Genome Biology

RESEARCH

Open Access

Evidence for the role of transcription factors in the co-transcriptional regulation of intron retention

Fahad Ullah[1], Saira Jabeen[1], Maayan Salton[2], Anireddy S. N. Reddy[3] and Asa Ben-Hur[1*]

*Correspondence: asa@colostate.edu

[1] Department of Computer Science, Colorado State University, Fort Collins, CO, USA
[2] Department of Biology, Colorado State University, Fort Collins, CO, USA
[3] Biochemistry and Molecular Biology Department, The Hebrew University Faculty of Medicine, Jerusalem, Israel

Abstract

Background: Alternative splicing is a widespread regulatory phenomenon that enables a single gene to produce multiple transcripts. Among the different types of alternative splicing, intron retention is one of the least explored despite its high prevalence in both plants and animals. The recent discovery that the majority of splicing is co-transcriptional has led to the finding that chromatin state affects alternative splicing. Therefore, it is plausible that transcription factors can regulate splicing outcomes.

Results: We provide evidence for the hypothesis that transcription factors are involved in the regulation of intron retention by studying regions of open chromatin in retained and excised introns. Using deep learning models designed to distinguish between regions of open chromatin in retained introns and non-retained introns, we identified motifs enriched in IR events with significant hits to known human transcription factors. Our model predicts that the majority of transcription factors that affect intron retention come from the zinc finger family. We demonstrate the validity of these predictions using ChIP-seq data for multiple zinc finger transcription factors and find strong over-representation for their peaks in intron retention events.

Conclusions: This work opens up opportunities for further studies that elucidate the mechanisms by which transcription factors affect intron retention and other forms of splicing.

Availability: Source code available at https://github.com/fahadahaf/chromir

Keywords: Alternative splicing, Intron retention, Deep learning

Introduction

Alternative splicing is a widespread regulated phenomenon that enables a single gene to encode structurally and functionally different transcripts [1, 2]. The primary forms of alternative splicing are exon skipping, intron retention (IR), and alternative 3′ and 5′ splicing. While exon skipping is well studied, IR remains an under-appreciated phenomenon [3]. IR is the primary form of alternative splicing in plants [4, 5], and recent studies

have shown it to have a high prevalence in human [6, 7]. Many disease-causing mutations are pathogenic through their effect on splicing, often leading to IR [6, 8, 9]. For example, IR is associated with genetic variants with deleterious effect on the function of tumor suppressor genes [10].

In recent years, efforts have been made to understand the regulation of IR and the factors that contribute to it. Braunschweig et al. [7] recently published a draft IR splicing code: a predictive model that uses a total of 136 features thought to be associated with IR in mammals. These features include base composition of an intron and its flanking exons, features that describe gene architecture, and splice site strength. This model is limited in that it does not model sequence elements that contribute to the regulation of IR. The discovery that splicing occurs co-transcriptionally suggests that chromatin state might be relevant to alternative splicing [11, 12]. Recent work provides evidence for the regulatory contribution of chromatin state to exon skipping [13], and our labs have provided preliminary evidence for its role in regulating IR in plants [14]. Open chromatin is one of the most important signatures for the study of chromatin structure. One of the primary tools for probing open chromatin is through exposure of DNA to deoxyribonuclease I (DNase I), which is an enzyme that cleaves DNA. Regions of the genome that are sensitive to its action—DNase I hypersensitive sites (DHSs)—have been used as an indicator of chromatin accessibility in vivo [15]. DHSs have been used extensively to identify several types of regulatory elements such as promoters, enhancers, silencers, and insulators [16, 17]. Furthermore, when a regulatory protein binds DNA, it protects it against the action of DNase I [18] and leaves a footprint which can be identified using DNase I-seq data [19, 20]. When it comes to alternative splicing, Mercer et al. [13] have shown an association between DHSs and exon-skipping, reporting that higher numbers of DHS-containing exons are alternatively spliced. Furthermore, this study reports that DHS exons with promoter and enhancer-like features have a higher fractional overlap with alternative splicing. Braunschweig et al. [7] explored the co-transcriptional regulation of splicing, reporting higher chromatin accessibility in retained introns and that polymerase II elongation speed affects IR and vice-versa. In another work, it has been reported that zinc finger transcription factors (TFs) have a regulatory role in exon skipping [21]. Recently, we studied the association between chromatin accessibility and intron retention in plants [14]. We identified potential regulatory elements occurring primarily in the 3′ flanking exons of IR events, several of which significantly match plant zinc finger binding site motifs. As further motivation for considering the role of TFs in splicing regulation, we provide evidence for extensive TF binding within human genes using ChIP-seq data. We collected ChIP-seq data in K562 for 11 different TFs and computed the number of peaks per Mb in intergenic regions and compared it to the number of peaks in intragenic regions. The results shown in Fig. 1 clearly demonstrate that for this selection of TFs the number of intragenic peaks is higher. A similar observation was made in plants [22]. This suggests a regulatory role of TFs beyond the regulation of gene expression.

Deep neural networks have become the tool of choice for exploring complex biological phenomena such as gene expression and chromatin state [23–28]. A remarkable advantage of these models is their ability to capture the underlying patterns in large noisy datasets directly from sequence with minimal pre-processing, learning motifs

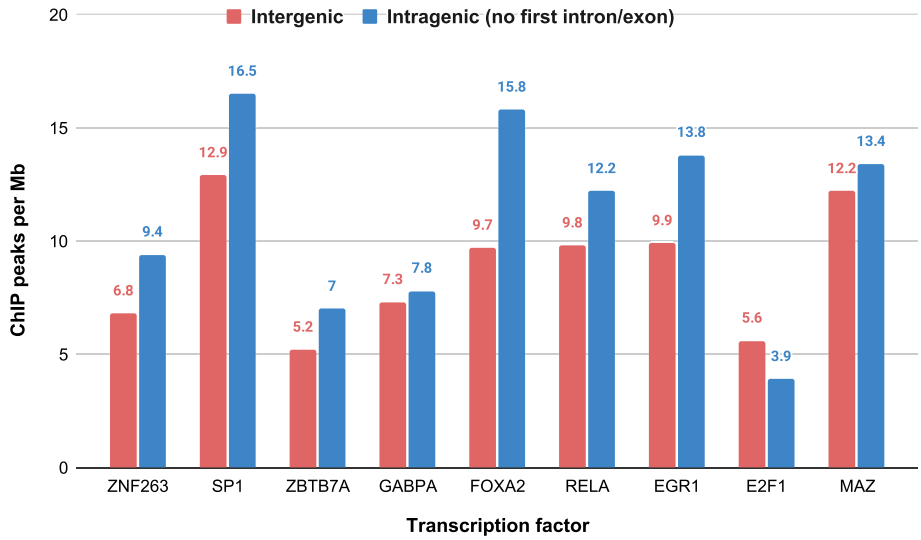

Fig. 1 TF binding across the genome. We computed the number of ChIP-seq peaks per Mb for a selection of TFs, distinguishing between intergenic and intragenic peaks. Intergenic counts excluded the promoter region while intragenic counts excluded the first exon and intron to remove the effect of the promoter region

of the regulatory proteins involved as part of the training process. Deep learning has been used in genomics for TF binding prediction [29–31], chromatin accessibility analysis [23–25], prediction of chromatin structure and its modifications [32, 33], identification of RNA-binding protein sites [28, 34, 35], and prediction of splice site usage from sequence [36, 37]. Several labs have developed sophisticated models of exon skipping on the basis of large collections of genomic features [38–40], but have not considered the role of chromatin.

In this study, we demonstrate that deep learning models can distinguish with good accuracy regions of open chromatin associated with IR from other intronic regions of open chromatin. By analyzing the motifs learned by the network, we find that specific families of TFs are associated with IR events, mostly members of the zinc finger family of TFs; results of ChIP-seq experiments for multiple zinc finger TFs in the K562 cell line, one of three tier 1 ENCODE cell lines, support our findings for this association. Analysis of knockdown experiments of some of these TFs suggest they function as splicing enhancers by binding the flanking exons of IR events. Our work provides convincing evidence for a novel role of TFs in the regulation of IR, proposing a promising direction for further research.

Results

DHSs associated with IR can be accurately predicted from their sequences

In order to discover the sequence elements that regulate IR via its coupling with chromatin state, we trained and evaluated deep learning models to distinguish DHSs associated with IR events from non-IR DHSs in human and assessed and compared their performance. As IR DHSs, we used regions in which a DHS overlapping IR event was detected in at least one DNase I-seq experiment in a compendium of 164 samples; IR events were extracted from the Ensembl gene models as described in the "Methods"

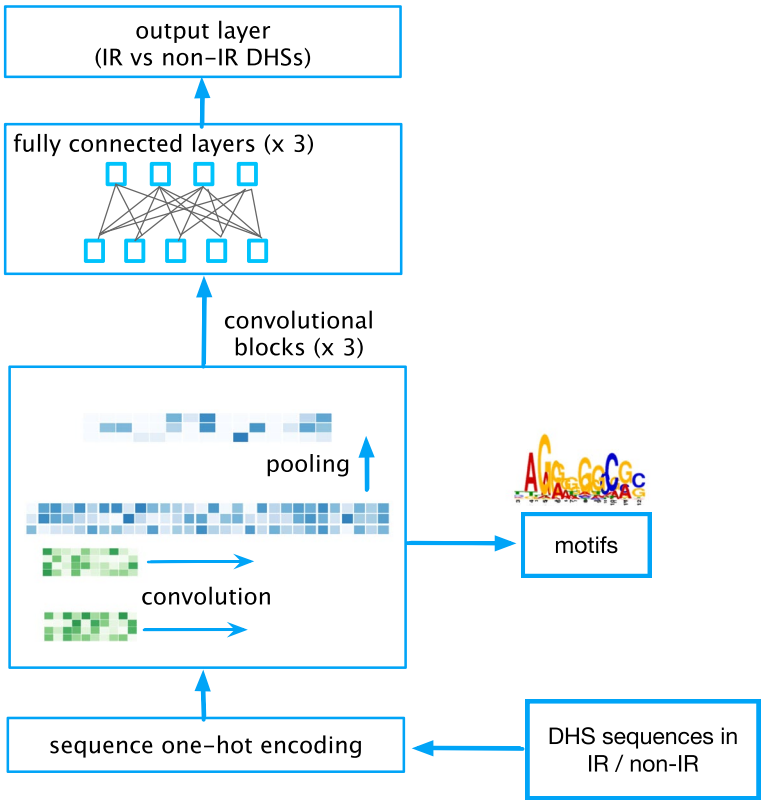

Fig. 2 A deep learning model for predicting whether a region of open chromatin exhibits IR. The model receives as input the sequences of intragenic DHSs labeled as associated with IR or non-IR; the one hot encoding is processed through three layers of convolution, followed by three fully connected layers and the output layer that predicts a binary response that indicates whether a DHS exhibits IR or not. The convolutional filters of the first layer are used to extract position weight matrices (PWMs) that are searched against a database of known TFs

section. For non-IR DHSs, we used intronic regions exhibiting a DHS where no IR is known to occur. In this work, we chose to focus on the purely convolutional architecture shown in Fig. 2, that has demonstrated its effectiveness for predicting chromatin accessibility by Kelley et al. [23]. The model hyperparameters were tuned for our problem as described in the "Methods" section. Using this model we obtained accuracy of 0.546 as measured using the area under the precision-recall curve (AUC-PRC) (see Fig. 3a). A more sophisticated model that uses a combination of convolutional and recurrent layers with multi-head attention achieved a similar level of accuracy (see Fig. 3a and Additional file 1: Fig. S1). We note that both deep learning architectures outperformed a baseline approach that uses the gkm-SVM method [41]. This method achieved an AUC-PRC of 0.503. ROC curves are provided in Additional file 1: Fig. S1.

Our results were generated using a one-hot encoding of the sequence of DHS regions. We note that word2vec embeddings provided a small improvement in accuracy, as shown in Additional file 1: Fig. S1. However, this came at a cost of reduced interpretability of the models, leading to reduced ability to infer motifs associated with the learned convolutional filters (see discussion in the Additional file 1). Therefore, we chose to focus on models that used one-hot encoding as input. During the revision of the manuscript, we discovered 72 duplicate DHSs out of the 7500 training examples in the

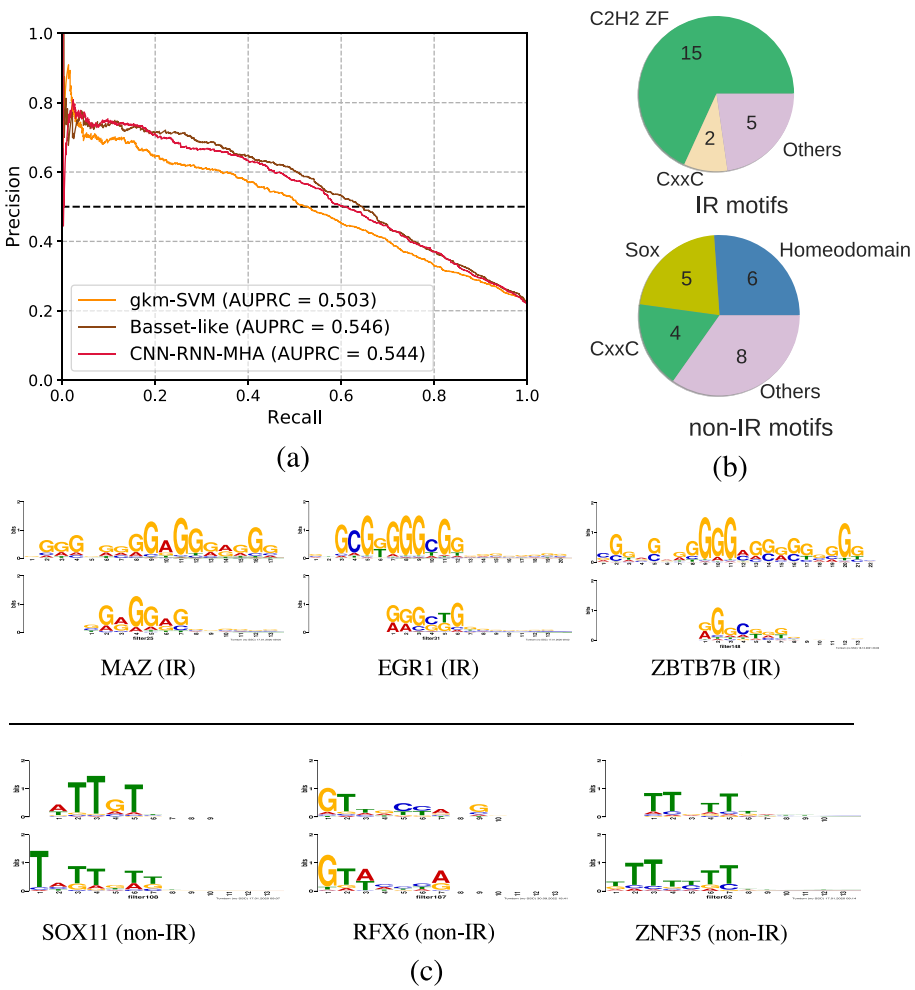

Fig. 3 Classification accuracy and motifs detected by the network. **a** Precision-recall curves for the two deep learning architectures and the gkm-SVM. The AUC-PRC values are also provided in the legend. **b** The distribution of TF families enriched in IR vs non-IR events. **c** The top three matches for the IR and non-IR convolutional layer filters against the CISBP database. In each match, the known TF motif is shown in the top row and the bottom row shows the CNN filter/motif. The motifs shown above the line are associated with IR DHSs, and those below the line are associated with non-IR DHSs

original dataset provided by Kelley et al. [23]. Some of these occurred across the train and test set. At a threshold of 80% sequence identity, we also found using CD-HIT [42] 39 sequences whose similarity is above that threshold. We re-trained the classifier without the duplicates and similar sequences and found that the accuracy was unchanged.

The zinc finger family of TFs are enriched in IR events

The filters of convolutional networks can be readily interpreted as motifs. To do so, we implemented the strategy described elsewhere [23, 29] (see "Methods" section for details). We analyzed the motifs that were derived from the convolutional filters for both the top positive and the top negative examples and searched both sets of motifs against the Human CIS-BP TF database [43] using TomTom [44]. We found that 22 IR-associated motifs had significant hits against multiple known human TFs at a q-value < 0.01.

In comparison, 23 of the non-IR motifs had significant matches. Figure 3c shows some of the top hits for both IR and non-IR motifs, and a complete list is found in the github repository of the project. The median information content of the IR motifs was 4.21, and 4.26 for the non-IR motifs. The other architectures provided motifs with similar information content (see Additional file 1: Tables S4 and S5). Furthermore,when comparing the TF hits for the IR motifs with an adjusted p-value of 0.01 or better for the three different architectures (the purely convolutional network and variants that include attention with and without a recurrent layer), we found that 21 out of the 25 motif hits discovered by the purely convolutional architecture were common across the three architectures.

Most of the IR motifs had significant hits in the C2H2 zinc finger family of TFs (C2H2 ZF). Non-IR motifs on the other hand, were predominantly matched to the Homeodomain and Sox families of TFs (see Fig. 3b). Zinc finger TFs have previously been implicated in the regulation of alternative splicing [21], particularly exon skipping. Here we report a role of this family in the regulation of IR. We note that some of our filters do not match a unique transcription factor. For example, the filter that matched MAZ, was also a good match for ZNF263. This is not surprising due to the similarity of the binding sites of zinc finger TFs. Below we provide additional evidence for the role of zinc finger TFs in regulating IR.

We also searched for motif matches to RNA-binding proteins in the CISBP-RNA database [45] using the same methodology employed for TFs. At the same p-value threshold used for searching for TF hits, we found three significant hits in the IR motifs, and two significant hits in non-IR motifs. This is compared with 22 TF hits in IR motifs and 23 TF hits in non-IR motifs. Details of those matches are found in the github repository of this project. This suggests that TFs play a major role in IR in comparison to RNA-binding proteins, perhaps as a result of our focus on regions of open chromatin.

TF ChIP-seq analysis supports model predictions

To validate our findings using experimental data, we downloaded K562 ENCODE ChIP-seq datasets for all the zinc finger TFs identified by our model, resulting in six datasets. Using these datasets, we tested TF binding enrichment in IR vs. non-IR events, following a strategy similar to our previous work [14]: for each TF, we measured the overlap of its ChIP-seq peaks with IR and non-IR events and tested its significance using the Fisher exact test. All the TFs demonstrated highly significant enrichment in IR events (see Table 1), validating our in silico findings that the C2H2 ZF family plays a role in the regulation of IR.

RNA-seq of TF Knockdowns suggest IR TFs function as splicing enhancers

To obtain a better understanding of the way the TFs predicted to be associated with IR function to regulate IR, we analyzed RNA-seq datasets of the K562 cell line with knockdown/silencing of several IR-associated TFs: MAZ, SP1, SP2, and E2F4. To evaluate the effect of the knockdown of each TF, we looked for differential IR events in the knockdown samples with respect to baseline K562 using iDiffIR [46]. In all cases, there were many more up-regulated IR events, i.e., events with increased IR with respect to the wild-type that are statistically significant at a p-value of 0.05 and above (see Table 2).

Table 1 Enrichment of C2H2 ZF TF binding in IR compared to non-IR events quantified using ChIP-seq peaks of the corresponding TF. We note that for the bottom three TFs, the *p*-value is for the significance of enrichment in non-IR events

TF	IR TF occupancy (%)	Non-IR TF occupancy (%)	*p*-value
EGR1	12.51	7.16	1.27E−45
MAZ	11.42	6.06	9.75E−51
ZBTB7A	10.6	5.52	3.15E−49
SP1	3.04	1.53	7.64E−16
SP2	1.32	0.77	7.21E−06
ZNF263	1.14	0.67	2.81E−05
FOXK2	5.2	6.12	3.01E−04
GATA1	0.77	1.07	5.0E−03
JUN	2.68	4.86	3.21E−23

Table 2 RNA-seq results for TF knockdown experiments in K562. For each TF, we provide the number of statistically significant IR events that are up-regulated (down-regulated), i.e., exhibit increased (decreased) retention with respect to the wild-type. Within the up-regulated events we provide the number of events with occurrences of the motif compared to the background set composed of IR events. This is done in both the intron and the flanking exons. The significance of the difference in the rates of occurrence of the motif is provided in the last column. In all cases the significance was exhibited in the flanking exons, except for E2F4 which exhibited similar levels of significance in both the introns and flanking exons

TF	Up-regulated	Down-regulated	Motif occurrences		*p*-value
			Intron	Exon	
MAZ	69	34	23% (25%)	31% (16%)	0.001
SP1	86	18	69% (64%)	86% (51%)	1.3E−9
SP2	99	25	31% (27%)	41% (16%)	1.4E−9
E2F4	174	58	18% (12%)	18% (11%)	0.008

This suggests that these TFs predominantly function as splicing enhancers. Furthermore, we searched for the hits for the motifs of each TF in the differentially retained introns compared to introns that are not differentially retained. We found that in introns that showed a statistically significant increase in IR levels, there was a much higher number of hits for the motif of each TF compared to IR events where no significant difference in retention was observed; this difference was statistically significant (see Table 2), further support for the role of these TFs as splicing enhancers.

Regulatory interactions between TFs in IR events

It is well known that TFs often function in tandem with each other to regulate their targets. To extract such regulatory interactions, we have recently developed a method called SATORI to interpret deep architectures that use *attention* layers and extract statistically significant interactions between its convolutional filters [47]. SATORI uses the so-called attention matrix, which encodes relations between different positions of the sequence; subsequent analysis of the convolutional filters that are active provides a profile of interactions between pairs of TFs that are associated with those filters. By comparing those profiles to those in a background set of sequences, we obtain interactions that are statistically significant. Using SATORI, with the negative examples as a background

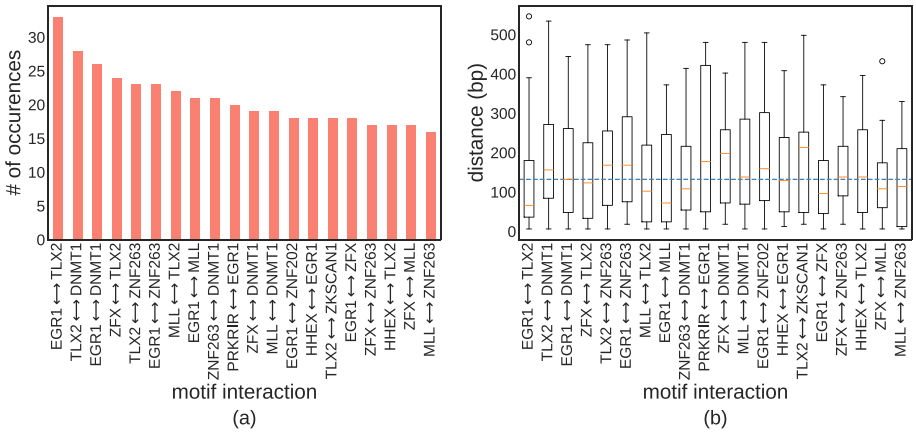

Fig. 4 TF interactions. **a** The most frequent TF interactions in IR events. **b** The distribution of distances between detected TF interactions. The dotted blue line represents the median distance across all significant interactions

set to assess statistical significance we detected over 400 TF interactions in DHSs associated with IR at a significance level of 0.05. The top 20 predictions are shown in Fig. 4, and the complete list is provided in the results directory of the project's github repository. A histogram for the number of interactions between TF families is provided in Additional file 1: Fig. S2. A majority of the interactions involve the C2H2 ZF family, which is expected since C2H2 ZF TFs have the most hits from our model. To validate these interactions, we searched for matches in annotated interactions in the TRRUSTv2 [48] database that annotates TF regulatory roles and their interactions by text-mining the biomedical literature. Of the interactions detected by our model, we found 23 overlapping interactions in TRRUSTv2, which currently contains 8324 interactions. This is highly significant, with a p-value equal to 0 in a hypergeometric test. We also obtained significant overlap with protein-protein interactions from the HIPPIE database [49]: 17 of the detected interactions had support in HIPPIE, with a hypergeometric p-value of 1E−52. The interactions overlapping with TRRUSTv2 and HIPPIE database are listed in Additional file 1: Tables S2 and S3, respectively. As further support for predicted interactions, we looked at ChIP-seq data in K562 for the interactions EGR1-MAZ and EGR1-ZNF263 and evaluated the overlap between the peaks in intragenic regions. For the EGR1-MAZ interaction, we found 21,592 intragenic peaks for EGR1 and 16,613 for MAZ. Out of those peaks, 9065 were within 150 bp of each other. Using Locus Overlap Analysis [50, 51] to evaluate the significance of the overlap, we obtained a p-value of 0. For the EGR1-ZNF263 interaction we found 21,592 intragenic ChIP-seq peaks for EGR1 and 1619 for ZNF263. Out of those peaks, 715 were within a window of 150 bp of each other with a p-value of 1E-17. Finally, we looked at the average distance between motifs predicted to interact and found that TF motifs preferentially interact in proximity, with a median distance of 120 bp, which is significantly less than what we would expect by chance (p-value of 3.65E−13 in the Mann-Whitney U test). These results suggest that regulation of IR is orchestrated by complex interactions among TFs, predominantly from the C2H2 ZF family.

Discussion

In our motif analysis, we found that the C2H2 zinc finger family of TFs has a strong association with IR events: Over 65% of the motifs associated with IR have significant hits to C2H2 ZF TFs. This is consistent with previous work reporting that zinc finger TFs influence exon skipping [21], and suggests that the C2H2 ZF family plays an important role in the regulation of alternative splicing in general.

To validate our predictions on the association of these TFs with IR, we used ChIP-seq data for multiple zinc finger TFs: MAZ, EGR1, SP1, ZBTB7A, SP2, and ZNF263. We observed much higher occupancy of these TFs in IR events in the K562 human cell line, validating the model's predictions. Robson et al. [52] have reported that MAZ4 elements that contain four copies of the MAZ binding sequence influence alternative splicing. More recently, it was demonstrated that MAZ acts in conjunction with CTCF to remodel chromatin to affect changes in alternative splicing [53]. They have also demonstrated that like CTCF, MAZ can slow the elongation of RNAPII and affect splicing outcome.

There are multiple potential mechanisms by which TFs can affect co-transcriptional splicing. First, TFs are known to be critical in establishing chromatin state, which in turn can regulate alternative splicing by a purely kinetic model of the coupling between transcription and splicing whereby higher speeds of transcription in regions of accessible chromatin give less time for the spliceosomal machinery to recognize and splice those introns co-transcriptionally [11, 54, 55]. An alternative explanation of this phenomenon is that accessible chromatin is a mark of binding of TFs or other regulatory proteins that recruit splicing factors directly or indirectly through chromatin modifications to affect the outcome of splicing [7]. Wet-lab experiments are required to explore these hypotheses and provide more mechanistic details on how TFs regulate IR and other forms of alternative splicing.

Our model of retained introns considered only chromatin accessibility. There are other aspects of chromatin organization that can be considered: histone modifications and DNA methylation. Through their effect on chromatin organization, histone modifications impact the speed of RNAPII elongation and thereby alternative splicing [54]. Luco et al. [56] proposed the *adaptor system* model whereby DNA-binding proteins recognize a histone modification and recruit a splicing regulator that affects the splicing outcome (see also [57]). Methylation-dependent alternative splicing has been shown to be widespread [58], and its patterns have been observed to delineate exons and their boundaries [59, 60]. Histone modifications and methylation patterns can thus provide another layer of information relevant to the regulation of IR.

In this work, we focused on the local coupling of accessible chromatin and IR. We expect that non-local interactions through chromatin loop anchors like those that allow enhancers to affect promoter activity [61] can affect IR; evidence for their impact on exon skipping has recently been reported in human [62]. Recent work has demonstrated the role of a specific enhancer within a chromatin loop and its role in regulating alternative splicing [63]. Future work can incorporate them in the context of a comprehensive model of alternative splicing.

Conclusions and future work

Using deep learning to model intragenic DHSs allowed us to explore the regulatory elements that are predictive of IR in an unbiased fashion and identify TFs as key contributors to the regulation of IR. Further experimental work is required in order to validate the role of TFs in IR regulation. This will be supported by extensions of the model that allow tissue-specific prediction of the IR state of regions of open chromatin, and create the chromatin-mediated IR code. Furthermore, the modularity of deep learning will allow the extension of the model to incorporate other sources of data indicative of chromatin state such as histone modifications. Much in the same way chromatin loop anchors allow enhancers to affect the activity of promoter regions and affect gene expression [61], there is recent evidence for their impact on exon skipping [62]. Therefore, we expect that chromatin interaction information captured by Hi-C or Micro-C data is likely to improve the model and provide a more holistic view of IR regulation. Such data can be incorporated in a deep learning model with modules that use graph convolution; recent work has shown the effectiveness of this approach for modeling various aspects of chromatin state [64].

Methods

Data collection, processing, and representation

We used DNase I-seq data from 125 human immortalized cell-lines and tissues from the ENCODE database [65] and 39 cell types from the Roadmap Epigenetics consortium [66] as processed by [23]: every DNase I-seq peak was extended to a length of 600 bp around its midpoint and adjacent peaks are greedily merged until no two peaks overlap by more than 200 bp. For our analysis we focused on over a million DHSs that occur within genes.

Next, we extracted IR events from the Ensembl GRCh37 (hg19) reference annotations, utilizing code from SpliceGrapher [67] and iDiffIR [46]. In total, we identified 58,305 unique IR events out of which, 15,400 had overlapping DHSs. These constitute our positive examples. We used a strict criterion requiring a DHS to overlap the retained intron, i.e., DHSs overlapping only the flanking exons did not qualify. All other intragenic DHSs that did not overlap an IR event were labeled as negative examples. The number of negative examples was roughly twice the size of the positive set.

We used two methods to transform the sequences into input for our neural networks: one-hot encoding and sequence embedding. For one-hot encoding, a sequence is represented as a $4 \times N$ matrix where N is the length of the sequence. Each position in the sequence is represented by the columns of the matrix with a non-zero value at a position corresponding to one of the four DNA nucleotides. To represent a sequence using embedding, we first decomposed it into overlapping k-mers of length k and then used a word2vec model [68] to map each k-mer into an m-dimensional vector space. This gave us an embedding matrix of dimensions $(N - k + 1) \times m$. This representation is designed to preserve the context of the k-mers by producing similar embedding vectors for k-mers that tend to co-occur.

Network architecture

In this work, we investigated several network architectures. The primary network element, a one-dimensional convolutional layer, scans a set of filters against the matrix representing the input sequence. Formally, we can express the convolution operation as:

$$x_{i,j} = \sum_{a=0}^{A-1} \sum_{b=0}^{B-1} \boldsymbol{W}^{j}_{a,b} \boldsymbol{X}_{i+a,b}, \tag{1}$$

where $\boldsymbol{X}$ is the input matrix, i is the current output index, and j is the index of the filter. $\boldsymbol{W}$ is the weight matrix with size $A \times B$ where A is the length of the filter (window size) and B is the number of input channels: 4 for DNA one-hot encoding, d in case of word-2vec embeddings, and *number of previous layer filters* in case of higher convolutional layers. The output of a convolutional layer is produced by applying a non-linear activation function to the result of the convolution operation. We use the rectified linear unit (ReLU) which is given by:

$$f(x) = max(0, x). \tag{2}$$

Next, the size of the output is reduced by max-pooling where the maximum value in a window of a pre-determined size is selected. This reduces the input size for the next layer and also leads to invariance to small shifts in the input sequence.

We also incorporated a multi-head self-attention layer as the basis for an alternative deep learning model. Attention is a powerful feature that is able to model dependencies within an input sequence regardless of their distances [69]. By doing so, it guides the network to focus on relevant features within the input and ignore irrelevant information. Our implementation uses the same architecture used for the SATORI method [47], and consists of a single convolutional layer followed by a max-pooling layer and a multi-head attention layer. We also used a recurrent layer as an option in conjunction with the multi-head attention layer, since it provided improved performance in other datasets [47]. RNNs have an internal state that enables them to capture distant feature interactions in the input sequence. Specifically, we employed a bi-directional RNN with Long Short-Term Memory (LSTM) units [70]. In a bi-directional RNN, a forward and a backward layer are used that traverse the input in both directions, improving the model's performance. The bi-directional LSTM layer was used between the convolutional layer and the multi-head attention layer. Code for all the architectures is available in the project's github repository.

Network training and evaluation

First, the data was split into training, validation, and test sets with 80%, 10%, and 10% of the total data, respectively. Next, using the training and validation sets, we tuned the network hyperparameters by employing a semi-randomized grid search that uses a 5-fold cross-validation strategy. For the Basset-like model variant, we started with the hyperparameters reported in [23] and fine-tuned their values. The hyperparameters are summarized in Additional file 1: Table S1. All the models were evaluated

using the test set using the area under the ROC curve (AUC-ROC) and the area under the precision-recall curve (AUC-PRC).

Gapped kmer SVM

As a baseline we used the large-scale gapped kmer SVM (gkm-SVM), called the LS-GKM [41]. This version can handle bigger datasets (50k–100k examples) and exhibits better scalability. We run the package with the following parameters: $-m$ 20000 and $-T$ 16 which specify the size of the memory cache in MB and number of processing threads, respectively.

Motif extraction and analysis

To interpret the deep learning models, we extracted sequence motifs using the weights (filters) of the first convolutional layer, similar to the methodology described by Kelley et al. [23]. We selected the positive examples (DHSs overlapping IR events) in the test set with prediction probability greater than 0.65. This cutoff was chosen as a trade-off between the number of qualified examples and confidence in the prediction. For the negative examples, we used a cutoff value of less than 0.35. Next, for each filter, we identified regions in the set of sequences that activated the filter with a value greater than half of the filter's maximum score over all sequences. The highest scoring regions from all the sequences are stacked and for each filter, a position weight matrix is calculated using the nucleotide frequency and background distribution. We generated the sequence logos using the WebLogo tool [71]. The resulting PWMs were searched against the human CIS-BP database [43] using TomTom [44] with distance metric set to Euclidean. This was performed separately for motif hits in IR and in non-IR events, allowing us to associate motifs with IR or non-IR. For filters which yielded significant hits in both IR and non-IR, we chose the more significant hit.

TF ChIP-seq analysis

We downloaded ChIP peaks of all the TFs that were detected as enriched in IR events from the ENCODE database [65]. Next, we used our previously published pipeline [14] to test the enrichment of a given TF ChIP peaks in IR events. Briefly, we quantified the overlap of ChIP peaks with IR events and compared them to the overlap with non-IR events. The significance of overlap was tested using the Fisher exact test. The accession numbers for the ENCODE K562 ChIP-seq datasets used in our analysis are as follows: EGR1: wgEncodeEH001646, MAZ: wgEncodeEH002862, ZBTB7A: wgEncodeEH001620, SP1: wgEncodeEH001578, SP2: wgEncodeEH001653, ZNF263: wgEncodeEH000630, FOXK2: GSE91647, GATA1: wgEncodeEH000638, JUN: wgEncodeEH000620.

TF knockdown RNA-seq analysis

We downloaded RNA-seq data for knockdown of the following TFs in K562 from the ENCODE database: MAZ, SP1, SP2, E2F4 (accession numbers GEO:GSE88056, GEO:GSE127134, GEO:GSE127145, and GEO:GSE88612). In addition, we used wild-type K562 (accession number GEO:GSE33480). We computed differential IR events in the knockdown samples with respect to baseline K562 using iDiffIR [46]. Analysis of

motif hits in differentially retained introns was performed using BioPython [72] using the motif of each TF retrieved from Jaspar [73].

Discovering interactions between TFs

To discover regulatory interactions between TFs we used SATORI [47], which takes advantage of the self-attention matrix to infer possible interactions between sequence motifs. When running SATORI, we used the default parameters with exception to the following: `--attncutoff 0.08` and `--usevalidtest True`. The postprocessing was performed using Jupyter notebooks provided with SATORI.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02885-1.

Additional file 1. Supplementary Material. The Supplement includes additional tables and figures.

Additional file 2. Review History.

Peer review information
Kevin Pang was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 2.

Authors' contributions
This work was conceived by AB and FU in collaboration with ASNR. All the experiments were performed by FU with additional validation work by SJ. All the authors contributed to the writing of the manuscript. The author(s) read and approved the final manuscript.

Availability of data and materials
Code, training data, and additional results are available through the project's github repository [74]. A copy of the repository is available through zenodo [75].

Declarations

Ethics approval and consent to participate
N/A.

Competing interests
The authors declare that they have no competing interests.

Received: 3 January 2022 Accepted: 16 February 2023
Published online: 22 March 2023

References

1. Kalsotra A, Cooper T. Functional consequences of developmentally regulated alternative splicing. Nature Rev Genet. 2011;12:715–29.
2. Reddy AS. Alternative splicing of pre-messenger rnas in plants in the genomic era. Annu Rev Plant Biol. 2007;58(1):267–94.
3. Monteuuis G, Wong JJ, Bailey CG, Schmitz U, Rasko JE. The changing paradigm of intron retention: regulation, ramifications and recipes. Nucleic Acids Res. 2019;47(22):11497–513.
4. Reddy AS, Rogers MF, Richardson DN, Hamilton M, Ben-Hur A. Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements. Front Plant Sci. 2012;3:18.
5. Chaudhary S, Khokhar W, Jabre I, Reddy AS, Byrne LJ, Wilson CM, Syed NH. Alternative splicing and protein diversity: plants versus animals. Front Plant Sci. 2019;10:708.
6. Wong JJ-L, Au AY, Ritchie W, Rasko JE. Intron retention in mRNA: No longer nonsense. Bioessays. 2016;38(1):41–49.

7. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. Widespread intron retention in mammals functionally tunes transcription. Genome Res. 2014;24:1774–86.
8. Ge Y, Porse BT. The functional consequences of intron retention: alternative splicing coupled to NMD as a regulator of gene expression. Bioessays. 2014;36:(3):236–243.
9. Vanichkina DP, Schmitz U, Wong JJ-L, Rasko JE. Challenges in defining the role of intron retention in normal biology and disease. In: Seminars in Cell & Developmental Biology. Elsevier; 2017.
10. Jung H, Lee D, Lee J, Park D, Kim YJ, Park W-Y, Hong D, Park PJ, Lee E. Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat Genet. 2015;47(11):1242.
11. Naftelberg S, Schor IE, Ast G, Kornblihtt AR. Regulation of alternative splicing through coupling with transcription and chromatin structure. Ann Rev Biochem. 2015;84:165–98.
12. Oesterreich FC, Herzel L, Straube K, Hujer K, Howard J, Neugebauer KM. Splicing of nascent RNA coincides with intron exit from RNA polymerase II. Cell. 2016;165(2):372–81.
13. Mercer TR, Edwards SL, Clark MB, Neph SJ, Wang H, Stergachis AB, John S, Sandstrom R, Li G, Sandhu KS, Ruan Y, Nielsen LK, Mattick JS, Stamatoyannopoulos J. DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. Nat Genet. 2013;45:852–9.
14. Ullah F, Hamilton M, Reddy AS, Ben-Hur A. Exploring the relationship between intron retention and chromatin accessibility in plants. BMC Genomics. 2018;19(1):21.
15. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee B.K, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. The accessible chromatin landscape of the human genome. Nature. 2012;489:(7414):75–82.
16. Felsenfeld G, Groudine M. Controlling the double helix. Nature. 2003;421:448–53.
17. Gross DS, Garrard WT. Nuclease hypersensitive sites in chromatin. Annu Rev Biochem. 1988;57:159–97.
18. Galas DJ, Schmitz A. DNase footprinting: A simple method for detection of protein-DNA binding specificity. Nucleic Acids Res. 1978;5:3157–70.
19. Hesselberth JR, Chen XY, Zhang ZH, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, Fields S, Stamatoyannopoulos JA. Global mapping of protein-DNA interactions in-vivo by digital genomic footprinting. Nat Methods. 2009;6:283–9.
20. Boyle AP, Song LY, Lee B-K, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. High-resolution genome-wide in-vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011;21:456–64.
21. Han H, Braunschweig U, Gonatopoulos-Pournatzis T, Weatheritt RJ, Hirsch CL, Ha KC, Radovani E, Nabeel-Shah S, Sterne-Weiler T, Wang J, et al. Multilayered control of alternative splicing regulatory networks by transcription factors. Mol Cell. 2017;65(3):539–53.
22. Burgess SJ, Reyna-Llorens I, Stevenson SR, Singh P, Jaeger K, Hibberd JM. Genome-wide transcription factor binding in leaves from C3 and C4 grasses. Plant Cell. 2019;31(10):2297–314.
23. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26(7):990–9.
24. Banovich NE, Li YI, Raj A, Ward MC, Greenside P, Calderon D, Tung PY, Burnett JE, Myrthil M, Thomas SM, et al. Impact of regulatory variation across human iPSCs and differentiated cells. Genome Res. 2018;28(1):122–31.
25. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18(1):67.
26. Atak ZK, Taskiran II, Demeulemeester J, Flerin C, Mauduit D, Minnoye L, Hulselmans G, Christiaens V, Ghanem G-E, Wouters J, Aerts S. Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning. Genome Res. 2021;31(6):1082–96.
27. Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet. 2022;54(7):940–9.
28. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18(10):1196–203.
29. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831.
30. Qin Q, Feng J. Imputation for transcription factor binding predictions based on deep learning. PLoS Comput Biol. 2017;13(2):1005403.
31. Trabelsi A, Chaabane M, Ben-Hur A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. Bioinformatics. 2019;35(14):269–77.
32. Koh PW, Pierson E, Kundaje A. Denoising genome-wide histone chip-seq with convolutional neural networks. Bioinformatics. 2017;33(14):225–33.
33. Schreiber J, Libbrecht M, Bilmes J, Noble W. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. bioRxiv. 2018;103614.
34. Pan X, Shen H-B. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. BMC Bioinformatics. 2017;18(1):136.
35. Zhang S, Zhou J, Hu H, Gong H, Chen L, Cheng C, Zeng J. A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Res. 2015;44(4):32.
36. Jaganathan K, Panagiotopoulou SK, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, Gao H, Kia A, Batzoglou S, Sanders SJ, Farh KK-H. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535–548.
37. Zeng T, Li YI. Predicting rna splicing from dna sequence using Pangolin. Genome Biol. 2022;23(1):1–18.

38. Leung MK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. Bioinformatics. 2014;30(12):121–9.
39. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, Krainer AR, Jojic N, Scherer S, Blencowe BJ, Frey BJ. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347(6218):1254806.
40. Jha A, Gazzara MR, Barash Y. Integrative deep models for alternative splicing. Bioinformatics. 2017;33(14):274–82.
41. Lee D. LS-GKM: a new gkm-svm for large-scale datasets. Bioinformatics. 2016;32(14):2196–8.
42. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
43. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook KB, Zheng HY, Goity A, van Bakel H, Lozano JF, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJM, Bouget F-Y, Ratsch G, Larrondo LF, Ecker JR, Hughes TR. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158:1431–43.
44. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2006;8:24.
45. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecenas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LOF, Lei EP, Fraser AG, Blencowe BJ, Morris QD, Hughes TR. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013;499(7457):172–7.
46. Filichkin SA, Hamilton M, Dharmawardhana PD, Singh SK, Sullivan C, Ben-Hur A, Reddy AS, Jaiswal P. Abiotic stresses modulate landscape of poplar transcriptome via alternative splicing, differential intron retention, and isoform ratio switching. Front Plant Sci. 2018;9:5.
47. Ullah F, Ben-Hur A. A self-attention model for inferring cooperativity between regulatory features. Nucleic Acids Res. 2021;49(13):77.
48. Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46(D1):380–6.
49. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic Acids Res. 2017;45(D1):D408–14. https://doi.org/10.1093/nar/gkw985.
50. Sheffield NC, Bock C. Lola: enrichment analysis for genomic region sets and regulatory elements in r and bioconductor. Bioinformatics. 2016;32(4):587–9.
51. Simovski B, Kanduri C, Gundersen S, Titov D, Domanska D, Bock C, Bossini-Castillo L, Chikina M, Favorov A, Layer RM, et al. Coloc-stats: a unified web interface to perform colocalization analysis of genomic features. Nucleic Acids Res. 2018;46(W1):186–93.
52. Robson-Dixon ND, Garcia-Blanco MA. MAZ elements alter transcription elongation and silencing of the fibroblast growth factor receptor 2 exon IIIb. J Biol Chem. 2004;279(28):29075–84.
53. Xiao T, Li X, Felsenfeld G. The Myc-associated zinc finger protein (MAZ) works together with CTCF to control cohesin positioning and genome organization. Proc Natl Acad Sci. 2021;118(7):e2023127118.
54. Saldi T, Cortazar MA, Sheridan RM, Bentley DL. Coupling of RNA polymerase II transcription elongation with pre-mRNA splicing. J Mol Biol. 2016;428(12):2623–35.
55. Dvinge H. Regulation of alternative mRNA splicing: old players and new perspectives. FEBS Lett. 2018;592:2987-3006. https://doi.org/10.1002/1873-3468.13119.
56. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. Science. 2010;327(5968):996–1000.
57. Schor IE, Allo M, Kornblihtt AR. Intragenic chromatin modifications: A new layer in alternative splicing regulation. Epigenetics. 2010;5(3):174–9.
58. Wan J, Oliver VF, Zhu H, Zack DJ, Qian J, Merbs SL. Integrative analysis of tissue-specific methylation and alternative splicing identifies conserved transcription factor binding motifs. Nucleic Acids Res. 2013;41(18):8503–14.
59. Gelfman S, Cohen N, Yearim A, Ast G. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. Genome Res. 2013;23(5):789–99.
60. Lev Maor G, Yearim A, Ast G. The alternative role of DNA methylation in splicing regulation. Trends Genet. 2015;31(5):274–80. https://doi.org/10.1016/j.tig.2015.03.002.
61. Greenwald WW, Li H, Benaglio P, Jakubosky D, Matsui H, Schmitt A, Selvaraj S, D'Antonio M, D'Antonio-Chronowska A, Smith EN, Frazer KA. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. Nat Commun. 2019;10(1):1–17.
62. Zhang Y, Cai Y, Roca X, Kwoh CK, Fullwood MJ. Chromatin loop anchors predict transcript and exon usage. Briefings in Bioinformatics. 2021;22(6):bbab254. https://doi.org/10.1093/bib/bbab254.
63. Dahan S, Sharma A, Cohen K, Baker M, Taqatqa N, Bentata M, Engal E, Siam A, Kay G, Drier Y, Elias S, Salton M. VEGFA's distal enhancer regulates its alternative splicing in CML. NAR Cancer. 2021;3(3):029.
64. Lanchantin J, Qi Y. Graph convolutional networks for epigenetic state prediction using both sequence and 3D genome data. Bioinformatics. 2020;36:659–67.
65. ENCODE-Project-Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.
66. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30. https://doi.org/10.1038/nature14248.
67. Rogers MF, Thomas J, Reddy ASN, Ben-Hur A. SpliceGrapher: Detecting patterns of alternative splicing from RNA-seq data in the context of gene models and EST data. Genome Biol. 2012;13:1–17.
68. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. Curran Associates, Inc; 2013. pp. 3111–3119.
69. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in Neural Information Processing Systems. Curran Associates, Inc; 2017. pp. 5998–6008.
70. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.

71. Crooks GE, Hon G, Chandonia JM, Brenner SE. Weblogo: A sequence logo generator. Genome Res. 2004;14:1188–90.
72. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–3.
73. Fornes O, Castro-Mondragon JA, Khan A. Van der Lee R, Zhang X, Richmond P.A, Modi B.P, Correard S, Gheorghe M, Baranašić D, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020;48(D1):87–92.
74. Ullah F, Ben-Hur A. chromIR: a Deep Learning Method for Detecting Retained Introns in Accessible DNA. GitHub. https://github.com/fahadahaf/chromir.
75. Ullah F, Ben-Hur A. chromIR: a Deep Learning Method for Detecting Retained Introns in Accessible DNA. Zenodo. https://doi.org/10.5281/zenodo.7626606.

Publisher's Note