The Epigenomic Landscape of Splicing in Cancer

Donghoon Lee[1], Jing Zhang[1,2], Mark B Gerstein[1,2,3]

[1] Program in Computational Biology and Bioinformatics [2] Department of Molecular Biophysics and Biochemistry
[3] Department of Computer Science, Yale University, New Haven, CT 06520, USA

Nearly all protein-coding genes in eukaryotes undergo alternative splicing, which provides important means to expand transcriptome diversity. It is implied that splicing occurs co-transcriptionally and splicing is both spatially and temporally linked to DNA elements and epigenetic modifications.

Chromatin structures have been correlated with transcription dynamics, and thus, epigenetic regulation may play a significant role in alternative splicing. Recent studies have revealed splicing regulation is characterized by increased levels of nucleosome density and positioning, DNA methylation, and distinct histone modification patterns.

Cancer cells show very distinct exon-level expression profile compared to normal cells and change in transcriptome can be attributed to epigenetic dysregulation (Figure 1). Yet, the extent, nature, and effects of epigenomic dysregulation in splicing remain unsolved.

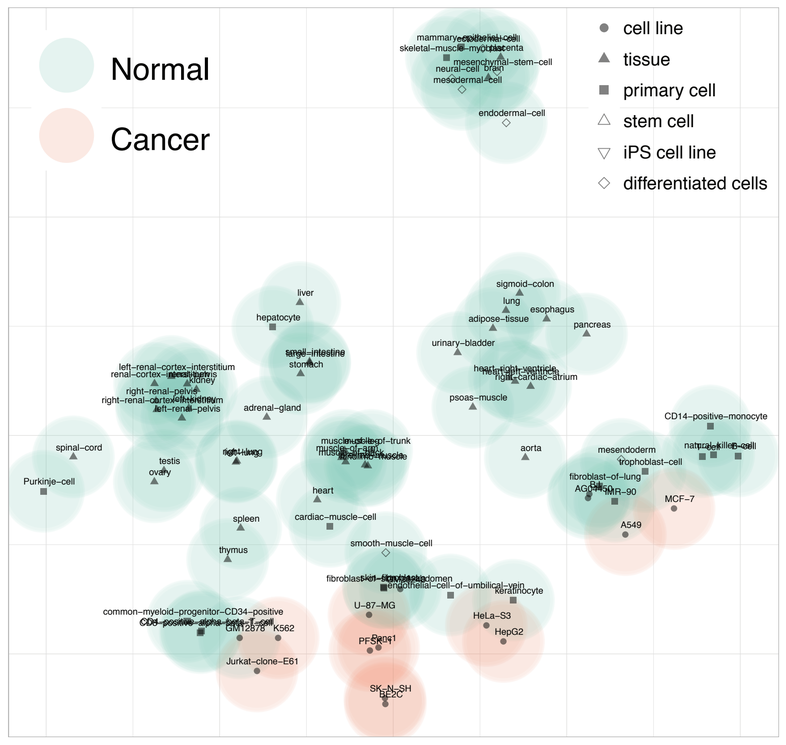

Figure 1. Clustering of exon-level expression profile using t-SNE

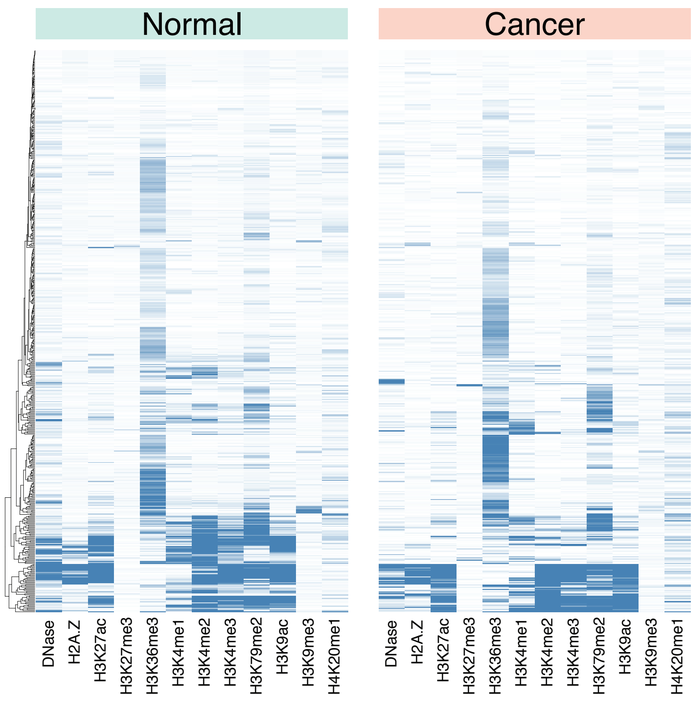

Figure 2. Chromatin signatures at 3' acceptor splice sites for cassette exons expressed in both tumor and normal samples

In this study, we aimed to learn complex regulatory vocabulary of splicing and identify distinct epigenetic signatures that characterize splicing in cancer by systematically profiling the epigenomic landscape of splicing using transcriptomic and epigenomic data from the ENCODE and the Epigenome Roadmap projects.

Alternative splicing is believed to be highly context-dependent and we found very complex histone signatures at both 3' and 5' splice sites. In general, H3K36me3 mark was enriched at both splice sites, as previously reported, but this mark was not always required to splice an exon. For cassette exons spliced and expressed, tumor and normal samples showed different levels of histone enrichments (Figure 2).

To explore the epigenomic contribution to the regulation of alternative splicing, we built a machine learning classifier to predict alternate usages of cassette exons using DNA sequences and epigenomic signals adjacent to splice sites (Figure 3).

We applied a deep recurrent neural network (RNN) model to solve this problem, which has proven successful in various sequential information processing and predictions such as natural language processing and translation. The rationale for applying RNN to our model is that (1) RNN is optimized for processing sequential information such as genomic sequences (2) RNN has a time direction resembling how RNA is always transcribed by RNA polymerase in the 5' to 3' direction (3) temporal memory cells of RNN allow the model to learn about complex context-dependent relationships among features, and (4) RNN is very flexible with the types of input and output data and therefore can easily integrate heterogeneous types of sequential data.

Compared to conventional machine learning classifiers such as Random Forest and KNN, the RNN model consistently outperformed in terms of classification accuracy (Figure 4). On average, the prediction accuracy was 84.72% for the LSTM-based model (87.57% for the GRU-based model, Figure 5)). We trained a model from healthy primary tissues (adult liver and lung) and predicted splicing outcomes in matching cancerous cell lines (HepG2 and A549). We observed a moderate drop in the prediction accuracies, as expected. This implies that normal cells and cancerous cells may use different chromatin context to determine splicing.

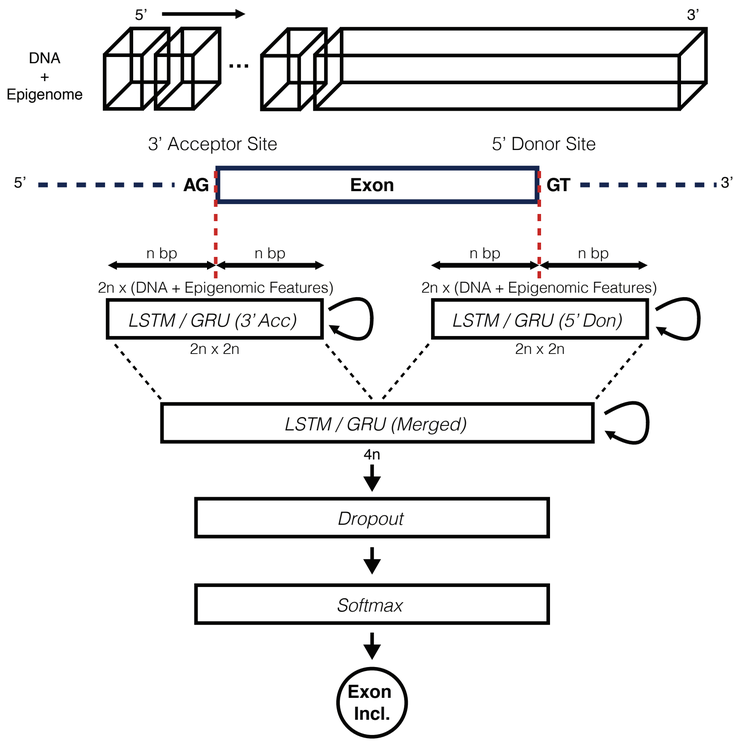

Figure 3. Schematic of splicing prediction model

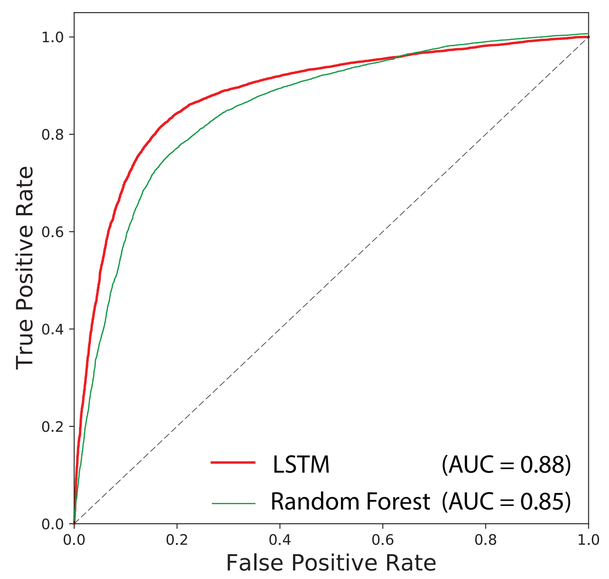

Figure 4. Performance comparison of machine learning algorithms

		Breast Myoepithelial Primary Cells	Breast vHMEC Mammary Epithelial Cells	Liver Primary Tissues	Lung Primary Tissues	A549 EtOH 0.02pct Lung Carcinoma Cell Line	GM12878 Lymphoblastoid Cells	HepG2 Hepatocellular Carcinoma Cell Line	HMEC Mammary Epithelial Primary Cells	K562 Leukemia Cells
CORE DNA + 5 Histone Marks	**LSTM** 2x200-400	84.65% (89.03%)	87.81% (93.00%)	80.18% (86.55%)	82.96% (85.26%)	87.65% (93.08%)	83.86% (90.77%)	86.17% (91.78%)	84.28% (89.23%)	84.93% (91.67%)
	GRU 2x200-400	84.97% (89.22%)	**93.22% (97.49%)**	80.06% (86.04%)	83.69% (88.73%)	88.60% (94.37%)	**87.14% (93.20%)**	**89.59% (95.50%)**	**92.83% (97.49%)**	**88.04% (94.81%)**
	Simple RNN 2x200-400	82.44% (85.69%)	82.70% (85.52%)	74.56% (84.11%)	79.49% (80.85%)	84.89% (88.67%)	68.83% (75.07%)	77.82% (66.95%)	82.26% (86.79%)	83.40% (89.49%)
FULL Core + All Available Epigenomic Datasets	**LSTM** 2x200-400	85.24% (89.93%)	88.54% (93.78%)	**81.39% (88.25%)**	81.57% (84.50%)	88.58% (94.16%)	86.87% (92.95%)	87.25% (92.68%)	84.85% (89.85%)	87.32% (93.51%)
	GRU 2x200-400	**86.61% (91.98%)**	86.17% (90.60%)	78.28% (85.76%)	**85.20% (88.77%)**	**94.12% (98.11%)**	86.10% (91.71%)	87.91% (93.20%)	88.62% (94.36%)	85.70% (92.25%)
	Simple RNN 2x200-400	82.59% (85.65%)	77.92% (85.15%)	77.10% (74.17%)	79.84% (79.95%)	83.81% (89.10%)	84.14% (89.64%)	85.52% (90.16%)	81.85% (87.08%)	83.62% (90.26%)

Figure 5. Splicing prediction model performance