Epigenome-based Splicing Prediction using a Recurrent Neural Network

Donghoon Lee[1,2], Jing Zhang[1,2], Jason Liu[2], and Mark B Gerstein[1,2,3,4*]

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

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

3 Department of Computer Science, Yale University, New Haven, CT 06520, USA

4 Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA

* Corresponding author

E-mail: pi@gersteinlab.org

Abstract

Alternative RNA splicing provides an important means to expand metazoan transcriptome diversity. Contrary to what was accepted previously, splicing is now thought to predominantly take place during transcription. Motivated by emerging data showing the physical proximity of the spliceosome to Pol II, we surveyed the effect of epigenetic context on co-transcriptional splicing. In particular, we observed that splicing factors were not necessarily enriched at exon junctions and that most epigenetic signatures had a distinctly asymmetric profile around known splice sites. Given this, we tried to build an interpretable model that mimics the physical layout of splicing regulation where the chromatin context progressively changes as the Pol II moves along the guide DNA. We used a recurrent-neural-network architecture to predict the inclusion of a spliced exon based on adjacent epigenetic signals, and we showed that distinct spatio-temporal features of these signals were key determinants of model outcome, in addition to the actual nucleotide sequence of the guide DNA strand. After the model had been trained and tested (with >80% precision-recall curve metric), we explored the derived weights of the latent factors, finding they highlight the importance of the asymmetric time-direction of chromatin context during transcription.

Author Summary

In humans, only about 2% of the genome is comprised of so-called coding regions and can give rise to protein products. However, the human transcriptome is much more diverse than the number of genes found in these coding regions. Each gene can give rise to multiple transcripts through a process during transcription called alternative splicing. There is a limited understanding of the regulation of splicing and the underlying splicing code that determines cell-type-specific splicing. Here, we studied epigenetic features that characterize splicing regulation in humans using a recurrent neural network model. Unlike feedforward neural networks, this method contains an internal memory state that learns from spatiotemporal patterns – like the context in language – from a sequence of genomic and epigenetic information, making it better suited for characterizing splicing. We demonstrated that our method improves the prediction of spicing outcomes compared to previous methods. Furthermore, we applied our method to 49 cell types in ENCODE to investigate splicing regulation and found that not only spatial but also temporal epigenomic context can influence splicing regulation during transcription.

Introduction

Alternative splicing of pre-messenger RNA plays an integral role in diversifying the transcriptome. This process is more pervasive in higher eukaryotes and is estimated to affect approximately 95% of protein-coding genes in humans [1,2]. Accurate characterization of the process by which multiple functional protein products are produced from a single gene is crucial for understanding the function of the transcriptome [3].

Recent discoveries have revealed that splicing occurs predominantly during transcription in humans [4–8]. Nascent RNA is almost immediately spliced upon transcription [9,10] and introns are mostly spliced out during transcript elongation. This timing suggests that the recruitment of splicing factors and spliceosome assembly, detection of exon-intron boundaries, and modulation of alternative splicing must occur at the same time scale as transcription [9].

Co-transcriptional splicing indicates a key observation that splicing takes place progressively in the direction of RNA transcription, rather than processed simultaneously after transcription. As a result, the contexts of guide DNA, nascent RNA, and its immediate folded structure progressively change as RNA polymerase II (Pol II) moves along the guide DNA strand [11] and may influence splicing regulation. Furthermore, co-transcriptional splicing signifies the physical proximity of the spliceosome assembly to Pol II and other transcriptional machinery [9]. Pol II physically interacts with nucleosomes and its histone modifications around them, modulating the transcription rate [12].

DNA sequence alone may not contain sufficient information to process alternative splicing deterministically [13]. Djebali et al. [4] and many others have shown that there is an enrichment of chromatin marks around spliced exons, suggesting the role of epigenetic modifications during context-dependent modulation of alternative splicing [14,15]. For example, exonic boundaries are characterized by increased levels of nucleosome density and positioning [16–18], DNA methylation [19,20], and strong enrichment of specific histone modifications including H3K36me3, H3K79me1, H2BK5me1, H3K27me1, H3K27me2, and H3K27me3 [16,17,21–23]. In addition, a recent genome-wide survey of alternative splicing showed that DNA methylation can either enhance or silence exon recognition in a context-dependent manner [24]. Furthermore, studies have shown that there is significant regulatory crosstalk between histone modifications during transcriptional elongation [12].

Despite many efforts to characterize the splicing regulatory code both experimentally and computationally, we have yet to understand how the cell type-specific epigenomic context is utilized during co-transcriptional splicing. Previous computational methods on splicing have largely focused on discovering novel splice junctions based on RNA sequencing (RNA-seq) alignments [25,26], utilizing machine learning approaches [27,28] including deep neural networks [29]. Only a limited set of tools can model splicing regulation based on genomic sequences and select RNA features [30–32]. Moreover, studies on splicing regulation have focused heavily on identifying mutations that land within splice sites (SSs), cis-acting splicing regulatory elements, and trans-acting splicing factors [30,33]. The extent, nature, and effects of the epigenetic context in splicing regulation remain unsolved.

In this study, we propose a new computational approach to characterize the role of epigenetic modifications during co-transcriptional splicing. To build an interpretable model, we adopted a recurrent neural network (RNN) architecture, which to some degree resembles the physical characteristics of co-transcriptional splicing (Figure 1). The model can learn from a temporal sequence of epigenetic contexts, similar to how epigenetic contexts progressively change as Pol II moves forward along the guide DNA strand during co-transcriptional splicing. The RNN model allows us to predict the inclusion of exons based on adjacent DNA sequences and epigenetic modifications. Moreover, the physical resemblance of the model allows us to interpret the trained model weight parameters and explore the spatio-temporal links between the guide DNA elements and the surrounding epigenetic modifications. In summary, we leveraged the mechanistic properties of co-transcriptional splicing to build an interpretable splicing model, and we explored the trained model to understand the underlying characteristics of the epigenetic context during co-transcriptional splicing.

Results

We first explore the epigenetic data context around known splice sites in depth. We then describe the model and rationale for applying the specific architecture. Finally, we use the model to further examine the effect of epigenetic context during co-transcriptional splicing.

Distinct epigenomic signatures characterize splicing regulation

We studied the epigenetic context of alternative splicing by examining the enrichment of multiple histone modifications and DNA methylations around the exon-intron boundary. We mapped the epigenomic signatures around SSs of cassette exons at a base-pair resolution. We aggregated multiple histone modifications across 49 cell types in ENCODE and observed their enrichment as a function of distance from SSs (Figure 2A, B, Supplementary Figure 1, 2A, B). We found the most interesting trend within 100 bp of SSs for both the 3' acceptor and 5' donor. A strong enrichment pattern of H3K36me3 and H3K27me3 appeared around the exon boundary. Although studies have demonstrated a role for H3K36me3 in defining the exon-intron boundary [22,34], the dynamic interplay between other histone modifications has been overlooked. From the 3' acceptor, peak enrichment occurred around 100 bp into the exon; at the 5' donor, it was closer, at around 50 bp into the exon. We also observed a slight depletion of H3K27ac and H3K4me3 marks within 100 bp of the intron at the 3' acceptor SS but not within the 5' donor SS. As this region contains a branch site, these histone marks may indicate a role in defining the branch point.

Enrichment of RNA-binding factors around splice sites

Alternative splicing regulation is an elaborate process that requires precise coordination of multiple splicing factors and enzymes. Studies have shown that RNA-binding proteins (RBPs) facilitate splicing regulation during transcription [35]. For example, the serine/arginine-rich splicing factor family member SRSF7 binds to poised exons and promotes the inclusion rate [36][37]. Another member of the serine/arginine-rich splicing factor family, U2AF1, is responsible for mediating the binding of U2 small nuclear ribonucleoprotein to the pre-mRNA branch site [38]. The recent release of the ENCODE project included enhanced CLIP experiments (eCLIP) datasets that span 112 RBPs from K562 and HepG2 cell types. As sequence-specific RBPs have been shown to facilitate splicing regulation in a context-specific manner [15], we investigated their spatial relationship to both the 5’ donor and 3’ acceptor splicing sites. Specifically, we investigated the enrichment of splicing factors (n=29) and their relative distance to these sites. We observed that, on average, splicing factors show preferential binding to the intronic side of the splicing site in both 3’ acceptor and 5’ donor SSs (Supplementary Figure 2C). Furthermore, we found that splicing factors may show slightly different patterns in their spatial binding preferences. In particular, hnRNP A1 and SRSF1 were enriched in the intronic region outside 3’ SSs whereas SF3B4 and hnRNP C were enriched in the exonic region (Figure 2C). At 5’ SSs, RBM22 and PRPF8 were bound at the exonic end, which has been shown to be critical for splicesome assembly [39,40].

Correlating epigenomic signatures to exonic expression

We tested whether histone modifications have any effect on inclusion and expression of alternative exons. We observed a trend where enrichment of H3K36me3 at the exon-intron boundary was positively correlated with exonic expression, whereas H3K27me3 marks showed the opposite trend (Figure 3A, B, Supplementary Figure 3). Compared to excluded or nominally expressed alternative exons, highly expressed spliced exons had statistically significant enrichment of H3K36me3 and depletion of H3K27me3 at their exon-intron boundary (Figure 3C). The contrasting trend and the correlation of these histone methylations to exonic expression suggest that the splicing code may be directly or indirectly encoded within the epigenomic context.

Clustering biosamples based on splicing patterns

Previous studies have shown that various epigenomic marks are correlated across similar tissues and cell types [41]. It is now widely accepted that the transcriptional regulatory circuitry of a particular cell type is reflected in its epigenetic landscape. To explore the potential linkage between epigenetic regulation and tissue-specific splicing, we examined splicing patterns across 49 ENCODE biosamples. Based on a similarity of percent-splice-in (PSI) values for all coding exons (n=185,405), we clustered biosamples into five categories using hierarchical clustering (Figure 3D). Splicing patterns were highly correlated among tissue types from the same cell-of-origin, reproducing similar clustering results based on epigenetic marks. For example, blood-lineage cell types formed cluster C2 whereas brain and neural cells were clustered in cluster C4.

In addition to using the PSI similarity matrix to cluster cell types into categories, we can project the cells onto a low-dimensional cell space using principal component analysis (PCA). We measured alternative splicing patterns in terms of exonic expression level (fragment per kilobase per million reads mapped, FPKM) across diverse ENCODE cell types and examined how cells are placed in the context of others. Interestingly, we observed that cancer-related cell lines were located proximal to each other in the PCA cell space (Supplementary Figure 4).

Modeling splicing regulation: key characteristics of an RNN architecture

To investigate the latent representation of splicing instruction encoded within the epigenomic context, we aimed to construct a predictive model of splicing. We opted for an RNN architecture, which has proven successful in various sequential information processing and prediction tasks such as natural language processing and translation [42–44], to explore the contribution of the epigenomic context to the regulation of alternative splicing.

We start by describing a simple RNN, which shares many of the features we intend to model. A simple RNN is made of many recurrent neurons that are sequentially linked to each other. A neuron at specific time point t is influenced by previous time point $t-1$, combining some relationship of the current input x_t with the previous hidden state h_{t-1}.

$$h_t = f(h_{t-1}, x_t)$$

where h_t is hidden state at time t and x_t is input variable at time t. If we suppose the activation function as a hyperbolic tangent for a simple RNN, the state at time t can be represented as

$$h_t = \tanh({W_h}^T h_{t-1} + {W_x}^T x_t + \mathrm{b})$$

where W_h and W_x are the weight of the hidden state and input variable, respectively, and b is the bias vector. The output can be expressed in terms of an output weight matrix, W_y, and a hidden state at time t, h_t:

$$\widehat{y_t} = \mathrm{S}\left({W_y}^T h_t\right)$$

where S is sigmoid function:

$$S(x) = \frac{e^x}{e^x + 1}$$

This time-dependency allows us to explore the complex contextual relationship between features. In particular, we adopted the long short-term memory (LSTM) [45] model to describe an RNN architecture. In principal, a simple RNN allows us to model a time-dependent task from sequential data. However, in practice, the simple model suffers from the problem of vanishing gradients, where the gradients responsible for updating weights with respect to the partial derivative of error function becomes negligible in a long sequence and hampers the model from learning long-term time dependencies. Therefore, we used both LSTM and gated recurrent unit (GRU), which have many of the same simple intuitive properties of the simple RNN but allow learning from longer sequences. The LSTM is an extension of the same idea that includes more sophisticated gates, which allows the cell to retain long-term memory between cells while avoiding the problem of vanishing gradients when training the network. The specific equations for the LSTM model we adopted is shown in the Methods.

Modeling splicing regulation: How the RNN architecture fits the problem

The rationale for applying an RNN to our model is that (1) an RNN is optimized for processing sequential information like genomic sequences and epigenomic profiles along genomic coordinates, (2) an RNN has a time-direction resembling how RNA is transcribed by RNA polymerase in the 5’ to 3’ direction, (3) temporal memory cells of an RNN allow the model to learn about complex context-dependent relationships among epigenomic features, such as the influence of features and input seen at t-1 on the neural cell at time t, and (4) an RNN is very flexible with the type of input and output data and therefore can easily integrate heterogeneous sequential information. Not surprisingly, researchers recently have applied RNN models to the area of genomics to predict non-coding DNA function [46] and to detect exon junctions [47]. Moreover, since the mechanics of the RNN calculation is somewhat parallel to the actual spatial and temporal dependency found in co-transcriptional splicing, the overall results from the trained model are more readily interpretable. The data processing and implementation of the predictive models are collected in a package named Epigenome-based Splicing Prediction using Recurrent Neural Network (ESPRNN; available at https://github.com/gersteinlab/esprnn). Using our method, we attempted to decipher context-dependent effects of various epigenomic features on splicing for both canonical (e.g., dinucleotide GT for 5' donors and AG for 3' acceptors) and non-canonical SSs. Our model is especially useful since splicing signals are not only enriched at the splice site but often found up and downstream of splice sites.

Modeling splicing regulation: Initial evaluation

We used ESPRNN to predict alternate usages of cassette exons (inclusion or exclusion of exons), the most common form of alternative splicing events [48], using DNA sequences and epigenomic signals adjacent to SSs (Figure 4A). We used the exon definition of splicing, which is considered to be the dominant mechanism in higher eukaryotes [49]. Our model had an average F1 score (harmonic mean of the precision and recall) of 0.8472 for the LSTM-based model across cell types [0.8757 for the GRU-based model] using five core histone modification tracks (Figure 4B). The average F1 score marginally increased to 0.8573 when using 17 histone, chromatin accessibility, DNA methylation, and nucleosome density profiles.

We performed the splicing prediction with or without the RBP profile and measured how much predictive performance is gained from additional information. We observed a marginal improvement in predictive performance when RBP binding profiles were added to the baseline model (measured in improvement of F1 score from 0.84 to 0.86) (Supplementary Figure 9A, B). This suggests RBP binding information may be redundant and already represented in the epigenetic features. We also compared prediction results from normal cell types to those from cancerous cell lines. Since previous studies on cancer-specific alternative splicing [50,51] have suggested potential linkage of aberrant splicing events to the disease risk [52–55], we expected to see differences in splicing regulation between normal and cancerous cell types. However, we did not observe a significant difference in prediction performance between normal and cancerous cell types (average F1 score for normal biosamples: 0.8465, cancerous biosamples: 0.8765). We also cross-tested a model trained from one cell type to another. After we fit our model to one cell type, we transferred the fitted weights and model parameters to predict splicing on other cell types. When we tested between cell types from the same cell-of-origin (e.g., train on adult liver model and test on HepG2 data, train on lung model and test on A549 data), we did not observe a significant difference in predictive performance. However, we observed a moderate reduction in splicing prediction performance when we cross-tested cells from different cell-of-origin (Supplementary Figure 5B, F1 score is better metric for comparing cross-cell testing due to class imbalance across cell types). Thus, the epigenomic regulatory landscape around SSs appears to be generally conserved across cell types. Moreover, we compared the classification performance to other models based on random forest and k-nearest neighbors and found that our model was superior in terms of classification accuracy (Figure 4D, Supplementary Figure 7).

We tried to measure the contribution of each individual epigenetic feature to splicing in a number of ways. (1) We performed an empirical analysis via a leave-one-out strategy. Using GM12878 as an example, we first built a reference model based on all available epigenetic features. By removing one variable at a time, we then measured the mean decrease in F1 score and area under the receiver operating characteristic curve (ROC AUC), as an indicator of variable importance (Figure 4C). (2) Alternatively, we trained a DNA-only model using DNA sequence features only and compared to a "baseline model." The baseline model was trained using DNA sequence features plus additional chromatin accessibility (DHS) and 6 histone marks. Here, we observed a significant loss of predictive performance in the DNA-only model (13% reduction in F1 score) (Supplementary Figure 6A). (3) Next, starting from the DNA-only model, we added one epigenetic feature at a time to measure the information gain from each feature (Supplementary Figure 6B). While the addition of some epigenetic features like H3K27ac increased the variability in prediction performance, an active mark H3K36me3 or a repressive mark H3K27me3 was the most informative at predicting splicing. Moreover, the combination of both H3K36me3 and H3K27ac further improved the prediction performance compared to other pairs (Supplementary Figure 6C). We observed that the combination of H3K36me3 and H3K27ac features together contributed more than when they were used individually (Supplementary Figure 6D).

Overall, we found H3K36me3 to be the most important variable in predicting splicing. This observation coincides with previous studies reporting that H3K36me3 recruits the splicing factors PTB [34] and SRSF1 [56] to facilitate splicing. Interestingly, one of the top predictors of splicing was H3K79me2, which was previously shown to associate with H3K36me3 at gene bodies [57]. H3K9me3, a histone modification that can recruit adaptor proteins like HP1 to facilitate splicing factors [24], was also ranked among the top predictors.

Interpretation of weights of the splicing model

Since the model follows the physical layout of splicing regulation, one can examine the trained model and learn from the trained weights how each epigenetic feature contributes to splicing regulation. To interpret the splicing model, we designed an LSTM-based model composed of only one hidden state and trained for a longer period (400 epochs). We made sure that this simplified model performs nearly as well at predicting splicing as our main model (usually after >20 epochs of training, Supplementary Figure 8A). We also made sure that the overall predictive performance of the simplified model is stable after approximately 100 epochs (Supplementary Figure 8B, C). When we analyzed the simplified model, we found that the trained weights of various gates at the recurrent unit showed that open chromatin (DHS), H3K27ac, K3K36me3, and H3K4me1 are weighted more highly than other epigenetic features -- as expected (Supplementary Figure 8D). We also noticed that H3K27me3 and K3K9me3 were negatively weighted at the input gate, suggesting that these features have a negative impact on exon inclusion, consistent with our previous findings.

Influence of temporal epigenetic context on splicing regulation

We specifically designed our splicing model to represent the physical layout of splicing regulation, where a sequence of chromatin contexts is fed progressively to the model. Therefore, the model takes into account the temporal direction (progression from 5’ to 3’ in direction). To show that model has learned this asymmetric temporal relationship of epigenetic features, we first trained a baseline model (in the normal 5’ to 3’ direction) and then fed a series of epigenetic signals in a “reverse” order (3’ to 5’ in direction) as input to it. We observed how the model prediction behaved in this context. If the model was agnostic to the temporal direction of features, both forward and reverse input features should give the same predictive power. In fact, we saw a moderate decrease in prediction performance (Supplementary Figure 9A, B), with an F1 score decreasing from 0.871 to 0.865 and ROC AUC decreasing from 0.886 to 0.865.

Discussion

Our prediction model revealed that the epigenomic signature of an SS plays a large role in determining the splicing outcome. In addition, the positive results suggest that our model can be extended to predict the full transcriptomic composition from a genomic and epigenomic context. We expect that we could further improve the proposed model by adding more deep hidden layers and increasing the number of training samples by utilizing the full set of available epigenomic data in the ENCODE project. Our approach does contain some limitations, as it is still challenging to visualize and evaluate the multi-dimensional context of the weight matrix in the trained model. We could apply dimensionality reduction techniques to probe the latent representation of relationships between various epigenomic signals.

In this study, we used ENCODE polyA RNA-seq assays to measure splicing and exon-level expression; we note that this is an indirect measure of what is actually happening during transcription. RNAs are often unstable and may be subjected to many post-transcriptional modifications. RNA-seq measures the steady-state level of the transcript, accounting for both mRNA synthesis and decay. Future studies with a more direct measure of transcriptional rates, such as nuclear run-on assays like global run-on (GRO-seq) or bromouridine sequencing (Bru-seq), will allow us to accurately measure the effect of epigenomic context on splicing and, ultimately, on the transcriptional rate.

Future studies should focus on comparing splicing models from normal and cancer samples in the hope of illuminating the differences in the epigenomic landscapes of splicing regulation.

Although splicing is an elaborate process, it could become pathogenic when misregulated [58,59]. Unsurprisingly, aberrant splicing events, which collectively referred to splicing events that could confer the risk of a disease, are often implicated in systemic diseases like cancer [51,60]. Aberrant splicing events based on mutations are relatively well characterized [54,60–62]; however, a large fraction of aberrant splicing events that have no direct mutational cause still remain unknown. Although our understanding of epigenomic context on splicing regulation is incomplete, our prediction model highlights that splicing is elaborately regulated via various epigenomic signatures. This suggests that epigenomic dysregulation may be closely linked to the onset of aberrant splicing. Thus, even though aberrantly spliced RNAs in healthy cells may be degraded by the mRNA surveillance system, epigenomic dysregulation may render this checkpoint system useless. Further studies on cell-type-specific and context-dependent splicing regulation will reveal whether epigenetic modulation can serve as a therapeutic method of complex disease in the future.

Methods

Dataset

The current release of the ENCODE dataset provides an unprecedented number of functional assays across broad biosample types, including primary cells and tissues. In this study, we leveraged both the breadth and depth of ENCODE, including assays for histone modification (chromatin immunoprecipitation sequencing, ChIP-seq), chromatin accessibility (DNase I hypersensitive sites sequencing, DNase-seq), RBPs (eCLIP), methylations (WGBS and RRBS) and gene expression (RNA-seq), to systematically probe the data-rich context of alternative splicing and its regulation. The list of accessions for experiments used in this study is found in Supplementary Table 1.

Processing of RNA-seq data

To quantify levels of exon expression from RNA-seq data, we collected all raw sequencing reads from experiments tagged as reference epigenome series from the ENCODE portal. These reads were polyA plus long RNA-seq (200 bp or larger) from whole-cell fractions rather than nuclear or cytosolic fractions. To minimize potential batch effects and sample bias, we carefully selected untreated experiments from the reference epigenome series. As of November 2019, there are 81 cell and tissue types (covering 49 unique biosamples) in the reference epigenome series, including both RNA-seq and ChIP-seq of H3K4me1, H3K4me3, H3K36me3, H3K27ac, H3K27me3, and H3K9me3. We first aligned all RNA-seq data to the GRCh38 genome using RNA STAR (v 2.7.0). Since the model requires splice site annotation, we constructed exon annotation from GENCODE version 24 (to synchronize with ENCODE annotation) by extracting all unique exons with known protein-coding transcripts. We excluded exons that could ambiguously map to both chromosome X and Y. This analysis included 597,937 exons (185,405 unique exons after removing duplicates from isoforms) that averaged 28.01 exons per gene and 296.49 bp in length (150.92 bp in length for unique exons). We obtained read counts at each exon using HTSeq (v0.11.2) [63]. Based on read counts, we used a custom script (esprnn/preproc_calcExonFPKM.py) to calculate normalized exonic expression levels in FPKM. Our rationale for using the exonic expression was to intentionally make the model agnostic to the overall transcript level. Each exon was evaluated independently from other exons, and we counted the number of sequencing reads supporting the inclusion of a particular exon. The counts were normalized similar to how a gene's expression is normalized by size of annotation and total number of mapped reads (FPKM). We binarized the exonic expression level (FPKM) using a threshold of one. Therefore, we only considered whether an exon has enough evidence supporting exon inclusion.

In addition to the exonic expression level, alternatively, we calculated a metric, PSI, to measure the level of splicing. PSI represents the fraction of the reads supporting exon inclusion from the split reads at the splice junction. We used a custom script (esprnn/scripts/calcPSI.sh) based on equations from Schafer et al. [64] to calculate PSI normalized by the size of read and exon annotation.

$$\widetilde{F_i}^{incl} = \frac{F_i^{incl}}{L_i + L_f}$$

$$\widetilde{F_i}^{excl} = \frac{F_i^{excl}}{L_f}$$

$$PSI\ (\Psi) = \frac{\widetilde{F_i}^{incl}}{\widetilde{F_i}^{incl} + \widetilde{F_i}^{excl}} \%$$

F_i^{incl} number of reads or fragments supporting the inclusion of i-th exon; F_i^{excl} number of reads or fragments supporting the exclusion of i-th exon; L_f fragment length; L_i size of i-th exon.

RNA-binding proteins

RBP enrichment was calculated based on the peaks identified from the eCLIP experiments. We downloaded the ENCODE eCLIP uniformly processed peaks from K562 and HepG2 cell types (see Supplementary Table 1 for eCLIP data accession). The peak was called using CLIPPER software [65] and filtered for peaks having a score of 1,000. We then counted numbers of RBP binding events at a base-pair resolution, agnostic to cell type.

To examine preferential binding patterns of splicing factors around SSs, RBP peaks were annotated as splicing-related factors if they belong to hnRNP- and SR-families (n=29). We extended both 3' acceptor and 5' donor SS by 1,000 bp in both up and downstream direction and binned the region into 100 bp intervals. We defined the position relative to the distance to the SS, in the 5' to 3' direction. For each interval, we calculated the frequency of splicing factor binding normalized to the size of the interval. The value of RBP enrichment means the normalized binding frequency of splicing-related factors.

LSTM model

We adopted the following equations for the modeling of splicing using LSTM. σ function denotes sigmoid function. $\otimes$ denotes Hadamard product where two matrices are multiplied in a pair-wise fashion. x_t denotes input vector and h_t denotes output vector, f_t denotes forget gate vector, i_t denotes input or update gate vector, o_t denotes output gate vector, c_t denotes cell state vector.

$$f_t = \sigma({W_{hf}}^T h_{t-1} + {W_{xf}}^T x_t + b_f)$$

$$i_t = \sigma({W_{hi}}^T h_{t-1} + {W_{xi}}^T x_t + b_i)$$

$$o_t = \sigma({W_{ho}}^T h_{t-1} + {W_{xo}}^T x_t + b_o)$$

$$g_t = \tanh({W_{hg}}^T h_{t-1} + {W_{xg}}^T x_t + b_g)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t$$

$$h_t = o_t \otimes \tanh(c_t)$$

GRU model

We adopted the following equations for the modeling of splicing using GRU. x_t denotes input vector and h_t denotes output vector, z_t denotes update gate vector and r_t denotes reset gate vector.

$$z_t = \sigma\left({W_{hz}}^T h_{t-1} + {W_{xz}}^T x_t + b_z\right)$$

$$r_t = \sigma({W_{hr}}^T h_{t-1} + {W_{xr}}^T x_t + b_r)$$

$$h_t = (1 - z_t)h_{t-1}\otimes + z_t\otimes\tanh({W_{hh}}^T(r_t\otimes h_{t-1}) + {W_{xh}}^T x_t + b_h)$$

Pre-processing of data for the training model

We selected six normal and three cancer samples from the reference epigenome series. The dataset contains consolidated epigenomes from the Roadmap Epigenomics Consortium [41] and the ENCODE Consortium. All datasets were uniformly processed and mapped to the GRCh38 human reference genome. All samples contained a core set of histone modification tracks (H3K4me1, H3K4me3, H3K36me3, H3K27ac, H3K27me3, and H3K9me3) as well as RNA-seq data. We used additional histone modification tracks, as well as DNase I hypersensitivity, DNA methylation, and nucleosome positioning tracks, to predict alternative splicing upon availability. Detailed information on datasets used can be found in Supplementary Table 1. For each exon, we obtained DNA sequences at intron-exon boundaries (3’ acceptors) and exon-intron boundaries (5’ donors), as well as 100 bp upstream and downstream of SSs. Splice junctions included both canonical and non-canonical SSs. We processed all sequences to read in the 5’ to 3’ direction using strand information from each gene. Each 400 bp DNA sequence was encoded into a 1,000 by 4 binary array using one-hot encoding. We used RNA-seq expression profiles to indicate tissue-specific alternative splicing patterns. Genes having fewer than two exons were discarded and the first and last exons were excluded from the analysis. We classified an exon as being expressed if its FPKM was greater than or equal to 1. We normalized all ChIP-seq histone modification tracks and DNase-seq tracks over corresponding input signal tracks using MACS v2.0.10 (https://github.com/taoliu/MACS) [66]. We used negative log10 of the Poisson p-value to measure the enrichment level over the background. Due to the wide dynamic range observed, we used a p-value threshold of 1e-2 for the upper limit. We processed all feature tracks including DNA methylation and nucleosome signal tracks to read in the 5' to 3' direction and scaled them to a range of 0 to 1.

Performance evaluation of the model

There is no single metric that can give you a measure of performance in a binary classification problem. Relying on one metric can be misleading especially when there is high class imbalance. Therefore, we employed various metrics to measure the performance of the predictive model. ROC curve explains the tradeoff between true-positive rate (TPR) and false-positive rate (FPR). PR curve visualizes the tradeoff between positive predictive value (PPV) and true-positive rate (TPR).

$$Precison = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

In addition, we used F1-score, which is the harmonic mean of precision and recall, to measure the performance of the splicing model.

$$F_1 = 2 \cdot \frac{precison \cdot recall}{precison + recall}$$

Hyperparameter tuning of splicing model and training

We tested a range of dimensions and depths of RNN models and network design hyperparameters to optimize the alternative splicing model. We chose optimal hyperparameters by tuning one parameter at a time while fixing the rest. Hyperparameters included but were not limited to the number of recurrent layers, size of neurons in each layer, pooling strategy, dropout rate, choice of activation function and loss function, optimizer, and number of the epoch. We shuffled the order of the data and split the dataset into training and test sets using an 80 to 20% ratio. 20% of test data was set aside for the performance evaluation. 80% of training data was split again between 80 to 20% (64 and 16% of the original data) for fitting the model and validating the model fit during the training phase. We fed a range of sequences from 50 to 1,000 bp within each SS and found the 400 bp span to be the ideal size for the model. For the neural network architecture, we achieved the best result when two RNN units were stacked together, which allowed the model to learn higher-level temporal representations. We used a hidden state size of two by default and we recommend not using a hidden state size greater than 128 to avoid overfitting problems (Supplementary Figure 8A). We applied three variants of the RNN model, LSTM [45], GRU [67], and simple RNN. To compare the performance of memory-based units

(LSTM and GRU), we implemented a simple RNN model using the same network architecture. We found that both LSTM and GRU were capable of learning long-term dependencies and were effective in learning high-dimensional contextual relationships between epigenomic features around the SSs. We split the input sequences into two parts where the first half represented a 3' acceptor SS and the latter half represented a 5' donor SS. We fed these sequences into two separate RNN units of size 200 and merged them into another RNN unit of size 400. The last RNN layer was followed by a dropout layer to prevent overfitting of the training dataset. The last fully-connected layer contained the softmax activation function for classifying exons as either spliced or unspliced. To train the model, we used a binary cross-entropy objective function with the Adam optimizer [68]. For each dataset, we trained the model for 20 epochs. We tested the implementation of ESPRNN using TensorFlow v2.0 (https://www.tensorflow.org). Our implementation also works with Keras v1.0.3 or v2.2.4 (https://github.com/fchollet/keras) with either TensorFlow v1.15 and Theano v0.8.2 [69] backend with a minor tweak. We used various Nvidia GPUs (Titan K20m, K80, GTX 1080ti, RTX2080, P100, and Titan V) to train the model.

Acknowledgements

We would like to acknowledge Steve Weston from the Yale Center for Research Computing for technical support in setting up our GPU computing infrastructure.

References

1. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456: 470–6. doi:10.1038/nature07509
2. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40: 1413–5. doi:10.1038/ng.259

3. Graveley BR. Alternative splicing: Increasing diversity in the proteomic world. Trends in Genetics. 2001. pp. 100–107. doi:10.1016/S0168-9525(00)02176-4
4. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489: 101–108. doi:10.1038/nature11233
5. Listerman I, Sapra AK, Neugebauer KM. Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. Nat Struct Mol Biol. 2006;13: 815–822. doi:10.1038/nsmb1135
6. Wada Y, Ohta Y, Xu M, Tsutsumi S, Minami T, Inoue K, et al. A wave of nascent transcription on activated human genes. Proc Natl Acad Sci. 2009;106: 18357–18361. doi:10.1073/pnas.0902573106
7. Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, et al. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat Struct Mol Biol. 2011;18: 1435–1440. doi:10.1038/nsmb.2143
8. Girard C, Will CL, Peng J, Makarov EM, Kastner B, Lemm I, et al. Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. Nat Commun. 2012;3: 994. doi:10.1038/ncomms1998
9. Carrillo Oesterreich F, 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: 372–381. doi:10.1016/j.cell.2016.02.045
10. Alpert T, Herzel L, Neugebauer KM. Perfect timing: splicing and transcription rates in living cells. Wiley Interdisciplinary Reviews: RNA. Blackwell Publishing Ltd; 2017. doi:10.1002/wrna.1401
11. Herzel L, Ottoz DSM, Alpert T, Neugebauer KM. Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. Nature Reviews Molecular Cell Biology. Nature Publishing Group; 2017. pp. 637–650. doi:10.1038/nrm.2017.63
12. Tanny JC. Chromatin modification by the RNA polymerase II elongation complex. Transcription. 2014;5. doi:10.4161/21541264.2014.988093
13. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. 2012;22: 1616–1625. doi:10.1101/gr.134445.111
14. Motta-Mena LB, Heyd F, Lynch KW. Context-Dependent Regulatory Mechanism of the Splicing Factor hnRNP L. Mol Cell. 2010;37: 223–234. doi:10.1016/j.molcel.2009.12.027
15. Fu X-DD, Ares M. Context-dependent control of alternative splicing by RNA-binding proteins. Nat Rev Genet. 2014;15: 689–701. doi:10.1038/nrg3778
16. Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Res. 2009;19: 1732–1741. doi:10.1101/gr.092353.109
17. Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. Nat Struct Mol Biol. 2009;16: 990–995. doi:10.1038/nsmb.1659
18. Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcárcel J, et al. Nucleosome positioning as a determinant of exon recognition. Nat Struct Mol Biol. 2009;16: 996–1001. doi:10.1038/nsmb.1658
19. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature. 2011;479: 74–9. doi:10.1038/nature10442

20. Lev Maor G, Yearim A, Ast G. The alternative role of DNA methylation in splicing regulation. Trends Genet. 2015;31: 274–280. doi:10.1016/j.tig.2015.03.002

21. Hon G, Wang W, Ren B. Discovery and annotation of functional chromatin signatures in the human genome. Segal E, editor. PLoS Comput Biol. 2009;5: e1000566. doi:10.1371/journal.pcbi.1000566

22. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet. 2009;41: 376–381. doi:10.1038/ng.322

23. Spies N, Nielsen CB, Padgett RA, Burge CB. Biased Chromatin Signatures around Polyadenylation Sites and Exons. Mol Cell. 2009;36: 245–254. doi:10.1016/j.molcel.2009.10.008

24. Yearim A, Gelfman S, Shayevitch R, Melcer S, Glaich O, Mallm JP, et al. HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. Cell Rep. 2015;10: 1122–1134. doi:10.1016/j.celrep.2015.01.038

25. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25: 1105–1111. doi:10.1093/bioinformatics/btp120

26. Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res. 2010;38: 4570–4578. doi:10.1093/nar/gkq211

27. Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. 2001;29: 1185–90. doi:10.1093/nar/29.5.1185

28. Sonnenburg S, Schweikert G, Philips P, Behr J, Rätsch G. Accurate splice site prediction using support vector machines. BMC Bioinformatics. 2007;8: S7. doi:10.1186/1471-2105-8-S10-S7

29. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;0. doi:10.1016/j.cell.2018.12.015

30. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature. 2010;465: 53–59. doi:10.1038/nature09000

31. Xiong HY, Barash Y, Frey BJ. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. Bioinformatics. 2011;27: 2554–2562. doi:10.1093/bioinformatics/btr444

32. Barash Y, Vaquero-Garcia J, González-Vallinas J, Xiong HY, Gao W, Lee LJ, et al. AVISPA: a web tool for the prediction and analysis of alternative splicing. Genome Biol. 2013;14: R114. doi:10.1186/gb-2013-14-10-r114

33. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet. 2002;3: 285–298. doi:10.1038/nrg775

34. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. Science (80-). 2010;327: 996–1000. doi:10.1126/science.1184208

35. Witten JT, Ule J. Understanding splicing regulation through RNA splicing maps. Trends Genet. 2011;27: 89–97. doi:10.1016/j.tig.2010.12.001

36. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature. 2007;446: 926–929. doi:10.1038/nature05676

37. Pervouchine D, Popov Y, Berry A, Borsari B, Frankish A, Guigó R. Integrative transcriptomic analysis suggests new autoregulatory splicing events coupled with nonsense-mediated mRNA decay. Nucleic Acids Res. 2019;47: 5293–5306. doi:10.1093/nar/gkz193

38. Ruskin B, Zamore PD, Green MR. A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. Cell. 1988;52: 207–219. doi:10.1016/0092-8674(88)90509-0

39. Rasche N, Dybkov O, Schmitzová J, Akyildiz B, Fabrizio P, Lührmann R. Cwc2 and its human homologue RBM22 promote an active conformation of the spliceosome catalytic centre. EMBO J. 2012;31: 1591. doi:10.1038/EMBOJ.2011.502

40. Wickramasinghe VO, Gonzàlez-Porta M, Perera D, Bartolozzi AR, Sibley CR, Hallegger M, et al. Regulation of constitutive and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5′ splice site strength. Genome Biol. 2015;16: 201. doi:10.1186/s13059-015-0749-3

41. Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518: 317–330. doi:10.1038/nature14248

42. Graves A, Mohamed A, Hinton G. Speech Recognition with Deep Recurrent Neural Networks. IEEE Int Conf Acoust Speech Signal Process. 2013; 6645–6649. doi:10.1109/ICASSP.2013.6638947

43. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Proc 2014 Conf Empir Methods Nat Lang Process. 2014; 1724–1734. doi:10.3115/v1/D14-1179

44. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation By Jointly Learning To Align and Translate. Iclr 2015. 2014; 1–15. doi:10.1146/annurev.neuro.26.041002.131047

45. Hochreiter S, Schmidhuber J, Hochreiter S, Schmidhuber J, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9: 1735–80. doi:10.1162/neco.1997.9.8.1735

46. Quang D, Xie X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44: gkw226. doi:10.1093/nar/gkw226

47. Lee B, Lee T, Na B, Yoon S. DNA-Level Splice Junction Prediction using Deep Recurrent Neural Networks. arXiv e-prints. 2015; 1–6. Available: http://arxiv.org/abs/1512.05135

48. Koscielny G, Texier V Le, Gopalakrishnan C, Kumanduri V, Riethoven JJ, Nardone F, et al. ASTD: The Alternative Splicing and Transcript Diversity database. Genomics. 2009;93: 213–220. doi:10.1016/j.ygeno.2008.11.003

49. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 2010;11: 345–55. doi:10.1038/nrg2776

50. Liu S, Cheng C. Alternative RNA splicing and cancer. Wiley Interdiscip Rev RNA. 2013;4: 547–566. doi:10.1002/wrna.1178

51. Oltean S, Bates DO. Hallmarks of alternative splicing in cancer. Oncogene. 2014;33: 5311–5318. doi:10.1038/onc.2013.533

52. Jiang P, Freedman ML, Liu JS, Liu XS. Inference of transcriptional regulation in cancers. Proc Natl Acad Sci U S A. 2015. doi:10.1073/pnas.1424272112

53. Ntziachristos P, Abdel-Wahab O, Aifantis I. Emerging concepts of epigenetic dysregulation in hematological malignancies. Nature Immunology. 2016.

doi:10.1038/ni.3517

54. Jung H, Lee D, Lee J, Park D, Kim YJ, Park W-Y, et al. Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat Genet. 2015;47: 1242–1248. doi:10.1038/ng.3414

55. Obeng EA, Ebert BL. Charting the “Splice” Routes to MDS. Cancer Cell. 2015. doi:10.1016/j.ccell.2015.04.016

56. Pradeepa MM, Sutherland HG, Ule J, Grimes GR, Bickmore WA. Psip1/Ledgf p52 Binds Methylated Histone H3K36 and Splicing Factors and Contributes to the Regulation of Alternative Splicing. Reik W, editor. PLoS Genet. 2012;8: e1002717. doi:10.1371/journal.pgen.1002717

57. Huff JT, Plocik AM, Guthrie C, Yamamoto KR. Reciprocal intronic and exonic histone modification regions in humans. Nat Struct Mol Biol. 2010;17: 1495–1499. doi:10.1038/nsmb.1924

58. Venables JP. Aberrant and alternative splicing in cancer. Cancer Research. 2004. pp. 7647–7654. doi:10.1158/0008-5472.CAN-04-1910

59. Tazi J, Bakkour N, Stamm S. Alternative splicing and disease. Biochimica et Biophysica Acta - Molecular Basis of Disease. 2009. pp. 14–26. doi:10.1016/j.bbadis.2008.09.017

60. Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene. 2016;35: 2413–2427. doi:10.1038/onc.2015.318

61. Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes and Development. Cold Spring Harbor Lab; 2003. pp. 419–437. doi:10.1101/gad.1048803

62. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science (80-). 2014;347: 1254806-. doi:10.1126/science.1254806

63. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics. 2015. doi:10.1093/bioinformatics/btu638

64. Schafer S, Miao K, Benson CC, Heinig M, Cook SA, Hubner N. Alternative Splicing Signatures in RNA□seq Data: Percent Spliced in (PSI). Curr Protoc Hum Genet. 2015;87. doi:10.1002/0471142905.hg1116s87

65. Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol. 2013;20: 1434–1442. doi:10.1038/nsmb.2699

66. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008;9: R137. doi:10.1186/gb-2008-9-9-r137

67. Cho K, van Merrienboer B, Bahdanau D, Bengio Y. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. Proc SSST-8, Eighth Work Syntax Semant Struct Stat Transl. 2014; 103–111. Available: http://arxiv.org/abs/1409.1259

68. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. Int Conf Learn Represent. 2014; 1–13. Available: http://arxiv.org/abs/1412.6980

69. Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints. 2016; 19. Available: http://arxiv.org/abs/1605.02688

Supporting Information Legend

Supplementary Table 1

List of datasets and the accession numbers used for the study.

Supplementary Table 2

Overview of dataset used for training the ESPRNN model. The model was trained using the CORE (highlighted in red) and FULL set based on the availability of data. The CORE set was used to compare the predictive performance across cell types.

Supplementary Table 3

ESPRNN model prediction performance measured by F1 score. Predictive performance was compared between the CORE and FULL set of genomic features. For each set, performance was compared using LSTM, GRU, and simple RNN models. Predictive performance was measured by F1 score.

Supplementary Table 4

Comparison of models trained with 50 bp span and 100 bp span data. Each model was trained using genomic features derived from 50 bp span or 100 bp span data from splice sites using the LSTM model. Performance was measured using F1 score and ROC AUC.

Supplementary Figure 1

(Shadow figure of the main Figure 2A) Enrichment of various epigenomic marks of HepG2 at the exon-intron boundary. High PSI indicates exon inclusion, mid PSI indicates exons with 40-60% PSI, and low PSI indicates exon skipping.

Supplementary Figure 2

(Shadow figure of the main Figure 2B) Comparison of epigenetic enrichment around different segments of the 3’ acceptor site for **(A)** K562 and **(B)** HepG2. High PSI indicates exon inclusion, mid PSI indicates exons with 40-60% PSI, and low PSI indicates exon skipping. Mann-Whitney-Wilcoxon two-sided test, ns: $0.05 < p <= 1$; *: $0.01 < p <= 0.05$; **: $0.001 < p <= 0.01$; ***: $0.0001 < p <= 0.001$; ****: $p <= 0.0001$. **(C)** Fold enrichment of splicing-related RBPs to non-splicing-related RBPs around the 3’ acceptor splice site and 5’ donor splice site.

Supplementary Figure 3

Correlation of exonic expression (FPKM) and histone enrichment of **(A)** HepG2 H3K36me3, **(B)** HepG2 H3K27me3, **(C)** liver H3K36me3, and **(D)** liver H3K27me3. PCC: Pearson Correlation Coefficient.

Supplementary Figure 4

Splicing patterns based on exonic expression level (FPKM) for diverse ENCODE cell types are projected on a PCA cell space.

Supplementary Figure 5

(A) Difference in splicing prediction performance when RBP binding profiles were added as an additional feature of the base model containing chromatin accessibility and histone marks. **(B)** Cross-cell testing of model. Model was trained on HepG2 data and tested on K562 data, and vice versa.

Supplementary Figure 6

(A) Comparison of the baseline model trained using chromatin accessibility and 6 histone marks to a model using DNA sequence feature only **(B)** Measure of information gain from additional epigenetic feature based on DNA sequence only model **(C)** Comparison of splicing prediction performance using a pair of epigenetic features.

Supplementary Figure 7

Comparison of LSTM-based model with other machine learning algorithms. Four different algorithms, k-Nearest neighbor (kNN), decision tree, random forest, and support vector machine (SVM), were compared to the LSTM-based model across four different tissue types (A549, HepG2, GM12878, K562).

Supplementary Figure 8

(A) Comparison of splicing prediction performance across different sizes of hidden state. **(B)** Loss of training an LSTM model with 1 hidden layer for 400 epochs. **(C)** Accuracy of training an LSTM model with one hidden layer for 400 epochs. **(D)** Trained weights of LSTM recurrent cells.

Supplementary Figure 9

Comparison of splicing prediction performance when input features are reversed in time-direction via a **(A)** precision-recall curve and a **(B)** ROC curve.

Figure Legend

Figure 1

Overview of the co-transcriptional splicing model. Depiction of co-transcriptional splicing in terms of **(A)** biological context, **(B)** genomic and epigenomic data context, and how it relates to the **(C)** RNN model.

Figure 2

(A) Enrichment of various epigenomic marks of K562 at the exon-intron boundary. We aggregated histone modifications up to 500 bp upstream and downstream of intronic and exonic regions flanking 3’ and 5’ SSs for cassette exons across ENCODE cell types. High PSI indicates exon inclusion, mid PSI indicates exons with 40-60% PSI, and low PSI indicates exon skipping. **(B)** Statistical significance testing of epigenetic mark enrichment. Average histone modification enrichment at four exonic segments were compared based on PSI values. Mann-Whitney-Wilcoxon two-sided test, ns: 0.05 < p <= 1; *: 0.01 < p <= 0.05; **: 0.001 < p <= 0.01; ***: 0.0001 < p <= 0.001; ****: p <= 0.0001. **(C)** RBP enrichment across the exon-intron boundary.

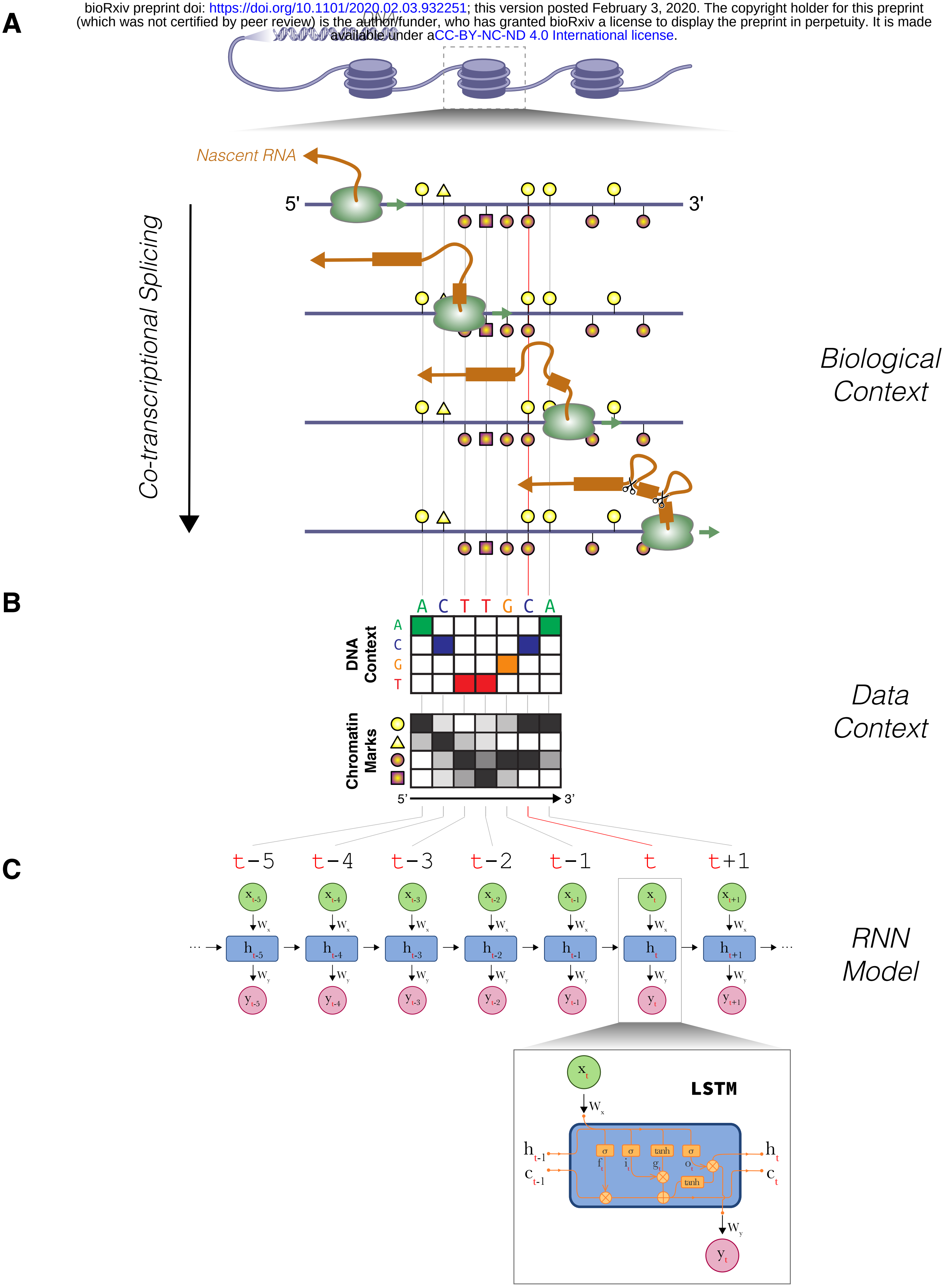
Figure 3

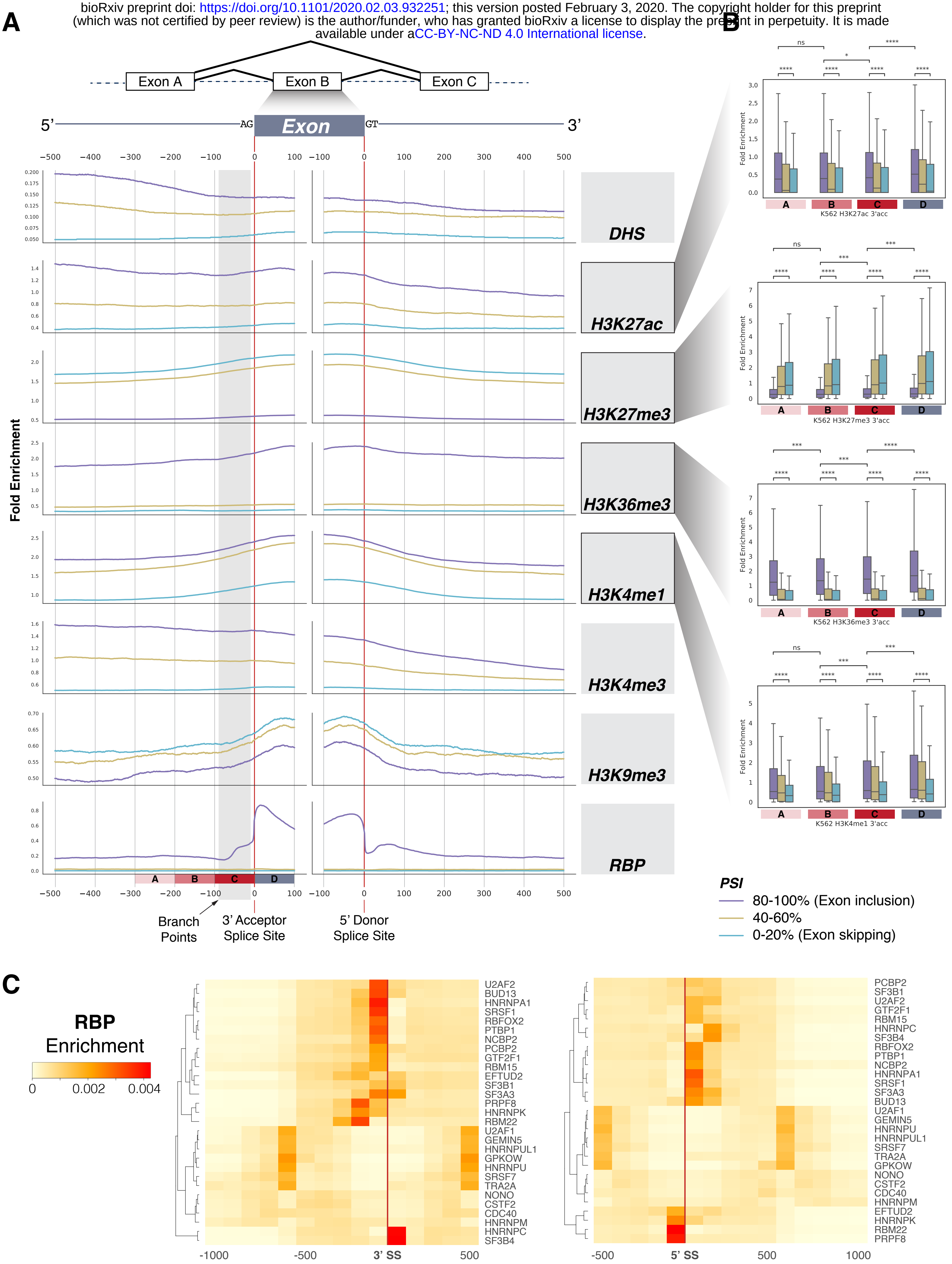
Correlation of exonic expression to **(A)** H3K36me3 and **(B)** H3K27me3. The line represents a linear regression model fit, and the shaded band represents 95% confidence interval. **(C)**

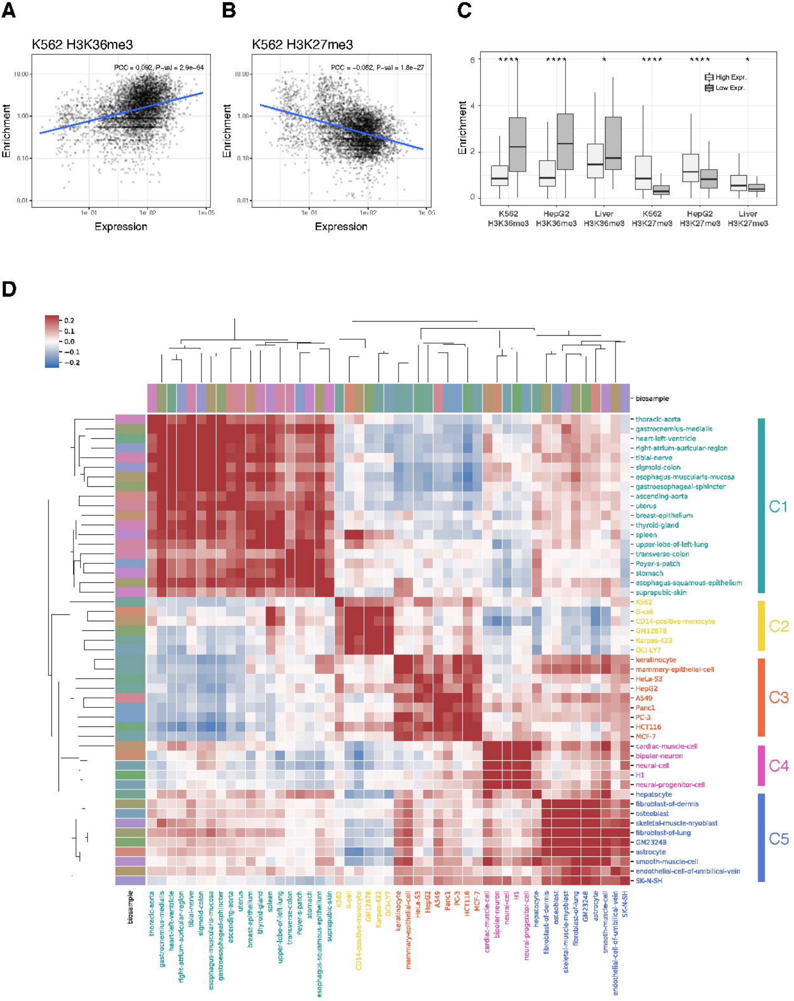
Alternative exons were grouped by expression level and their relative histone enrichment was compared near the SSs. Asterisks represents statistical significance using the Wilcoxon rank sum test; (*) P <= 0.05, (**) P <= 0.01, (***) P <= 0.001, (****) P <= 0.0001. **(D)** Hierarchical clustering of similarity based on PSI across 49 ENCODE biosamples. The results are clustered into five categories of cell types.

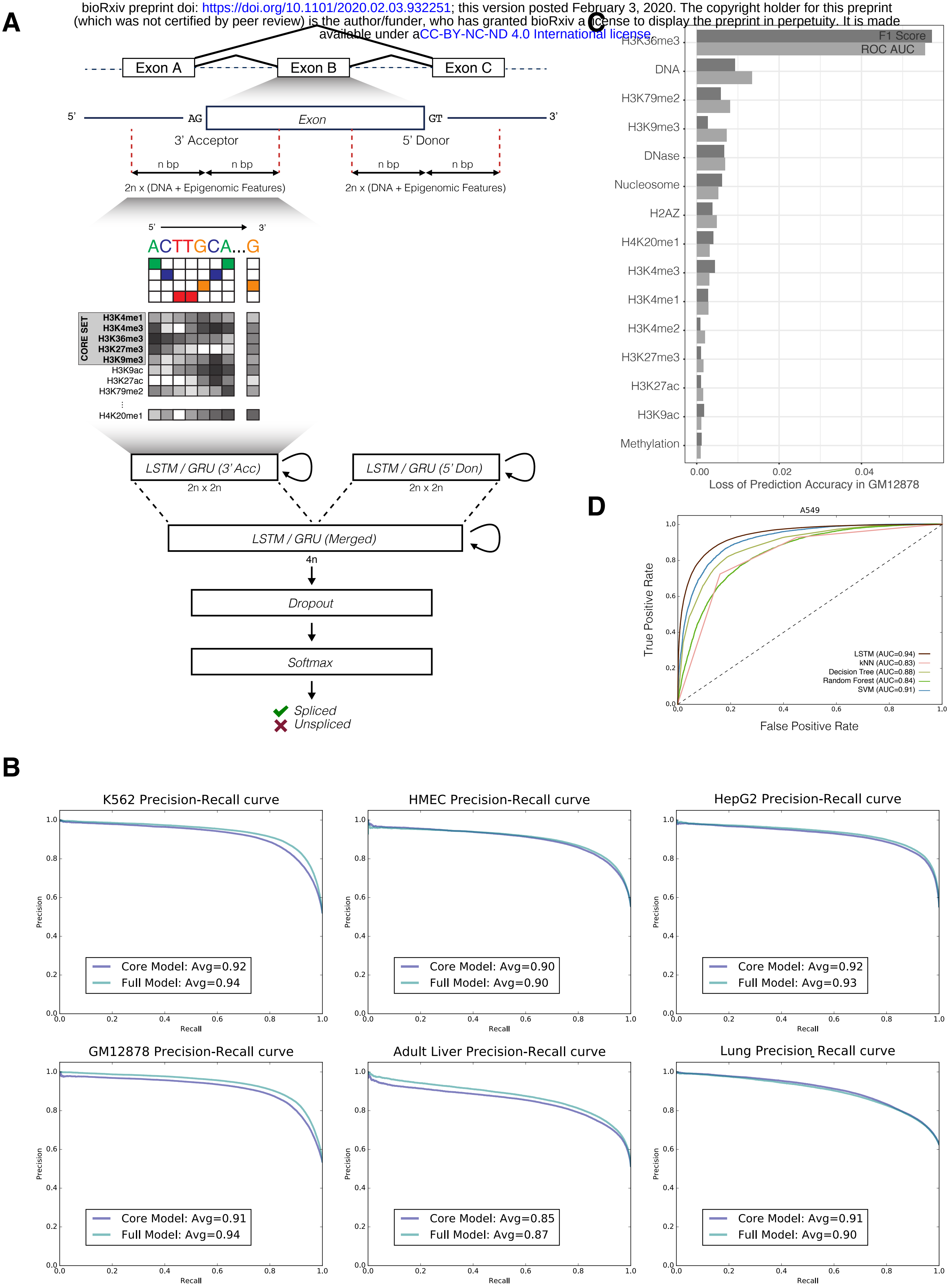
Figure 4

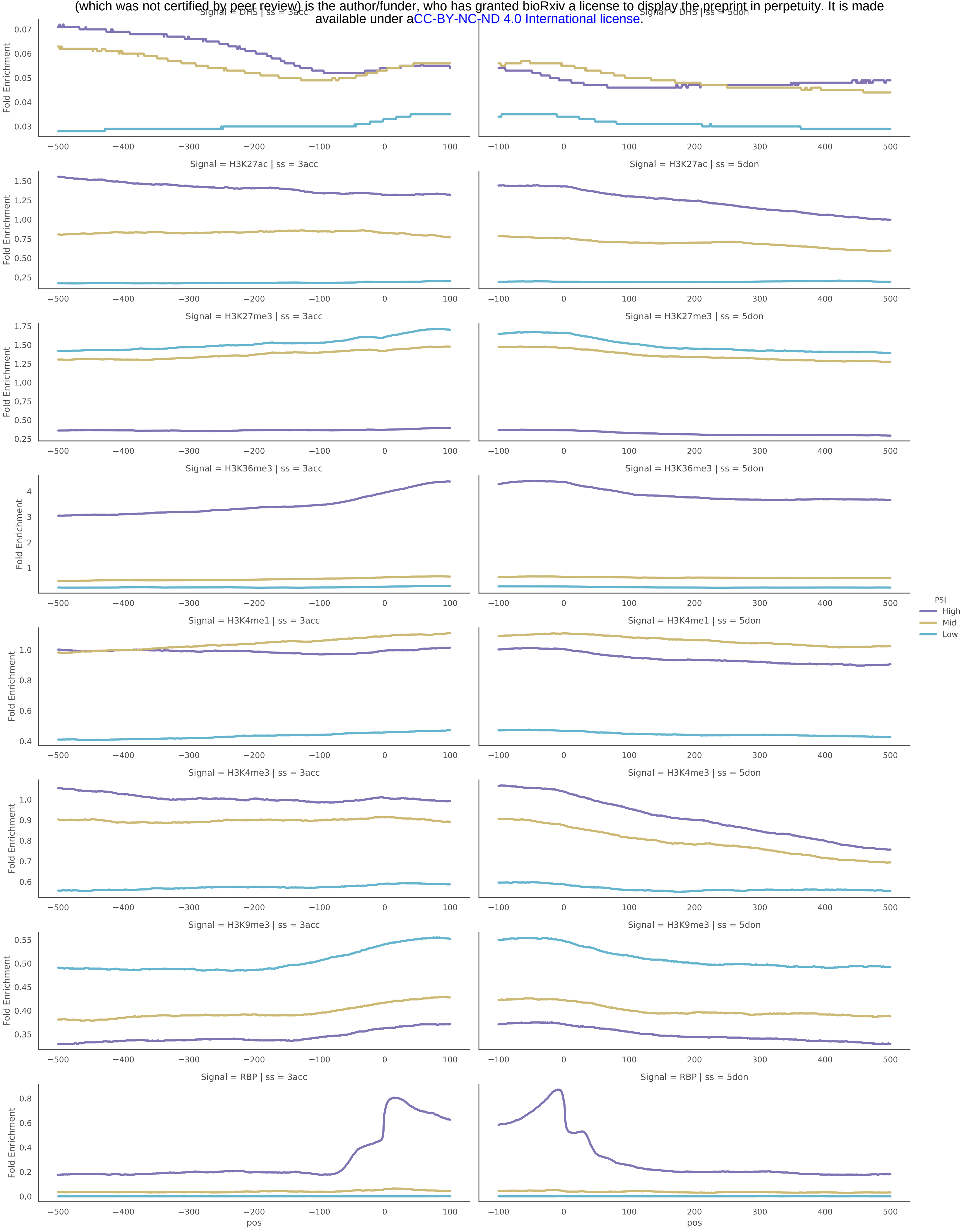
(A) Overview of the ESPRNN model. The model is composed of two recurrent layers. Inputs from 3’ and 5’ SSs are separately processed in the first recurrent layer and then merged in the next recurrent layer. A softmax classifier is used to determine the inclusion of the exon. Using genomic sequences and epigenomic contexts as input, the alternative usage of the exon is predicted. **(B)** Precision-recall curves from six different ENCODE cell types. **(C)** Epigenetic features that contribute to splicing regulation. The order and magnitude of importance was determined using leave-one-out analysis and loss of the ROC AUC was calculated when training the model lacking a particular feature. **(D)** Comparison of LSTM model with other models based on k-nearest neighbor, support vector machine, decision tree, and random forest algorithms.

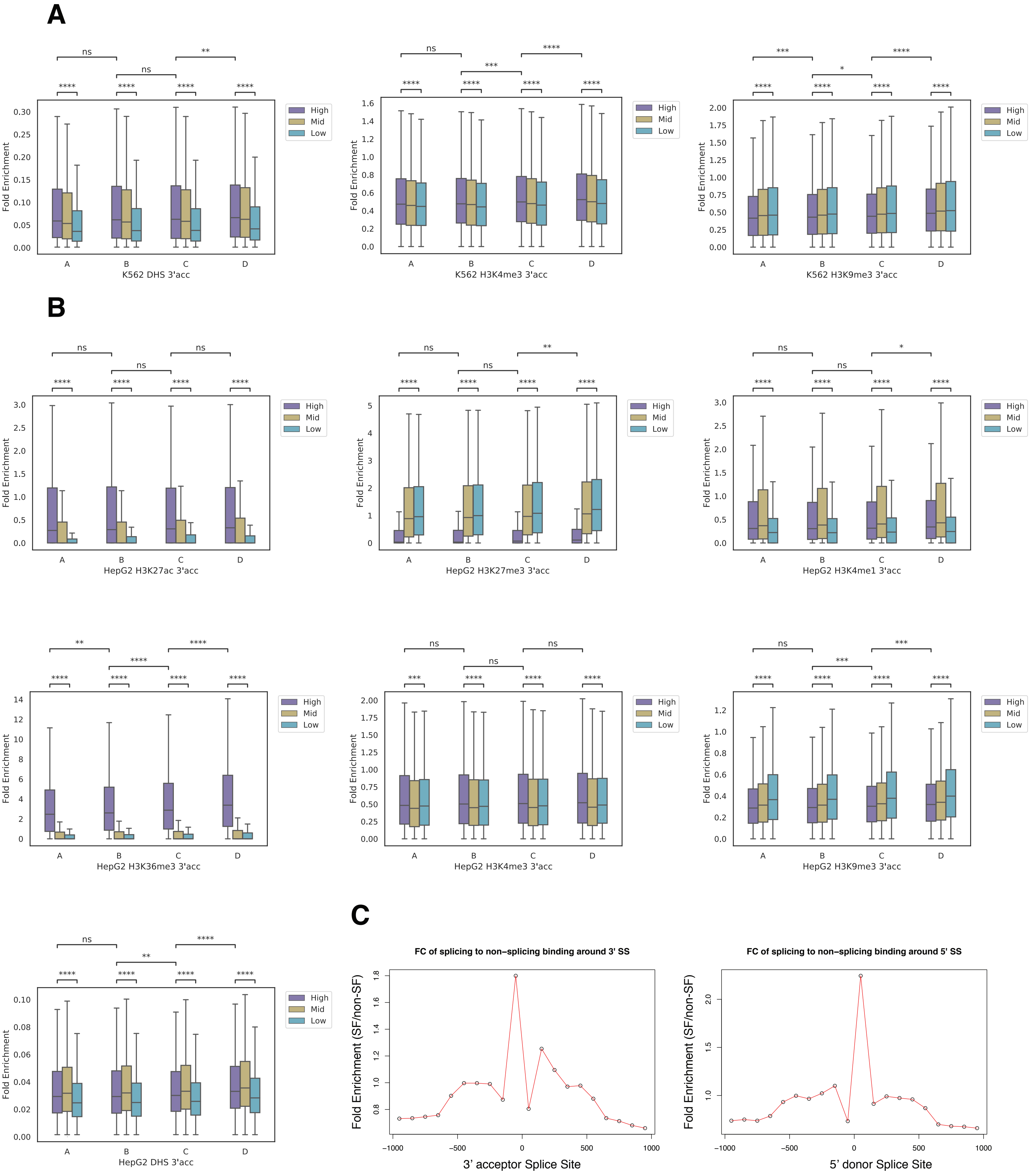

A
DNA
Nascent RNA
5'
3'
Co-transcriptional Splicing
Biological Context
B
A C T T G C A
DNA Context
A
C
G
T
Chromatin Marks
5'
3'
Data Context
C
t-5
t-4
t-3
t-2
t-1
t
t+1
x t-5
x t-4
x t-3
x t-2
x t-1
x t
x t+1
W x
h t-5
h t-4
h t-3
h t-2
h t-1
h t
h t+1
W y
y t-5
y t-4
y t-3
y t-2
y t-1
y t
y t+1
RNN Model
LSTM
h t-1
c t-1
σ
σ
tanh
σ
f t
i t
g t
o t
tanh
h t
c t

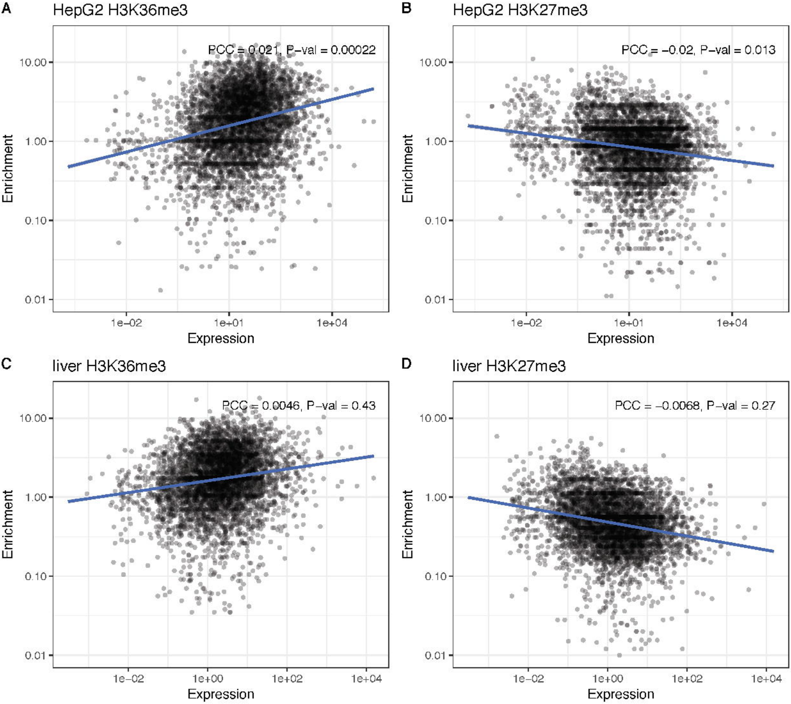

A
Exon A
Exon B
Exon C
5'
AG
Exon
GT
3'
−500 −400 −300 −200 −100 0 100 −100 0 100 200 300 400 500
Fold Enrichment
DHS
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K9me3
RBP
A B C D
Branch Points
3' Acceptor Splice Site
5' Donor Splice Site
B
Fold Enrichment
ns
*
****
***
K562 H3K27ac 3'acc
K562 H3K27me3 3'acc
K562 H3K36me3 3'acc
K562 H3K4me1 3'acc
PSI
80-100% (Exon inclusion)
40-60%
0-20% (Exon skipping)
C
RBP Enrichment
0
0.002
0.004
U2AF2
BUD13
HNRNPA1
SRSF1
RBFOX2
PTBP1
NCBP2
PCBP2
GTF2F1
RBM15
EFTUD2
SF3B1
SF3A3
PRPF8
HNRNPK
RBM22
U2AF1
GEMIN5
HNRNPUL1
GPKOW
HNRNPU
SRSF7
TRA2A
NONO
CSTF2
CDC40
HNRNPM
HNRNPC
SF3B4
-1000
-500
3' SS
500
PCBP2
SF3B1
U2AF2
GTF2F1
RBM15
HNRNPC
SF3B4
RBFOX2
PTBP1
NCBP2
HNRNPA1
SRSF1
SF3A3
BUD13
U2AF1
GEMIN5
HNRNPU
HNRNPUL1
SRSF7
TRA2A
GPKOW
NONO
CSTF2
CDC40
HNRNPM
EFTUD2
HNRNPK
RBM22
PRPF8
-500
5' SS
500
1000

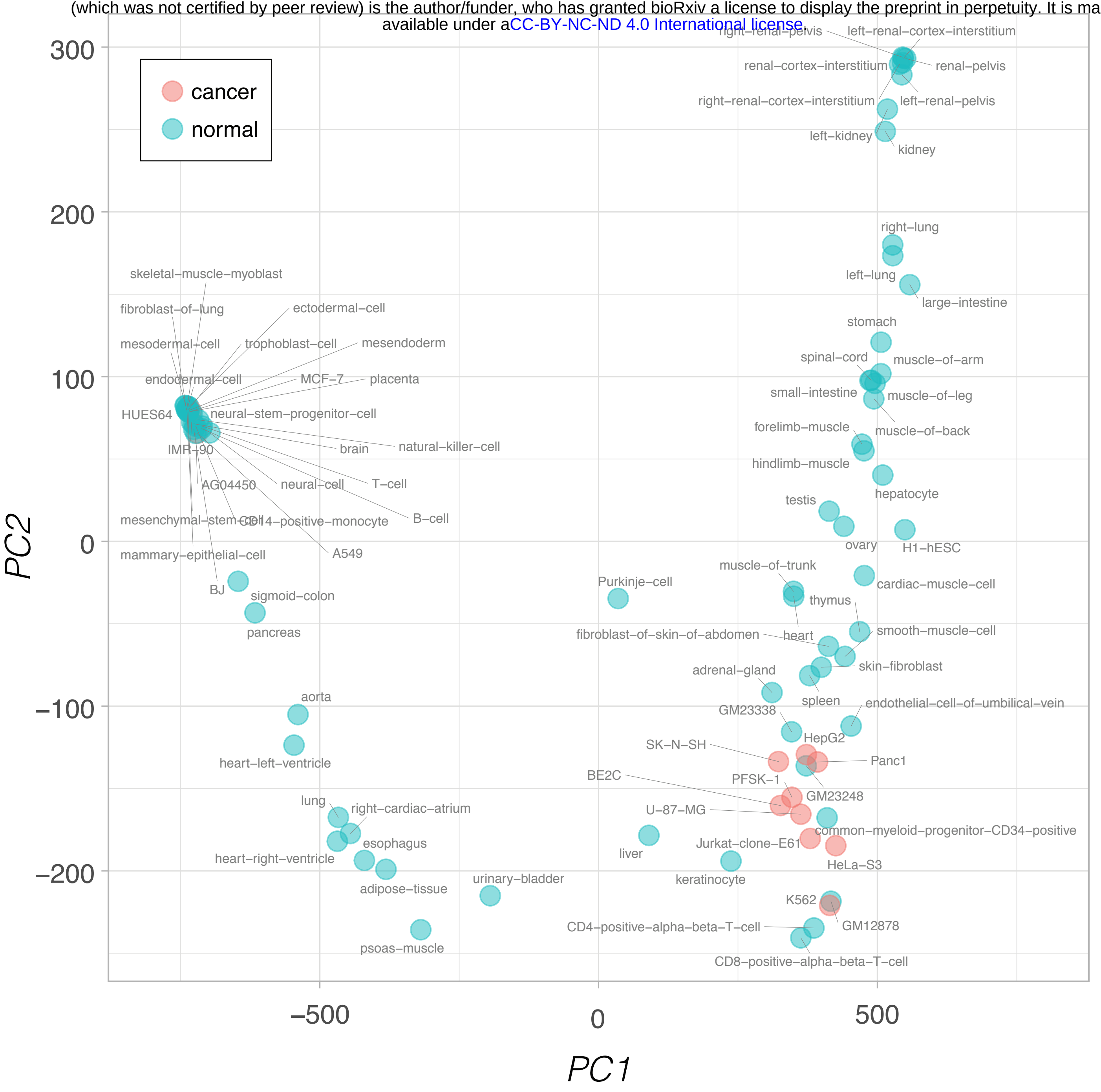

A
K562 H3K36me3
PCC = 0.092, P-val = 2.9e-64
Enrichment
Expression
10.00
1.00
0.10
0.01
1e-01
1e+02
1e+05
B
K562 H3K27me3
PCC = -0.062, P-val = 1.8e-27
Enrichment
Expression
C
Enrichment
High Expr.
Low Expr.
****
****
*
****
****
*
K562 H3K36me3
HepG2 H3K36me3
Liver H3K36me3
K562 H3K27me3
HepG2 H3K27me3
Liver H3K27me3
0
2
4
6
D
0.2
0.1
0.0
-0.1
-0.2
biosample
thoracic-aorta
gastrocnemius-medialis
heart-left-ventricle
right-atrium-auricular-region
tibial-nerve
sigmoid-colon
esophagus-muscularis-mucosa
gastroesophageal-sphincter
ascending-aorta
uterus
breast-epithelium
thyroid-gland
spleen
upper-lobe-of-left-lung
transverse-colon
Peyer-s-patch
stomach
esophagus-squamous-epithelium
suprapubic-skin
K562
B-cell
CD14-positive-monocyte
GM12878
Karpas-422
OCI-LY7
keratinocyte
mammary-epithelial-cell
HeLa-S3
HepG2
A549
Panc1
PC-3
HCT116
MCF-7
cardiac-muscle-cell
bipolar-neuron
neural-cell
H1
neural-progenitor-cell
hepatocyte
fibroblast-of-dermis
osteoblast
skeletal-muscle-myoblast
fibroblast-of-lung
GM23248
astrocyte
smooth-muscle-cell
endothelial-cell-of-umbilical-vein
SK-N-SH
C1
C2
C3
C4
C5

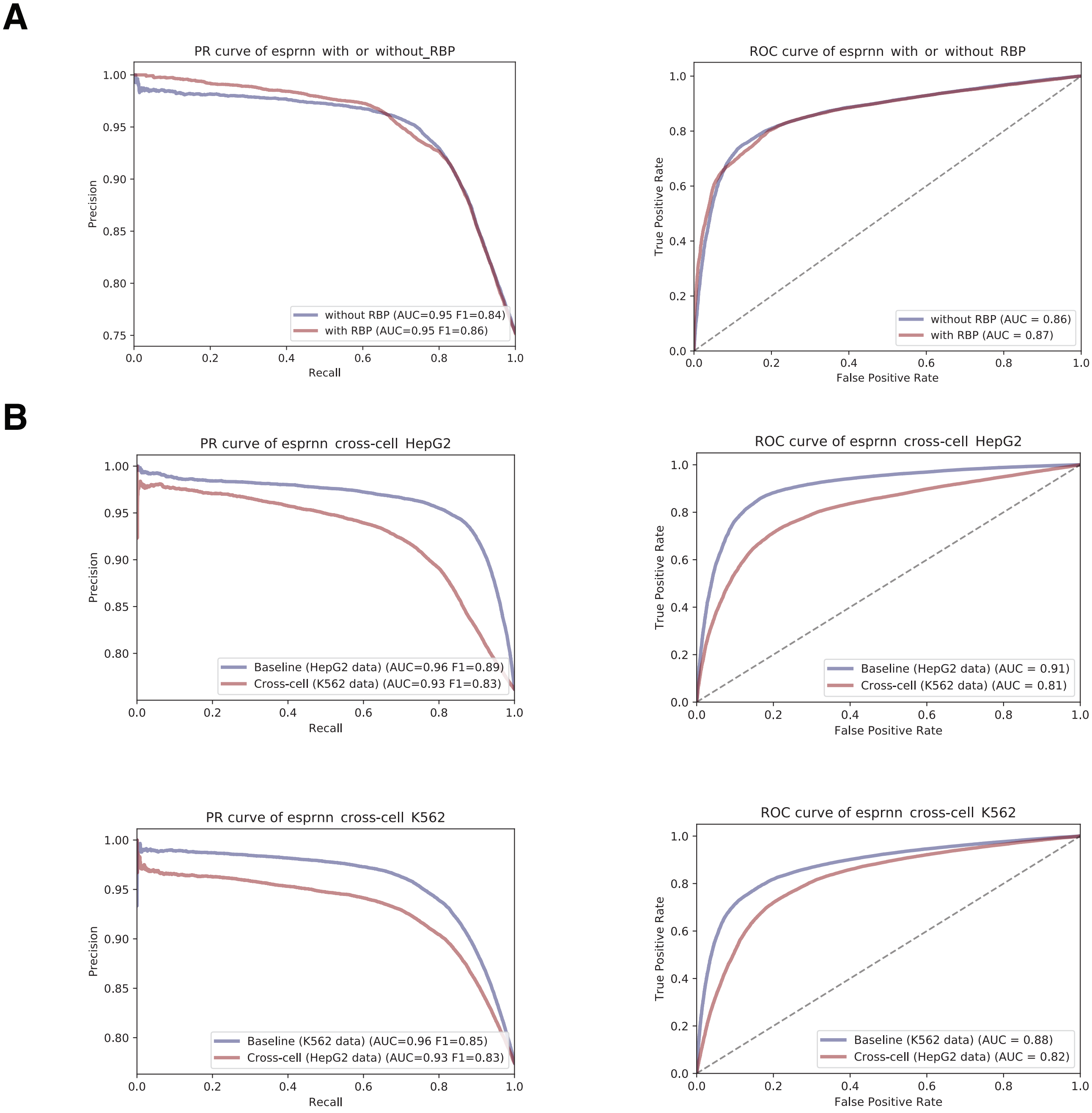

A
Exon A
Exon B
Exon C
5'
AG
Exon
GT
3'
3' Acceptor
5' Donor
n bp
n bp
n bp
n bp
2n x (DNA + Epigenomic Features)
2n x (DNA + Epigenomic Features)
5'
3'
ACTTGCA...G
CORE SET
H3K4me1
H3K4me3
H3K36me3
H3K27me3
H3K9me3
H3K9ac
H3K27ac
H3K79me2
⋮
H4K20me1
LSTM / GRU (3' Acc)
2n x 2n
LSTM / GRU (5' Don)
2n x 2n
LSTM / GRU (Merged)
4n
Dropout
Softmax
Spliced
Unspliced
C
F1 Score
ROC AUC
H3K36me3
DNA
H3K79me2
H3K9me3
DNase
Nucleosome
H2AZ
H4K20me1
H3K4me3
H3K4me1
H3K4me2
H3K27me3
H3K27ac
H3K9ac
Methylation
0.00
0.02
0.04
Loss of Prediction Accuracy in GM12878
D
A549
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LSTM (AUC=0.94)
kNN (AUC=0.83)
Decision Tree (AUC=0.88)
Random Forest (AUC=0.84)
SVM (AUC=0.91)
B
K562 Precision-Recall curve
HMEC Precision-Recall curve
HepG2 Precision-Recall curve
GM12878 Precision-Recall curve
Adult Liver Precision-Recall curve
Lung Precision_Recall curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Core Model: Avg=0.92
Full Model: Avg=0.94
Core Model: Avg=0.90
Full Model: Avg=0.90
Core Model: Avg=0.92
Full Model: Avg=0.93
Core Model: Avg=0.91
Full Model: Avg=0.94
Core Model: Avg=0.85
Full Model: Avg=0.87
Core Model: Avg=0.91
Full Model: Avg=0.90

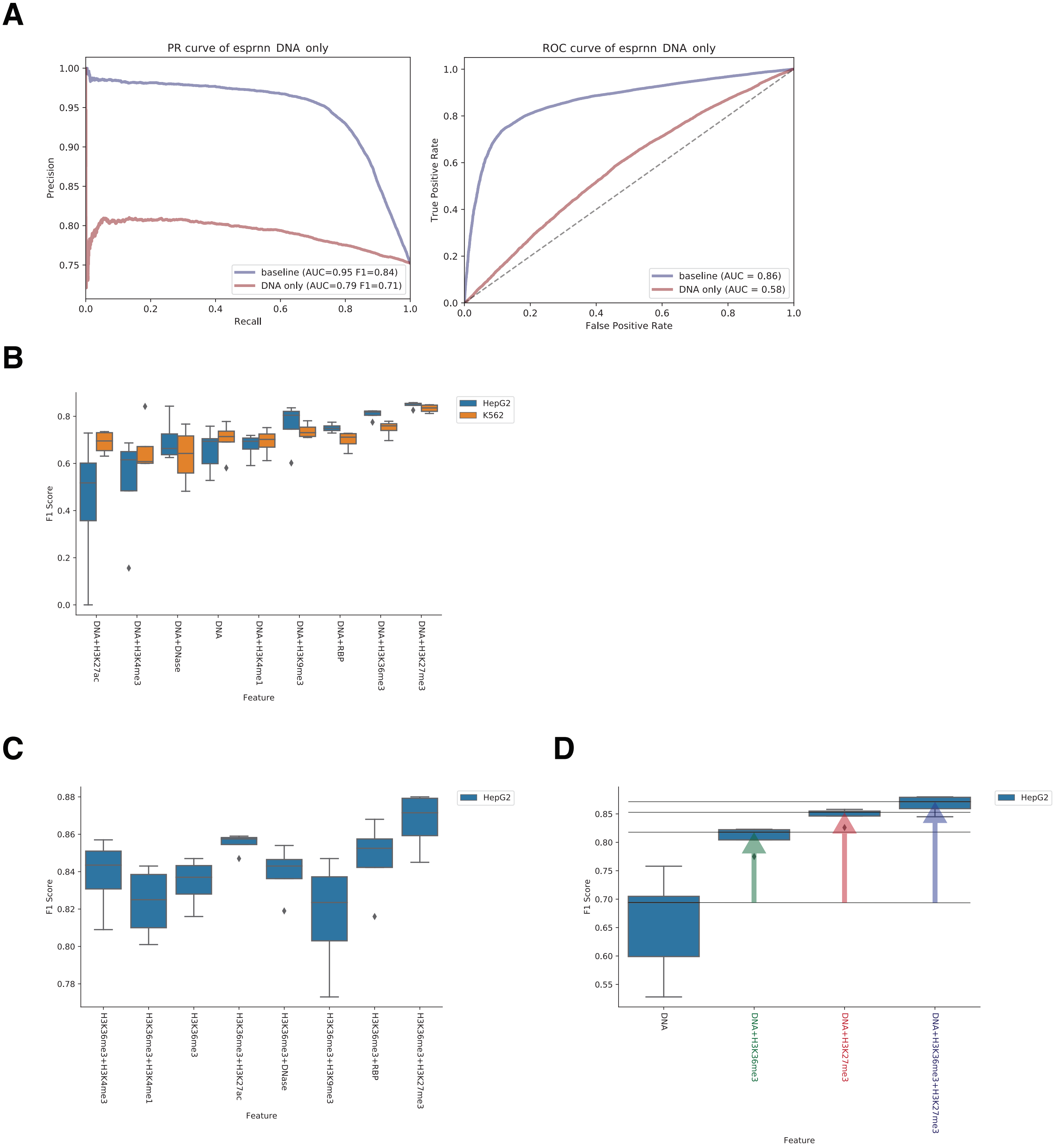

Signal = DHS | ss = 3acc
Signal = DHS | ss = 5don
Signal = H3K27ac | ss = 3acc
Signal = H3K27ac | ss = 5don
Signal = H3K27me3 | ss = 3acc
Signal = H3K27me3 | ss = 5don
Signal = H3K36me3 | ss = 3acc
Signal = H3K36me3 | ss = 5don
Signal = H3K4me1 | ss = 3acc
Signal = H3K4me1 | ss = 5don
Signal = H3K4me3 | ss = 3acc
Signal = H3K4me3 | ss = 5don
Signal = H3K9me3 | ss = 3acc
Signal = H3K9me3 | ss = 5don
Signal = RBP | ss = 3acc
Signal = RBP | ss = 5don
Fold Enrichment
0.07
0.06
0.05
0.04
0.03
1.50
1.25
1.00
0.75
0.50
0.25
1.75
4
3
2
1
1.0
0.8
0.6
0.4
0.9
0.7
0.55
0.50
0.45
0.40
0.35
0.2
0.0
−500
−400
−300
−200
−100
0
100
200
300
400
500
pos
PSI
High
Mid
Low

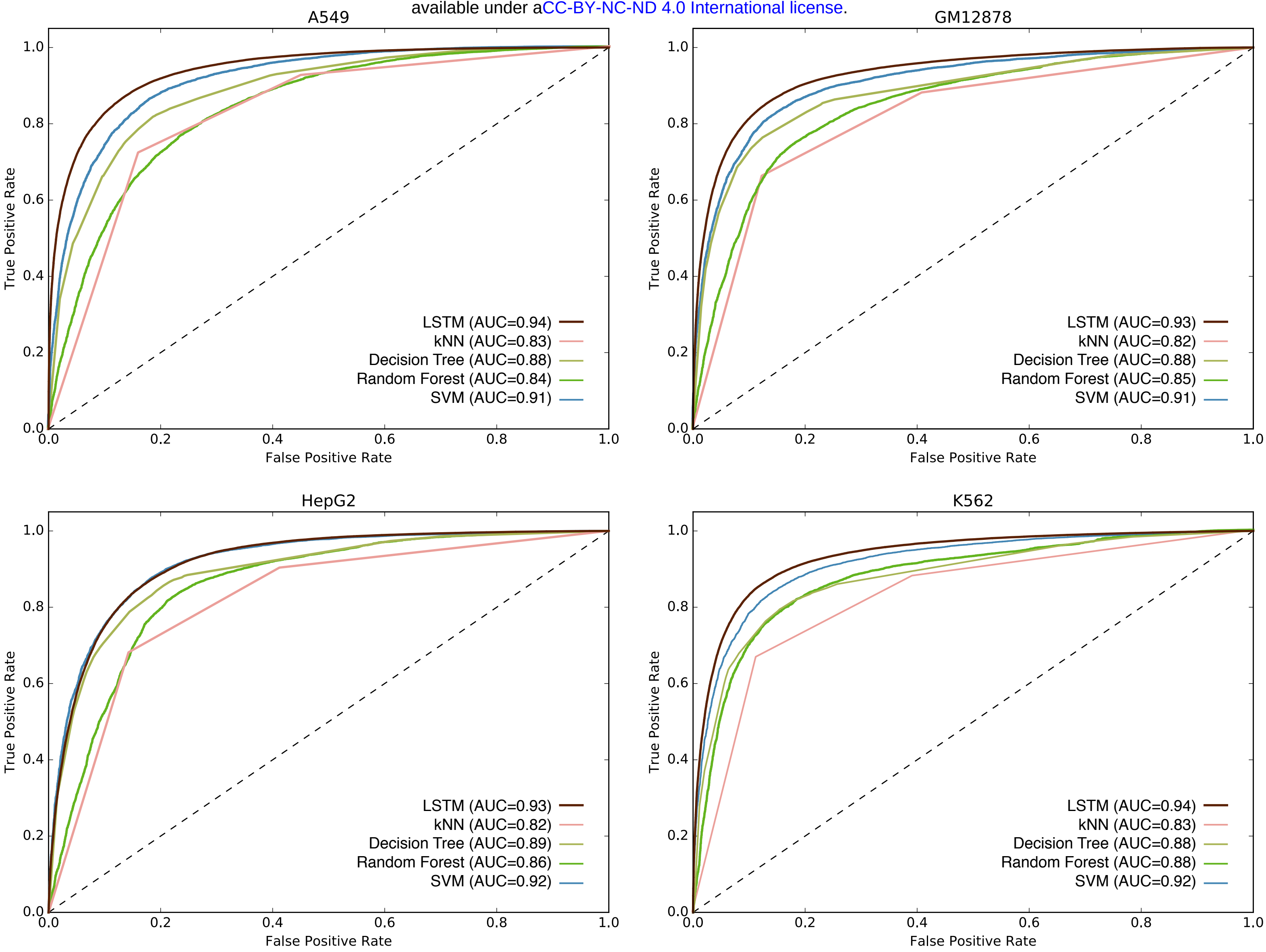

A
ns
**
ns
****
Fold Enrichment
High
Mid
Low
A
B
C
D
K562 DHS 3'acc
ns
****
***
K562 H3K4me3 3'acc
***
****
*
K562 H3K9me3 3'acc
B
HepG2 H3K27ac 3'acc
**
HepG2 H3K27me3 3'acc
*
HepG2 H3K4me1 3'acc
HepG2 H3K36me3 3'acc
HepG2 H3K4me3 3'acc
HepG2 H3K9me3 3'acc
HepG2 DHS 3'acc
C
FC of splicing to non−splicing binding around 3' SS
Fold Enrichment (SF/non-SF)
3' acceptor Splice Site
FC of splicing to non−splicing binding around 5' SS
5' donor Splice Site
−1000
−500
0
500
1000

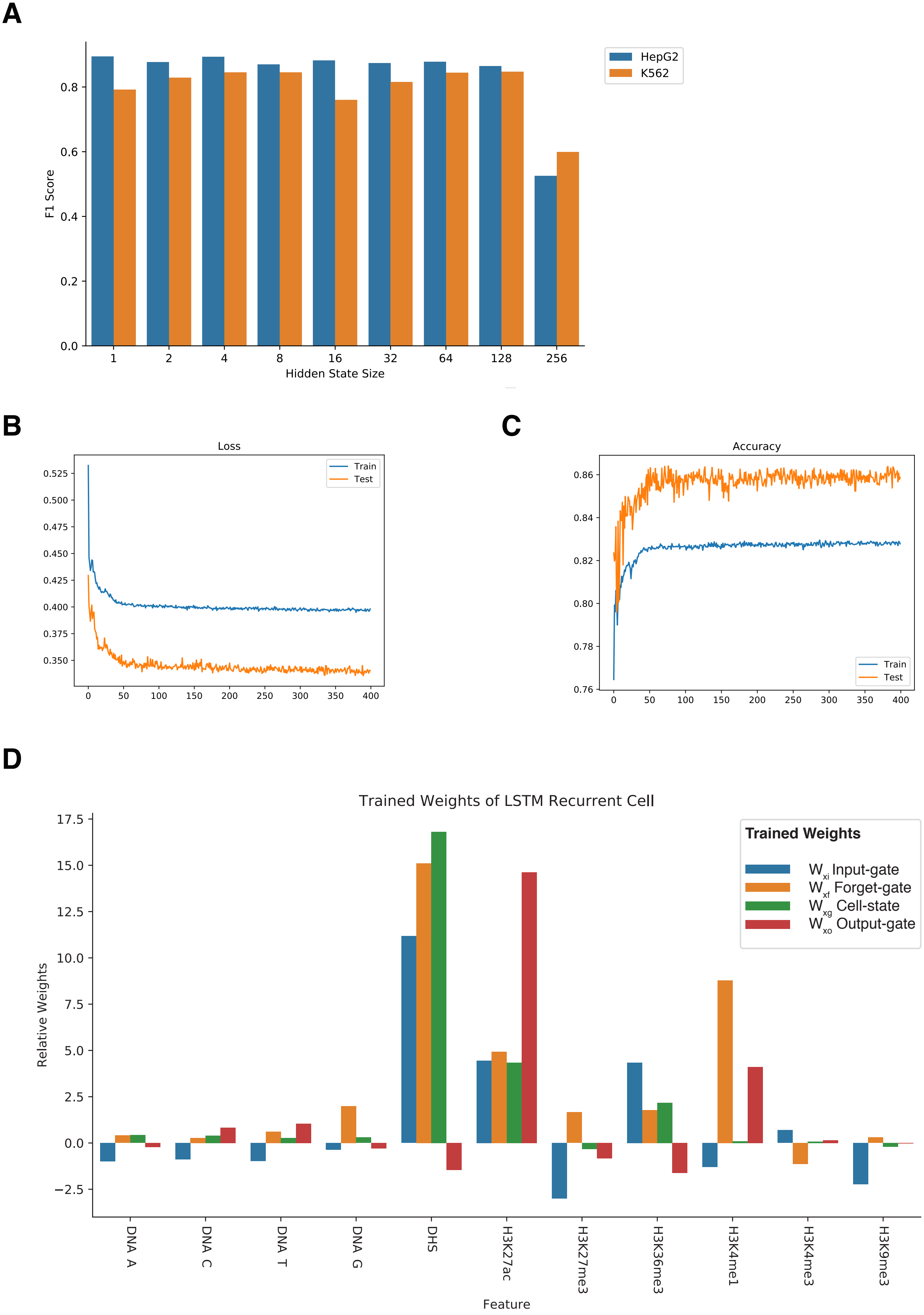

A
HepG2 H3K36me3
PCC = 0.021, P−val = 0.00022
B
HepG2 H3K27me3
PCC = −0.02, P−val = 0.013
C
liver H3K36me3
PCC = 0.0046, P−val = 0.43
D
liver H3K27me3
PCC = −0.0068, P−val = 0.27
Enrichment
10.00
1.00
0.10
0.01
Expression
1e−02
1e+01
1e+04
1e+00
1e+02

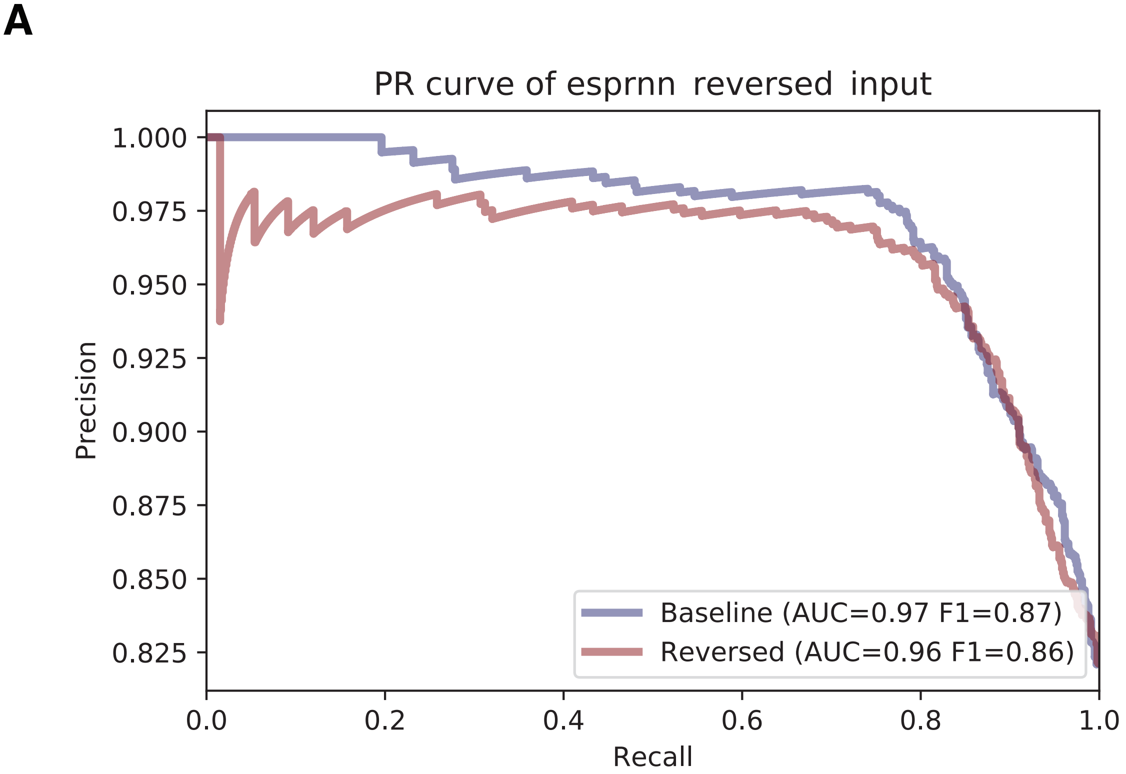

cancer
normal
PC2
PC1
300
200
100
0
−100
−200
−500
0
500
right−renal−pelvis
left−renal−cortex−interstitium
renal−cortex−interstitium
renal−pelvis
right−renal−cortex−interstitium
left−renal−pelvis
left−kidney
kidney
right−lung
left−lung
large−intestine
stomach
spinal−cord
muscle−of−arm
small−intestine
muscle−of−leg
forelimb−muscle
muscle−of−back
hindlimb−muscle
hepatocyte
testis
ovary
H1−hESC
muscle−of−trunk
cardiac−muscle−cell
Purkinje−cell
thymus
fibroblast−of−skin−of−abdomen
heart
smooth−muscle−cell
skin−fibroblast
adrenal−gland
spleen
GM23338
endothelial−cell−of−umbilical−vein
HepG2
SK−N−SH
Panc1
BE2C
PFSK−1
GM23248
U−87−MG
common−myeloid−progenitor−CD34−positive
liver
Jurkat−clone−E61
HeLa−S3
keratinocyte
K562
GM12878
CD4−positive−alpha−beta−T−cell
CD8−positive−alpha−beta−T−cell
skeletal−muscle−myoblast
fibroblast−of−lung
ectodermal−cell
mesodermal−cell
trophoblast−cell
mesendoderm
endodermal−cell
MCF−7
placenta
HUES64
neural−stem−progenitor−cell
IMR−90
brain
natural−killer−cell
AG04450
neural−cell
T−cell
mesenchymal−stem−cell
CD14−positive−monocyte
B−cell
mammary−epithelial−cell
A549
BJ
sigmoid−colon
pancreas
aorta
heart−left−ventricle
lung
right−cardiac−atrium
esophagus
heart−right−ventricle
adipose−tissue
urinary−bladder
psoas−muscle

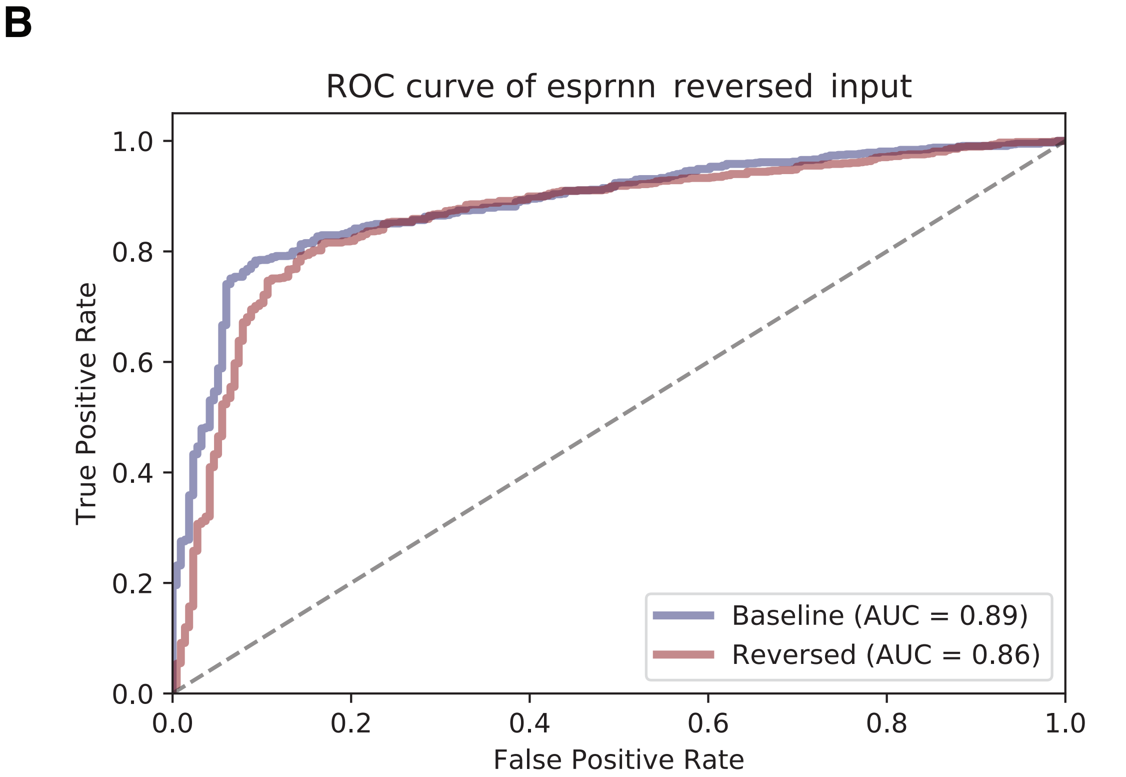

A
PR curve of esprnn with or without_RBP
Precision
Recall
without RBP (AUC=0.95 F1=0.84)
with RBP (AUC=0.95 F1=0.86)
ROC curve of esprnn with or without RBP
True Positive Rate
False Positive Rate
without RBP (AUC = 0.86)
with RBP (AUC = 0.87)
B
PR curve of esprnn cross-cell HepG2
Baseline (HepG2 data) (AUC=0.96 F1=0.89)
Cross-cell (K562 data) (AUC=0.93 F1=0.83)
ROC curve of esprnn cross-cell HepG2
Baseline (HepG2 data) (AUC = 0.91)
Cross-cell (K562 data) (AUC = 0.81)
PR curve of esprnn cross-cell K562
Baseline (K562 data) (AUC=0.96 F1=0.85)
Cross-cell (HepG2 data) (AUC=0.93 F1=0.83)
ROC curve of esprnn cross-cell K562
Baseline (K562 data) (AUC = 0.88)
Cross-cell (HepG2 data) (AUC = 0.82)
0.0
0.2
0.4
0.6
0.8
1.0
0.75
0.80
0.85
0.90
0.95
1.00


A
PR curve of esprnn DNA only
Precision
1.00
0.95
0.90
0.85
0.80
0.75
Recall
0.0
0.2
0.4
0.6
0.8
1.0
baseline (AUC=0.95 F1=0.84)
DNA only (AUC=0.79 F1=0.71)
ROC curve of esprnn DNA only
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
baseline (AUC = 0.86)
DNA only (AUC = 0.58)
B
F1 Score
0.0
0.2
0.4
0.6
0.8
HepG2
K562
DNA+H3K27ac
DNA+H3K4me3
DNA+DNase
DNA
DNA+H3K4me1
DNA+H3K9me3
DNA+RBP
DNA+H3K36me3
DNA+H3K27me3
Feature
C
F1 Score
0.78
0.80
0.82
0.84
0.86
0.88
HepG2
H3K36me3+H3K4me3
H3K36me3+H3K4me1
H3K36me3
H3K36me3+H3K27ac
H3K36me3+DNase
H3K36me3+H3K9me3
H3K36me3+RBP
H3K36me3+H3K27me3
Feature
D
F1 Score
0.55
0.60
0.65
0.70
0.75
0.80
0.85
HepG2
DNA
DNA+H3K36me3
DNA+H3K27me3
DNA+H3K36me3+H3K27me3
Feature


A549
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LSTM (AUC=0.94)
kNN (AUC=0.83)
Decision Tree (AUC=0.88)
Random Forest (AUC=0.84)
SVM (AUC=0.91)
GM12878
True Positive Rate
False Positive Rate
LSTM (AUC=0.93)
kNN (AUC=0.82)
Decision Tree (AUC=0.88)
Random Forest (AUC=0.85)
SVM (AUC=0.91)
HepG2
True Positive Rate
False Positive Rate
LSTM (AUC=0.93)
kNN (AUC=0.82)
Decision Tree (AUC=0.89)
Random Forest (AUC=0.86)
SVM (AUC=0.92)
K562
True Positive Rate
False Positive Rate
LSTM (AUC=0.94)
kNN (AUC=0.83)
Decision Tree (AUC=0.88)
Random Forest (AUC=0.88)
SVM (AUC=0.92)


A
F1 Score
0.0
0.2
0.4
0.6
0.8
1
2
4
8
16
32
64
128
256
Hidden State Size
HepG2
K562
B
Loss
0.350
0.375
0.400
0.425
0.450
0.475
0.500
0.525
0
50
100
150
200
250
300
350
400
Train
Test
C
Accuracy
0.76
0.78
0.80
0.82
0.84
0.86
0
50
100
150
200
250
300
350
400
Train
Test
D
Trained Weights of LSTM Recurrent Cell
Relative Weights
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
DNA A
DNA C
DNA T
DNA G
DHS
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K9me3
Feature
Trained Weights
W_{xi} Input-gate
W_{xf} Forget-gate
W_{xg} Cell-state
W_{xo} Output-gate


A
PR curve of esprnn reversed input
Precision
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Baseline (AUC=0.97 F1=0.87)
Reversed (AUC=0.96 F1=0.86)


B
ROC curve of esprnn reversed input
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Baseline (AUC = 0.89)
Reversed (AUC = 0.86)