL1 (also called HUR) antagonizes the repression effect of the miRNA miR-331-3p by binding to a U-rich element (URE) near the miRNA target region called miR-331b [22]. With a mutant URE, we have observed that the new ELAVL1 binding sites predicted by DeBooster shift to a position more distant from the miR-331b region, which is largely consistent with the previous experimental studies. At last, we have used DeBooster to predict the effects of the single nucleotide variant (SNV) mutations on the RBP binding sites related to splicing events, which may provide useful hints for identifying pathogenic mutations and investigating their connections to the pathogenesis of complex diseases. Based on these test results, we expect that DeBooster will have great application potentials and be widely used by the community to analyze more CLIP-seq experimental data and discover more biologically relevant findings on the functional roles of RBPs in post-transcriptional gene regulation.

2 Results

2.1 The DeBooster framework

We have developed a deep boosting based approach, called DeBooster, to predict the sequence specificities of RNA-binding proteins (RBPs) from high-throughput CLIP-seq data (Fig 1). As RNA primary sequence can be viewed as a string over the alphabet {A, U, C, G}, we mainly use the basic bag-of-words model [23] as in the nature language processing field to encode the features of a given RNA sequence (Fig 1a). In particular, for each word of fixed length k, we count how many times it appears in the RNA sequence and store its frequency information in a vector of length 4^k. We extract the word frequency information for both an RBP target region and its upstream and downstream flanking regions of 150 nucleotides each. We consider words of lengths 1, 2, 3, which results in $2 \times (4 + 4^2 + 4^3) = 168$ features in total.

Note that the bag-of-words model mainly focuses on the occurrences of words and reflects little about the order of the letters in a sequence. In other words, if we swap the first half and the second half

of an RNA sequence, the features provided by the bag-of-words model would roughly remain the same. To better incorporate the order of letters into the model, we futher use the following scheme to extract the 'second-order' word count information. For a fixed stride m and a given RNA sequence $a_1a_2\cdots a_t$, we count the words a_1a_{m+1}, a_2a_{m+2}, $\ldots$, $a_{t-m}a_t$ and use a vector to record the corresponding count information. As before, we also consider both an RBP target region and the flanking regions of 150 nucleotides both upstream and downstream. We consider the stride lengths 4, 5 and 6, which generates $2 \times 3 \times 4^2 = 96$ more features in total. Moreover, we consider five additional features, such as the length of the target region, whether the word length is a multiple of 3, whether the target region contains the stop codons UAG, UAA and UGA. Thus, overall we extract $168 + 96 + 5 = 269$ features for a given RNA sequence.

We then apply a deep boosting based method, to learn a classification model from the above encoded features (Fig 1b). The deep boosting method [24] is a generalization of several well-established learning approachs, such as AdaBoost [25] and logistic regression [26]. It uses decision tree models as base classifiers and partitions the decision trees of different depths into different sets, denoted by $H_1, \cdots, H_k$, respectively, where H_i stands for the set of decision trees of depth k. In our deep boosting method, we aim to learn a classifier from a family of convex ensembles $f = conv(\bigcup_{i=1}^{k} H_i)$. That is, f can be written in the form of $f(x) = \sum_{t=1}^{n} \alpha_t h_t(x)$, where $\alpha_t \geq 0$, and $h_t \in H_{p_t}$ for some $p_t \in [1, k]$. During the training process, the deep boosting method seeks to minimize the following objective function:

$$E = \frac{1}{m}\sum_{i=1}^{m} \Phi(1 - y_i \sum_{t=1}^{n} \alpha_t h_t(x_i)) + \sum_{t=1}^{n} (\lambda r_t + \beta)\alpha_t,$$

where (x_i, y_i) denotes the i-th training sample, m strands for the total number of training samples, Φ stands for the loss function (e.g., the exponential cost function as in AdaBoost [25] or the logitstic function [26]), r_t stands for the Rademacher complexity of set H_{p_t}, and λ and β are two hyperparameters to be chosen. The above objective function can be optimized as in other boosting algorithms [25, 27]. After training, the learned model can be used to predict the sequence specificities and investigate the corresponding binding motifs of the RBP targets (Fig 1b, Methods).

2.2 DeBooster captures the sequence preferences of RBP binding

We first ran a 10-fold cross-validation procedure for each of 24 CLIP-seq datasets (Methods) to evaluate the overall prediction performance of DeBooster. The hyperparameters in the deep boost-

ing framework were determined using an independent dataset (Methods). We also compared the performance of DeBooster with the state-of-the-art approaches for predicting RBP target sites, including GraphProt [17] and the deep belief net (DBN) method [19]. The comparison results (Figs 1a-1c) showed that DeepBooster can significantly outperform both GraphProt and the DBN method, with the increase of the area under receiver operator characteristic curve (AUROC) by up to 10.1%. Note that GraphProt and the DBN method integrate both RNA sequence and structural information (i.e., RNA secondary structural information [17] or both RNA secondary and tertiary structural profiles [19]) into the prediction framework, while DeBooster requires only RNA sequence information. The performance improvement in DeBooster was probably attributed to our new feature encoding scheme (see Fig 1a and Section 2.1) and the better predictive power of the underlying deep boosting model.

Through a transcriptome-wide analysis on RBP binding targets, we also found that the difference in the predicted binding scores of DeBooster over different characterized genomic regions mostly reflected the known functions of individual RBPs (Supplementary Notes). In addition, we examined the sequence motifs of the RBP binding sites generated from training data (Methods). Our results indicated that the sequence motifs resulting from DeBooster agreed well with those reported in the literature (Fig 2d). For example, the binding sequence motif of AGO2 computed by DeBooster was enriched with A, U and C but depleted of G, which was consistent with the previous study [28]. PTB, as indicated by its name (polypyrimidine tract-binding protein), mainly binds to the U/C-rich regions [29], which was also reflected in the sequence motif derived from DeBooster. EWSR1, FUS and TAF15 belong to the FET family. Although several works showed that they bind to the GU-rich motif [30, 31], recent studies found that the FET protein family prefers binding to the AU-rich stem loops, and the AU-rich sequences achieve higher binding affinities than those enriched with G and U [32]. Such an AU-rich pattern was also observed in the sequence motif generated by DeBooster. It has been found that the binding targets of QKI usually contain a core sequence NACUAAY (where Y stands for a pyrimidine) and a half-site UAAY [33]. The biniding motif of QKI identified by DeBooster also agreed well with such a pattern. DeBooster yielded a U-rich sequence motif for the binding sites of HNRNPC, which can also be supported by a known fact that HNRNPC generally binds to the poly-U tracts [34]. According to the DeBooster prediction results, SFRS1 prefers binding to a GA-rich motif, which aligned well with the previous result [35]. As shown in the previous study [7], PUM2 binds to a consensus motif UGUANAUA, which shared high similarity with the corresponding binding motif

predicted by DeBooster. The majority of the TDP43 binding sites predicted by DeBooster contained the $(UG)_n$ motif and was relatively less enriched with A and C. Such an observation agreed well with the previous known result [36]. Taken together, most of the sequence motifs of RBP binding sites captured by DeBooster were consistent with the previous known results in the literature.

2.3 The predictions of DeBooster can be validated through cross-platform datasets

It is well-known that different CLIP-seq experiments can yeild a large fraction of non-overlapping results and individual experiments may miss a vast number of true RBP binding sites [37,38]. Here, we showed that the prediction results of DeBooster can be validated through cross-platform datasets and thus effectively alleviate the false negative problem in current high-throughput CLIP-seq data (Fig 3). In particular, we tested DeBooster on different cross-platform ELAVL1 datasets, which displayed a large degree of discrepancy between the original RBP binding targets measured from CLIP-seq experiments (Fig 3a). Such a large variation indicated that in general a single CLIP-seq experiment cannot cover all RBP binding sites and individual datasets may have high false negative rates in current experimental measurement. The tests on the cross-platform ELAVL1 datasets showed that the predictions of DeBooster from one dataset can be well validated by another one collected from a different platform, achieving both high AUROC scores and similar sequence motifs (Fig 3b). In addition, most of the sequence features encoded in DeBooster displayed highly correlated weights except the outliers G and UNNNNU (Figs 3c-3d), which was probably due to experimental bias introduced from the original CLIP-seq data. Thses results implied that the predictions of DeBooster can be well validated through cross-platform CLIP-seq datasets.

We also investigated the agreement of the DeBooster prediction results between different RBPs from the same family. In particular, we examined the consistency between the DeBooster prediction scores of 8-mers for TAF15, FUS and EWSR1, all belonging to the FET family. Consistent with the previous results that these three RBPs have a large overlap in binding sites [32], our tests showed that the 8-mers from different RBPs exhibited highly correlated prediction scores (Figs 3e-3f). Such observations further supported the above argument that the prediction results of DeBooster can be verified from cross-platform CLIP-seq datasets, even for different RBPs from the same family. These results suggested that DeBooster was not prone to overfitting, and may provide a practically useful tool to analyze high-throughput CLIP-seq data and recover false negatives that are common in current

CLIP-seq data.

2.4 The binding scores predicted by DeBooster match the experimentally measured binding affinity data

To investigate whether the prediction results of DeBooster can truely reflect the RBP binding preferences, we further checked the agreement between the binding scores predicted by DeBooster and the experimentally determined binding affinity data. In particular, we compared the predicted binding scores with both *in vivo* determined K_d values [39,40] and *in vitro* measured binding affinities from RNAcompete assays [10,11].

We first checked the agreement between the prediction scores of DeBooster, which was trained using the *in vivo* CLIP-seq data, and the experimentally determined K_d values for two RBPs, including SFRS1 and TDP43 (Figs 4a-4b). Our comparison showed that for the 8-mers as the potential RNA targets of SFRS1, the prediction scores of DeBooster closely matched the *in vivo* measured K_d values [39] (Fig 4a). In addition, for the RNA nucleotides as the potential binding targets of TDP43, the prediction scores of DeBooster aligned well with the K_d values experimentally measured from the electrophoretic mobility shift assay (EMSA) [40] (Fig 4b).

Next, we examined the prediction performance of DeBooster on the *in vitro* binding data derived from the RNAcompete experiments [10,11]. The original version of DeBooster took binary (i.e., positive or negative) samples as input. To enable the model to consider the real-valued RNAcompete data, we also made a simple extention and proposed a "regression" version of DeBooster (Methods). A cross-validation test (Methods) showed that the prediction scores of DeBooster and the *in vitro* binding scores from the RNAcompete assays [10, 11] for the synthesized oligonucleotides displayed high correlations. In particular, the comparisons to the earlier version of the RNAcompete data [10] for the synthesized 7-mers of nine RBPs exhibited high consistency, in which most of Pearson correlation efficients were above 0.7 (Fig 4c). In addition, the comparisons to the latest version of the RNAcompete data [11] for three RBPs, including FUS, QKI and HNRNPC, showed good agreement between the prediction scores and the *in vitro* binding affinity scores, with Pearson correlation coefficients above 0.65 (Figs 4d-4f). Overall, these comparison results implied that the binding scores predicted by DeBooster may be regarded as a useful indicator of RBP binding affinities for both *in vivo* and *in vitro* scenarios.

2.5 The predicted targets of RNA helicases may be associated with the regulation of mRNA degradation

RNA helicases, such as MOV10, regulate the life cycle of mRNAs and thus gene expression by remodeling RNA secondary structures and RNA-protein interactions [41]. Here, we showed that the RNA targets of MOV10 predicted by DeBooster can be connected to the regulation of mRNA half-lives and thus may provide useful hints for understanding the funcitonal roles of MOV10 in controlling gene expression. Our analysis was performed on a set of 7000 mRNAs, in which the fold changes of their half-lives had been measured after MOV10 knockdown [42]. These mRNAs were basically divided into four groups according to the fold changes of their half-lives, i.e., top 25%, 25%-50%, 50%-75% and bottom 25%, which corresponded to Group 1, Group 2, Group 3 and Group 4, respectively. Only the bottom group (i.e., Group 4) contained those genes whose expression levels were unchanged or up-regulated after MOV10 knockdown.

Compared to the results derived directly from the original CLIP-seq data (Fig 5a), the fraction of UTRs with MOV10 binding resulting from DeBooster prediction displayed a more evident decreasing trend (Fig 5b). In addition, the sum of all positive prediction scores per UTR, which basically considered both binding strength and the number of hits for the MOV10 binding targets on individual genes, also exhibited the same decreasing order for four groups of genes that were divided and ranked according to the fold changes of mRNA half-lives (Fig 5c). Moreover, when we grouped all transcripts according to the DeBooster prediction scores, the resulting fold changes of mRNA half-lives also presented a similar decreasing trend (Fig 5d). Furthermore, the DeBooster predition scores for seven genes also showed good agreement with the fold changes of mRNA half-lives experimentally measured by qRT-PCR (Fig 5e). Taken together, the above results demonstrated that the binding targets of the RNA helicase MOV10 predicted by DeBooster were associated with the changes of mRNA half-lives. Thus, the prediction results from DeBooster may provide useful clues for further understanding the regulatory machanisms of RNA helicases on the life cycle of mRNAs.

2.6 Applying DeBooster to study the difference between productive and non-productive ADAR binding patterns

ADARs are a family of homologous enzymes catalyzing adenosine-to-inosine (A-to-I) editing in the RNA, and have similar double-stranded RNA binding domains (dsRBDs) and a common deaminase domain [43]. Despite their major role as RNA-editing enzymes, a fraction of ADAR binding events might be "non-productive", that is, these bindings might not trigger any RNA editing [21]. On the contary, those ADAR binding events that indeed produce RNA editing were considered "productive". To investigate the difference of the binding behaviors between productive and non-productive ADAR binding targets, we compared the prediction results from three DeBooster models, which were trained using all, productive and non-productive ADAR binding sites, respectively.

We first introduced the concept of the binding-editing distance, which was defined as the genomic distance between an ADAR binding position and its closest editing site. The known RNA editing sites were obtained from the RADAR database [44]. Our first model, also called the all-binding model, was trained using all ADAR1 binding sites measured from CLIP-seq experiments [45] as the postive samples. The negative samples were defined as those unbound regions that were adjacent to the positive samples in transcripts and had the lengths equal to those of the corresponding positve samples. In our second model, also called the productive binding model, the CLIP-seq sites (i.e., the ADAR1 binding sites measured from CLIP-seq experiments) with small binding-editing distances ($0-100$ nt) were used as the positive samples, while the CLIP-seq sites with large binding-editing distances (>1000 nt) together with the adjacent unbound regions were used as the negative samples. In our third model, also called the non-productive binding model, the CLIP-seq sites with large binding-editing distances (>1000 nt) were used as the positive samples, while the CLIP-seq sites with small binding-editing distances ($0-100$ nt) together with the adjacent unbounded regions were used as the negative samples. The median of the binding-editing distances resulting from the all-binding model was 814 nt (Fig 6a), which was roughly on the same scale as from the original CLIP-seq data (606 nt). The median of the binding-editing distances from the productive binding model was zero (i.e., the ADAR binding region contained at least one editing site), which was significantly different from that of the non-productive binding model (4665 nt, Fig 6a). The above results implied that indeed DeBooster may be able to distinguish productive and non-productive ADAR binding sites based on current high-throughput CLIP-seq data.

We also examined the sequence motifs of the ADAR binding sites identified by three different

DeBooster models (Fig 6b). Although all three sequence motifs showed high GC content, the motif generated by the productive-binding model had relatively higher frequencies of As and Us than those from the other two models. This observation indicated that those ADAR binding sites with relatively lower GC content might be more prone to being edited. This result was also in agreement with the known evidence that the published motif of the ADAR binding sites [45] contained relatively higher GC content than that of the genomic regions near the editing sites [46].

As in previous studies [21], our results showed that some ADAR binding sites were close to the editing sites, while many others were thousands of nucleotides away from the editing sites. Although it was possible that this phenomenon was attributed to the lack of the complete editing records in the database, the clear discrepancy of the binding-editing distances and the sequence motifs of the ADAR binding sites between productive and non-productive models derived from DeBooster indicated that there may exist different binding modes to accomplish diverse regulatory effects of ADAR binding.

2.7 The shift of the predicted RBP binding scores from mutations may predict the antagonizing effect of RBP binding on miRNA repression

RBPs and miRNAs are two classes of essential regulators controlling mRNA degradation and expression, and they often interplay with each other to display co-regulatory effects [47]. For example, in the 3′ UTR of an oncogene *ERBB2*, the RBP ELAVL1 (also called HUR) antagonizes the repression effect of the miRNA miR-331-3p by binding to a U-rich element (URE) near the miRNA target region called miR-331b [22]. With a mutant URE, the repression effect of ELAVL1 binding on miR-331-3p is weakened, since the new ELAVL1 binding sites shift to a position that is more distant from the miR-331b region (Fig 7a), and also reduces the binding affinity of ELAVL1 (the magnitude of the experimentally measured K_d values change from 10^{-8} M to 10^{-7} M) [22]. Here, we showed that DeBooster can successfully identify this mutational effect that was consistent with the previous experimental observation.

We used the CLIP-seq dataset of ELAVL1 measured from the Hela cells [48] as training data (those overlapping records about the measured binding sites in the 3′ UTR of gene *ERBB2* were removed) and performed a comparative study on the predicted binding scores of four cases, i.e., WT-URE/WT-331b, MT-URE/WT-331b, WT-URE/MT-331b and MT-URE/MT-331b, which represented the wild-type sequence, a URE mutant with the wild-type miR-331b region, the wild-type URE with a miR-331b

mutant, and a sequence with mutations in both URE and miR-331b regions, respectively (Fig 7b). All the binding scores predicted by DeBooster showed obvious peaks near the URE, indicating the high-affinity binding of ELAVL1 in this region. More importantly, the prediction results of DeBooster displayed a clear position-shifted and affinity-decreased pattern of ELAVL1 binding on a URE mutant (Fig 7b). The curves of the predicted binding scores for WT-URE/WT-331b (i.e., wild-type) and WT-URE/MT-331b (i.e., only mutations in the miR-331b region) had similar shapes, which was consistent with the previous experimental result that the mutations in the miR-331b region rarely affect ELAVL1 binding [22]. In addition, the peaks of these two curves were approximately located in positions 50-90 along the $3'$ UTR of ERBB2, while the peaks of the other two curves with mutations in the URE region (i.e., MT-URE/WT-331b and MT-URE/MT-331b) were located around positions 45-60. Such a position shift of the ELAVL1 binding sites identified by DeBooster in fact agreed with the previous experimental RNA footprinting results (see Fig 7b in [22]). Morever, the decrease of the binding scores predicted by DeBooster was also consistent with the loss of the experimentally-determined K_d values with respect to the same mutations [22]. Taken together, these results indicated that DeBooster can successfully identify the changes of the RBP binding scores caused by the mutations in binding targets which may be used to predict the antagonizing effect of RBP binding on miRNA repression.

2.8 The prediction scores of DeBooster may provide a useful index to study pathogenic mutations affecting RNA splicing

Recent studies revealed that abnormal splicing play a vital role in development of many human diseases, such as cancer and neurological disorders [49–51]. The mutations near splice sites or on splicing regulatory elements, such as exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs), may influence RNA splicing and cause human diseases by disrupting RBP binding [2]. Here, we were particularly interested in whether DeBooster can identify pathogenic mutations and thus be used as a useful tool to study the mutational effects of sequence variants related to splicing events. We first examined the overall changes of the predicted binding scores of individual RBPs with respect to the sequence variants of their binding targets near $5'$ and $3'$ splice sites (Methods) and checked whether DeBooster was able to distinguish pathogenic mutations from neutral sequence variants. Our comparisons showed that the changes of the binding scores predicted by DeBooster for a majority of pathogenic sequence variants in regions near $5'$ and $3'$ splice sites were significantly different from those

of neutral mutations (Fig 8). In addition, almost all of these pathogenic mutations displayed relatively larger changes in the predicted binding scores than neutral variants. On the other hand, most of the neutral mutations near $5'$ and $3'$ splice sites displayed similar effects (with only 4 among 20 RBPs showing significant difference with $p<0.001$ in the Student's t test). Furthermore, the pathogenic mutations near splice sites generally showed a greater extent of difference in the predicted binding scores than those pathogenic mutations randomly chosen from the COSMIC records [52] (Supplementary Fig 2). For instance, among 20 RBPs, 15 and 18 proteins exhibited significantly different mutational effects on the pathogenic variants near $5'$ and $3'$ splice sites, respectively, compared to only 7 RBPs in those pathogenic mutations randomly selected from COSMIC (Fig 8 and Supplementary Fig 1). Such an observation implied that the seqeuence disruptions of the RBP binding targets around splice sites may generally play a more important role in the pathogenisis of a disease. Overall, our studies indicated that the binding scores derived from DeBooster may provide an effective indicator for distinguishing pathogenic mutations from neutral variants in RBP binding targets near splice sites.

Next, we further analyzed the mutational effects predicted by DeBooster for a number of known pathogenic single-nucleotide variants (SNVs) obtained from COSMIC [52]. Below we describe several examples (Fig 9). First, a synonymous substitution of the last base in Exon 7 (G to A) of gene *CDH1* (which encodes the E-cadherin protein) led to an increase in the SFRS1 binding scores predicted by DeBooster near a 5 splice site (Fig 9a), which may be related to the dysregulation of *CDH1* that causes tumor metastasis [53]. Such an observation may also be supported by a previous experimental validation study that this mutation can actually alter splicing by causing intron retention to various extents [54].

As a second example, a mutation from G to A in a *TCF7L2* exon [55] disrupted the ESE motifs (which are 6 nt motifs located in exons and bound by SR proteins to promote exon splicing [56]) and suppressed SFRS1 binding (Fig 9b), while a mutation from U to A in a *THRAP3* exon [55] enriched the ESE motifs and thus enhanced SFRS1 binding (Fig 9c). Such disruptions in those disease-relevant genes may influence the binding behaviors of the important splicing regulator SFRS1, and thus may be related to the tumorgenesis associated with aberrant splicing [39].

In our third example, the mutation from U to C near a 3 splice site of gene *TRRAP* [55] weakened TIA1 binding (Fig 9d). TRRAP interacts with oncoproteins MYC and E2A [57], and its mis-regulation can be heavily related to various types of cancers [58]. On the other hand, another mutation from C to U near a 3 splice site of gene *KTN1* [55] strengthened TIA1 binding (Fig 9e). *KTN1* encodes kinectin

1, and has been shown to display different splicing patterns in cancers [59]. Thus, these two sequence variants in the binding sites of TIA1 may be associated with cancer pathogenesis by changing the alternative splicing modes of its target genes.

Another interesting example is the intronic mutation near a 5 splice site of gene *ATM* [55], which increased the binding scores of both FUS and QKI (Fig 9f). Such a mutation may influence the splicing result of this tumor suppressor (i.e., ATM) [60] by creating new potential binding sites for both splicing regulators (i.e., FUS and QKI).

In addition to the above cases, there were other examples to demonstrate that the prediction scores of DeBooster may reflect the pathogenic effects of sequence disruptions in RBP binding. For instance, a substitution from C to U near a 5 splice site of gene *NF1* [55] enhanced HNRNPC binding (Supplementary Fig 3a), which may be associated with the known related neurologic disorders [61]. On the other hand, a mutation from U to C near a splice site of the proto-oncogene *BRAF* [55] decreased the HNRNPC binding score (Supplementary Fig 3b). In addition, a mutation from A to G [55] near a splice site of gene *TET2* may help form a novel GU-repeat region for strong TDP43 binding (Supplementary Fig 3c), and thus influence the splicing process. Moreover, the *SMAD4* splicing site may be disrupted by the mutation from G to U [55] that may increase the PTB binding score (Supplementary Fig 3d) and thus alter the corresponding splicing result. Both *TET2* and the *SMAD4* genes act as tumor suppressors [62, 63], so the inhibition of their normal splicing may thus facilitate cancer formation.

Taken together, the above examples illustrated that the RBP binding scores predicted by DeBooster may offer a useful index to investigate the pathogenic effects of sequence disruptions related to RNA splicing.

3 Conclusion

We developed DeBooster, a deep boosting based framework to model the sequence binding specificities of RNA-binding proteins (RBPs) from high-throughput CLIP-seq data. Compared to the state-of-the-art methods which usually require both sequence and structure profiles, DeBooster uses only sequence information as input. Tests on 24 CLIP-seq datasets demonstrated that DeBooster can achieve better prediction performance than previous methods. Through a validation test on several

cross-platform CLIP-seq datasets of ELAVL1, we showed that DeBooster can be useful for addressing the false negative problem that is prevalent in current CLIP-seq data. In addition, the prediction scores of DeBooster agreed with the experimentally-determined binding affinity scores, such as *in vivo* measured K_d values and the *in vitro* binding affinities measured from RNAcompete.

We further showed the great application potentials of DeBooster by applying it to study the regulatory roles of several important RBPs. In particular, we demonstrated that the predicted targets of the RNA helicase MOV10 can better explain its binding effects on the regulation of mRNA degredation than the original CLIP-seq data. In addition, the predicted RBP binding sites may help understand the difference between productive and non-productive binding patterns of the RNA-editing enzymes ADAR. We also showed that a shift of the predicted ELAVL1 binding scores from wild-type to mutant in a U-rich element (URE) region of gene *ERBB2* can effectively predict the antagonizing effect of RBP binding on miRNA regulation. Moreover, DeBooster may be used as an effective index to identify pathogenic mutations from normal sequence variants and study the effects of potential disease-causing mutations in RBP binding sites related to splicing. Based on these test results and analyses, we expect that DeBooster will provide a promising tool to analyze more large-scale CLIP-seq data and gain more biological insights related to RBP regulation.

4 Methods

4.1 Datasets

We used 24 sets of CLIP-seq based data about RBP binding sites to train and validate our prediction model. These datasets were preprocessed in [17] to generate both positive and negative samples. The list of all RBP names in these 24 CLIP-seq datasets can also be found in Fig 2a. Among these datasets, AGO1-3 and IGF2BP1-3 contained the binding targets of several RBPs from the same protein family, while ELAVL1 HITS-CLIP, ELAVL1 PAR-CLIP(A), ELAVL1 PAR-CLIP(B) and ELAVL1 PAR-CLIP(C) included the binding sites of RBP ELAVL1 measured from different experimental platforms.

4.2 Determination of hyperparameters

We use an independent validation dataset of RBP C22ORF28 to determine the optimal setting of the hyperparameters of DeBooster, including the type of the loss function (denoted by Φ), the number of the base decision tree classifiers (denoted by n), the maximun depth of these decision trees (denoted by k), and parameters λ, β controlling the relative importance of the complexity penalty. This process yields the following optimal setting of the hyperparameters: the exponential function as the loss function Φ, $n = 200$, $k = 5$, $\lambda = 0.3$ and $\beta = 0$.

4.3 Motif generation

We use the following procedure to generate representive motifs of the RBP binding sites predicted by DeBooster. First, we use the set of the weighted decision trees resulted from the deep boosting algorithm to evaluate the relative importance of each encoded feature. In particular, for each decision tree with weight ω in the model, we identify the feature ψ and the corresponding threshold τ used to split the root node for this feature. Suppose that at the root node a fraction p_1 of all examples in the training set are postive, and at the right child of the root node (in which the value of feature ψ is larger than τ), a proportion p_2 of all examples in the training set are positive. We then use $(p_1 - p_2)\omega$ to represent the importance of feature ψ. By doing so, we score each feature based on its contribution to RBP binding. A higher absolute value of a positive score means higher contribution to RBP binding, while a higher absolute value of a negative score means less contribution to RBP binding. We use a vector s to store the importance scores of all encoded features. Next, we go through all 8-mers and extract the feature vector v_i for each of them. We then rank these 8-mers according to the inner product of v_i and s, and we select the top 500 8-mers with the highest ranking scores. As the top 8-mers may come from shifts around the best one, we align all 8-mers with respect to the top one such that the largest number of base matchings is achieved. After that, we generate the binding motif based on this alignment step and visualize it using the WebLogo site [64].

4.4 An extension of DeBooster to real-value labeled data

Although the original version of DeBooster is designed to take only binary (i.e., positive or negative) labeled examples in training data, its output value is nevertheless real-valued, which offers us the possibility to adapt it to handle a regression problem.

To be specific, suppose the training set S for the regression task is $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where each target value (i.e., the label) y_i is real-valued. We now construct a new training set $\widetilde{S}$ as follows: for each $(x_i, y_i) \in S$, we add an example $(x_i, \tilde{y_i})$ into $\widetilde{S}$, where $\tilde{y_i}$ takes value 1 with probability p_i, -1 with probability $1 - p_i$ and $p_i = \exp(2y_i)/(1 + \exp(2y_i))$.

In this way, the new training set $\widetilde{S}$ only consists of examples with binary labels (i.e., either 1 or -1), and thus is suitable for classification. In addition, we assert that the ground truth value y_i should minimize the loss in expectation. Therefore, ideally the output value predicted by the model for x_i after training using dataset $\widetilde{S}$ should be equal to y_i.

To see that, recall that our model uses the exponential function as the loss function Φ. Hence, if the original deep boosting model outputs t_i for the ith example, the expected loss is given by:

$$\begin{aligned} E[loss(t_i, \tilde{y_i})] &= E[\Phi(1 - \tilde{y_i} t_i)] \\ &= p_i \cdot \Phi(1 - t_i) + (1 - p_i) \cdot \Phi(1 + t_i) \\ &= p_i \cdot \exp(1 - t_i) + (1 - p_i) \cdot \exp(1 + t_i) \end{aligned}$$

We seek a variable t_i that minimizes the above expected loss by setting the derivative to 0, i.e.,

$$\frac{\partial E[loss(t_i, \tilde{y_i})]}{\partial t_i} = -p_i \cdot \exp(1 - t_i) + (1 - p_i) \cdot \exp(1 + t_i) = 0$$

Plugging in $p_i = \exp(2y_i)/(1+\exp(2y_i))$, we can see that $t_i = y_i$ satisfies the above equation, which means that $t_i = y_i$ minimizes the expected loss. Therefore, in an ideal case the model will output y_i on an input x_i after training based on $\widetilde{S}$.

We also consider several practical issues for the above regression version of DeBooster. First, in order to handle different training sets whose distributions of the target values y's may be largely different, we first preprocess the y's and transform them into a Gaussian distribution. To do so, all y's are sorted and the ith y is then transformed into $(2 * i - 1)/(2n)$, where n is the number of training examples. In this way, these transformed y's will have a uniform distribution in $[0, 1]$. Hence it can be further transformed into a Gaussian distribution in a standard way.

Besides, as the output range of the deep boosting model is only within $[-1, 1]$, it would only make sense to guarantee the transformed y's to mainly fall in this range. Therefore, we choose to transform the y into a Gaussian distribution with zero mean and standard deviation of 0.4. Notice that $1/0.4 = 2.5$, which means that we treat those target values y's that are outside 2.5 standard

deviations as outliers and force them into $[-1, 1]$ by assigning them with value 1 or -1.

To evaluate the performance of the regression, we first split the dataset into 70% for training and 30% for testing. A transformation function f is first learned from the training set, and then used to predict the value t for each example in the test set. After that, the Pearson coefficient between t's and and $f(y)$'s is used to evaluate the regression performance.

4.5 Predicting ELAVL1 binding scores along the 3′ UTR of gene *ERBB2*

Both wild-type and mutant 3′ UTR sequences were obtained from [22] (Supplementary Notes). The lengths of these sequences are all 119 nt. For each sequence, we took a window of length 41 nt (the average length of the ELAVL1 target regions over training samples) and slided this window along the 3′ UTR of mRNA ERBB2 with a stride length of 1 nt. For each sliding window, we assigned the resulting prediction score to the central nucleotide of this window. Overall, we obtained the prediction scores along positions 21-99 for each sequence (Fig 7), and the first and last 20 nucleotides were not included in our analysis.

4.6 Studying the effects of mutations in RBP binding targets

The mutation data related to splicing events were derived from COSMIC [52]. Sequences with mutation sites in the middle and lengths equal to those of the corresponding RBP binding targets were prepared as input samples to DeBooster. For both pathogenic or neutral mutations near 5′ or 3′ splice sites, we selected those single-nucleotide variant (SNV) mutations within 10 nt from splice sites. The lengths of RBP binding targets are usually larger than 20 nt, so generally splice sites were covered by samples centered at mutation positions. In total, we collected 7,000 neutral mutations in both regions near 5′ and 3′ splice sites, and 4,000 and 20,000 mutations in regions near 5′ and 3′ splice sites, respectively. In Fig 8, the change of the prediction score resulting from a mutation was calculated as "(prediction score for the mutant sequence)−(prediction score for the wild-type sequence)".

In Fig 9 and Supplementary Fig 3, the prediction scores for regions around the mutation sites along both wild-type and mutant sequences were shown. For each selected mutation, we showed the prediction scores for 41 positions, including the mutation site and the flanking regions of 20 nucleotides both upstream and downstream. For each site, its prediction score was calculated using the window

centered at this position and of length equal to the average length of the corresponding RBP targets in the training data.

Acknowlegements

This work was supported in part by the National Basic Research Program of China Grant 2011CBA00300 and 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003 and 61472205, US National Science Foundation grants DBI-1262107 and IIS-1646333, China's Youth 1000-Talent Program and the Beijing Advanced Innovation Center for Structural Biology. The authors are grateful to Dr. Qiangfeng Zhang and Mr. Hailin Hu, Mr. Bin Zhou, and Mr. Xuan He for their helpful discussions about this work.

Author contributions

S.L., F.D, Y.W. and J.Z. conceived the research project. J.Z. supervised the research project. F.D. and Y.W. designed and implemented DeBooster, and carried out model training and validation tasks. S.L., F.D., S.Z., C.Z., T.J., X.L. and J.Z. performed the computational and statistical analysis. S.L., F.D. and J.Z. wrote the manuscript. All the authors discussed the test results and commented on the manuscript.

Competing financial interests

The authors declare no competing financial interests.

References

[1] Tina Glisovic, Jennifer L Bachorik, Jeongsik Yong, and Gideon Dreyfuss. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, 582(14):1977–1986, 2008.

[2] Marina M Scotti and Maurice S Swanson. RNA mis-splicing in disease. *Nature Reviews Genetics*, 17(1):19, 2016.

[3] Albert E Dahlberg, C Wesley Dingman, and Andrew C Peacock. Electrophoretic characterization of bacterial polyribosomes in agarose-acrylamide composite gels. *Journal of Molecular Biology*, 41(1):139, 1969.

[4] J Czworkowski, O W Odom, and B Hardesty. Fluorescence study of the topology of messenger RNA bound to the 30S ribosomal subunit of escherichia coli. *Biochemistry*, 30(19):4821–4830, 1991.

[5] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464, 2008.

[6] Julian Konig, Kathi Zarnack, Gregor Rot, Toma Curk, Melis Kayikci, Bla Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, 17(7):909, 2010.

[7] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Annacarina Jungkamp, Mathias Munschauer, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010.

[8] Andrew D Ellington and Jack W Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346(6287):818, 1990.

[9] Regina Stoltenburg, Christine Reinemann, and Beate Strehlitz. SELEX-A (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomolecular Engineering*, 24(4):381, 2007.

[10] Debashish Ray, Hilal Kazan, Esther T Chan, Lourdes Pena Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J Blencowe, Quaid Morris, and Timothy R Hughes. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*, 27(7):667, 2009.

[11] Debashish Ray, Hilal Kazan, Kate Cook, Matthew T Weirauch, Hamed Shateri Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Ally Yang, Hong Na, et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172, 2013.

[12] Paula H Reyesherrera and Elisa Ficarra. Computational methods for CLIP-seq data processing. *Bioinformatics and Biology Insights*, 2014(8):199–207, 2014.

[13] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin C Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William Stafford Noble. MEME suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37, 2009.

[14] Barrett C Foat, Alexandre V Morozov, and Harmen J Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Nucleic Acids Research*, 22(14), 2006.

[15] Michael Hiller, Rainer Pudimat, Anke Busch, and Rolf Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Research*, 34(17), 2006.

[16] Hilal Kazan, Debashish Ray, Esther T Chan, Timothy R Hughes, and Quaid Morris. RNAcontext: A new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLOS Computational Biology*, 6, 2010.

[17] Daniel Maticzka, Sita J Lange, Fabrizio Costa, and Rolf Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biology*, 15(1), 2014.

[18] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831, 2015.

[19] Sai Zhang, Jingtian Zhou, Hailin Hu, Haipeng Gong, Ligong Chen, Chao Cheng, and Jianyang Zeng. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Research*, 2015.

[20] Emanuele Buratti and Francisco E Baralle. Influence of RNA secondary structure on the pre-mRNA splicing process. *Molecular and Cellular Biology*, 24(24):10505, 2004.

[21] Yvonne Klaue, Annika M Kallman, Michael Bonin, Wolfgang Nellen, and Marie Ohman. Biochemical analysis and scanning force microscopy reveal productive and nonproductive ADAR2 binding to rna substrates. *RNA*, 9(7):839–846, 2003.

[22] Michael R Epis, Andrew Barker, Keith M Giles, Dianne J Beveridge, and Peter J Leedman. The RNA-binding protein HuR opposes the repression of ERBB-2 gene expression by microRNA mir-331-3p in prostate cancer cells. *Journal of Biological Chemistry*, 286(48):41442–41454, 2011.

[23] Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[24] Corinna Cortes, Mehryar Mohri, and Umar Syed. Deep boosting. *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[25] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. 1995.

[26] Strother H Walker and David B Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54:167–179, 1967.

[27] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[28] Jingjing Li, Taehyung Kim, Razvan Nutiu, Debashish Ray, Timothy R Hughes, and Zhaolei Zhang. Identifying mRNA sequence elements for target recognition by human argonaute proteins. *Genome Research*, 24(5):775–785, 2014.

[29] Yuanchao Xue, Yu Zhou, Tongbin Wu, Tuo Zhu, Xiong Ji, Youngsoo Kwon, Chao Zhang, Gene W Yeo, Douglas L Black, Hui Sun, et al. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Molecular Cell*, 36(6):996, 2009.

[30] Ana Lerga, Marc Hallier, Laurent Delva, Christophe Orvain, Isabelle Gallais, J Marie, and Francoise Moreaugachelin. Identification of an RNA binding specificity for the potential splicing factor TLS. *Journal of Biological Chemistry*, 276(9):6807, 2001.

[31] Kentaro Takahama, Shigeki Arai, Riki Kurokawa, and Takanori Oyoshi. Identification of Ewing's sarcoma protein as a G-quadruplex DNA- and RNA-binding protein. *FEBS Journal*, 278(6):988–998, 2011.

[32] Erik G Larsson, Simon Runge, Jeffrey D Nusbaum, Sujitha Duggimpudi, Thalia A Farazi, Markus Hafner, Arndt Borkhardt, Chris Sander, and Thomas Tuschl. RNA targets of wild-type and mutant FET family proteins. *Nature Structural and Molecular Biology*, 18(12):1428, 2011.

[33] Andre Galarneau and Stephane Richard. Target RNA motif and target mRNAs of the Quaking STAR protein. *Nature Structural and Molecular Biology*, 12(8):691, 2005.

[34] Zuzana Cienikova, Fred F Damberger, J Hall, Frederic H T Allain, and Christophe Maris. Structural and mechanistic insights into poly(uridine) tract recognition by the hnRNP C RNA recognition motif. *Journal of the American Chemical Society*, 136:14536–14544, 2014.

[35] Jeremy R Sanford, Xin Wang, Matthew Mort, Natalia Vanduyn, D N Cooper, Sean D Mooney, Howard J Edenberg, and Yunlong Liu. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Research*, 19(3):381, 2009.

[36] Claudia Colombrita, Elisa Onesto, Francesca Megiorni, Antonio Pizzuti, Emanuele Buratti, and Antonia Ratti. TDP-43 and FUS RNA-binding proteins bind distinct sets of cytoplasmic messenger RNAs and differently regulate their post-transcriptional fate in motoneuron-like cells. *Journal of Biological Chemistry*, 287(19):15635–15647, 2012.

[37] Benjamin J Blencowe, Sidrah Ahmad, and Leo J Lee. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes and Development*, 23(12):1379–1386, 2009.

[38] Thomas Derrien, Jordi Estelle, Santiago Marco Sola, David G Knowles, Emanuele Raineri, Roderic Guigo, and Paolo Ribeca. Fast computation and applications of genome mappability. *PLOS ONE*, 7(1), 2012.

[39] Olga Anczukow, Martin Akerman, Antoine Clery, Jie Wu, Chen Shen, Nitin H Shirole, Amanda Raimer, Shuying Sun, Mads A Jensen, Yimin Hua, et al. SRSF1-regulated alternative splicing in breast cancer. *Molecular Cell*, 60(1):105–117, 2015.

[40] Amit Bhardwaj, Michael P Myers, Emanuele Buratti, and Francisco E Baralle. Characterizing TDP-43 interaction with its RNA targets. *Nucleic Acids Research*, 41(9):5062, 2013.

[41] Cyril F Bourgeois, Franck Mortreux, and Didier Auboeuf. The multiple functions of RNA helicases as drivers and regulators of gene expression. *Nature Reviews Molecular Cell Biology*, 17(7):426, 2016.

[42] Lea Haarup Gregersen, Markus Schueler, Mathias Munschauer, Guido Mastrobuoni, Wei Chen, Stefan Kempa, Christoph Dieterich, and Markus Landthaler. MOV10 is a 5' to 3' RNA helicase contributing to UPF1 mRNA target degradation by translocation along 3' UTRs. *Molecular Cell*, 54(4):573–585, 2014.

[43] Kazuko Nishikura. Functions and regulation of RNA editing by ADAR deaminases. *Annual Review of Biochemistry*, 79:321–349, 2010.

[44] Gokul Ramaswami and Jin Billy Li. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research*, 42, 2014.

[45] Jae Hoon Bahn, Jaegyoon Ahn, Xianzhi Lin, Qing Zhang, Jaehyung Lee, Mete Civelek, and Xinshu Xiao. Genomic analysis of ADAR1 binding and its involvement in multiple RNA processing pathways. *Nature Communications*, 6, 2015.

[46] Tom Greene and Brenda L Bass. Predicting sites of ADAR editing in double-stranded RNA. *Nature Communications*, 2:319, 2011.

[47] Silvia Anna Ciafrè and Silvia Galardi. microRNAs and RNA-binding proteins: a complex network of interactions and reciprocal regulations in cancer. *RNA biology*, 10(6):934–942, 2013.

[48] Svetlana Lebedeva, Marvin Jens, Kathrin Theil, Björn Schwanhäusser, Matthias Selbach, Markus Landthaler, and Nikolaus Rajewsky. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Molecular cell*, 43(3):340–352, 2011.

[49] Sebastian Oltean and D O Bates. Hallmarks of alternative splicing in cancer. *Oncogene*, 33(46):5311–5318, 2014.

[50] Maria J Pajares, Teresa Ezponda, Raul Catena, Alfonso Calvo, Ruben Pio, and Luis M Montuenga. Alternative splicing: an emerging topic in molecular and clinical oncology. *Lancet Oncology*, 8(4):349–357, 2007.

[51] B Kate Dredge, Alexandros D Polydorides, and Robert B Darnell. The splice of life: Alternative splicing and neurological disease. *Nature Reviews Neuroscience*, 2(1):43–50, 2001.

[52] Simon Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari Ward, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43, 2015.

[53] Tamer T Onder, Piyush B Gupta, Sendurai A Mani, Jing Yang, Eric S Lander, and Robert A Weinberg. Loss of E-cadherin promotes metastasis via multiple downstream transcriptional pathways. *Cancer research*, 68(10):3645–3654, 2008.

[54] Tatsuya Oda, Yae Kanai, Tsukasa Oyama, Kenta Yoshiura, Yutaka Shimoyama, Walter Birchmeier, Takashi Sugimura, and Setsuo Hirohashi. E-cadherin gene mutations in human gastric carcinoma cell lines. *Proceedings of the National Academy of Sciences*, 91(5):1858–1862, 1994.

[55] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad Ozenberger, Kyle Ellrott, Ilya Shmulevich, C Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113, 2013.

[56] Jun Zhu and Adrian R Krainer. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Molecular Cell*, 8(6):1351–1361, 2001.

[57] Steven B Mcmahon, Heather A Van Buskirk, Kerri A Dugan, Terry D Copeland, and Michael D Cole. The novel ATM-related protein TRRAP is an essential cofactor for the c-Myc and E2F oncoproteins. *Cell*, 94(3):363–374, 1998.

[58] Rabih Murr, Thomas Vaissiere, Carla Sawan, Vivek Shukla, and Zdenko Herceg. Orchestration of chromatin-based processes: mind the TRRAP. *Oncogene*, 26(37):5358–5372, 2007.

[59] Hongcheng Wang, Yanrong Su, Kejun Han, Xuewen Pang, Jirun Peng, Bin Liang, S G Wang, and Weifeng Chen. Multiple variants and a differential splicing pattern of kinectin in human hepatocellular carcinoma. *Biochemistry and Cell Biology*, 82(2):321–327, 2004.

[60] Susan E Morgan and M B Kastan. p53 and ATM: cell cycle, cell death, and cancer. *Advances in Cancer Research*, 71:1–25, 1997.

[61] David H Viskochil. Review article : Genetics of neurofibromatosis 1 and the NF1 gene. *Journal of Child Neurology*, 17(8):562–570, 2002.

[62] Francois Delhommeau, Sabrina Dupont, Chloe James, Aline Masse, Jean Pierre le Couedic, Veronique Della Valle, Antonio Alberdi, Philippe Dessen, Michaela Fontenay, Nicole Casadevall, Jean Soulier, Bernard, Olivier, and William Vainchenker. TET2 is a novel tumor suppressor gene inactivated in myeloproliferative neoplasms: identification of a pre-JAK2 V617F event. *Blood*, 112(11), 2008.

[63] Irmgard Schwartewaldhoff, Olga V Volpert, Noel Bouck, Bence Sipos, Stephan A Hahn, Susanne Kleinscory, J Luttges, Gunter Kloppel, Ulrich Graeven, Christina Eilertmicus, et al. Smad4/DPC4-mediated tumor suppression through suppression of angiogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 97(17):9624, 2000.

[64] Gavin E Crooks, Gary C Hon, Johnmarc Chandonia, and Steven E Brenner. Weblogo: A sequence logo generator. *Genome Research*, 14, 2004.

[65] Rufang Yeh, Phillip A Sharp, and Christopher B Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–1013, 2002.

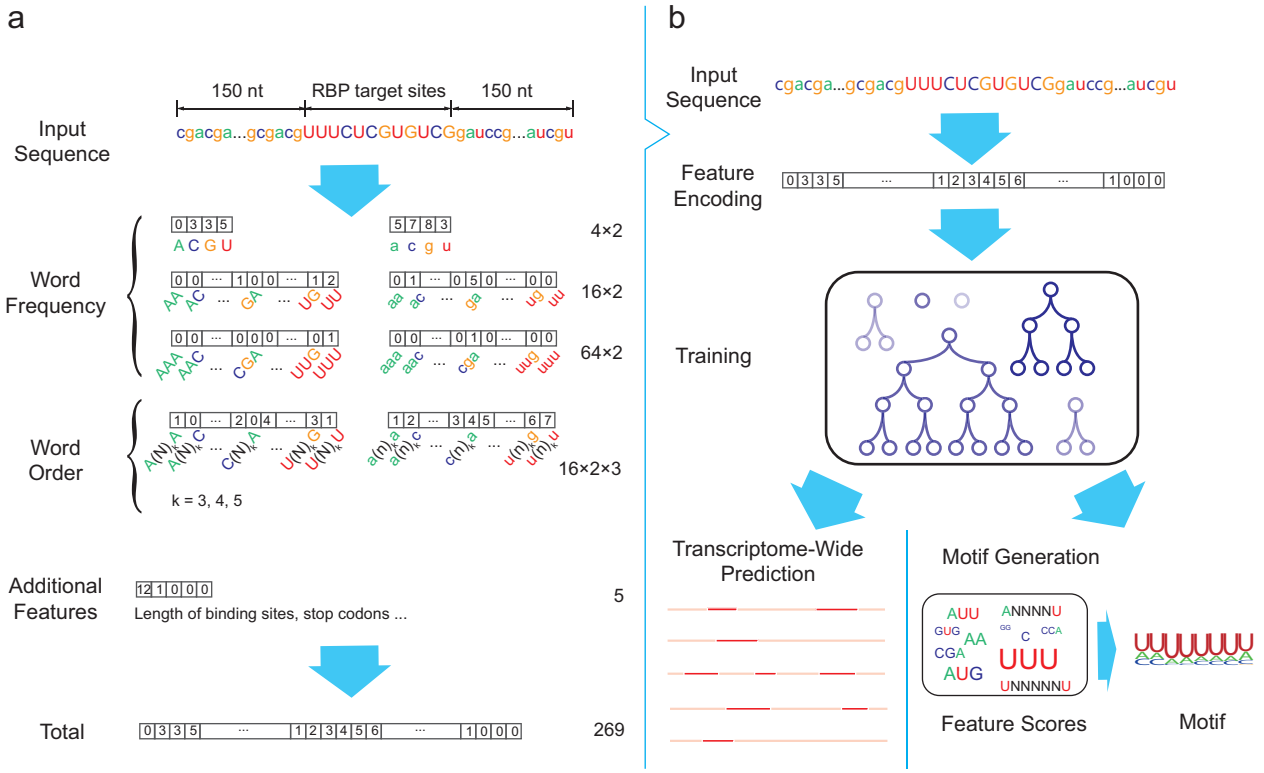
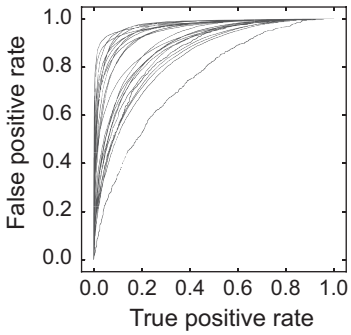


Figure 1 Schematic overview of DeBooster, a deep boosting approach for identifying the sequence specificities of RNA-binding proteins (RBPs). **(a)** Schematic illustration of the strategy for encoding the sequence features of RBP binding targets. The nucleotides in the target region of an input sequence are represented by capitalized letters while the extended regions on both sides are represented by lowercase letters. Each number within a box stands for the value of the corresponding feature. The numbers on the right side represent the total number of feaures in individual catagories. **(b)** Schematic illustration of the prediction pipeline. More details can be found in the main text.

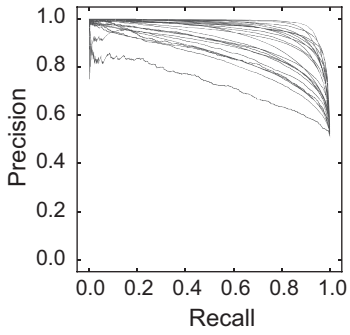
a

Dataset	GraphProt	DBN	DeBooster
ALKBH5	0.680	0.714	**0.752**
C17ORF85	0.800	0.820	**0.866**
C22ORF28	0.751	0.792	**0.840**
CAPRIN1	0.855	0.834	**0.903**
AGO2	0.765	0.809	**0.863**
ELAVL1	0.955	0.966	**0.972**
SFRS1	0.898	0.931	**0.950**
HNRNPC	0.952	**0.962**	0.958
TDP43	0.874	0.876	**0.909**
TIA1	0.861	0.891	**0.902**
TIAL1	0.833	0.870	**0.886**
AGO1-4	0.895	0.881	**0.908**
ELAVL1(B)	0.935	0.961	**0.976**
ELAVL1(A)	0.959	0.966	**0.972**
EWSR1	0.935	0.966	**0.969**
FUS	0.968	0.980	**0.982**
ELAVL1(C)	0.991	0.994	**0.995**
IGF2BP1-3	**0.889**	0.879	0.883
MOV10	0.863	0.854	**0.918**
PUM2	0.954	0.971	**0.973**
QKI	0.957	**0.983**	0.980
TAF15	0.970	0.983	**0.988**
PTB	0.937	**0.983**	0.943
ZC3H7B	0.820	0.796	**0.897**

b



c

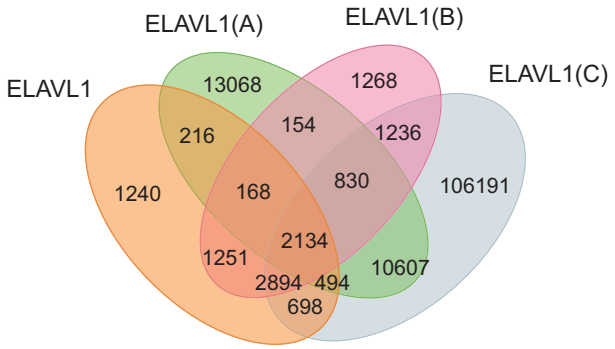


d

RBP	Motif	RBP	Motif
AGO2		PTB	
EWSR1		QKI	
HNRNPC		SFRS1	
PUM2		TDP43	

Figure 2 Performance evaluation of DeBooster on 24 CLIP-seq datasets. **(a)** The comparisons of the area under receiver operator characteristic curve (AUROC) scores between different prediction approaches via a 10-fold cross-validation procedure. The best prediction result for each dataset is highlighted in bold. **(b)** and **(c)** The receiver operator characteristic (ROC) and precision-recall (PR) curves achieved by DeBooster for all 24 CLIP-seq datasets in the cross-validation results, respectively. **(d)** Examples of the sequence motifs of the RBP binding targets predicted by DeBooster.

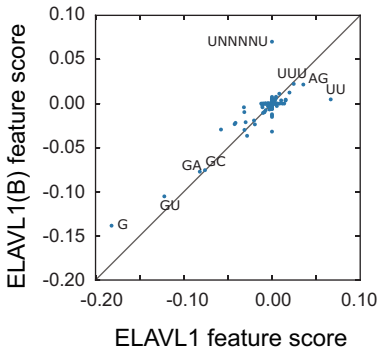
a



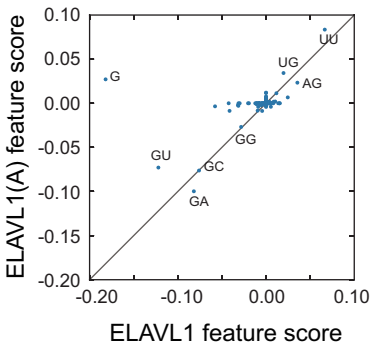
b

Training\Test Dataset	ELAVL1	ELAVL1(A)	ELAVL1(B)	ELAVL1(C)	Motif
ELAVL1	**0.972**	0.975	0.955	0.943	
ELAVL1(A)	0.972	**0.972**	0.950	0.940	
ELAVL1(B)	0.941	0.944	**0.976**	0.937	
ELAVL1(C)	0.967	0.983	0.929	**0.995**	

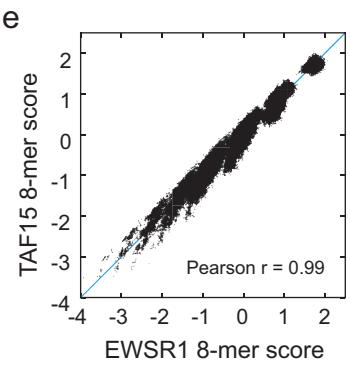
c



d



e



f

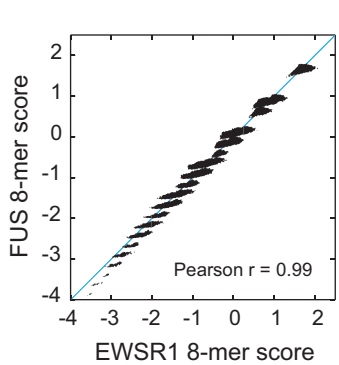


Figure 3 Performance validation of DeBooster through cross-platform CLIP-seq datasets. **(a)** The Venn diagram of four ELAVL1 CLIP-seq datasets collected from different experimental platforms. If binding region A from a dataset has at least one nucleotide overlap with binding region B from

another dataset, we regarded A and B as a common element of these two datasets. The datasets ELAVL1, ELAVL1(A) and ELAVL1(C) were from the HEK293 cells, while the dataset ELAVL1(B) was from the Hela cells. **(b)** The AUROC scores and binding sequence motifs computed by DeBooster using different combinations of training and test datasets. The diagonal scores shown in bold correspond to the cross-validation results in which both training and test datasets were collected from the same experimental platform. **(c, d)** The plots of the relative weights of individual sequence features computed by DeBooster for the ELAVL1 datasets collected from different experimental platforms, including ELAVL1(B) vs. ELAVL1 **(c)** and ELAVL1(A) vs. ELAVL1 **(d)**. **(e, f)** The plots of the DeBooster prediction scores for all 8-mers across different RBPs within the same family, including TAF15 vs. EWSR1 **(e)** and FUS vs. EWRS1 **(f)**. TAF15, FUS and EWSR1 all belong to the FET family and generally share similar binding preferences.

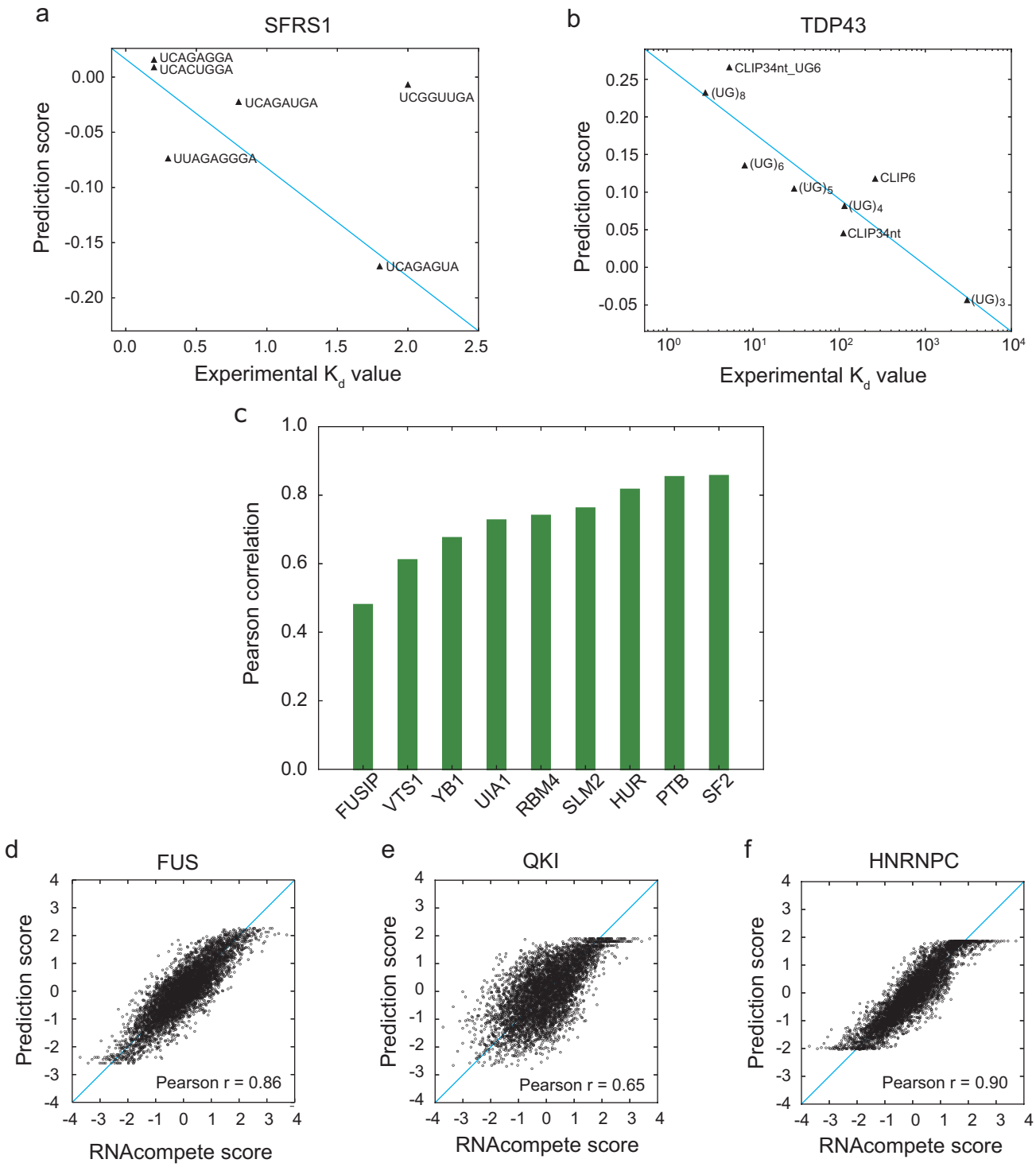


Figure 4 The comparisons between the prediction scores derived by DeBooster and the experimentally determined binding affinity data. **(a, b)** The plots of the prediction scores derived by DeBooster (which was trained based on CLIP-seq data) vs. the experimentally determined K_d values of different 8-mers or RNA oligonucleotides for SFRS1 and TDP43, respectively. The K_d values of SFRS1 were

measured *in vivo* [39], while the K_d scores of TDP43 were acquired from the electrophoretic mobility shift assay (EMSA) [40]. The same terminology as in [40] for the names of RNA oligonucleotides was used for the binding targets of TDP43 (Supplementary Notes). **(c)** The Pearson correlation coefficients between the prediction scores derived by DeBooster vs. the *in vitro* binding affinity scores of 7-mers derived from the RNAcompete data in [10]. **(d-f)** The plots of the prediction scores derived by DeBooster vs. the *in vitro* exprimentally measured binding affinity scores of 7-mers derived from the RNAcompete data in [11] for FUS, QKI and HNRNPC, respectively. In **(c-f)**, an extended "regression" version of DeBooster was used and a cross-validation procedure was applied to evaluate the prediction performance (Methods).

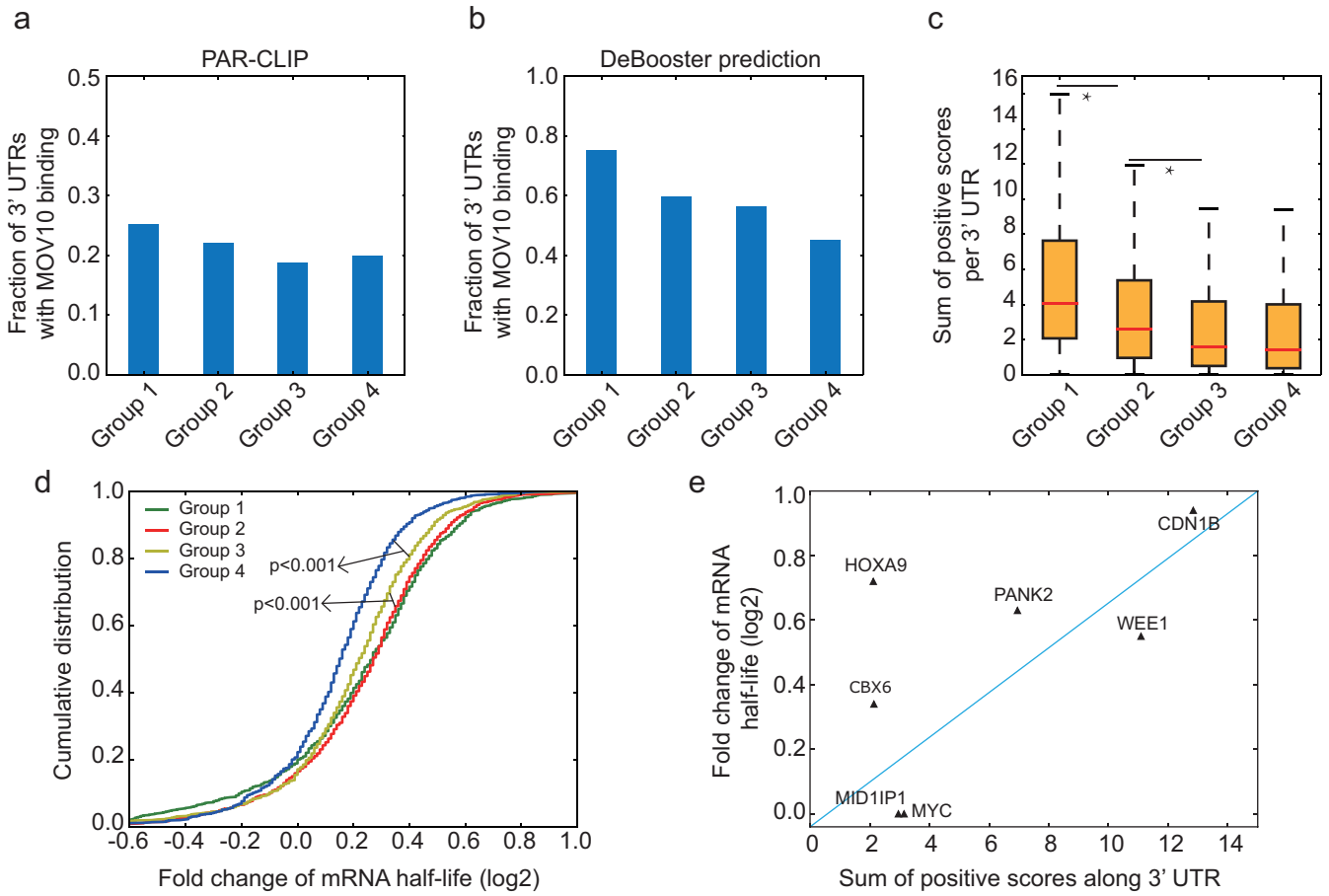


Figure 5 Understanding the predicted binding effects of MOV10 on mRNA dagradation. **(a, b)** Fractions of 3′ UTRs with MOV10 binding for four groups classified according to the original CLIP-seq data **(a)** and the DeBooster prediction results **(b)**, respectively. Genes were evenly separated into four groups according to the fold changes of their mRNA half-lives. Groups 1, 2, 3 and 4 corresponded to top 25%, 25%-50%, 50%-75% and bottom 25%, respectively. In the DeBooster prediction results, we only considered those robust binding targets with prediction scores > 0.2 (The default threshold was zero and the range of prediction scores was in [-1,1]). **(c)** The sum of positive prediction scores per UTR for four groups of genes, which were classified and ranked according to the fold changes of their mRNA half-lives in a descending order. *: p value < 0.001, Wilicoxon rank sum test. **(d)** The cumulative distribution on the fold changes of mRNA half-lives for four groups of genes, classified and ranked according to the DeBooster prediction scores in a descending order. That is, Groups 1, 2, 3 and 4 corresponded to genes with top 25%, 25%-50%, 50%-75% and bottom 25% predicted scores, respectively. The p values were computed using the Wilicoxon rank sum test. **(e)** The plot of the DeBooster prediction scores vs. the fold changes of mRNA half-lives meansured by qRT-PCR for seven genes.

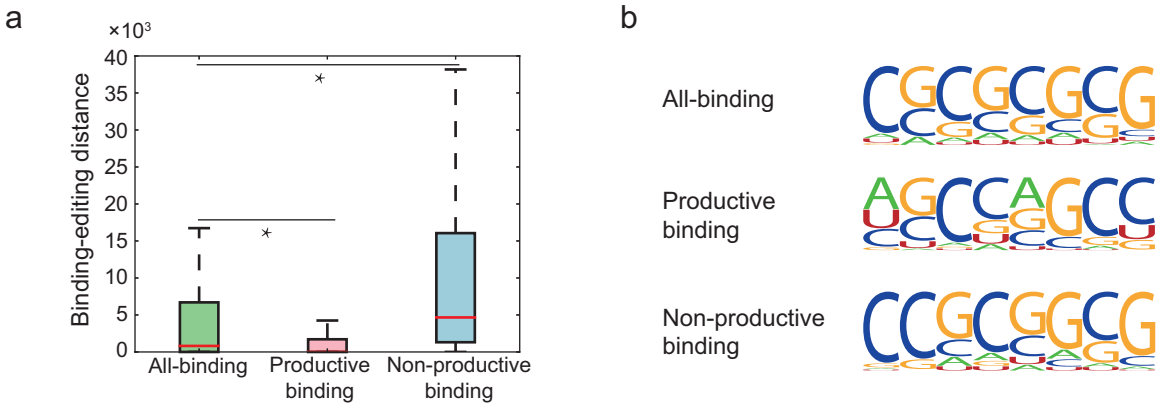


Figure 6 The comparison results on three different DeBooster models, which were trained using all ADAR binding sites identified by CLIP-seq experiments, productive ADAR binding sites (i.e, triggering A-to-I editting), and non-productive ADAR binding sites (i.e, without triggering A-to-I editting), respectively. **(a)** The boxplot of the binding-editing distances, which were defined as the genomic distances between the ADAR binding sites and the closet editing sites, for three different DeBooster models. *: p value < 0.001, Wilicoxon rank sum test. **(b)** The sequence motifs of the ADAR binding sites identified by three different DeBooster models. More details can be found in the main text.

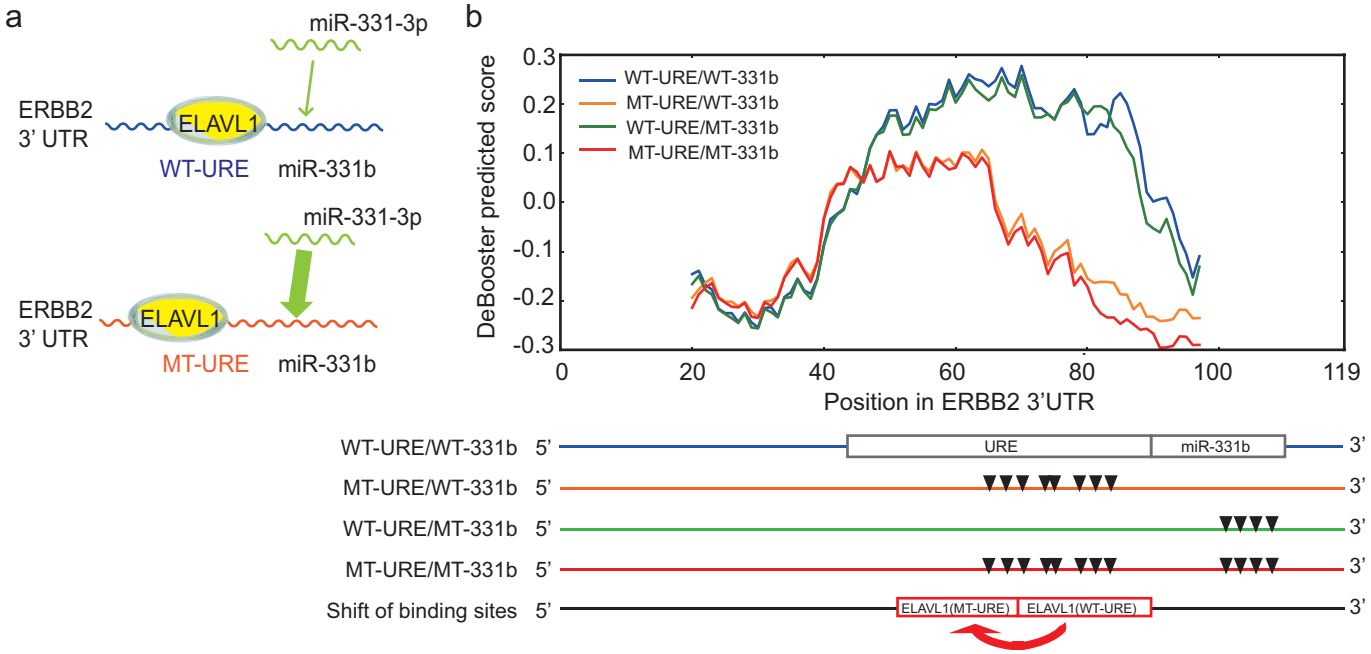


Figure 7 The predicted influence of ELAVL1 binding on the repression effect of miRNA miR-331-3p. **(a)** An illustrative model of the co-binding of RBP ELAVL1 and miRNA miR-331-3p on the 3′ UTR of gene *ERBB2*. miR-331b represents the binding region of miRNA miR-331-3p. The width of the arrow represents the relative strength of miR-331-3p binding. **(b)** The change of the predicted binding scores corresponded to the shift of ELAVL1 binding sites from the wild-type to the URE mutant on the 3′ UTR of gene *ERBB2*. The bottom shows the locations of URE and miR-331b regions, mutation positions in the URE region, mutation positions in the miR-331b region, mutation positions in both URE and miR-331b regions, and the experimentally detected shift of ELAVL1 binding resulting from the URE mutant, respectively. All mutation sites are represented by the inverted triangles. Abbreviation: WT, wild-type; MT, mutant; URE, U-rich element.

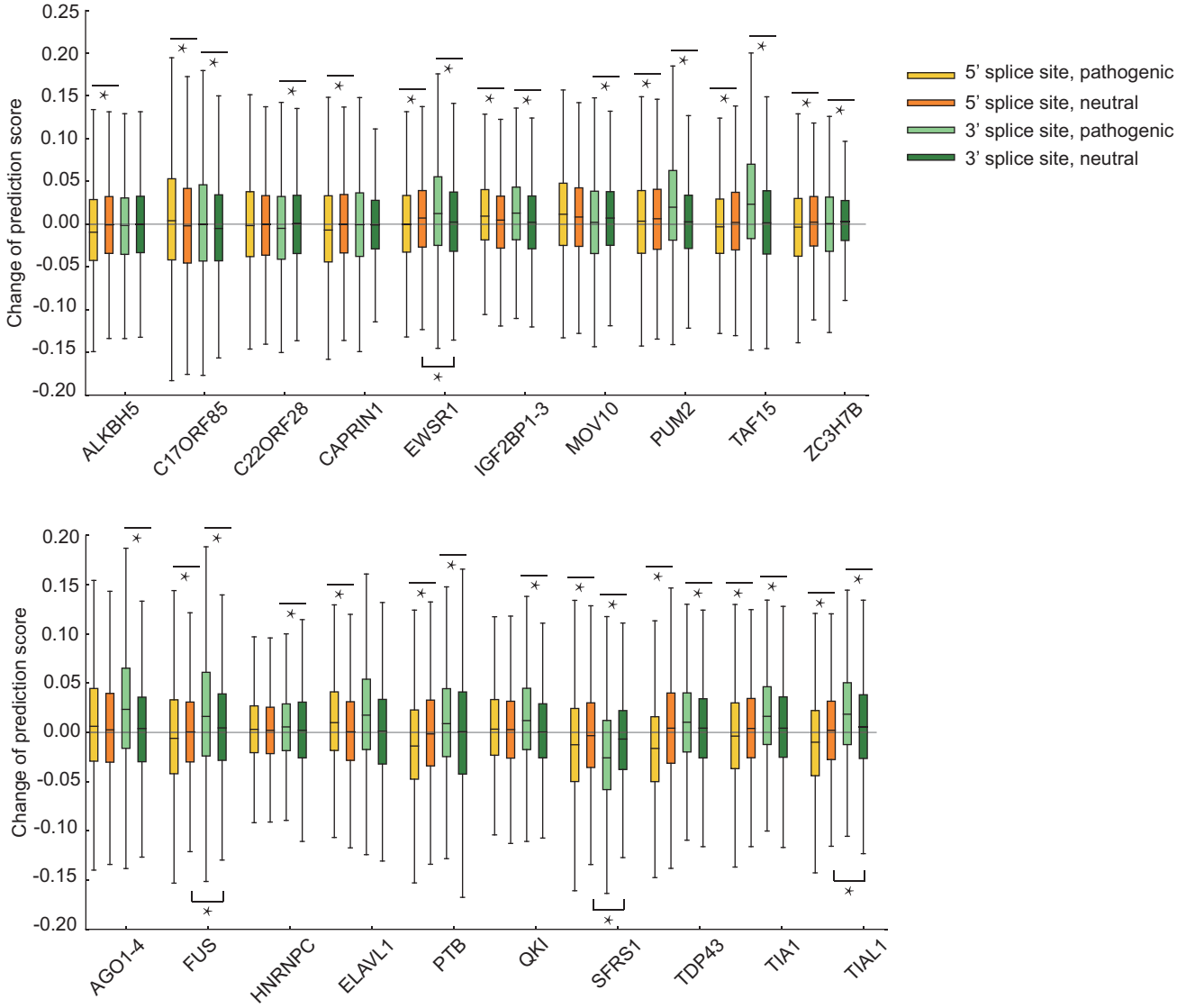


Figure 8 The comparisons between the overall changes of the predicted binding scores of individual RBPs after pathogenic or neutral mutations in regions near 5′ and 3′ splice sites. *: p<0.001, Student's t test.

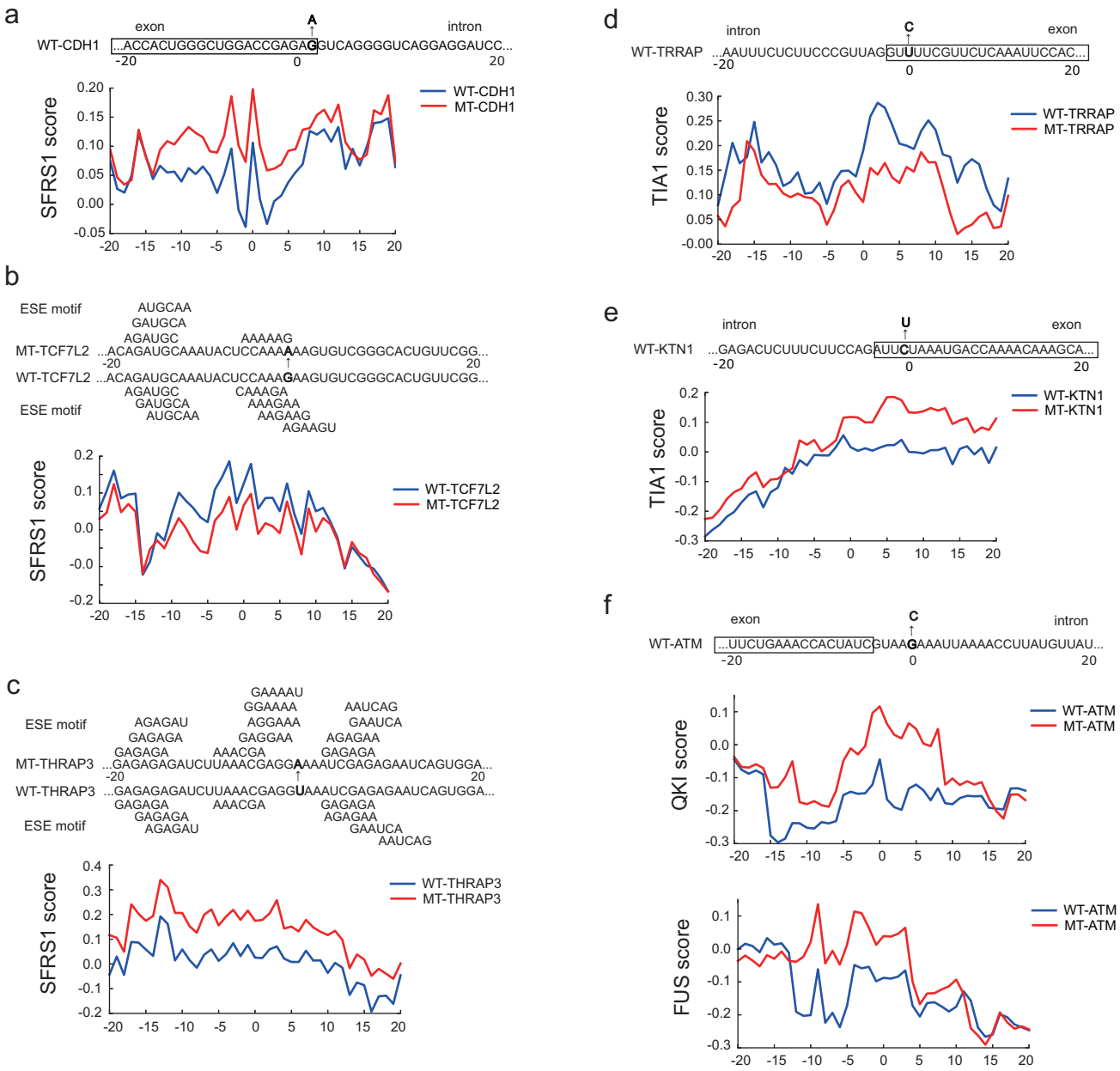


Figure 9 Examples of the predicted effects on the potentially disease-causing mutations near splice sites or on exonic splicing enhancers (ESEs). **(a)** The exonic mutations of the SFRS1 binding sites near a 5′ splice site for gene *CDH1*. **(b, c)** The mutations of SFRS1 binding sites disrupting or creating exonic splicing enhancer (ESE) motifs for genes *TCFIL2* and *THRAP3*, repectively. The ESE motifs were obtained from [65]. **(d, e)** The exonic mutations of the TIA1 binding sites near the splice sites for genes *TRRAP* and *KTN1*, respectively. **(f)** A mutation near a 5′ splice site of gene *ATM* that changed the predicted binding scores of both QKI and FUS. Abbreviation: WT, wild-type; MT, mutant; URE, U-rich element.