ABSTRACT

Title of Dissertation:	Mixture Models for Nucleic Acid Sequence Feature Analysis
	Bixuan Wang, Doctor of Philosophy, 2023
Dissertation directed by:	Professor Stephen M. Mount, CBMG

Signals in nucleotide sequences play a crucial role in interactions among macromolecules and the regulation of biological functional processes such as transcription, the splicing of messenger RNA precursors and translation. Recognition of signals in nucleotide sequences is the first step in functional annotation, which is critical for the identification of deleterious mutations and the identification of targets for disease treatment. One of the essential steps in gene expression, RNA splicing removes introns from newly transcribed RNA, ligating exons to generate mature RNA. Splicing involves the formation and recycling of the spliceosome, a large macromolecular complex whose assembly requires complex coordination by splicing factors through the recognition of RNA-protein binding sites. One potential method to reveal unknown subtypes of samples and identify distinctively distributed features is by applying a mixture model called the admixture model or Latent Dirichlet Allocation (LDA), which allows samples to have partial memberships of different clusters that can be interpreted for functional motif identification. By applying mixture models to RNA sequences, I found splicing signals such as the polypyrimidine tract and the branch point in intron sequences. Mixture models also showed motifs associated

with reading frames from coding sequences, which further revealed potential coding regions from 5’ untranslated regions and long non-coding RNAs.

Dynamic single-molecule imaging of nascent RNAs coupled with multiple genome-wide assays reveals that splicing happens far more often than expected, and partial intron removal can be captured prior to completion of the entire transcript. I hypothesize that the spliceosome progressively removes large introns in small pieces through 'recursive splicing' instead of removing the whole intron at once. However, the sequence features that distinguish sites of recursive splicing from canonical splice sites remain to be discovered. Here, I applied mixture models to sequences from human introns to identify sequence features associated with recursive splicing. This method helped me to recognize and visualize splicing signals from annotated intron sequences and identify potential coding sequences from human 5' untranslated regions and long non-coding RNA. After applying mixture models to the sequences surrounding recursive and canonical splicing sites, I found that transcripts where large introns can be recursively spliced can be distinguished from those without recursive splicing by the presence of CG-rich motifs flanking 5' splice sites upstream of first introns, and the absence of DNA methylation at these sites.

In addition to applications of mixture models, I also explored RNA Bind-N-Seq data reflecting the binding activities of the splicing factor U2AF and found that the recursive 3' splice sites have higher U2AF binding affinities than the downstream canonical 3'SS.

The observations suggest that, first, mixture models have the potential to identify functional motifs, including subtle signals in sequences such as the branch sites that only occur in a subgroup of introns. Second, the usage of recursive splicing sites is associated with sequence

features in the first exons of the transcripts, suggesting a testable model for the regulation of recursive splicing in human introns.

MIXTURE MODELS FOR NUCLEIC ACID SEQUENCE FEATURE ANALYSIS

by

Bixuan Wang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Dr. Stephen Mount, Chair
Dr. Daniel Larson, Co-chair
Dr. Najib El-Sayed
Dr. Philip Johnson
Dr. Antony Jose
Dr. Robert Patro, Dean's Representative

Acknowledgments

I would like to express my sincere gratitude to my advisors, Dr. Steve Mount and Dr. Dan Larson, for their invaluable guidance and support over the past few years. Dr. Mount consistently encourages me to explore new ideas and provides constructive feedback that enhances my work. He fosters an environment that nourishes creativity, celebrates every small achievement, and offers mental support during hard times. I am truly fortunate to have had Dr. Mount as a mentor. Dr. Larson is a role model for me as a scientist. His passion for science and research has always been a guiding light throughout my PhD journey. I believe it was Dr. Larson's dedication to our project, even facing a prolonged period without major discoveries, that finally made our recursive splicing project work. I deeply appreciate his mentorship, which helped me shape a scientific mindset when facing difficulties.

I am also thankful to my committee members. I want to thank Dr. Antony Jose for consistently providing new ideas and solutions for my project. I benefited a lot from Dr. Rob Patro's computational biology algorithm course, and I want to thank him for giving me insightful guidance about statistical inference in my dissertation. Dr. Philip Johnson always reminds me to think about every aspect of the statistical model I am working on and encourages me to learn from other alternative models. Dr. Najib El-Sayed provided insights about how to interpret my results when I had no idea about them. His inquiries about my well-being and happiness made me feel the caring concern from a professor.

I am grateful to my lab mates from Mount Lab and Larson Lab. I want to thank all the students from Mount Lab for being so supportive and providing invaluable feedback about my research projects. I believe everyone in Larson Lab is a passionate scientist with great potential, and I enjoy talking to each of them. I want to thank Dr. Ben Donovan, Dr. Gloria Garcia, and Dr. Jee Min Kim. I learned a lot by collaborating with them, and every time I meet with them, I can feel their passion for research, which has been a consistent resource of inspiration for me.

Studying in the BISI PhD program has been a memorable journey, and I am so fortunate to have such great friends to support me during difficult times. I want to thank Dr. Zhongchi Liu for giving me a spot in her lab, where I could work with my friends Dr. Muzi Li and Dr. Fuxi Wang together. Dr. Muzi Li joined BISI two years before I did after we both graduated from the bioinformatics Master's program at Georgetown University. She is an amazing scientist with a fantastic talent for catching the core of a problem quickly and accurately, both in research and in real life. She always helps me to find the best solution and saves me from mental breakdowns. I met Dr. Fuxi Wang on the third day after I joined BISI. Fuxi has been a very supportive and caring friend, and we laugh and cry together. When we played Overcooked together, we laughed at almost every moment, especially when we made ridiculous mistakes. We even researched how to pass the final challenge as scientists, and we did it! I also want to thank my friends Dongze He, Rui Yin, and Yunliang Zhang. We joined BISI the same year. Dongze and I took hard-core computer science courses and solved challenging problems together. Rui Yin and Yunliang Zhang gave me valuable advice about my research projects. I am grateful to my roommates, Dr.

Xin Qiao, Elizabeth Chen, and Audrey Kuei, for their encouragement and support. I want to express my gratitude to my friend Dr. Shujin Zhong for listening to me and encouraging me when I need it the most.

I extend my sincere appreciation to the BISI program and NCI-UMD partnership for their crucial support throughout the duration of my Ph.D.

Last but not least, I would like to thank my parents. My parents were very supportive when I made the decision to quit my job and join a PhD program. They always encourage me to learn and explore. I could not have reached this point without their support.

Table of Contents

Chapter 1: Introduction

Biological Sequence Feature Recognition

Biological sequences convey messages as intricate encodings that capture the order and complexity of life. Information decoded from the sequences guides the regulation of various biological processes, influencing every step of gene expression. For example, the tri-nucleotide codon features define genes by signaling open reading frames, the expression of which is regulated by multiple signal recognition processes, including the interaction between transcription factors and the binding sites and the assembly of spliceosome at splicing sites that are composed by various sequence features (Hayes, 1998; Latchman, 1997; Will & Luhrmann, 2011). Interruptions in the signal recognition processes alter the destinies of the products, causing dysregulation of gene expression and diseases (T. I. Lee & Young, 2013; Seiler et al., 2018). Recognizing sequence motifs is essential for studying the sequence functions and macromolecule interactions, which requires comprehensive studies experimentally and computationally and provides insights into understanding biological processes, disease development, and drug target identification (Rittiner et al., 2022).

Tools for sequence feature description and identification

A sequence motif is a functional subsequence from the biological sequences that only exists or is over-expressed at specific regions. Functional motifs can be invariant,

such as the human 5'SS GU (Reyes et al., 1996). Sometimes, motifs have disconnected parts disconnected by unrelated sequence building blocks of various lengths. For example, the signals of RNA splicing require multiple features, including the polypyrimidine tract of undetermined length, which follows the branch site after a few nucleotides (Will & Luhrmann, 2011).

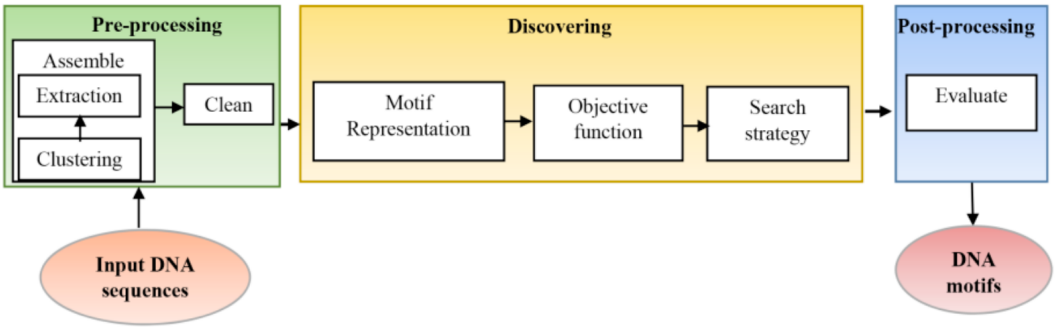


Figure 1.1 A block diagram of motif discovery pipeline (Hashim et al., 2019).

Motif discovery has three fundamental steps: input sequence preprocessing, motif discovery, and motif evaluation (Hashim et al., 2019). The candidate sequences for detailed investigation require carefully designed experiments targeting the sequence-specific molecular recognitions. For example, RNA bind-N-seq characterizes the RNA binding protein's preferences by scoring the enrichment of k-mers from the binding sequences in vitro, while RNA isolation by CLIP followed by high-throughput sequencing recognizes in vivo binding sites (Lambert et al., 2014). Moreover, recent studies suggest utilizing ATAC-Seq with single-cell sequencing technologies can also aid motif searching (Heumos et al., 2023). A successful motif discovery process requires two essential gears: motif representation and motif searching through subsequence differential expression.

Positional weight matrices

Compared with consensus sequences, the positional weight matrix (PWM) provides more details by pre-defining the length of a motif from aligned sequences and scoring every possible symbol at each position (Stormo, 2000; Xia, 2012).

Site	A	C	G	U	χ^2	P	A	C	G	U
1	83	30	49	84	10.10	0.0177	0.0525	−0.6332	−0.0260	0.3143
2	103	44	46	53	10.04	0.0182	0.3613	−0.0878	−0.1162	−0.3434
3	121	36	38	51	30.01	0.0000	0.5920	−0.3739	−0.3886	−0.3981
4	122	38	33	53	32.16	0.0000	0.6038	−0.2969	−0.5893	−0.3434
5	81	40	81	44	28.33	0.0000	0.0177	−0.2238	0.6933	−0.6081
6	0	1	245	0	948.34	0.0000	−6.6464	−5.0056	**2.2841**	−6.6469
7	0	9	0	237	582.23	0.0000	−6.6464	−2.3190	−6.6480	**1.8032**
8	239	1	2	4	462.46	0.0000	**1.5693**	−5.0056	−4.3320	−3.8633
9	16	24	1	205	387.81	0.0000	−2.2655	−0.9496	−5.0680	**1.5946**
10	2	0	243	1	928.96	0.0000	−4.8476	−6.6483	**2.2723**	−5.3416
11	9	7	2	228	521.06	0.0000	−3.0427	−2.6612	−4.3320	**1.7475**
12	87	15	34	110	53.66	0.0000	0.1198	−1.6111	−0.5468	0.7006
13	84	49	30	83	11.71	0.0085	0.0696	0.0659	−0.7246	0.2971
14	111	39	33	63	19.09	0.0003	0.4684	−0.2599	−0.5893	−0.0969
15	106	38	31	71	17.24	0.0006	0.4024	−0.2969	−0.6781	0.0738
16	92	30	40	84	13.69	0.0034	0.1997	−0.6332	−0.3155	0.3143
17	80	38	36	92	14.32	0.0025	−0.0001	−0.2969	−0.4655	0.4445

Figure 1.2 PFM (left columns ACGU) and PWM (right columns ACGU) of 246 donor splice sites. Each sequence contains 5bp of exon and 12bp of intron. (Xia, 2012)

To build a PWM, sequences of interest, such as experimentally identified transcription factor binding sites from DNA, are aligned at the first site of the binding interactions. The aligned sequences are then used to build the position frequency matrix (PFM), showing the fraction of every nucleotide at each position as the fundamental of PWM calculation. The most popular graphical representation of PWM is the sequence logo that schemes the consensus sequences with information content at every position of interest.

As a motif representation tool, PWM serves as the initial step to assemble an initial motif for further improvement in multiple motif discovery algorithms, including the original and extended versions of MEME and some Markov models (Bailey et al., 2006; Hashim et al., 2019). Also, the motif description by PWM provides a target for

motif searching from a sequence pool, aiding the functional sites searching in biological sequences (Ambrosini et al., 2018).

Hidden Markov model

The hidden Markov model is a stochastic model that measures both visible observations and invisible states, assuming the system follows a Markov process (Eddy, 2004). Compared with Bayesian models, HMM is sequential, and the length and order of observations are essential to determining the hidden state and the observation of the next event.

One example of HMM is the generation process of an RNA sequence, which allows the nucleotides but not the states of the nucleotides, such as intron, exon, promoter, or UTRs, to be observed. Though the states of the observation are selected randomly, they are not independent. The state of the n-th nucleotide depends on the state of the (n-1)-th nucleotide, which determines the choice of the corresponding nucleotide (Yoon, 2009).

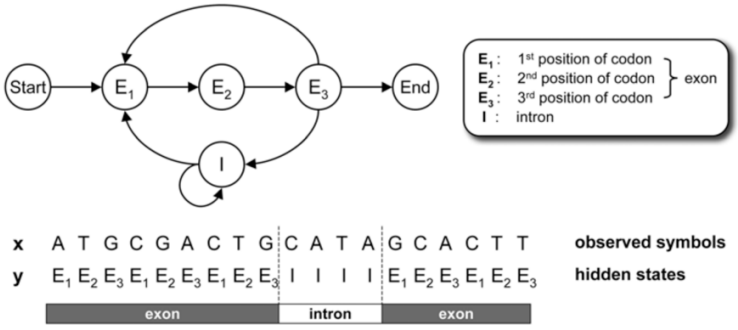


Figure 1.3 A simple HMM model for RNA sequences (Yoon, 2009).

When applying HMM to biological sequences, the hidden states can be used to imply the functional regions, such as the RNA-protein interaction sites. Thus, the

differential distribution of symbols between states can be further analyzed to discover the functional motifs (Yun et al., 2014).

MEME suite

MEME suite is an online collection of sequence motif discovery and analysis tools that cover a wide range of motif discovery solutions, such as continuous and gapped motifs, and various motif analysis methods, such as motif scanning and motif comparisons (Bailey et al., 2006).

The original MEME tool identifies the optimal PWM to represent motifs using expectation maximization (EM) algorithms. After determining an initial motif as PWM by estimating randomly selected subsamples, EM iterates all possible sites from the input sequences to evaluate and improve the PWM until it reaches convergence or iteration limit. Secondary motifs are also identifiable because MEME ignored discovered motifs in the next round of motif-searching steps.

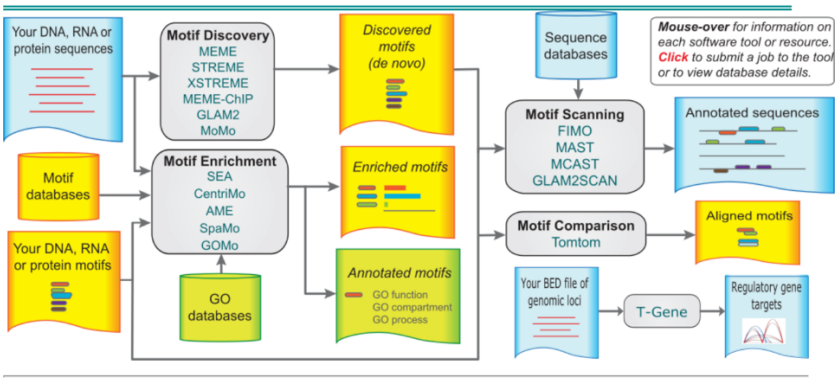


Figure 1.4 Diagram of motif analysis using MEME(https://meme-suite.org/meme/) tools.

MEME targets the discovery of motifs that are ungapped with fixed lengths. Multiple improvements have been applied to accelerate the algorithms and add features that are suitable for discovering more complicated motifs. STEME allows the search for ungapped motifs of a fixed length that are enriched in the input sequences with or

without another group of control/background sequences (Bailey, 2021). Another improvement, XSTREME, identifies a motif of any length, as well as multiple motifs of various sizes (Grant & Bailey, 2021).

Latent Dirichlet allocation

In natural language processing, latent Dirichlet allocation (LDA) is a topic model that describes the observations, which are the words from text documents, using unobserved features, which are the topics of the text documents, by statistically modeling the process of text document generation (Blei et al., 2003). LDA is also called mixture model since it takes the word counts from each document and transforms each document into a mixture of topics, which allows the comparison between multiple text documents. Though LDA is widely-used in natural language processing, the first admixture model was invented in 2000 as a method called STRUCTURE, which is one of the most famous frameworks in population genetics (Pritchard et al., 2000).

The first LDA model: STRUCTURE

Genetic compositions of populations reflect changes in genotypes and phenotypes, which are affected by evolutionary events. Population structures in geographic subgroups reveal traces of ancestral genetic features. In 2000, the admixture model was published and provided a novel approach to understanding population genetics by allowing individuals to have partial memberships of different ancestral populations. The method centered on a model with K populations illustrated by different allele frequencies at a pre-defined range of loci. Each individual should be

assigned to the population based on their genotype structure. Bayesian clustering is recruited to solve the best system of the populations and the population assignment for the individuals. The admixture model assumes that the features used in the population structure building are unlinked and can be applied to various genotype data, including SNPs and microsatellites (Pritchard et al., 2000).

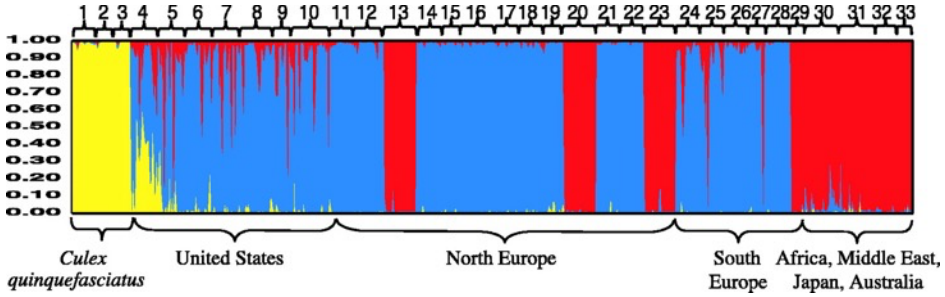


Figure 1.5 Bayesian clustering result of Culex pipens individuals by STRUCTURE. Individuals are single vertical lines with colors representing the probabilities of belonging to 3 types of genetic clusters (Fonseca et al., 2004).

LDA algorithm

The basic idea of LDA is that the documents are defined as random mixtures of latent topics. As in the population genetics analysis, the allele frequencies from individuals are drawn from random mixtures of populations. The model assumes a generative process of every document in a corpus. According to the original LDA paper, topic-word distributions can be modeled with a Dirichlet prior parameter to generate topics that only a few words can be observed frequently. Thus, LDA can be represented as the following figure.

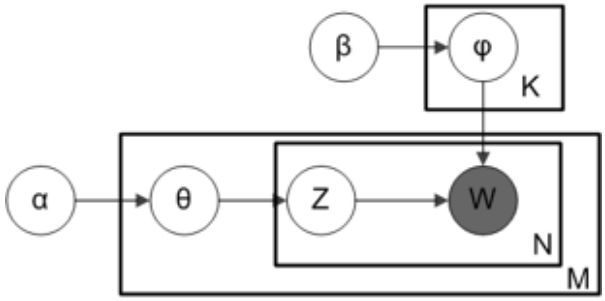


Figure 1.6 Plate notation of LDA (https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation).

As a probabilistic model, LDA can be described by plate notation, where the boxes represent the units that are repeated multiple times. In the LDA plate notation, the boxes M, N, and K denote the repeated calculations for the documents, words, and topics. The definitions of the variables are:

- M: the number of topics
- N: the number of words in a specific document
- α: the Dirichlet prior parameter for deciding the document-topic distributions
- β: the Dirichlet prior parameter for deciding the topic-word distributions
- θ_i: the topic distribution for the document i
- φ_k: the word distribution for the topic k
- z_{ij}: the topic of the j-th word in document i
- w_{ij}: the j-th word in document i

The generative process of a corpus D of M documents, each of which has length N_i words can be simplified as the following steps:

1. Choose $\theta_i \sim \text{Dir}(\alpha)$, $i \in \{1,\ldots,M\}$
2. Choose $\varphi_k \sim \text{Dir}(\beta)$, $k \in \{1,\ldots,K\}$
3. For each word $w_{i,j}$ where $i \in \{1,\ldots,M\}$ and $j \in \{1,\ldots,N_i\}$
    a. Choose topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$
    b. Choose word $w_{i,j} \sim \text{Multinomial}(\varphi_{z(i,j)})$

According to the model, the probability of observing the corpus is:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d$$

Solving the equation of multiple multinomial distributions requires statistical inference calculation. In the original STRUCTURE paper, Pritchard et al. proposed to get approximation through Monte Carlo simulations. Other popular alternatives include Gibbs sampling, variational Bayes, and likelihood maximization approaches (Griffiths & Steyvers, 2004).

One of the assumptions that the basic model requires is that the number of topics k is known and fixed. Defining the best k has always been a problem in various studies using LDA, which varies due to the features of the data and the goal of the analysis. One method to determine the best number is to measure the performance of the model based on how the model fits the input data, using likelihood and perplexity. However, the result from the performance measurement may not be the best fit: the topics generated may not provide semantic interpretations. Thus, human judgment is required to aid the process. Hierarchical Dirichlet process, which is the non-parametric generalization of LDA, is also helpful in determining the topic sizes because HDP assumes that the number of topics is unknown and is to be interpreted from the input data (Teh et al., 2006). Moreover, instead of using the model performance, topic coherence, which measures the similarity between items from the topic, is another approach to solving the topic number problem (Rosner et al., 2014).

Application of LDA in biology

LDA in gene expression analysis

Gene expression is the process by which the information coded by a gene is recognized and turned into functional molecules (Crick, 1970). Analyzing the regulation of gene expression is essential for understanding the functional biological processes. One of the critical techniques to measure gene expression is RNA sequencing, which shows the existence and quantity of RNA from biological samples (Z. Wang et al., 2009; Wilhelm & Landry, 2009).

CountClust is an R package that applies LDA to form clusters/topics to represent samples of RNA-Seq gene expression data. Interpreting clusters show the distinctively expressed genes, further implying the functional regulation behind the difference in gene expression data. Application of the CountClust to the GTex Data revealed that the clusters generated from samples can work as a dimension reduction, and the cluster memberships are associated with the tissue labels. Concurrently, cluster interpretation can discover the driving genes in each group, which corresponds to the functions of the tissues (Dey et al., 2017). For example, the whole blood sample has 3 significant clusters, and their driving genes represent immunoglobulins, beta-globins, and cytokines, which correspond to the functions of blood cells (Schroeder & Cavacini, 2010). Functional relationships between clusters and gene expression data indicate that LDA has the potential to reveal hidden connections in biological data.

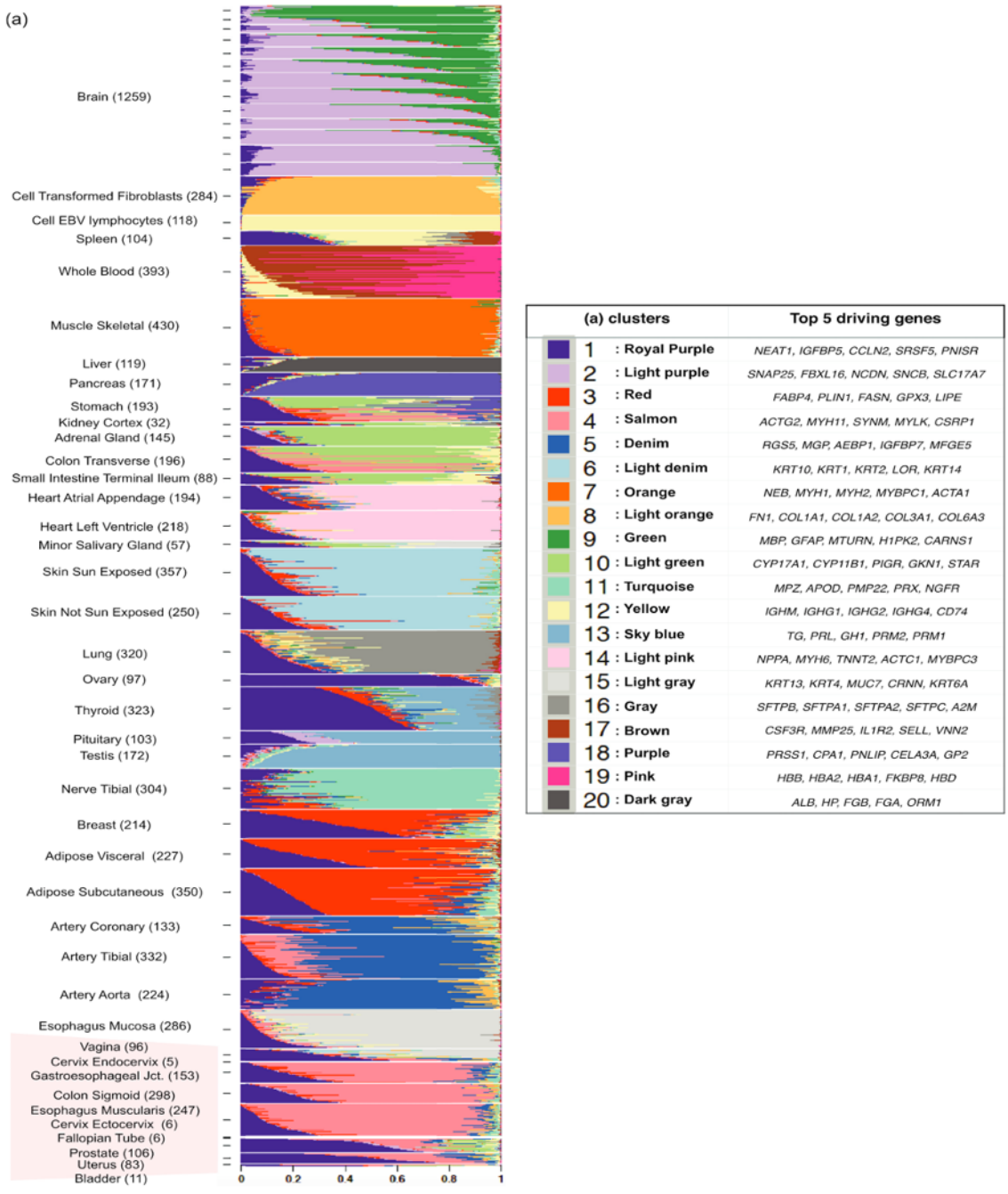


Figure 1.7 LDA analysis result of GTex tissues. Gene expression profiles of 8,555 samples from 53 tissues are modeled with 20 clusters (left). The driving genes in the clusters are also interpreted (right). (Dey et al., 2017)

LDA in mutation profile interpretation

Somatic mutations in cancer represent the consequences of various driving factors and mutational processes, providing a rich source to study features and potential

treatment for tumors (Zehir et al., 2017). Matsutani et al. utilized LDA with variational Bayes inference to investigate the signatures of cancer mutations in various cancer types. Applying the model to the COSMIC database successfully recovered known COSMIC mutation signatures and discovered novel signatures associated with multiple cancer types. Moreover, variational Bayes can identify the most suitable number of signatures based on the input dataset (Matsutani et al., 2019). The study suggested that with a careful and creative application, LDA can help recognize functional hidden features, which can explain the biological implications of diseases.

LDA in microbiome species identification

Investigating multi-omic data is challenging since the heterogeneous data requires various profiling methods (Subramanian et al., 2020). However, text documents in natural language processing also bring the problem of heterogeneity, which admixture models can overcome. Tataru et al. applied LDA to multi-omics data of microbial data to investigate the complexity of the human gut microbial communities and the differences between healthy and Autism individuals. Admixture modeling across different omics data revealed topics associated with diet and behavior, which were not in the information used in the modeling process. Interpretation of the topics also indicated that the features from topics are highly related to microbial functions. The application enabled the team to discover the metabolic differences between Autism and healthy individuals (Tataru et al., 2023). Also, the ability to model cross-omics topics shows that LDA has the potential to utilize various data types in biological analysis.

RNA splicing

RNA splicing transforms newly-made RNA into mature RNA by removing introns and ligating exons (Will & Luhrmann, 2011). Correctly processing intron signal recognition is critical to assembling the splicing machinery and maintaining the dynamic gene expression regulations (G.-S. Wang & Cooper, 2007). Moreover, all eukaryotic genomes have exon-intron structures, with various intron abundance and size distributions across species (Long & Deutsch, 1999). One pre-mRNA can go through different ways of assembly to make alternative mRNA isoforms, which can be achieved by 90% of human genes and makes the proteome more diverse than the variation of DNA sequences that code for functional products (Baralle & Giudice, 2017).

Mechanism of splicing catalyzed by the major spliceosome

The spliceosome is a large RNA-protein complex comprising five small nuclear ribonucleoproteins (snRNPs) and interacts with other proteins to carry out splicing in eukaryotes. The assembly of spliceosomes can happen co-transcriptionally and follows strict rules to splice introns with GT as 5'SS and AG as 3'SS (Will & Luhrmann, 2011).

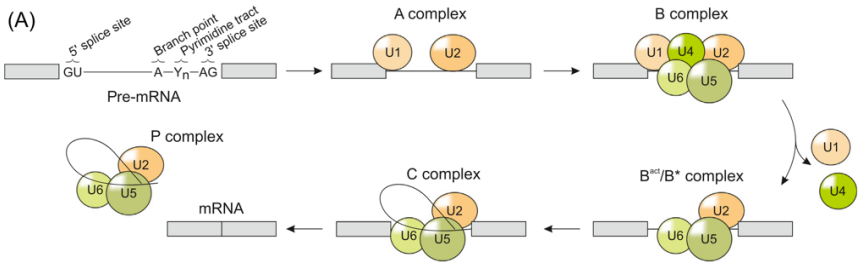


Figure 1.8 Stepwise assembly of spliceosome (Gehring & Roignant, 2021).

The initial step is the recognition of splicing sites, which is triggered by U1 snRNP binding to the 5'SS, splicing factor 1 binding to the branch site, U2AF heterodimer binding to the pyrimidine tract, and the 3'SS, and forms complex E. Then complex A is formed by U2 snRNP replacing SF1 and binding to the branch site. After that, the tri-snRNP U4/U6/U5 binds to the complex, of which U5 binds to the exon at 5'SS and U6 bins to U2, to make complex B. U1 is then released, followed by U5 moves to the intron region and U6 binds to U2. After U4 is released, transesterification is catalyzed by U2/U6, which brings 5'SS and the branch site together and brings U5 to the 3'SS, leaving 5'SS to be cleaved, and the lariat is formed. This is called the catalytic spliceosome or complex C. The last stage is the post-catalytic spliceosome (complex C*), where the 3'SS is cleaved with U2/U6/U5 remains on the lariat, and exons are ligated upon ATP hydrolysis. After splicing, RNA is released, and the spliceosome is recycled.

In addition to the spliceosome assembly, various RBPs can regulate splicing processes by recognizing enhancers and silencers in the flanking region (De Conti et al., 2013; Yan et al., 2019).

Alternative splicing

Alternative splicing is a gene regulation process that allows a single gene to code for multiple protein products by selecting different combinations of splicing sites. Differentially spliced RNAs have different exon combinations, which can represent different codon assortments and get translated into proteins of distinct functions (Baralle & Giudice, 2017; Y. Lee & Rio, 2015).

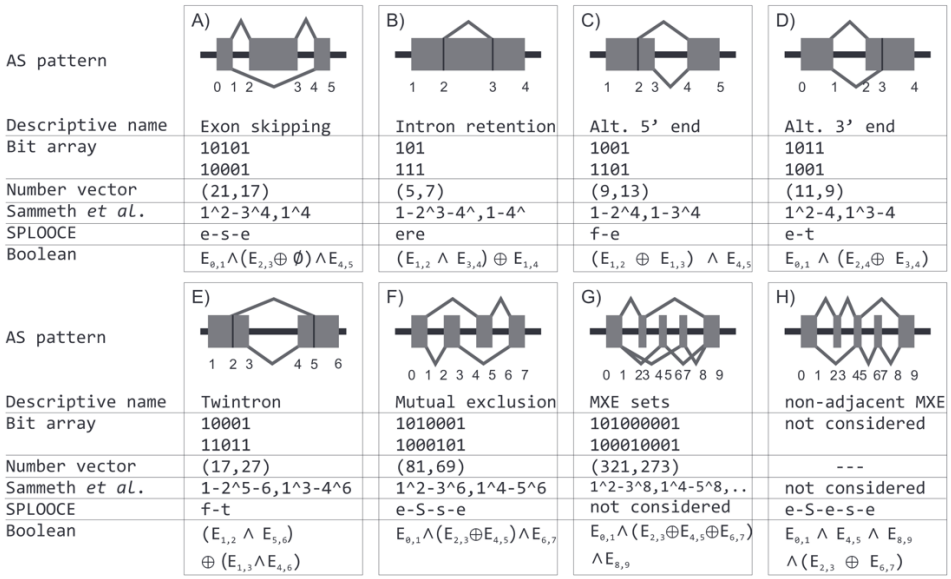


Figure 1.9 Graphical and symbolic representations of AS events (Pohl et al., 2013).

Common modes of alternative splicing include:

1. Exon skipping. Exon skipping is the most common alternative splicing mode, where an exon is spliced out from the pre-mRNA. One function of exon skipping is restoring the transcript's reading frames. Exon skipping has been applied in treating Duchene muscular dystrophy, where it fixes the truncated dystrophin protein (Takeda et al., 2021).
2. Intron retention. Intron retention is the rarest mode of alternative splicing, where a piece or the whole intron remains in the mature RNA and codes for amino acid cooperatively with adjacent exon sequences.
3. Mutually exclusive exons. Mutually exclusive exons are two exons that either exon can be included in the mature RNA but not both. Instead of being independent from each other, the expression of mutually exclusive exons is regulated coordinately. Mutually exclusive exons are not necessarily adjacent exons, and either part of the pair can be more than one exon (Pohl et al.,

2013). Two special cases of MXE are alternative first and alternative last exons.

4. Alternative 5'SS. The alternative 5'SS means two 5'SS competing to pair with a fixed 3'SS, which alters the 3' end of the upstream exon sequence.
5. Alternative 3'SS. The alternative 3'SS competes with the canonical 3’SS to be used with a fixed 5'SS, changing the boundary of the 5' end of the downstream exon.

Alternative splicing is regulated by both cis-acting sites on the pre-mRNA sequences and trans-acting proteins cooperatively. The effects of splicing factors can vary based on the binding site positions (Fu & Ares, 2014).

Two major cis-acting sequence features have been recognized. The splicing silencers are signals that can be recognized and bound by repressor proteins, resulting in decreased usage of adjacent splicing sites. Splicing silencers have different motifs to be recognized by various proteins. They can exist in the intron as intronic splicing silencer (ISS) or the adjacent exons as exonic splicing silencer (ESS) (Brooks et al., 2015; Y. Lee & Rio, 2015; Z. Wang & Burge, 2008). Two major splicing repressor proteins are heterogeneous nuclear ribonucleoproteins (hnRNPs) and polypyrimidine tract binding proteins (PTB) (Busch & Hertel, 2012; Clower et al., 2010).

The splicing enhancers are sequence features that can increase the splicing rate and the inclusion rate of the exons. They serve as binding sites of splicing activator proteins. Splicing enhancers also exist in both exon regions (exonic splicing enhancer, ESE) and nearby intron sites (intronic splicing enhancer, ISE) (Fu & Ares, 2014; Z. Wang & Burge, 2008). One of the essential members of the splicing

activators is the serine/arginine-rich protein family (SR proteins) that recruit spliceosome assembly at the splicing sites (Busch & Hertel, 2012; Jeong, 2017). Typically, the cis-acting factors are present within the vicinity of potential splicing junctions, indicating an interplay between the activators and repressors in determining the inclusion of the exon (Busch & Hertel, 2012). Studies have shown that the location and concentration of splicing factor proteins, as well as the density of the binding sites and the surrounding sequences, can affect their efficiency. Moreover, the same trans-acting proteins can have opposite effects when bound to different regions (Begg et al., 2020; Lim et al., 2011).

Methylation in alternative splicing regulations

Genome-wide studies of CpG methylation have shown that DNA methylation can affect the spliceosome assembly and splicing regulations (Lev Maor et al., 2015a). A study of honey bee DNA methylation found that decreasing DNA methylation by inhibiting DNA methyltransferase 3 caused many alternative splicing events (Li-Byarlay et al., 2013). Another study that manually altered the methylation status of a single gene revealed that increasing the methylation rate of the whole gene increased the chance of inclusion of alternative exons (Yearim et al., 2015).

In terms of exon-intron architectures, CG dinucleotides that are located near splicing sites tend to have higher methylation rates than those in surrounding regions, and methylation cytosines are found more abundant in constitutive exons than exons involved in alternative splicing events (Gelfman et al., 2013). However, high methylation does not guarantee a high inclusion of alternative exons. Studies have shown that the effect of methylation on splicing is context-dependent, and an

increased methylation rate can also cause lower inclusion of alternative exons (Yearim et al., 2015).

DNA methylation can regulate splicing by two major mechanisms. First, DNA methylation can inhibit the binding of DNA-binding protein CTCF, preventing the Pol II elongation rate from being slowed. Thus, the opportunity for spliceosome assembly is reduced, and the exon inclusion is decreased based on the 'window of opportunity' model (Ong & Corces, 2014; Shukla et al., 2011). Increased methylation can also upregulate the binding of MeCP2, which is a protein that binds to methylated-CpG sites. MeCP2 then promotes the Pol II pausing and increases the exon inclusion rate (Maunakea et al., 2013; Young et al., 2005). Second, DNA methylation can help recruit splicing factors to the pre-mRNA sequences. Studies have shown that HP1 proteins, part of the chromodomain super family proteins with lysine-methyl binding domain, colocalize with various splicing factors in different species, including SR proteins, hnRNPs, and siRNAs. Inhibition of HP1 expression changed alternative splicing. One explanation is that HP1 serves as a bridge between methylation and splicing regulation, recruiting splicing factors to alter the chances of the splicing sites being recognized (Alló et al., 2009; Llorian et al., 2010; Ule et al., 2006; Yearim et al., 2015).

Recursive splicing

Splicing a large intron (>50kb) is challenging for the spliceosome assembly because the distance between 3'SS and 5'SS makes the task of transesterification and lariat formation problematic (Shepard et al., 2009). One solution is to perform a multi-step form of splicing called recursive splicing (RS) by utilizing non-canonical splicing

sites to progressively remove the large intron by pieces (Sibley et al., 2015). When recursive splicing happens, recursive splicing sites can be reconstituted at the end of each step of RS. In 1998, Hatton et al. described the first observed RS in a 74kb intron of ultrabithorax gene in Drosophila melanogaster (Burnette et al., 2005; Duff et al., 2015a; Hatton et al., 1998).

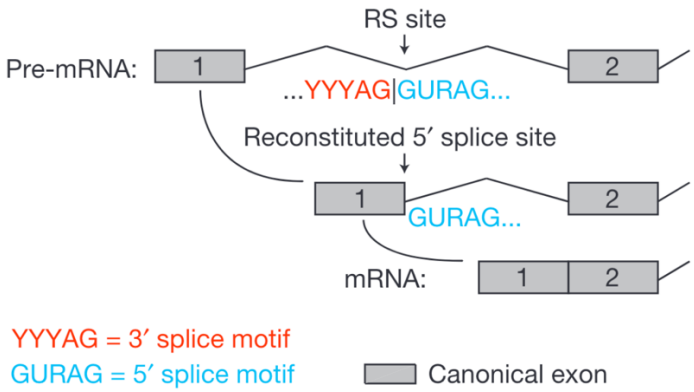


Figure 1.10 Graphical representation of the recursive splicing in D.melanogaster (Sibley et al., 2015).

Studies have identified plentiful cryptic splicing sites in the human genome that can potentially be used for recursive splicing. However, when the intron is removed, and the two sequences flanking the intron boundaries are ligated, a protein complex called exon junction complex (EJC) forms approximately 20-24 nucleotides upstream of the 5' of the splicing junction. EJC provides a position-specific memory of splicing that marks the completion of splicing and prevents surrounding cryptic splicing sites from being used. It is believed that EJC protects the integrity of RNA splicing, and consequently, using cryptic splicing sites in the human genome is very rare (Blazquez et al., 2018; Boehm et al., 2018).

Chapter 2: Mixture Models for Nucleotide Sequence Feature Analysis

Introduction

Nucleotide sequences carry information that encodes the complexity and diversity of life. Information in these sequences directs gene expression, from chromatin structure, transcription, splicing, and other steps of RNA processing, to mRNA translation, stability, and localization. In general, signals for these processes are strings of nucleotides that are recognized by RNA or protein molecules, either alone or as part of a multi-subunit complex. These signals are often represented by consensus sequences, nucleotide frequency matrices, position weight matrices (Stormo *et al.* 2000), or Hidden Markov models (Yoon *et al.* 2009), and there are many algorithms for identifying such signals (*e.g.* MEME and its many extensions, Bailey *et al.* 2015). The paradigm is that a signal can be represented by a sequence motif, which functions as the binding site for a single protein or RNA. However, many signals do not fit this paradigm. Some, (such as core splice sites) are recognized by multiple factors in a temporal sequence. Others (such as exonic splicing enhancers) may consist of a mixture of alternative motifs recognized by alternative factors, which can differ significantly, or may be completely unrelated. There are also signals that are best described by sequence composition, or by the presence, or absence, of very short motifs. Conversely, even when a single factor is necessary for a process and has a well-defined recognition motif, there are sometimes specific

sequences that function without that motif. One example is the general transcription factor TBP, which is thought to function at all transcription events and recognizes a well-defined TATA motif, but not all promoters have this motif (Akhtar & Veenstra, 2011; Cooper et al., 2006). Another example is the branch site, which is recognized by U2 snRNA during the splicing process and has a well-defined sequence motif complementary to U2 (Konarska et al., 1985). Here too, individual introns often lack this motif.

One sequence element that comes in distinct subtypes is the 3' splice site (Wilkinson et al., 2020). There are three components of the core 3' splice site: the site itself, including an AG dinucleotide; a pyrimidine tract upstream of, and the branch site, where lariat formation occurs during splicing. In addition, auxiliary sequences such as exonic splicing enhancers in the exon downstream of the splice site contribute to the recognition of 3' splice sites. The diversity of 3' splice sites is supported by a variety of observations. Individual human introns differ regarding whether they contain a recognizable branch site motif. While many contain a perfect match to the CTRAY consensus, others do not, and even the A residue is not invariant at actual sites of branching (Taggart et al., 2017; Wallace & Edmonds, 1983). In addition, 3' splice sites in short Drosophila introns are functionally distinct from 3' splice sites in long Drosophila introns (M. Guo & Mount, 1995).

Latent Dirichlet Allocation (LDA, (Blei et al., 2003)) uses the observation of words in documents to describe the composition of documents in terms of topics that use those words. LDA allows a single sample to have partial membership in each topic, making it possible to recognize similarities between parts of samples (Fig 1A). The topics are

matrices of words that are observed, and each individual document is described as a mixture of topics whose fractional representation sums to 1. Documents can be compared by analyzing their topics, and the distinguishing features of each document are interpretable because the driving features in topics are different. LDA has proven to be a useful and extremely popular model in population genetics (Pritchard et al., 2000), where genetic similarity among individuals, geographic isolation, and admixture can be recognized using multi-locus genotype data. Moreover, LDA has been applied to RNA-seq data analysis (Dey et al., 2017), where modeling gene expression data from tissue samples allowed similarities between tissues and estimates of the proportional representation of samples to be described in terms of robust interpretable features that could be easily understood as cell type-specific genes. Those applications of LDA focus on genotype- and gene- level data and show that the clustering power and interpretable features of LDA are useful for describing gene expression data. LDA has also been applied to mutational signatures in cancer (Matsutani et al., 2019), but not to nucleotide sequences more generally.

Here I apply the LDA mixture model to nucleotide sequences using k-mers, which are strings of k consecutive nucleotides, as features (Fig 1B). In the first approach, modeling positions in an alignment, sequences are aligned relative to a fixed position such as a splice site, transcription start site or polyadenylation site, and individual positions are used as samples. In a second approach, aligned sequences are used as samples, and positional k-mers, in which the location of the k-mer relative to the point of alignment is part of the feature, are used as features. Finally, a third approach uses k-mer counts in bulk (unaligned) sequences.

Applying LDA to splice sites from introns of different lengths allowed the identification of known core splice site motifs, including the branch site consensus, which is nearly impossible to derive from sequence alone using established methods such as MEME (Bailey et al., 2006). I also identify distinct intron subtypes distinguished by distinct sequence preferences throughout the intron. While these subtypes are preferentially associated with short or long introns, there are apparently some long introns of the short subtype and *vice versa*. I also find that LDA can reliably identify the reading frame within coding sequences and can also distinguish human and Drosophila coding sequences. In summary, the results show that LDA is a useful model for describing heterogeneous signals, for assigning individual sequences to subtypes and for identifying sequences that are unusual and do not fit the recognized subtypes.

Results

Modeling positions in an alignment identifies signals associated with RNA splicing.

To evaluate the ability of mixture models to identify position-related features in batched sequences, I aligned human and Drosophila intron sequences at annotated 3' splice sites and used k-mer compositions at each position in the alignment (Fig 2.2A). For the example shown in Fig. 2.1A, I used 6-mer features, applied a sliding window of six nucleotides, and fitted the LDA model with 6 topics to compare long and short human introns (Fig 2.2B). The driving k-mers of each topic were obtained based on K-L divergence (as in (Dey et al., 2017)). Examination of the driving k-mers (Fig.

2.2C) shows that topic 4 contains features similar to the branch site (CTNA), and topic 2 resembles the polypyrimidine tract.

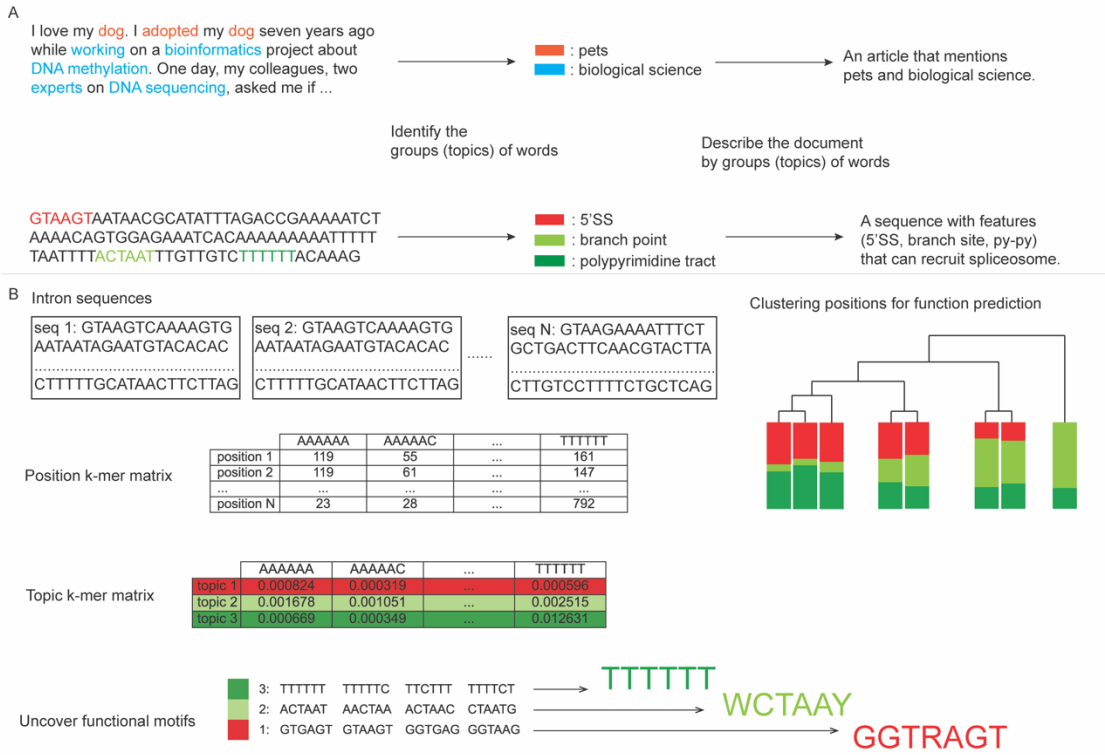


Figure 2.1 LDA applications in natural language and biological sequences. (A) Summarizing topics from input natural language sequences and biological sequences by LDA. LDA can analyze the building blocks from the input sequences (words or nucleotide k-mers) to recognize topics, which describe the features of the input sequences. (B) The pipeline of analyzing intron sequence features by LDA. After summarizing the k-mer counts at each position in a matrix, LDA calculates topic k-mer matrices and transforms sequences into topic memberships. Sequence clustering can be achieved by analyzing the topic distributions and the interpretation of topics can reveal functional motifs.

Both long and short human introns share a peak of topic 4 (branch site) membership at about -30 and a peak of topic 2 (pyrimidine tract) membership at about -15 relative to the annotated 3' splice site (3'SS; Fig 2.2B). Thus, both the relative positions of topic 2 and topic 4 peaks and the interpreted topic meanings are consistent with the known sequence feature of human introns. However, the fractional membership in topic 4 at some positions, such as -28 in short human introns, is close to 100%, much greater than the fraction of sequences with discernible branch site motifs. This is

despite the difficulty of deriving a branch site consensus from sequence alone, and the absence of significant information content in position weight matrices aligned at the 3'SS (Fig 2.2D, 2.2E). This implies that topic 4 includes k-mers that are characteristic of the branch site region that do not match the branch site consensus, and may not be branch sites.

The distribution of topics from the model fitted with long and short introns shows the potential of LDA to identify functional motifs and distinguish sequence types. However, fractional representation, like that shown in Figure. 2.2B and 2.2F, does not indicate how well the topics describe the sequence, only which topics do best. To assess the performance of these topics in describing the raw sequences, I calculated the likelihood of reproducing k-mer compositions at every position by every topic (Figure 2.2G). The highest likelihood values are associated with known motifs (the branch site and pyrimidine tract) in the correct positions, which confirms the potential of unsupervised mixture models to identify functional signals. Also, the significant decay of the total log-likelihood as the distance from the 3'SS increases indicates that, as expected, these known motifs are much better described by the model than is bulk intron sequence far from the splice site.

The analysis also reveals differences between long and short introns. Compared with long introns, the region upstream of short human intron 3’ splice sites has more complex features, including an enrichment of topic 6 between 40 and 80 bp upstream; topic 6 is characterized by GC-rich driving k-mers and is almost entirely missing from long introns. Topic 3 is enriched between 65 and 85 bp upstream of the 3'SS, and topic 1 is enriched upstream of that. In contrast, topic 5, characterized hexamers

containing AG or CG, is predominant in long introns. Because 48% of human intron sequences less than 150 bp are less than 100 bp (Figure 2.S1H), the region shown in Fig. 2.2B includes the 5'SS's, and most of the driving k-mers in topic 1 are subsets of the 5' splice consensus sequence CAG|GTRAG (Figure. 2.2C).

By analyzing the Drosophila genome with the same pipeline, topics corresponding to the branch site and the polypyrimidine tract were again identified, and at the same positions as in the human sequence analysis; but with different driving k-mers (Figure 2.S1A, 2.S1B). As with human sequences, the relationship between these signals, and the difference between short and long introns, is not apparent in positional weight matrices alone (Figure 2.S1C, 2.S1D). Again, one of the significant differences between long and short introns is the presence of the 5' splice site, represented here by topic 6, in the region modeled (Figure 2.S1A, 2.S1B). Again, all of the top 5 driving k-mers are subsets of the 5' splice site consensus, AG|GTRAGT. This identifies topic 2 as a feature of exon sequences (Figure 2.S1A, 2.S1B).

To confirm that the position of the signals from mixture models is consistent with the known signals, we compared the location of the branch site and polypyrimidine tract motifs with the location of the corresponding topics. We used the fraction of 6-mers containing CUNA as a marker for the human branch site, and CUAA for the Drosophila branch site, and 6-mers with only pyrimidines for the pyrimidine tract. While only a very small fraction of sequences contain these specific motifs and nearly all sequences are represented by the corresponding topics, both signals have peaks at the same positions as the consensus signal curves, with similar differences between long and short introns (Figure 2.2H, 2.S1G).

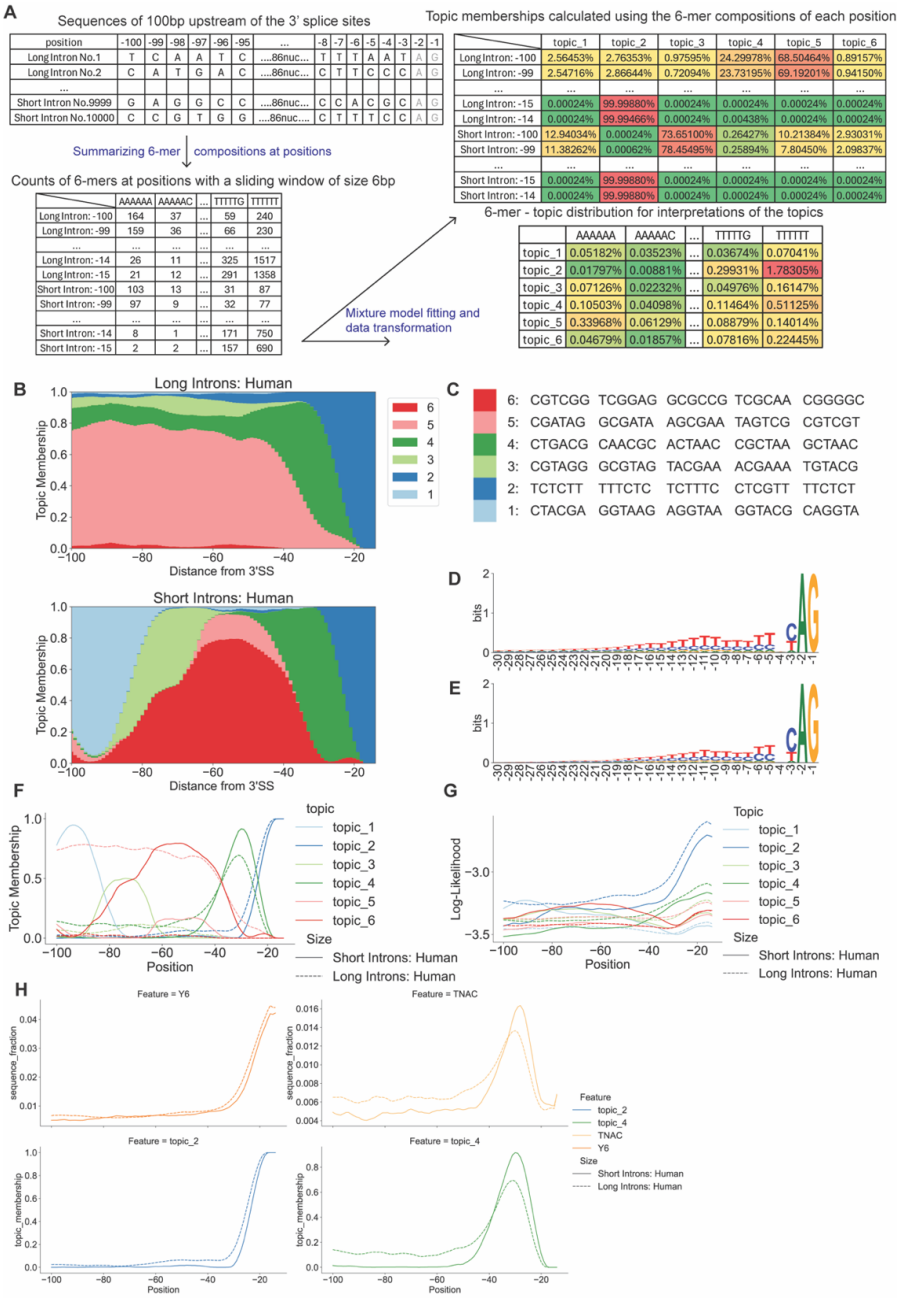


Figure 2.2 LDA characterization of short and long human introns using aligned positions as samples. LDA was applied to samples corresponding to positions in an alignment of 3' splice sites from long (>=150 nt.) and short (<150 nt.) introns. (A) Diagram of the LDA analysis of intron sequences aligned at the 3' splice sites. First, sequences are aligned at the 3' splice sites (up left), and the positions of A and G are -2 and -1. The region from position -100 to -3 are used for the analysis. Then the counts of 6-mers with a sliding window of size 6 at every position are summarized (down left). LDA fitting and data transforming will generate two new matrices. One shows the topic distributions at each position (up right), and the other shows the probability of observing 6-mers from topics (down right). (B) Structure plots of long (n=10,000) and short (n=10,000) human introns. Hexamer features in 5 nt. windows starting at positions -100 to -13 upstream of 3' splice sites (encompassing -100 to -3) were used as samples. (C). Table of top 5 driving hexamer features in the six topics in Fig 2B. (D) Sequence logo of 3'SS of long human introns. (E). Sequence logo of 3'SS of short human introns. (F) Line plot of topic distribution across positions relative to the 3'SS of human introns. (G) Line plot of the likelihood of observing the distribution of features at each position in human introns. (H) Line plots of the branch site and pyrimidine tract consensus sequence signals and corresponding topic signals.

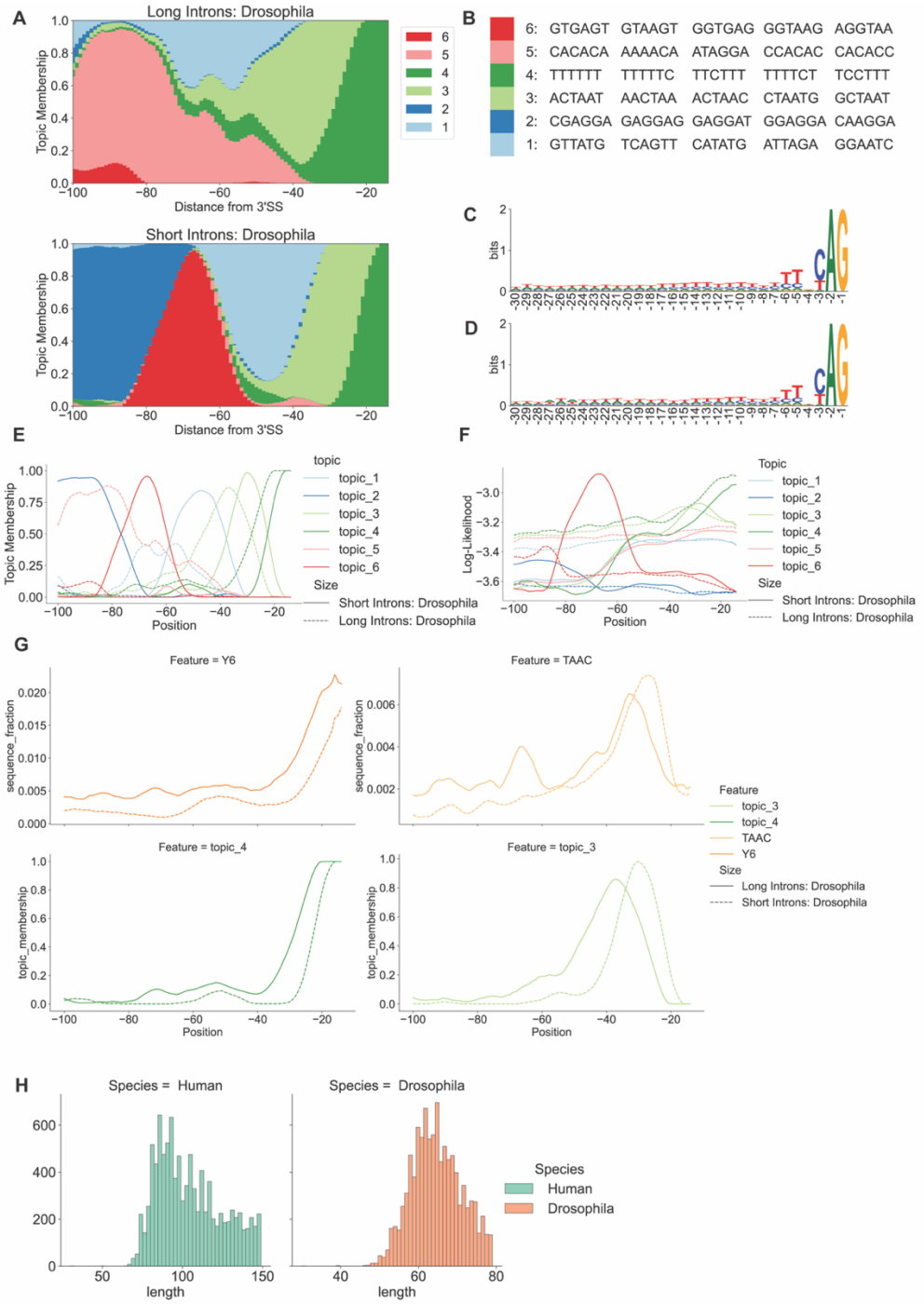


Figure 2.S1 LDA characterization of short and long introns using aligned positions as samples. LDA was applied to samples corresponding to positions in an alignment of 3' splice sites from long and short introns from Drosophila (long>=80 nt.; short<80 nt.). (A) Structure plots of long (n=10,000) and short (n=10,000) Drosophila introns. Hexamer features starting at positions -100 to -13 upstream of 3' splice sites (encompassing -100 to -3) were used as samples. (B) Table of top 5 enriched hexamer features in the six topics in (A). (C) Sequence logo of 3'SS of long Drosophila introns. (D) Sequence logo of 3'SS of short Drosophila introns. (E) Line plot of topic distribution across positions relative to the 3'SS of Drosophila introns. (F) Line plot of the likelihood of observing the distribution of features at each position in Drosophila introns. (G) Line plots of the branch site and pyrimidine tract consensus sequence signals and corresponding topic signals. (H) The distribution of short intron sizes in the analysis of Fig 2 and Fig S1.

Modeling sequences using positional k-mers distinguishes intron subtypes.

To distinguish intron subtypes on the single sequence level, I applied mixture models using blocks of sequence around individual splice sites, instead of positions, as experimental units. The k-mer compositions of introns with different lengths are similar, which makes bulk k-mers unsuitable for this classification task (Figure 2.S2A-D). However, mixture models indicated a difference between similar positions from introns of different lengths (Figure 2.2). Accordingly, I added positional information to the k-mer features, so that each feature is a k-mer at a specific position (e.g. CCAG_at_-4), and applied LDA to samples consisting of small packs of individual sequences (Figure 2.3A). I investigated the positional sequence features of 3'SS and 5'SS separately. To study the features of 3'SS from long (>=150bp) and short (<150bp) human introns, I selected 80 bp sequences that include 50bp from the intron and 30bp from the downstream exon. A model fitted with 60 packs of 250 single sequences (15,000 total), readily distinguished packs of 250 long or short introns (Figure 2.3B). Long intron sequence packs have topic 2 memberships around 90% while the short intron sequence packs have topic 2 memberships less than 15%. A separate set of sequences was used to test the model's performance on single sequences, which generated a classification accuracy of 67.95% (Figure 2.3C, Methods). Driving k-mers from these topics and sequence logos generated for the upstream region (Figure 2.3D, 2.3G, 2.3H) indicate that 1) the long intron group has a longer polypyrimidine tract, which is consistent with the observation of previous study (Yıldırım & Vogl, 2023); 2) topic 2, enriched in long introns, is characterized by runs of T in the pyrimidine tract (-20 to -5) while topic 1 is characterized by runs

of C in the pyrimidine tract (-6 to -3); 3) topic 1 is characterized by GC-rich sequences upstream (-50 to -44). These differences characterize the most distinctive k-mers; there are many other differences. Though long introns have more sequences that are enriched in topic 2 and can be better represented by topic 2, both groups have many individual introns with significant memberships of the topics of the other labels (Figure 2.3C, 2.3I, 2.3J). It is important to bear in mind that the classification of introns by subtype may be more accurate than the assignment to size classes if mechanistically or biochemically distinct subtypes do not perfectly correspond to size classes.

The same pipeline was applied to 80 bp sequences at the 5'SS, including 50bp from the intron and 30bp from the upstream exon (Figure 2.S3), yielding a classification accuracy of 66.6%. Long and short introns differ in base composition in the region 17bp-20bp downstream of the 5'SS (Fig. 2.S3F, 2.S3G), and some driving k-mers reflect this (e.g. TTTT at 18 in topic 2 and GAGG at 22 in topic 1). However, the top-ranked driving k-mers are in the core 5'SS, including the invariant GT dinucleotide (Figure 2.S3D, 2.S3E). Based on the similarity between long and short intron consensus sequences, I conclude that LDA has the potential to find subtle differences between sequences.

The difference between intron subtypes was also revealed from Drosophila introns, with accuracies of 71.75% from the 5'SS and 66.6% from the 3'SS (Figure 2.S4, 2.S5). Similar to the human sequence analysis, intron sequences at 3'SS from the Drosophila genome also indicated a difference in the polypyrimidine tract regions.

However, in Drosophila, the differences in the pyrimidine tract are more related to T-enrichment rather than the size of the signals. (Figure 2.S4F, 2.S4G).

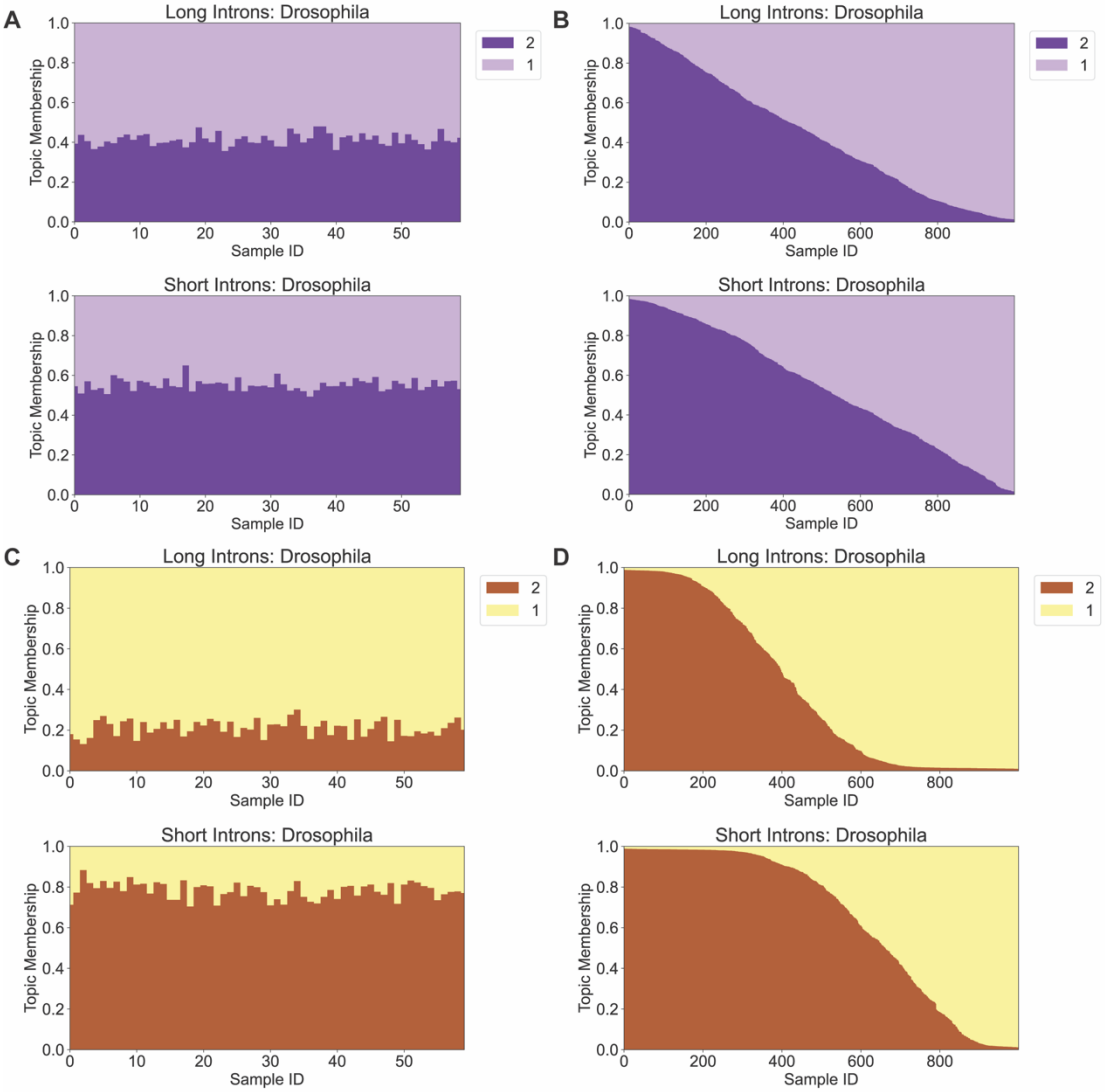


Figure 2.S2 Non-positional feature analysis can not distinguish intron subtypes from Drosophila introns. LDA was applied to Drosophila intron sequences to identify differences between long and short introns using single sequences. Unlike the result in Fig S1, non-positional features are not able to describe the differences between the two groups. (A) Structure plots of long (n=15,000) and short (n=15,000) Drosophila intron sequences (30 nt. of exon and 50 nt. of intron) near 3'SS. Every sample contains tetramer features from 250 sequences (15,000 sequences total). (B) Structure plots of long (n=2,000) and short (n=2,000) Drosophila intron sequences near 3'SS. Every sample is a single sequence. The topics are the same as the topics in Fig S2A. (C) Structure plots of long (n=15,000) and short (n=15,000) Drosophila intron sequences (30 nt. Of exon and 50 nt. Of intron) near 5'SS. Every sample contains tetramer features from 250 sequences (15,000 sequences total). (D) Structure plots of long (n=2,000) and short (n=2,000) Drosophila introns sequences near 5'SS. Every sample is a single sequence. The topics are the same as the topics in Fig S2C.

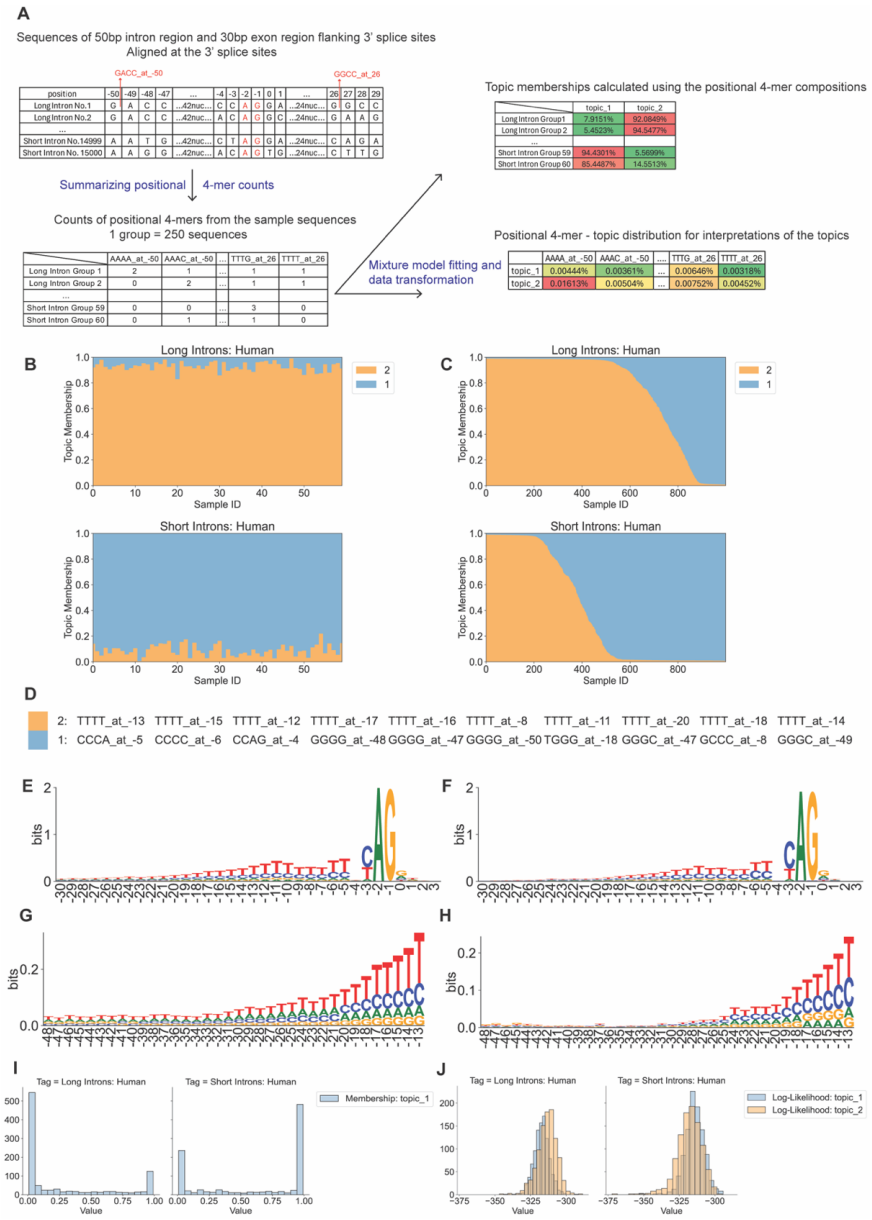


Figure 2.3 LDA characterization of short and long human 3'SS using individual sequences as samples. LDA was applied to samples corresponding to splice site regions from long and short introns from human (long>=150 nt.; short<150 nt.). (A) Diagram of the LDA analysis of intron sequences aligned at 5' or 3' splice sites. In this diagram, we introduce the analysis of sequences flanking 3' splice sites. First, sequences are aligned at the 3' splice sites (up left), and the positions of A and G are -2 and -1. Then every 250 sequences are grouped into one sample group and the counts of positional 4-mers are summarized (down left). Examples of positional 4-mers are shown in the up left table highlighted in red color. LDA fitting and data transformation act as dimension reduction method and represent each sample as topic distributions, which can be interpreted using the topic-kmer table (right). (B) Structure plots of long (n=15,000) and short (n=15,000) human intron sequences (30 nt. of exon and 50 nt. of intron) near 3'SS. Every sample contains positional tetramer features from 250 sequences (15,000 sequences total). (C) Structure plots of long (n=2,000) and short (n=2,000) human introns. Every sample is a single sequence. The topics are the same as the topics in Fig 3A. (D) Top 10 driving positional tetramers from topics 1 and 2 from the analysis of 3'SS of human introns. (E) Sequence logo of the 3'SS from long human introns. (F) Sequence logo of the 3'SS from short human introns. (G) Sequence logo of positions -48 to -13 from long human introns. (H) Sequence logo of positions -48 to -13 from short human introns. (I) Distribution of topic 1 memberships of single sequences in Fig 3B. (J) Distribution of the likelihood of generating sequences in Fig 3B by topics.

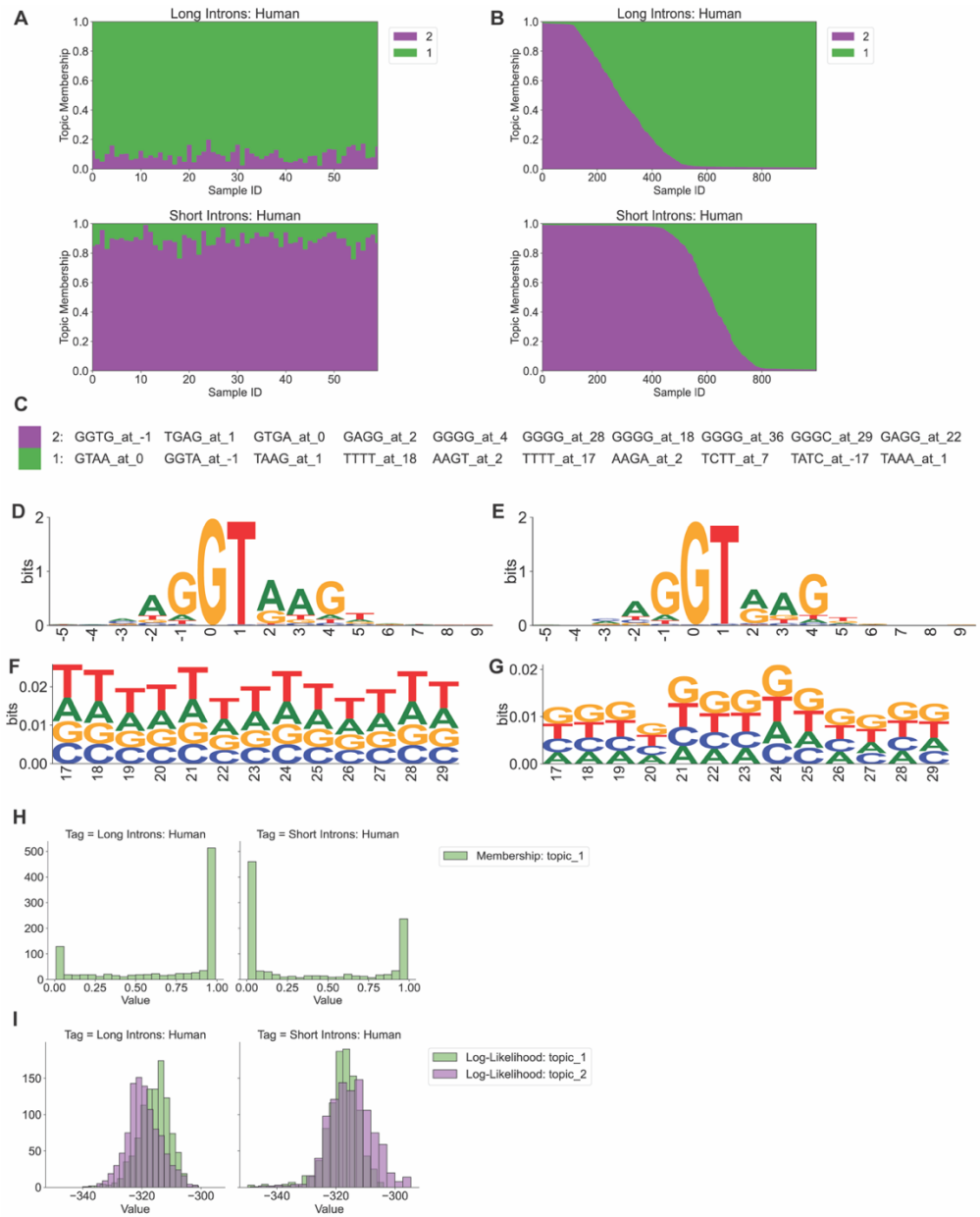


Figure 2.S3 LDA characterization of short and long human 5'SS using individual sequences as samples. LDA was applied to samples corresponding to splice site regions from long and short introns from human (long>=150 nt.; short<150 nt.). (A) Structure plots of long (n=15,000) and short (n=15,000) human intron sequences (30 nt. of exon and 50 nt. of intron) near 5'SS. Every sample contains positional tetramer features from 250 sequences (15,000 sequences total). (B) Structure plots of long (n=2,000) and short (n=2,000) human introns. Every sample is a single sequence. The topics are the same as the topics in Fig S3A. (C) Top 10 enriched positional tetramers from topics 1 and 2 from the analysis of 5'SS of human introns. (D) Sequence logo of the 5'SS from long human introns. (E) Sequence logo of the 5'SS from short human introns. (F) Sequence logo of positions 17 to 29 from long human introns. (G) Sequence logo of positions 17 to 29 from short human introns. (H) Distribution of topic 1 memberships of single sequences in Fig S3B. (I) Distribution of the likelihood of generating sequences in Fig S3B by topics.

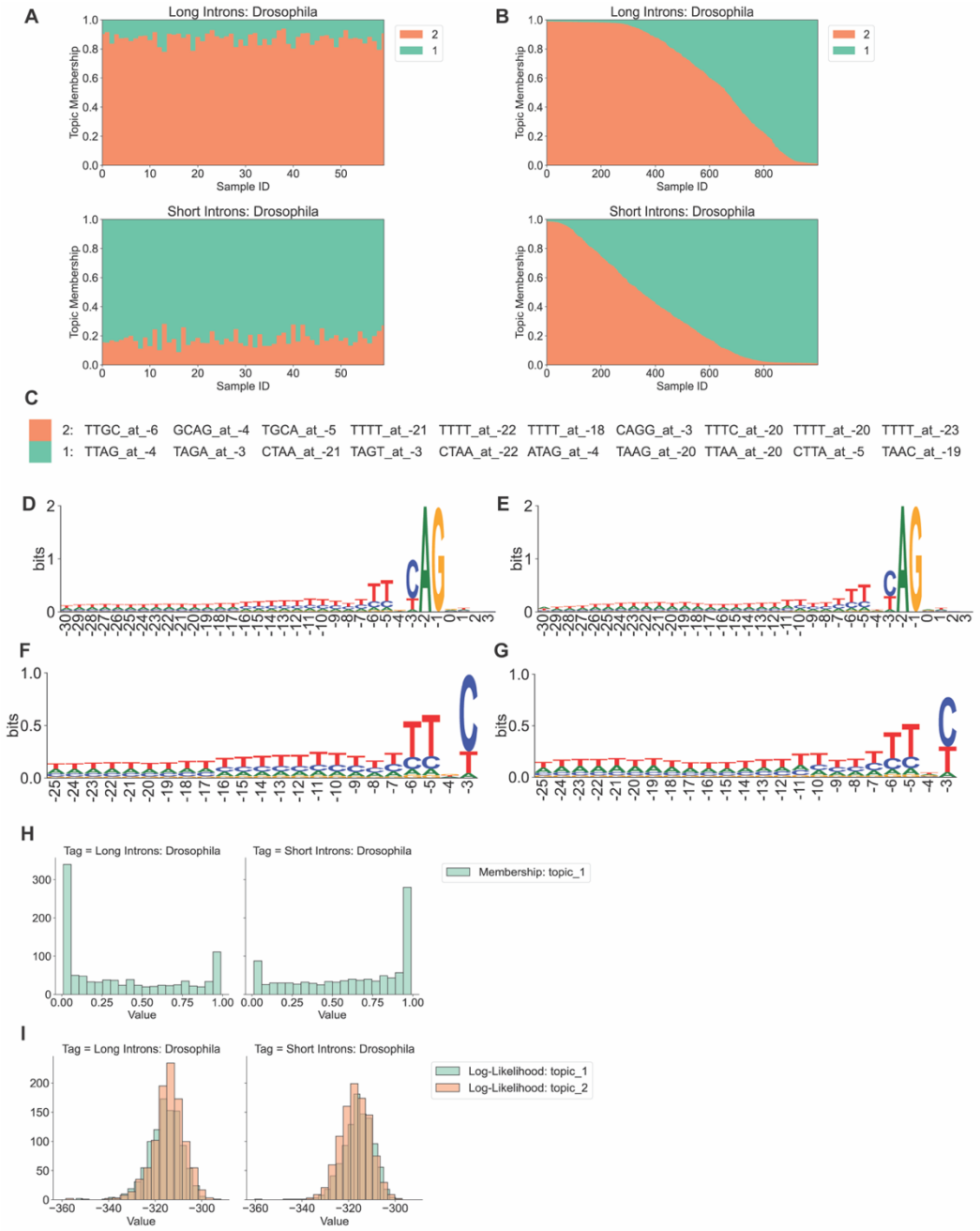


Figure 2.S4 LDA characterization of short and long Drosophila 3'SS using individual sequences as samples. LDA was applied to samples corresponding to splice site regions from long and short introns from Drosophila (long>=80 nt.; short<80 nt.). (A) Structure plots of long (n=15,000) and short (n=15,000) Drosophila intron sequences (30 nt. of exon and 50 nt. of intron) near 3'SS. Every sample contains positional tetramer features from 250 sequences (15,000 sequences total). (B) Structure plots of long (n=2,000) and short (n=2,000) Drosophila introns. Every sample is a single sequence. The topics are the same as the topics in Fig S4A. (C) Top 10 enriched positional tetramers from topics 1 and 2 from the analysis of 3'SS of Drosophila introns. (D) Sequence logo of the 3'SS from long Drosophila introns. (E) Sequence logo of the 3'SS from short Drosophila introns. (F) Sequence logo of positions -25 to -3 from long Drosophila introns. (G) Sequence logo of positions -25 to -3 from short Drosophila introns. (H) Distribution of topic 1 memberships of single sequences in Fig S4B. (I) Distribution of the likelihood of generating sequences in Fig S4B by topics.

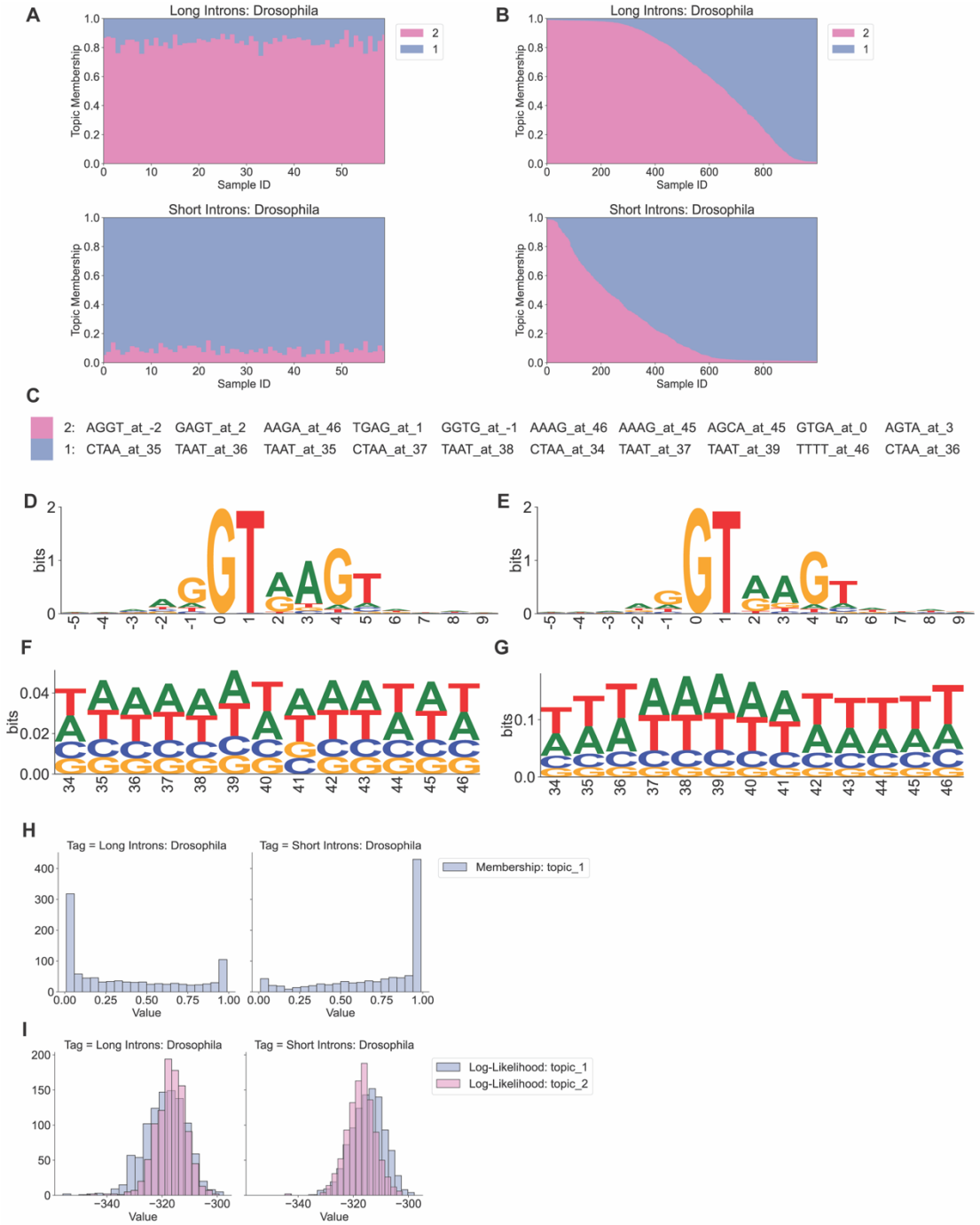


Figure 2.S5 LDA characterization of short and long Drosophila 5'SS using individual sequences as samples. LDA was applied to samples corresponding to splice site regions from long and short introns from Drosophila (long>=80 nt.; short<80 nt.). (A) Structure plots of long (n=15,000) and short (n=15,000) Drosophila intron sequences (30 nt. of exon and 50 nt. of intron) near 5'SS. Every sample contains positional tetramer features from 250 sequences (15,000 sequences total). (B) Structure plots of long (n=2,000) and short (n=2,000) Drosophila introns. Every sample is a single sequence. The topics are the same as the topics in Fig S5A. (C) Top 10 enriched positional tetramers from topics 1 and 2 from the analysis of 5'SS of Drosophila introns. (D) Sequence logo of the 5'SS from long Drosophila introns. (E) Sequence logo of the 5'SS from short Drosophila introns. (F) Sequence logo of positions 34 to 46 from long Drosophila introns. (G) Sequence logo of positions 34 to 46 from short Drosophila introns. (H) Distribution of topic 1 memberships of single sequences in Fig S5B. (I) Distribution of the likelihood of generating sequences in Fig S5B by topics.

Modeling unaligned protein-coding sequences reveals features associated with reading frames and species.

Many classic signals within genes are recurrent motifs over-represented in specific regions, such as the CpG islands in the promoter regions of human housekeeping genes. In addition, pervasive sequence differences characterize genomic regions, such as introns, and can be species-specific. To investigate the performance of mixture models in seeking over- and under-represented motifs from unaligned sequences, I analyzed the topic distribution of coding sequences in three reading frames with mixture models of 3 topics (Figure 2.4A). By randomly grouping 300 human CDS sequences as a single sample, I observed that each reading frame could be represented by a distinct topic, one with little distribution in other reading frames (Figure 2.4B, 2.4C). Since each topic is associated with one reading frame, I calculated the probability of generating each CDS using every topic. The expectation is consistent with the observation of the topic distributions (Figure 2.4D). The classification with a separate set of sequences shows an accuracy of 98.77% (Methods). However, some of the single sequences do show a misalignment of the topic and reading frame (Figure 2.S6A). A comparison of the predicted and true labels showed that the predicted frames of more than 45% of the misclassified sequences were shifted by one, suggesting potential frame-shifting events or annotation errors (Figure 2.4E).

To test whether the model fitted by human sequences can be applied to other species, I transformed the Drosophila coding sequence k-mer composition using the same model. Topic distribution showed that the Drosophila reading frames follow the

features identified from human sequences (Figure 2.S6B, 2.S6C), suggesting that the information learned by LDA may analyze sequences from novel species.

To explore if mixture models can distinguish the reading frames and species at the same time, I fitted a 6-topic model with both human and Drosophila CDS. The topic distribution suggests that each reading frame can be described by 2 topics, which have distinguishable distributions for different species (Figure 2.5A, 2.5B, 2.5C). I calculated the probability of generating each CDS using single topics to evaluate the association between topics and species. For all 3 pairs of topics from 3 reading frames, most sequences have higher expectations from topics associated with the source species (Figure 2.5D-F). Single sequence classification was performed, and 93.6% of the sequences were correctly predicted with regard to both the reading frame and species. Most of the misclassified sequences in this study were labeled as the wrong species (Figure 2.S6D, 2.S6E).

If LDA were to be used to identify novel smORFs as coding sequences, it is useful to know how long a subsequence is necessary to identify the correct reading frame. To investigate this, I collected k-mer counts for short subsequences from 5,400 full coding ORFs in the CDS test set. These subsequences were between 10 and 100 codons drawn from the middle of the original coding sequences and were transformed using the LDA model fitted by full coding sequences. This method achieves an accuracy of 80% with only 18 codons, 90% with less than 34 codons and 95% with less than 74 codons (Figure 2.5G).

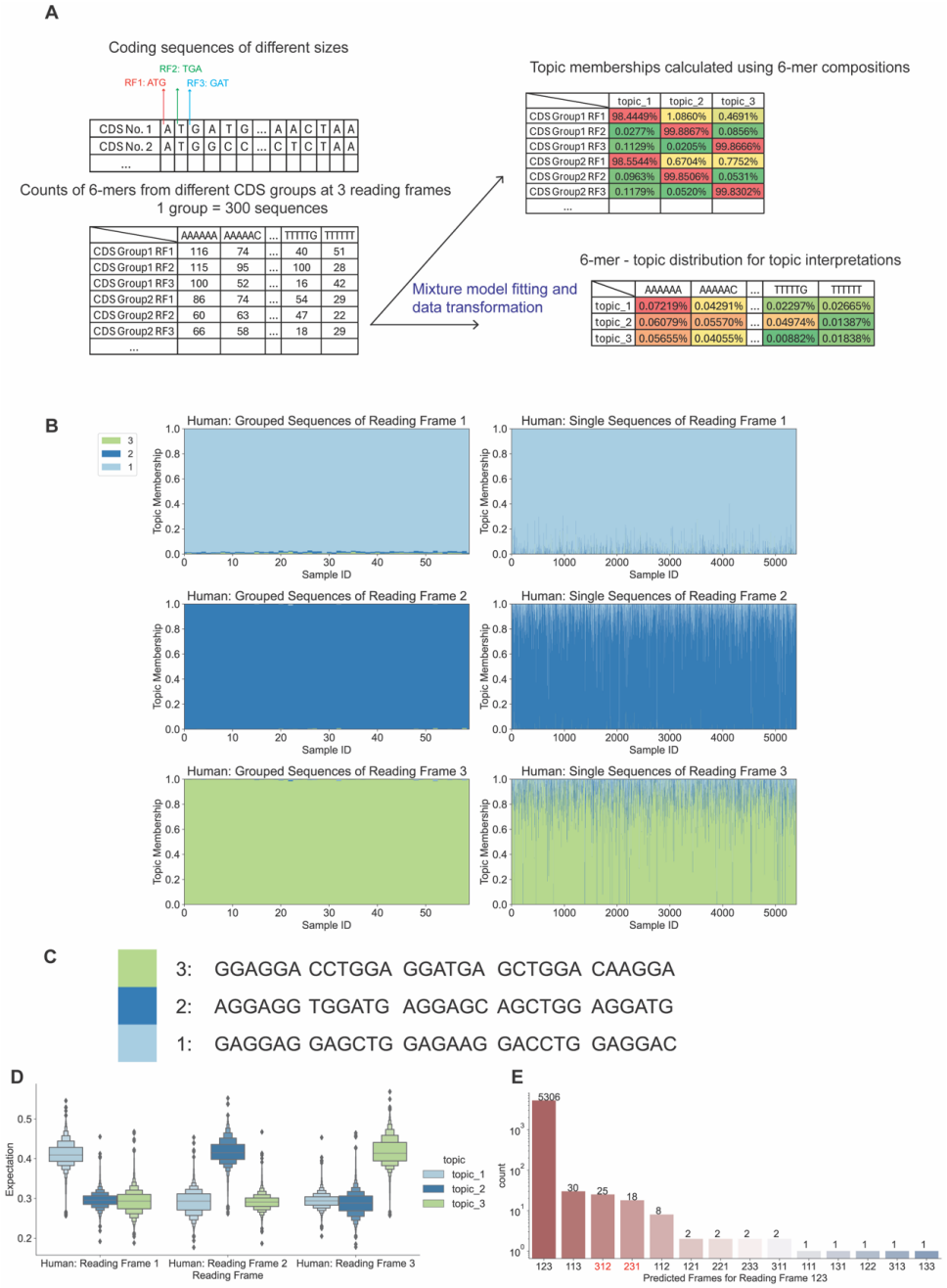


Figure 2.4 LDA characterization of reading frames using non-aligned coding sequences. LDA was applied to human CDS. (A) Diagram of the LDA analysis of coding sequences. First, the reading frames from CDS are labeled (up left). From the example of the No.1 CDS, the 3-mer starting from the first base is from reading frame 1 and the 3-mer starting from the second base is from reading frame 2. Similarly, 3-mers starting from the fourth and fifth bases are from reading frames 1 and 2. Then every 300 CDS are grouped as a sample group, and the counts of 6-mers in each reading frame are summarized (down left). After LDA is fitted and the data is transformed, the sequence groups are represented as topic distributions with interpretable topics (right). (B) Structure plots of human CDS sequences in reading frames 1, 2, and 3. The first column contains the model fitting data (18,000 CDS), in which each sample covers hexamer compositions of 300 sequences. The second column contains the topic distribution of single sequences (5,400 CDS). (C) Top 5 driving hexamer features of 3 topics in Fig. 4B. (D) Distribution of the expectation of the probabilities of generating each sample from the second column of Fig 4A by 3 topics. (E) Barplot of the counts predicted reading frames of sequences in Fig S6A. The correct order is 123.

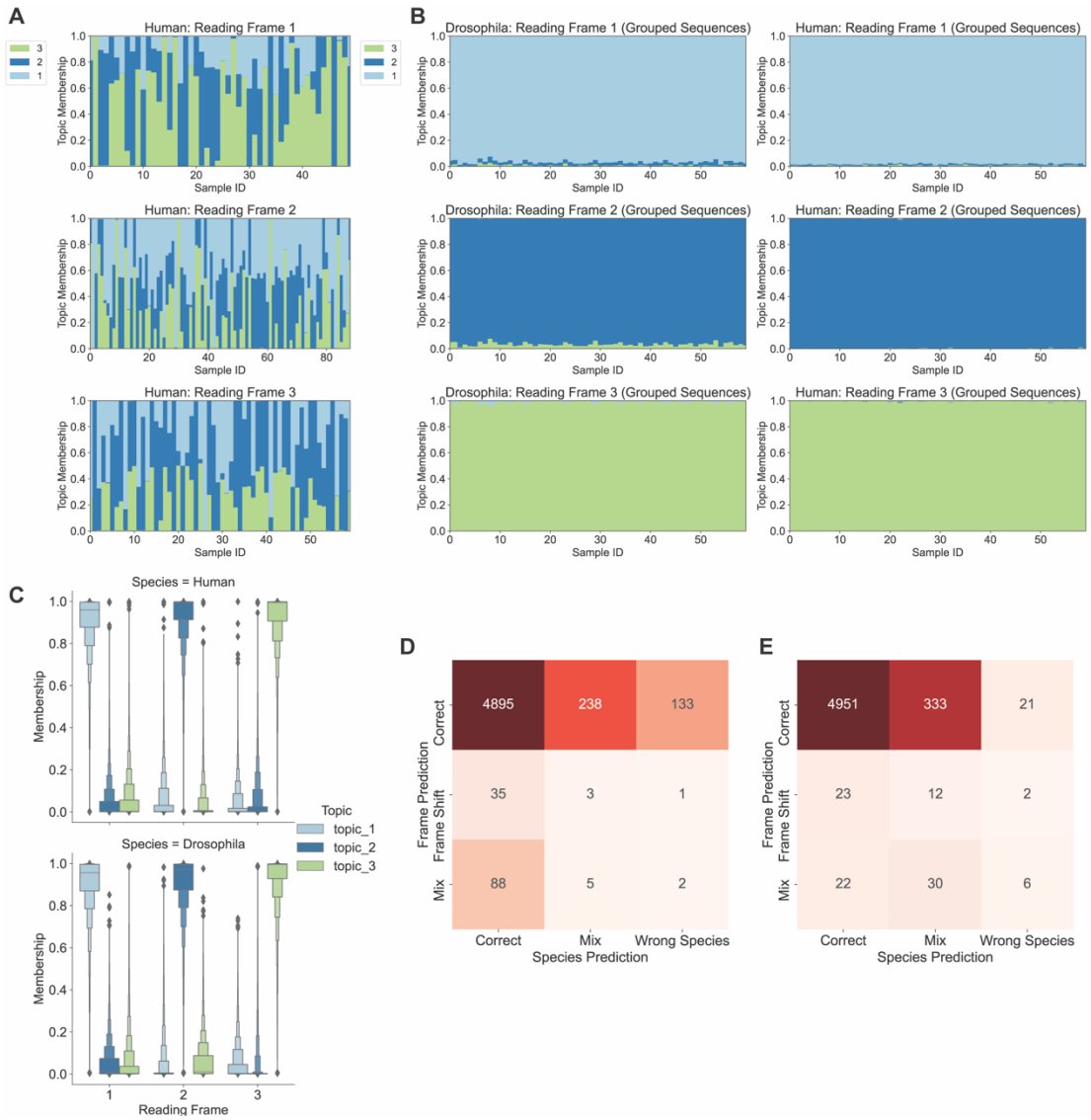


Figure 2.S6 Misclassification from LDA models and generalization of human data fitted LDA model in predicting Drosophila reading frames. Reading frame classification was carried out based on topic distributions. Also, the ability to distinguish reading frames by human data fitted LDA model was tested on the Drosophila sequences. (A) Topic memberships of human CDS sequences that are misclassified into incorrect reading frame tags. (B) Structure plots of transforming human and Drosophila CDS sequences with human CDS-fitted LDA model. (C) The topic distribution of sequences in Fig S6B. (D) The confusion matrix of predicting reading frames and species for human CDS using the model in Fig 5A. (E) The confusion matrix of predicting reading frames and species for Drosophila CDS using the model in Fig 5A.

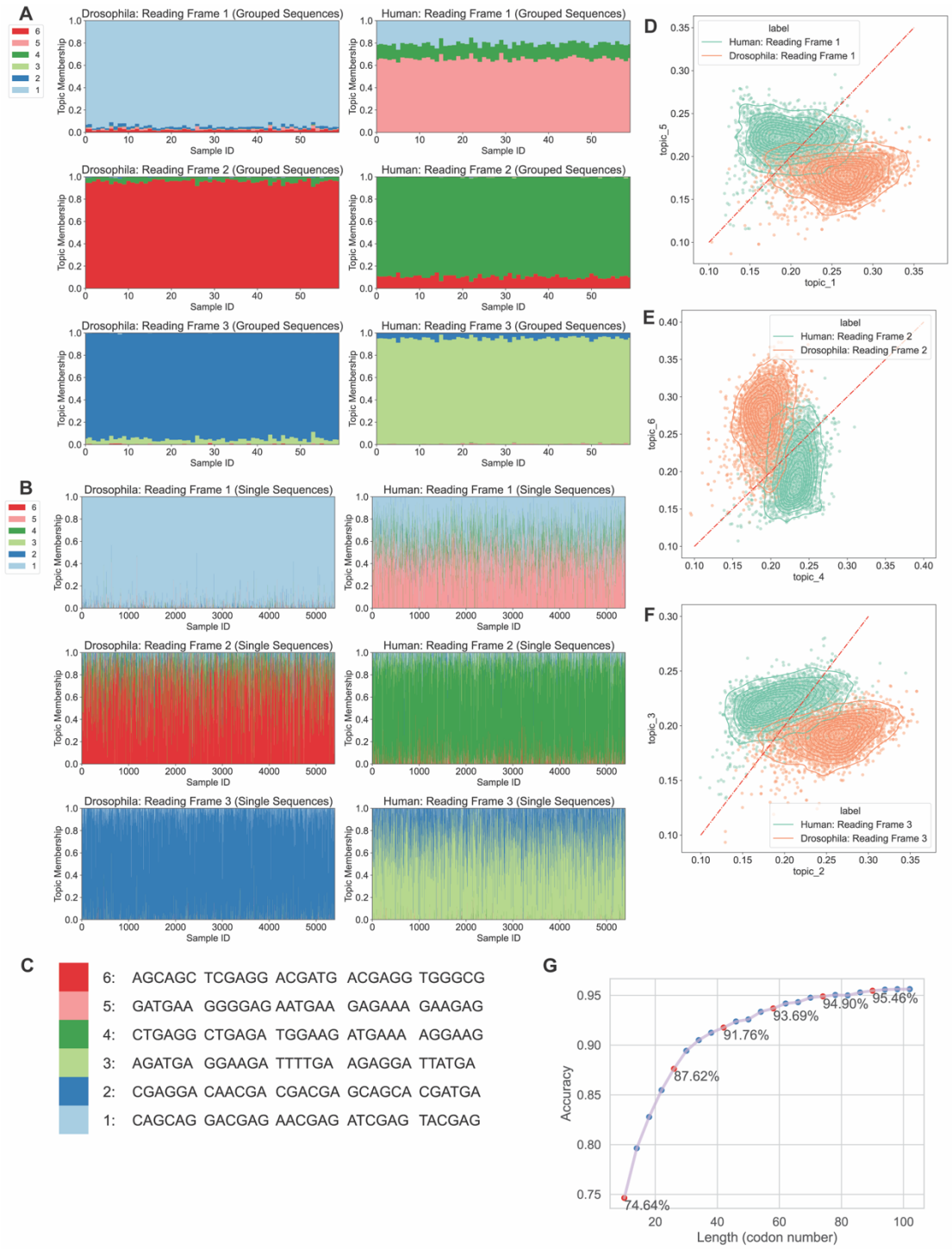


Figure 2.5 LDA characteristics of reading frames using non-aligned coding sequences. LDA was applied to both human and Drosophila CDS. (A) Structure plots of human (18,000) and Drosophila (18,000) CDS sequences in reading frame 1, 2, and 3. Each sample contains the hexamers of 300 sequences. (B) Structure plots of single sequences of human (5,400) and Drosophila (5,400) sequences. (C) Top 5 enriched hexamer features of 6 topics in Fig 5A. (D) The scatterplot of the probability expectations of generating reading frame 1 samples from Fig 5B using topic 1 or 5 from Fig 5C. (E) The scatterplot of the probability expectations of generating reading frame 2 samples from Fig 5B using topic 4 or 6 from Fig 5C. (F) The scatterplot of the probability expectations of generating reading frame 3 samples from Fig 5B using topic 2 or 3 from Fig 5C. (G) Line plot of the accuracies of different lengths of subsequences of CDS to predict the reading frames.

Models based on coding regions identify potential protein-coding information in annotated UTRs and long non-coding RNAs (lncRNA).

To confirm that the signals mixture models identified are abundant and accurate, I analyzed the k-mer compositions from small open reading frames (smORFs) with at least 25 potential codons between the start and stop codons (Method) in human 5'UTR and 3'UTR. The topic distribution distinguished 5'UTR, 3'UTR, and CDS, but no differences were observed between the three frames of smORFs in the UTRs, where different samples share similar topic memberships (Figure 2.S7A, 2.S7B). Often, 5'UTR smORFs can be translated into potentially functional protein products (Starck et al., 2016). The probability distribution of the UTR sequences showed that there are indeed some smORFs that have high expectations generated by the correct topics (Figure 2.6A, 2.6B). I analyzed single smORFs from 5'UTRs and recognized that 13,028 smORF CDS could be correctly classified into three reading frames using the classification model fitted by annotated CDS. Nevertheless, compared with real CDS, these UTR CDS showed different topic distributions of the correct topics associated with each reading frame (Figure 2.6A). To investigate the functions of genes with short 5' UTR ORFs having correctly classified ORFs, I performed a Gene Ontology enrichment (Thomas et al., 2022) analysis. Most GO terms discovered are immune responses, sensory perceptions, and metabolic processes (Figure 2.S7C).
I also analyzed human lncRNA sequences using the same method. The distribution of sequence-generating expectations by topics of lncRNA also indicates that 10,532 sequences have similar features of coding sequences. Compared with UTRs, lncRNA

showed patterns of topic distribution more similar to CDS. After reading frame prediction, we identified several lncRNAs with CDS features (Figure 2.7A).

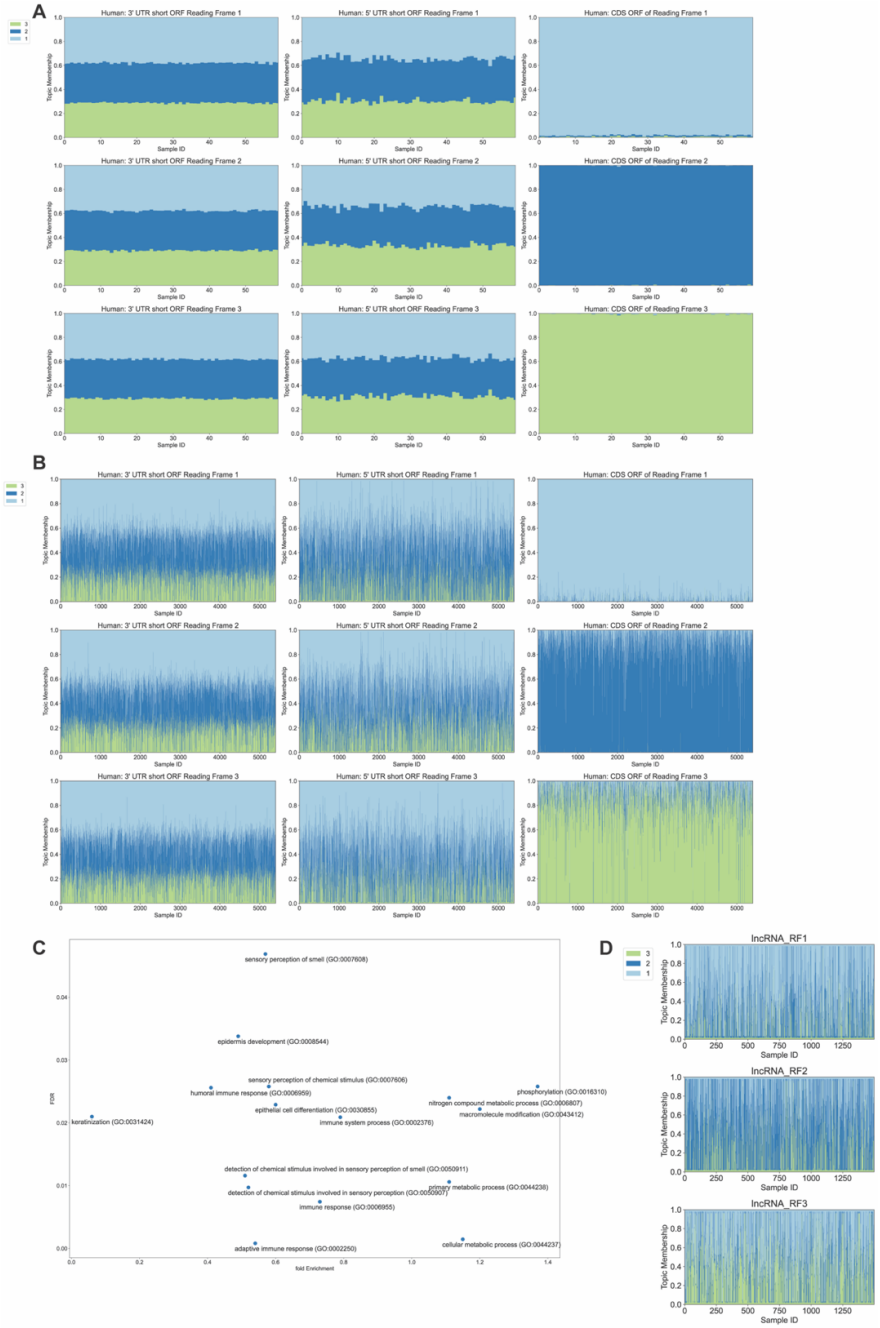


Figure 2.S7 Analyzing short ORF in human UTR5 and lncRNA by CDS LDA model. Human data fitted LDA model was used to analyze short ORF from non-coding sequences. (A) Structure plots of human small reading frames in UTR3, UTR5, and reading frames CDS. Data were transformed by the model Fig 4A. Each sample has hexamer compositions of 300 sequences (18,000 sequences total). (B) Structure plots of human small reading frames in UTR3, UTR5, and reading frames CDS. Each sample is a single sequence from Fig S7A. (C). Gene Ontology enrichment analysis of the smORF. (D) Structure plots of human lncRNA small reading frames. Data were transformed by the model in Fig 4A. Each sample is the hexamer composition of a single sequence (1,500 sequences are shown).

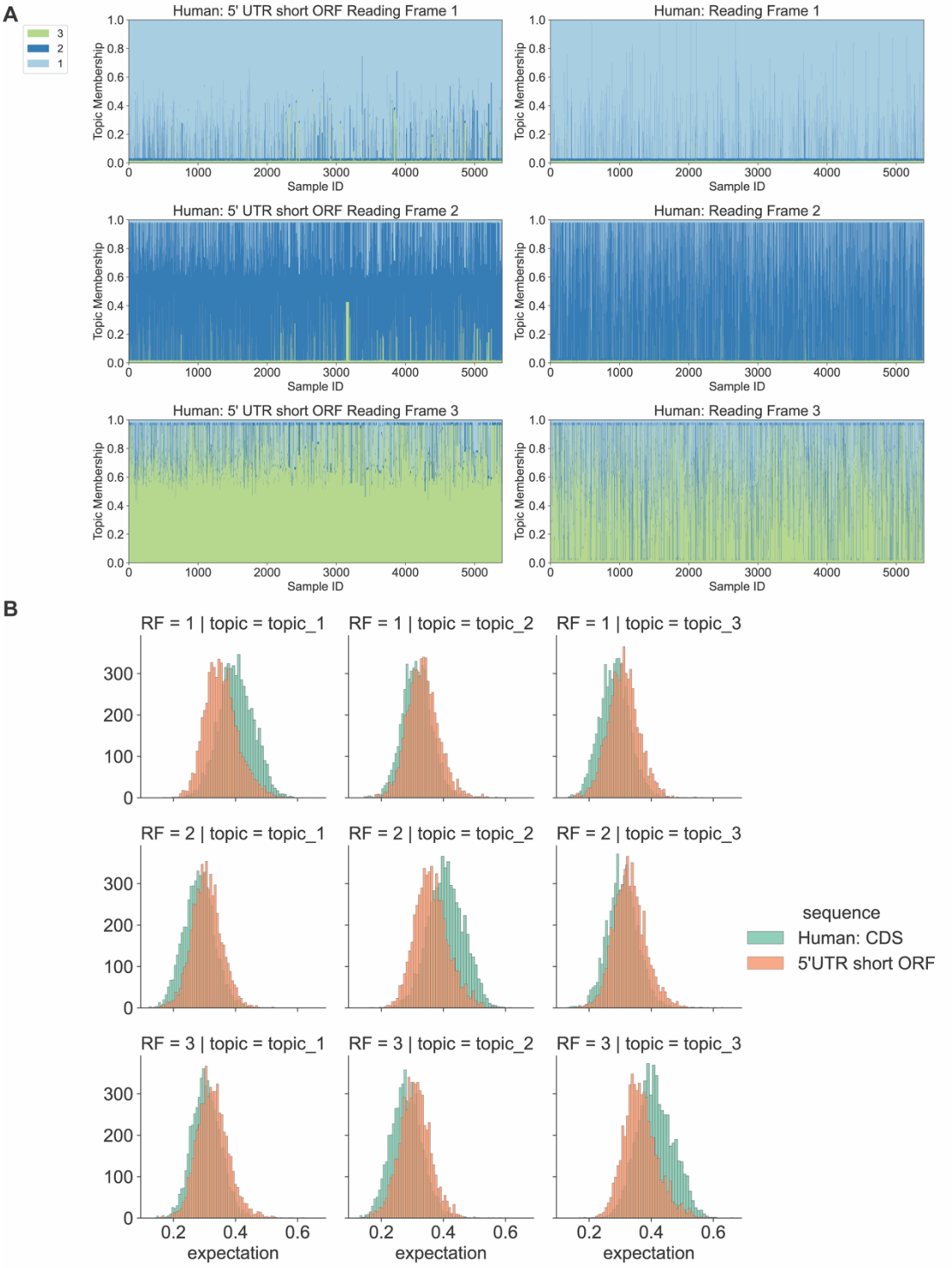


Figure 2.6 Identification of potential coding reading frames in human UTRs. The LDA model fitted by annotated CDS in Fig 4A was applied to small reading frames from human UTRs, each of which had a pair of start and stop codons and at least 25 potential codons. (A) Structure plots of single sequences from human small reading frames in UTR5 and reading frames from CDS. Data were transformed by the model fitted in Fig 4A. 5,400 of UTR5 sequences that follow the CDS topic distribution patterns were selected to show. (B) Distribution of the probability expectations of generating small reading frame samples from UTR5 and reading frames from human CDS using topics fitted in Fig 4A.

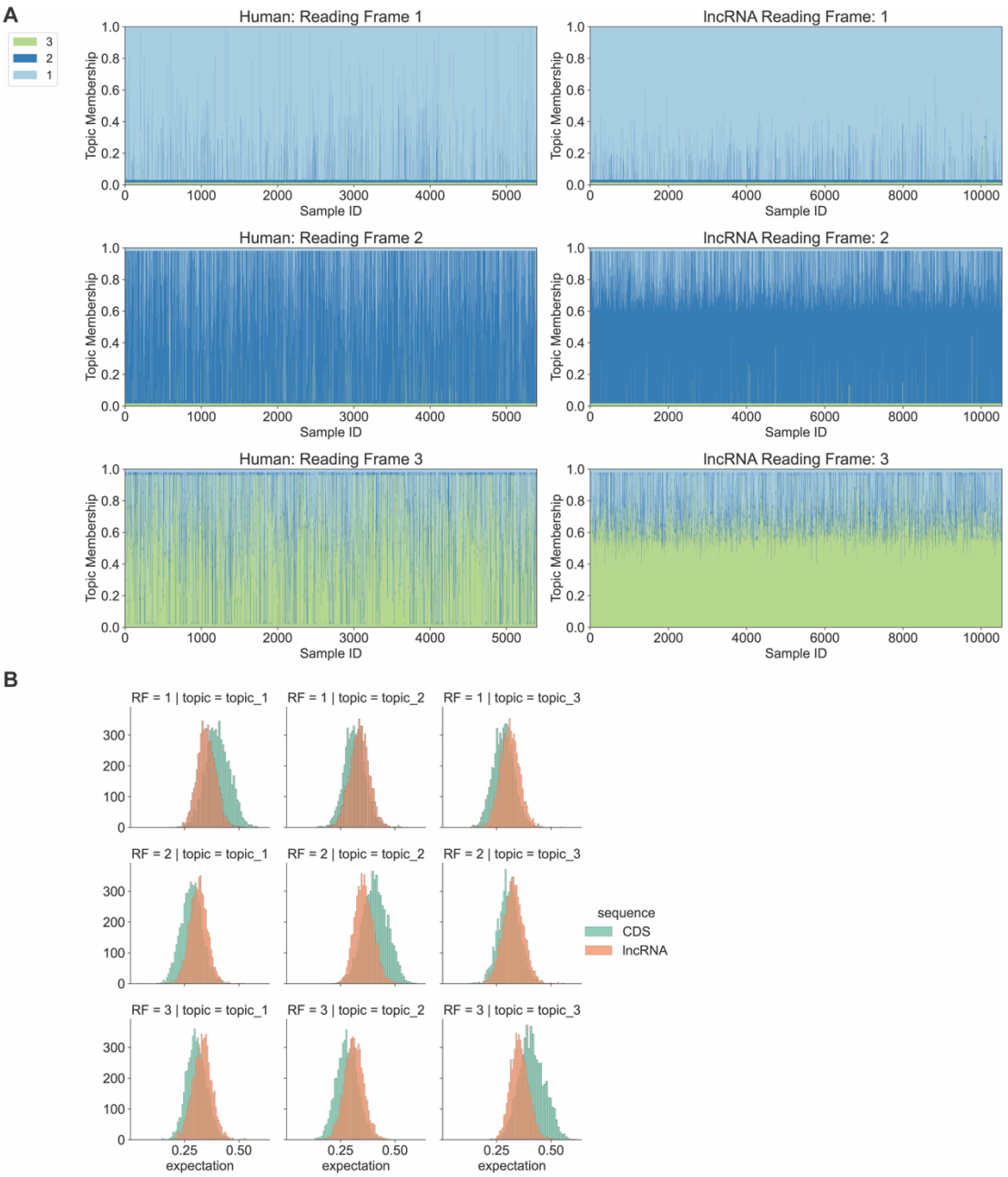


Figure 2.7 Figure 7. Identification of potential coding reading frames in human lncRNAs. The LDA model fitted by annotated CDS in Fig 4A was applied to small reading frames from human lncRNAs, each of which had a pair of start and stop codons and at least 25 potential codons. (A) Structure plots of single sequences from human small reading frames in lncRNA and reading frames from CDS. Data were transformed by the model fitted in Fig 4A. 5,400 of lncRNA sequences that follow the CDS topic distribution patterns were selected to show. (B) Distribution of the probability expectations of generating small reading frame samples from lncRNAs and reading frames from human CDS using topics fitted in Fig 4A.

Discussion

Here I have shown the utility of mixture models using LDA for visualizing and interpreting signals in nucleotide sequences. LDA readily identifies topics that correspond to sequence subtypes and reports the proportional membership of those topics in each sequence. Because topics are interpretable, it is possible to easily identify functional motifs that function in biological processes. Examples here include the branch site consensus (topic 2 in Figure 2.2C: RCTRAC), the 5' splice site (topic 3 in Figure 2.2C: AGGTA), and an unknown G-rich motif characteristic of short human introns (topic 6 in Figure 2.2C). In addition, LDA can uncover subtypes of sequences based on differences in topic distribution. Examples include subtypes of intron correlated with length, reading frame and species. Finally, the expectation that a sequence is derived from a topic can be used as a score to identify sequences with potential functions, such as potentially coding small ORFs in 5' UTRs.

Motif discovery

Because topics are interpretable, it is possible to identify motifs using LDA. A motif can be readily interpreted by viewing the enriched or driving k-mers from the enriched topics. While the basic probabilities of the observations for each topic could be used, top k-mers based on raw probabilities can be shared between topics, and alternative methods to identify the driving features are preferable. I identified distinctively distributed ("driving") k-mers using K-L divergence (following Dey et al., 2017). It is the application of this method (Fig. 2.2C) that yielded driving k-mers drawn from known consensus sequences for the branch site, 5' splice site and pyrimidine tract, as well as an unfamiliar G-rich motif characteristic of short introns.

Driving features that are k-mers with sequence overlap can be used to identify motifs that can be represented by sequence logos.

Sequence subtypes

LDA identifies sequence subtypes directly from topic membership, whether based on bulk k-mer composition or positional k-mer counts. However, subtype recognition can be improved in several ways. In this project, I have used sequences aligned at a splice site, or by reading frame. This creates a contrast in k-mer frequencies between samples based on that alignment. It is necessary to use positional k-mers to identify subclasses of sequence that have not been identified in advance. Sequences of variable length that are not readily aligned may be more amenable to bulk k-mer compositions.

Pooling sequences prior to model fitting is useful to overcome the problem that single sequences have 0 occurrences of many features, which is especially true for positional k-mer values. Thus, I improved model performances by randomly grouping sequences from the same known group as one single sample. This approach increased the distinction between known sequence groups. Single sequences transformed by the model fitted by grouped sequences have much higher variance in the topic distributions. The results revealed that grouping sequences can magnify the impact of enriched k-mers and push the model to generate topics to describe the generative process of two groups instead of single sequences. However, the enriched features in a group may not occur in every single sequence. One potential flaw of this method is that sequences with abundant repetitive patterns can introduce bias, potentially reducing the signal of true subtle features.

Overview, caveats, and prospects

Here I have used LDA and k-mer features to classify and characterize nucleotide sequences. One obvious extension of this work would be the application of LDA to protein sequences. Another would be to include features, such as accessibility in chromatin, that are not properties of the sequence *per se*.

As a generative statistical model, LDA derives topics that describe the data used to fit the model. These topics can then be used to score novel sequences using the expectation or likelihood that those novel sequences were generated from the topic. In this way, novel sequences that potentially share biological functions with the sequences in the fitted set sharing those topics can be identified. However, topics that are not observed by the model cannot be predicted.

Overall, LDA provides a solution that will be broadly useful for describing the properties of sequences in a functional class.

Methods

Overview of sequence feature analysis with LDA

For each analysis, matrices of k-mer feature counts were generated for each sample (see below). The latent Dirichlet allocation (LDA) algorithm was implemented using the sci-kit learn (Pedregosa et al., 2011) package. Driving k-mers from each topic are pulled by the calculation of distinctiveness through the Kullback-Leibler divergence (Dey et al., 2017). Structureplot visualization (Pritchard et al., 2000) is implemented on Matplotlib package (Hunter, 2007). Sequence logos were generated with the

Logomaker package (Tareen & Kinney, 2020). Matrix computations are performed with Numpy (Harris et al., 2020) and Pandas (McKinney, 2010).

Analysis using positions as samples

To evaluate intron sequence features associated with positions, I randomly selected human (hg19) and Drosophila (dm6) introns and aligned them at the annotated 3'SS (O'Leary et al., 2016). Introns were labeled as long or short if they were greater or less than 150 bp (human) or 80 bp (Drosophila). Feature k-mer length, window size, pack size (the number of individual sequences per sample) and the number of topics are all options. Here, we counted 6-mers starting at every position from 100bp to 7bp upstream of 3'SS, so that the region from -100 to -2 was represented. A sliding window of length 5 was applied to batch the counts at adjacent positions together. LDA was applied to fit 6 topics.

Branch site and polypyrimidine tract signal evaluations

I calculated the signal of the branch site by counting the fractions of k-mers that match the human (TNAC) and Drosophila (TAAC) branch site consensus. The polypyrimidine tract signals were evaluated by counting the fractions of k-mers with only Cs and Ts.

Positional k-mer counts as features

Positional features incorporate both sequence and position in the alignment. For example, the sequence AGTTAT, starting at position 1, would have AGTT_1, GTTA_2 and TTAT_3 as features. For both human (hg19) and Drosophila (dm6) splice sites from long and short introns, we randomly selected 16,000 sequences

comprising 30 exon and 50 intron nucleotides. 15,000 sequences were used for model fitting, and the remaining 1,000 sequences for model testing. Sequence feature analysis for 3'SS and 5'SS were performed separately for each species.

Model fitting by pooling sequences

Larger samples were obtained by pooling multiple sequences of the same type. For example, in the analysis of long and short introns, we used 60 samples from 15,000 introns (250 sequences per sample) to fit the model with 2 topics.

Scoring a sequence's fit to a topic

The intron subtype analysis was applied with 2 topic-LDA model. I hypothesized that the 2 topics generated are associated with long and short introns. By calculating the likelihood of generating each sequence from a topic, I could to evaluate the information obtained from the sequences.

$$\mathcal{L}(seq|topic) = \prod_{i=1}^{length(seq)-k} \mathrm{P}(seq[i:i+k]|topic)$$

K-mer composition as features and application to coding sequences

To evaluate the performance of LDA in differentiating sequence subtypes, I used protein-coding sequences (CDS) of human (hg19, CCDs) and Drosophila (dm6). For each sequence, I counted the occurrences of 6-mers in 3 reading frames. I used 60 samples from 15,000 CDS (300 sequences per sample) to fit the model with 6 topics. To evaluate the ability of this model to identify the correct reading frame, I randomly picked 5,400 sequences that were not used in the model fitting step to test the

performance on single sequences. Topics generated from the model fitting step were used to transform single sequence k-mer counts into topic memberships.

Sequence subtype prediction

I implemented a simple classifier for predicting sequence labels in two steps, associating topics with labels and calculating sequence scores for each label. To associate topics with labels, I built a scoring system. For a given label, the score from a topic is the log2 of the fraction between the average membership in samples from such label and the overall average membership.

$Score(topic|labels) = \log2(\frac{\frac{\sum_{i=1}^{N} membership(seq_i, topic)}{N}}{\frac{\sum_{j=1}^{M} membership(seq_j, topic)}{M}})$, where N is the number of sequences of the label of interest and M is the total number of sequences in the analysis.

When calculating the result of a sequence, I multiplied the sample-topic matrix with the topic-score matrix and picked the highest score as the prediction.

Frame-shift analysis

From the single sequences in the test set, I used the classifier to predict the reading frames based on each sequence's k-mer counts from 3 frames. If the predicted reading frames from the original counts of reading frames 1, 2, 3 were 2, 3, 1 or 3, 1, 2, the sequence would be labeled as a potential frame-shift event.

Prediction of functional smORFs in human UTRs and lncRNAs

I scanned 5' UTR and 3' UTR sequences for a pair of start codon and stop codon in the same frame with at least 25 potential codons between them. The subsequence of

the UTR from the start codon to the stop codon was identified as a small ORF reading frame 1 sequence, which was utilized to generate k-mer counts of 3 small reading frames. I predicted the reading frames of the UTR sequences and picked sequences with the correct prediction output.

Reading frame predictions using subsequences

Subsequences from the human CDS test set (n = 5,400) were extracted by fixing the center of each sequence and expanding upstream and downstream equally to obtain coding subsequences centered on the middle of the original ORF. These subsequences were then transformed into topic memberships using the LDA model fitted with the full-length CDS test set.

Chapter 3: Recursive Splicing Sequence Feature Analysis

Introduction

One of the essential programs altering gene expression regulation is RNA splicing, which removes introns from newly-made RNA and ligate exons to produce mature mRNA (G.-S. Wang & Cooper, 2007). The splicing process requires the spliceosome assembly upon recognition of various splicing signals and is regulated by various RNA-protein and protein-protein interactions (Gehring & Roignant, 2021). Though the primary splicing mechanism suggests that the spliceosome removes one intron sequence as a whole piece, a special multi-step splicing mechanism called recursive splicing (RS) has been observed in Drosophila and other vertebrate genes. Initially thought to be a rare event and restricted to large introns (Burnette et al., 2005; Duff et al., 2015b; M. Guo et al., 1993; Sibley et al., 2015), RS has recently been observed to occur pervasively in human transcripts (Wan et al., 2021) and may provide a mechanism which solves the problem of forming lariat structures with long sequences. In very large intron, the cryptic splicing sites can be used as ratchet points to let the spliceosome remove a sub-intron sequence, leaving constitutive splicing sites for use in the next step. Although cryptic splicing sequences are abundant in both Drosophila and human genomes the extent to which they are used in RS events is an area of active debate (Gehring & Roignant, 2021; Sibley et al., 2015)[include Wan REF]. Moreover, the cis-acting and trans-acting factors that decide whether an intron should be recursively spliced still remain to be discovered.

In addition, the usage of cryptic splicing sites in the human genome can be suppressed by exon junction complex (EJC), a protein complex that forms at about 20-24 nt of the 5' end of the splicing junction and interacts with the spliceosome to prevent the usage of adjacent splicing sites (Blazquez et al., 2018; Boehm et al., 2018). The reconstituted splicing sites from Drosophila are less sensitive to EJC, which could explain the RS observation in Drosophila (Blazquez et al., 2018). Recently, single-molecule imaging and genome-wide assays, including nascent RNA sequencing and lariat sequencing, have provided new insight into RS and raised the possibility of more in-depth analysis of sequence-encoded splicing signals (Wan et al., 2021; Zeng et al., 2022). For example, imaging of single large introns signaled that co-transcriptional splicing may happen before Pol II reaches the annotated 3'SS. A closer investigation of single introns with double labeling revealed that different RS sites could be selected for splicing reactions for the different molecules of the same intron sequence, suggesting a stochastic splicing model. In addition, nascent RNA-Seq revealed abundant unannotated splicing junctions in the intron region, most of which have sequence features that can be potentially used as splicing sites and pyrimidine tracts. Indeed, sequencing of the splicing lariats showed that a significant number of lariats are smaller than the annotated introns, indicating the usage of unannotated splicing sites. Together, the nascent RNA study suggests that RS is widely used in human cells (Wan et al., 2021).

Here we investigate the sequence features associated with RS events. By characterizing k-mer compositions of sequences with topic modeling, we showed that the most significant difference between RS and non-RS introns lies in the upstream

exon. The exons upstream of the RS introns are more enriched for CG dinucleotides. Further, compared with non-RS introns' upstream exons, CpG from exons upstream of RS introns are less methylated, suggesting a correlation between DNA methylation and RS regulation. Similar CG-rich features are also found in the short region downstream of the canonical 5'SS from the RS-intron, while both the canonical 5'SS of non-RS intron and the recursive 5'SS do not show the motif, implying a hypothesis that the similarity between recursive 5'SS and regular 5'SS promotes RS. Using both in vitro and in vivo binding measurements (RNA bind and seq and PAR-CLIP, respectively) we find that the recursive sites are bound by U2AF more tightly than the canonical 3'SS of the same introns. Finally, by analyzing public data of RBP in vitro binding assays, we made a list of candidate proteins that may recognize the sequence features in recursive splicing sequences and regulate the recursive splicing processes. Taken together, we identify a sequence signature of recursive splicing which suggests a potential biochemical mechanism for future experimental study.

Results

Four types of introns were discovered from nascent RNA-Seq.

Nascent RNA-Seq showed that all samples expressing nascent RNA have at least 30% splicing junctions marked as unannotated by STAR while regular RNA-Seq of the same cell-line showed less than 20% unannotated splicing junctions (Figure 3.1 A). The unannotated splicing junctions have similar lengths as annotated introns in nascent RNA-Seq, both of which are much smaller than the original introns to which unannotated junctions are mapped. (Figure 3.1 B). This observation is consistent with

previous findings that recursive splicing usually happens in large introns (Pai et al., 2018; Sibley et al., 2015).

We confirmed that over 280,000 unannotated junctions are parts of annotated introns. Combined with the observation of Par-CLIP U2AF binding signals in the intron region with sequences feature of potential 3' splice sites, we hypothesized that the spliceosome progressively removes the intron in small pieces through recursive splicing instead of splicing a whole intron at once (Figure 3.1 C).

Besides basic introns that overlap with annotated intron coordinates, we recognized three types of potentially recursively spliced introns (Figure 3.1 D): introns with unannotated recursive 5'SS and annotated 3'SS (RS5 introns), introns with unannotated recursive 3'SS and annotated 5'SS (RS3 introns), and introns with pairs of unannotated recursive 5' and 3'SS (nested introns). Potential recursive splicing junctions of at least 75bp long from annotated introns of at least 150bp indicated that recursive splicing removes, on average, 43.63% of the lengths of original intron sequences, with a standard deviation of 31.68% (Figure 3.1 E). More than 50% of RS5 introns have AG upstream of the recursive 5'SS, showing the signal of potential 3'SS for the upstream intron regions. However, most RS3 and nested introns do not have the GU downstream of the recursive 3’SS, indicating the following recursive splicing is conducted without a 5’SS signal (Figure 3.1 F). We annotated the ranks of introns in the nascent RNA-Seq as the order of introns from transcripts and observed a preference for the first introns by recursive splicing, while a much smaller fraction of introns without recursive splicing evidence are from the first introns in various genes (Figure 3.1 G).

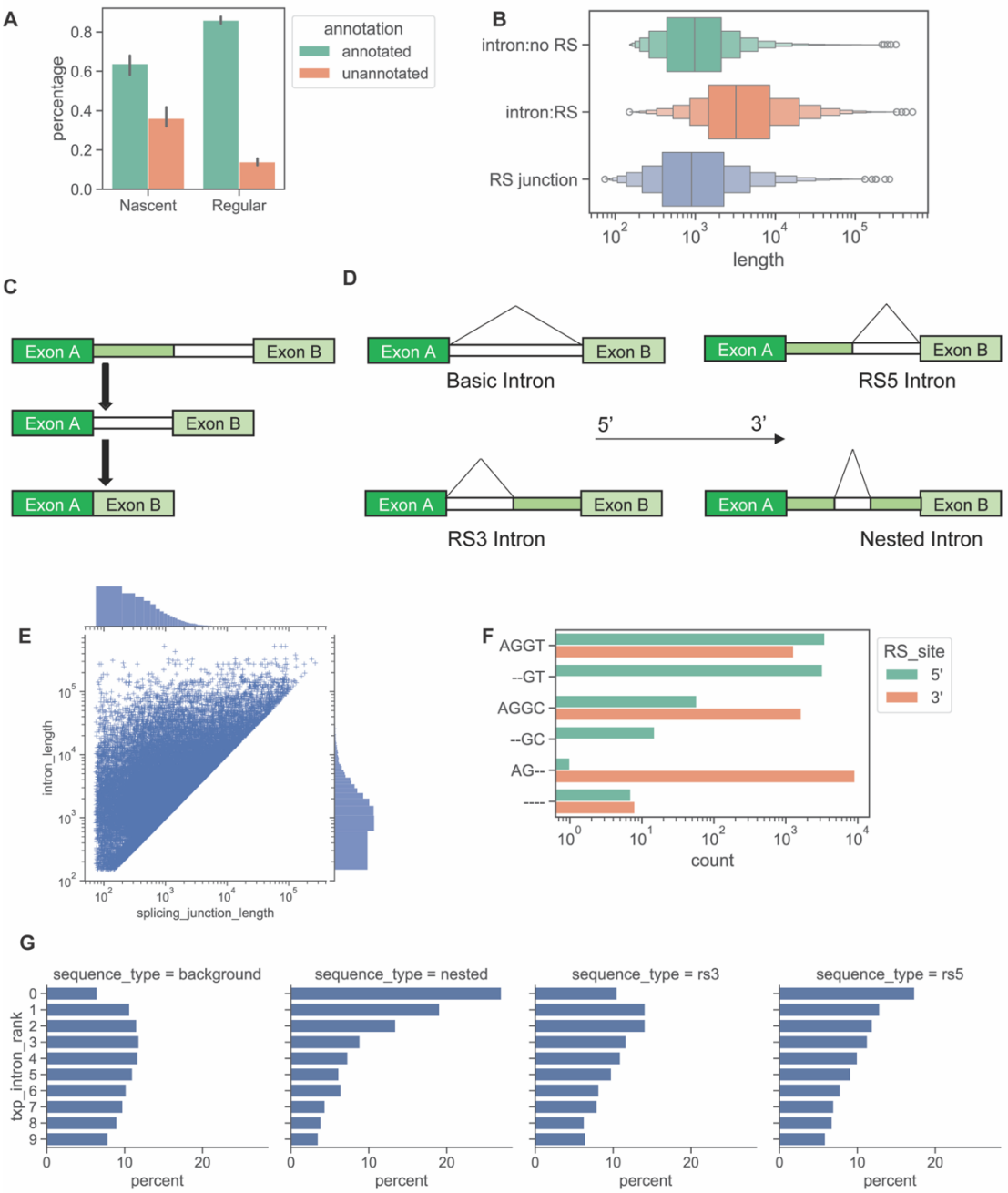


Figure 3.1 Nascent RNA-Seq revealed four types of introns. (A) The percentages of annotated and unannotated splicing junctions from three nascent RNA-Seq samples and two regular RNA-Seq samples. Both RNA-Seq experiments are based on HBEC cells. (B) The distribution of the sizes of annotated introns, potential recursive splicing junctions, and the original introns to which potential recursive splicing belongs. Both annotated and potential recursive splicing junctions are from nascent RNA-Seq data. (C) The recursive splicing model. The spliceosome utilizes a rachet point as a recursive 3'SS to pair with the annotated 5'SS to carry out the splicing of the first half of the intron. (D) Four types of introns are observed from the nascent RNA-Seq. (E) The scatter plot of the sizes of splicing junctions from nascent RNA-Seq and the sizes of annotated introns to which they belong. (F)The counts of different types of splicing sites used by potential recursive splicing events. The y-axis shows the tetramers surrounding the splicing site. (G) The distribution of the top 10 ranks of different types of introns. Introns are ranked based on the order in annotated transcripts. The first intron of a transcript is ranked 0 and the second intron of a transcript is ranked 1, etc..Exon sequences upstream of recursively-spliced introns are enriched in CG nucleotides.

Exon sequences upstream of recursively-spliced first-order introns are enriched in CG nucleotides.

To investigate the sequence features of recursive splicing events, we selected intron sequences that are the first introns of genes, are abundant in nascent RNA-seq, and met a list of filtering criteria to avoid false positive recursive splicing events. We first evaluated the differential tetramer compositions between the first introns with and without recursive splicing. Though the calculation at the first 50bp downstream of the canonical 5'SS suggested that recursively spliced introns have more CG-tetramers, the meta-gene plot indicated that it may be caused by the higher GC content in recursively spliced introns (Fig3.2 A-B). However, the tetramer evaluation in the upstream first exon region also showed higher fractions of CG-tetramers in exons upstream of recursively spliced introns while the GC content of two exon groups are on the same level (Fig3.2 B-C). To investigate motifs in the upstream exons, we prepared hexamer compositions of exon sequences of four types of introns by randomly grouping 10 sequences as an exon hexamer-composition sample (Fig3.2 D). Latent Dirichlet allocation was applied to the sample-hexamer count matrix to explore four possible features to build the sequence descriptive model. The application revealed that though three features exist in four types of introns, one feature, topic/feature 1, has relatively higher memberships in exons upstream of RS introns than non-RS introns (Fig3.2 E). We applied Kullback-Leiber divergence to the topic-tetramer distributions and identified that the distinctively distributed tetramers in this topic/feature 1 are enriched in CG-rich dinucleotides (Fig3.2 F). The topic/feature identified by LDA was a list of probabilities of observing k-mers from

such topic/feature. Thus, we calculated statistical inferences by revealing the probabilities of observing the original exon sequences by the different topics. The distribution of the expectation values of single sequences also indicated that the non-RS introns have lower expectations from topic/feature 1 than the other three RS introns. In summary, we identify a CG-rich topic in the sequences of upstream of introns which are recursively splice.

To investigate if the CG-rich feature has the potential to predict the recursive splicing events, we prepared separate model fitting and testing sequence groups. We used random forest classifiers to assess the performance of predicting recursive splicing using expectations calculated from 4 topics (Fig 3.2 E-G). The random forest model ranked topic/feature 1 as the most important feature and provided a training AUC of 69% and an overfitting testing AUC of 59% (Fig 3.2 H-I).

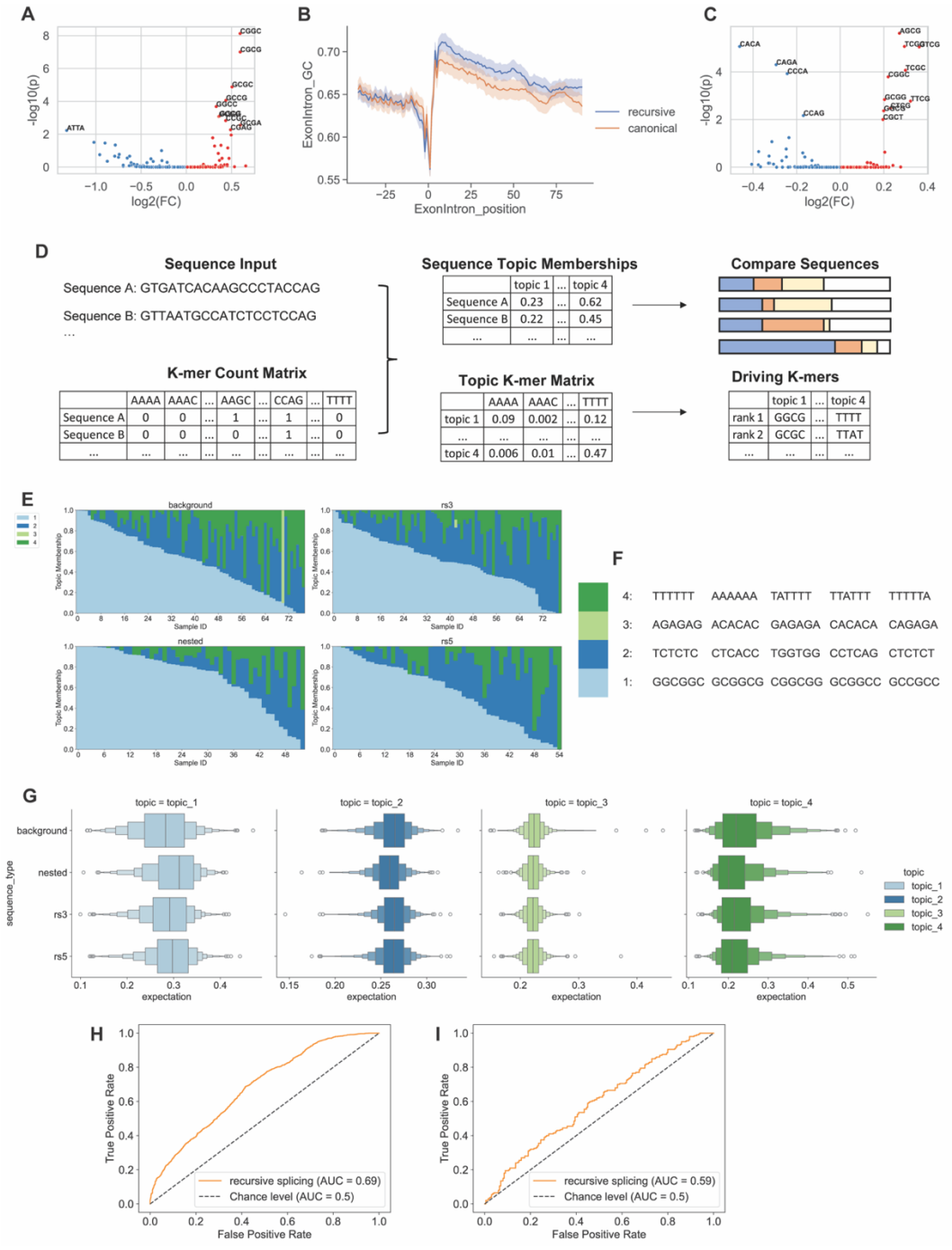


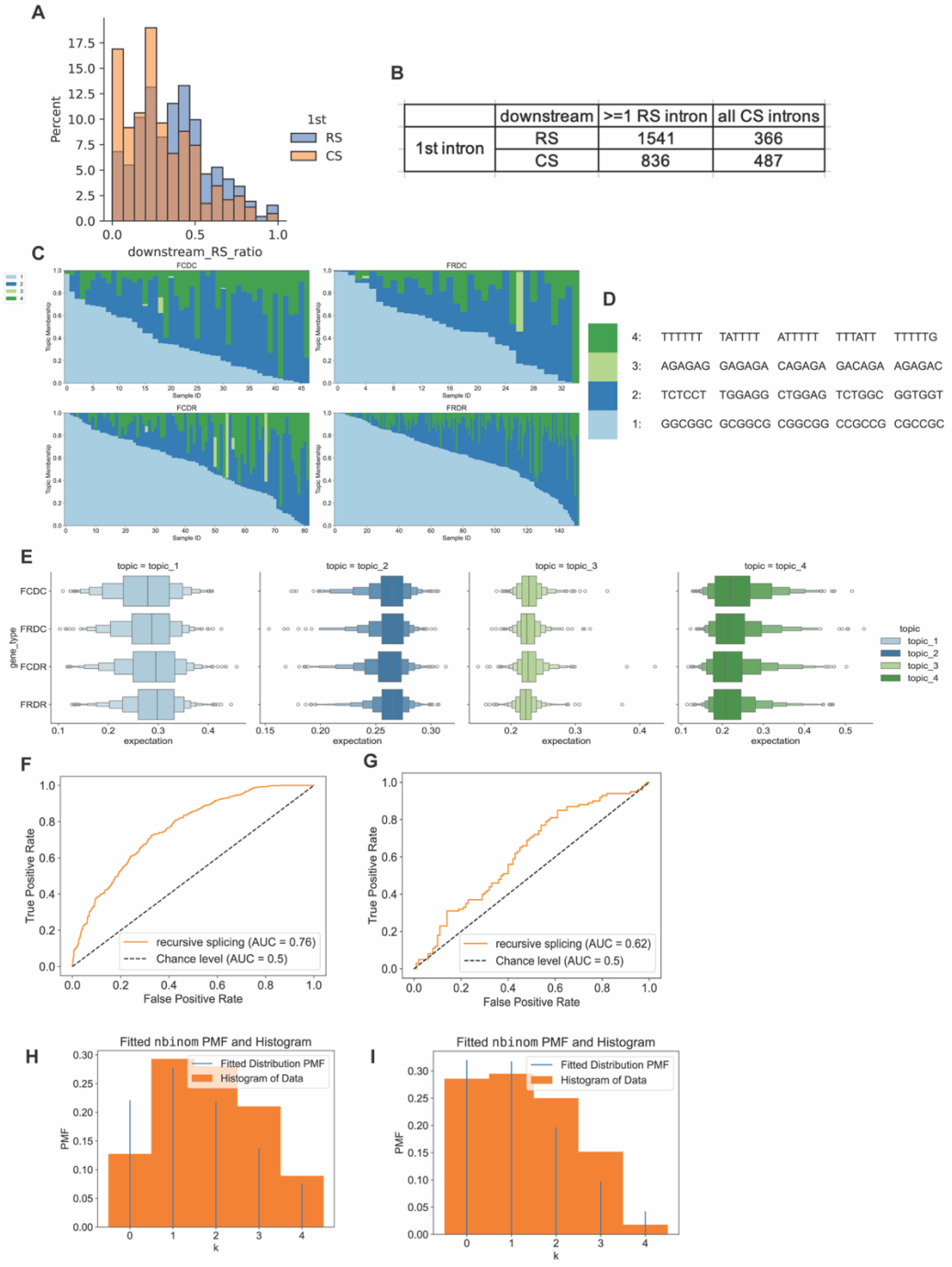
Figure 3.2 Sequence feature analysis of the first exons upstream of 4 types of first introns. (A) The volcano plot of tetramer fractions in the first 50bp of introns without recursive splicing (left) and introns with recursive splicing (right). Multi-test correction is applied. All the introns in the analysis are first introns of different genes. (B) The meta-gene plot of GC content of sequences including the 50bp at 3' of first exons and 100bp at the 5' of first introns. (C) The volcano plot off first exon sequences upstream of first intron without recursive splicing (left) and introns with recursive splicing (right). (D) The diagram of topic modeling of sequences of first exons. (E) The structure plot of first exon sequences. Each sample (x-axis) is hexamer composition of randomly selected 10 sequences. (F) The top 5 driving hexamers in (E). (G) The distribution of the expectation of using different topics to create single sequences. (H) ROC curve of using topic distributions and random forest to predict recursive splicing in the first introns. ROC curve is created with a training set. (I) ROC curve generated with a test set.

Transcripts with recursively spliced first introns are more likely to have recursive splicing in downstream introns.

The distribution of fractions of recursive splicing in different transcripts indicated that when the first intron can be recursively spliced, more recursively spliced downstream introns are observed (Fig 3.3 A). Thus, we hypothesized that recursive splicing in the first introns increases the probability of recursive splicing in downstream introns. Two-sided Fisher's exact test revealed a p-value of 2.27×10^{-28} with a prior odds ratio of 2.43, indicating the likelihood of observing RS in downstream introns with RS in the transcript's first intron is 2.43 times the likelihood of observing RS in transcripts with canonically spliced first introns (Fig 3.3 B). To evaluate the sequence features in the first exons that may be associated with recursive splicing in downstream introns, we assigned transcripts into four groups: first intron recursive splicing and at least one downstream intron recursive splicing (FRDR), first intron canonical splicing and at least one downstream intron recursive splicing (FCDR), first intron recursive splicing and all downstream introns canonical splicing (FRDC), first intron canonical splicing and all downstream introns canonical splicing (FCDC). We applied latent Dirichlet allocation to the first exons of four types of transcripts to model hexamer compositions. Topic distribution indicated that compared with transcripts without recursive splicing evidence, the first exons of transcripts with recursive splicing are more enriched in CG-rich feature (Fig 3.3 C-E). To predict downstream recursive splicing based on the first exons of transcripts, we fitted a random forest classifier using the expectations of single exon sequences calculated by four topics/feature and

observed 76% AUC from the training data set and overfitting 62% AUC from the test data set (Fig 3.3 F-G).

To evaluate if the splicing status in the first intron can affect downstream recursive splicing patterns, we assumed that non-first introns from the same transcript share an equal probability of being recursively spliced and fitted binomial distributions with transcript data. We selected all the transcripts with 5 introns expressed in the nascent RNA-Seq. Transcripts with first introns recursively spliced showed p(RS)=0.3137 in downstream introns, while transcripts with first introns canonically spliced have p(RS)=0.2477 (Fig 3.3 H-I).



	downstream	>=1 RS intron	all CS introns
1st intron	RS	1541	366
	CS	836	487

Figure 3.3 Recursive splicing in first introns makes it more likely for downstream introns to be recursively spliced. (A) Distribution of the fraction of recursively spliced non-first introns in transcripts with the first intron being recursively or canonically spliced. (B) The contingency table of first and downstream intron recursive splicing status. (C) Structure plot of first exon sequences from four types of transcripts. (D) The top 5 driving k-mers of (C). (E) Distribution of expectations calculated by four topics for single first exon sequences. (F) ROC of random forest classifier predicting downstream intron recursive splicing (training set). (G) ROC of random forest classifier for the test set. (H) Fitted binomial distribution using transcripts with first introns recursively spliced and 5 introns expressed in nascent RNA-Seq. (I) Fitted binomial distribution using transcripts with first introns canonically spliced and 5 introns expressed in nascent RNA-Seq.

The first exons from transcripts with recursive splicing events have low CpG methylation status.

The enrichment of CG-dinucleotide features from exons upstream of RS introns is one of the essential features of CpG islands. Moreover, several studies have shown that DNA methylation is associated with spliceosome assembly and splicing regulations (Lev Maor et al., 2015b; Yearim et al., 2015; Young et al., 2005). The genome-wide methylation profiling can be achieved by various sequencing methods, including whole genome bisulfite sequencing (WGBS) that identifies single-nucleotide resolution of cytosine methylation (W. Guo et al., 2014). Thus, we evaluated the methylation status of the CpGs in the exon region by calculating the average CpG methylation percentage of exons with WGBS of A549 cell line from ENCODE (The ENCODE Project Consortium, 2012) (Fig3.4 A). We chose A549 cells both because of both the quality of the dataset and the similarity of this human alveolar basal epithelial cell line to our human bronchial epithelial cell line. Compared with basic introns, a higher fraction of the exons upstream of recursively spliced introns have 0 average CpG methylation based on the WGBS data (Fig3.4 B). We also analyzed the CpG methylation by separating first exons based on the transcript types. Compared with transcripts without recursive splicing, the other transcript groups have higher fractions of no CpG methylation. Moreover, transcripts with both the first and downstream introns being recursively spliced have the highest fraction of 0 CpG methylation in the first exons (Fig 3.4 C).

The low profile of CpG island methylation is often observed near the transcription start site (Anastasiadi et al., 2018). To verify if the order of the exons causes the low

methylation status, we annotated the exon ranks based on the transcript references. From all four groups of exons, we observed that the first exons have much lower methylation profiles than others, and the average exon methylation rate increases as the exon rank increases (Fig3.3D). We analyzed the average exon methylation distribution for first and non-first exons separately, and both analyses showed that exons upstream of RS-introns are less methylated than exons of upstream canonical introns (Fig3.3E). Taken together, our analysis suggests that DNA methylation is correlated with splicing outcomes, with exons upstream of RS introns showing lower levels of DNA methylation.

We then added the methylation status in the first exons to our random forest classifiers. The methylation information slightly increased the performance of the recursive splicing predictions on the training set but had no impact on the test set (Fig 3.4 D-G), suggesting more complicated features affecting recursive splicing remain to be discovered.

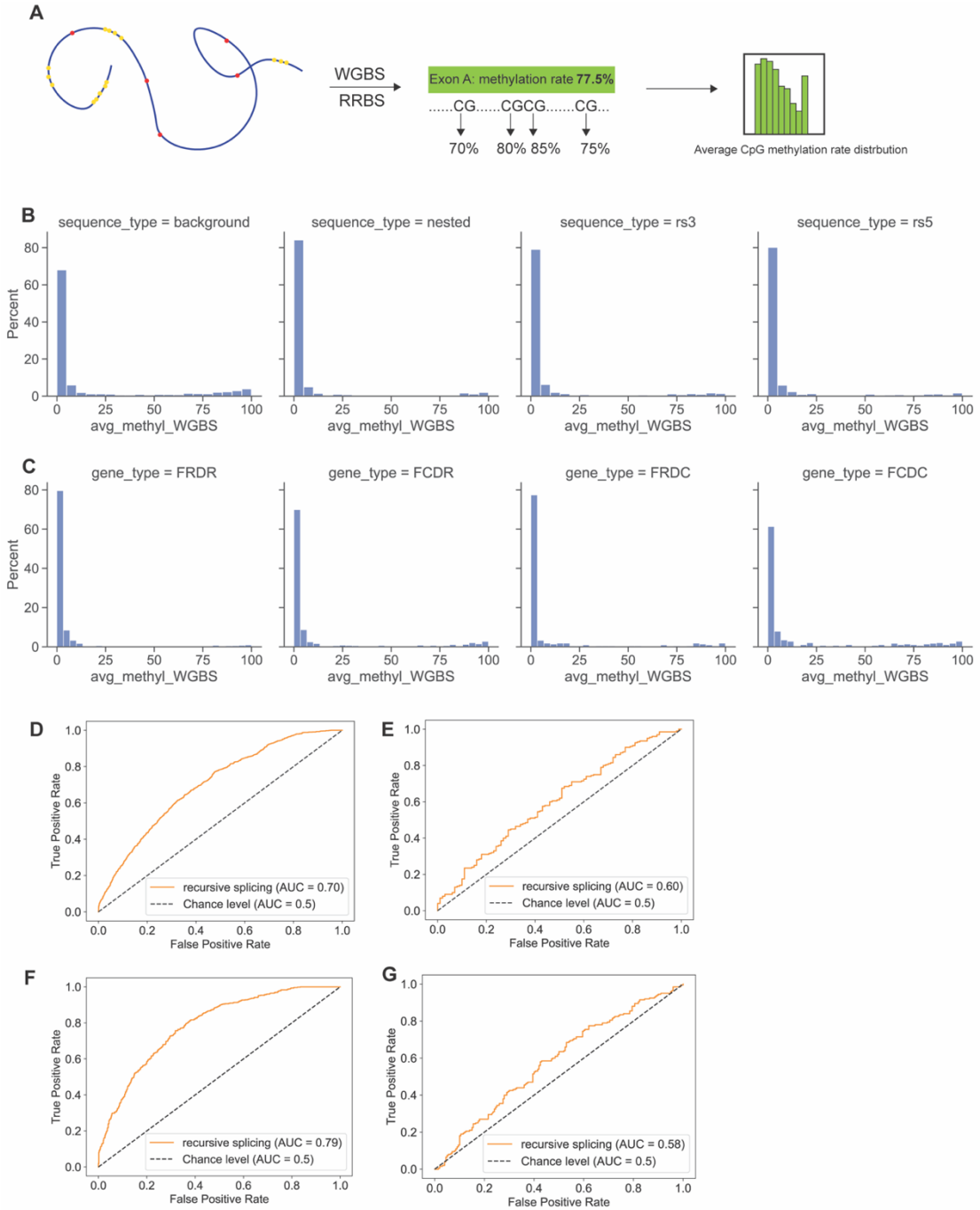


Figure 3.4 DNA methylation status analysis in the first exons. (A) Diagram of exon average CpG methylation calculation. The CpG methylation percentage of every CG in the exon region is obtained from ENCODE. The average CpG methylation percentage is calculated for every exon. (B) Distribution of the average CpG methylation percentage of first exons upstream of different types of first introns. Methylation status is calculated using A549 WGBS data. (C) Distribution of the average CpG methylation percentage of first exons of different types of transcripts. (D) ROC of random forest classifier predicting first intron recursive splicing using the topic distribution in Fig 3.2 and average methylation status (training set). (E) ROC of random forest classifier in (D) with the test set. (F) ROC of random forest classifier predicting downstream intron recursive splicing status using the topic distribution in Fig 3.3 and average methylation status (training set). (G) ROC of random forest classifier in (F) with the test set.

The recursive 3’ splice sites are stronger U2AF binding sites than the canonical 3’SS

Splicing occurs both co- and post-transcriptionally, and transcription elongation rate and pausing has been reported to affect the splice site availability and spliceosome activities (Herzel et al., 2017). This model can potentially explain our observation that there are more RS3 introns than RS5 and nested introns from the nascent RNA-Seq because the spliceosome may assemble at qualified recursive 3'SS before the canonical 3'SS can be transcribed, which is supported by the dwell time analysis of long introns splicing and transcription activities (Wan et al., 2021). This preponderance of RS3 introns prompted us to analyze the sequence feature in the region 50bp – 5bp upstream of recursive and canonical 3'SS (Fig3.5A). We first compared the enrichment of tetramers between recursive and canonical 3'SS. The volcano plot suggested that the flanking regions of recursive 3'SS have more AG- or GA-rich subsequences (Fig3.5B). To evaluate the sequence features of multiple types of 3'SS, we fitted mixture models with recursive 3'SS and canonical 3'SS of RS and non-RS introns (Fig3.5C). Studies have found that long and short introns vary near 3'SS, such as that longer introns tend to have longer polypyrimidine tracts (Yıldırım & Vogl, 2023). Thus, we generated two groups of basic introns: longer introns with sizes between 2,000bp to 11,000bp and shorter introns between 200bp to 2,00bp. We observed different topic distributions between the two groups of basic introns, indicating that shorter introns may have more GCs while longer introns are enriched in Ats (Fig3.5C). The canonical 3'SS from RS introns are similar to longer basic introns than shorter introns, which is expected since RS is believed to be a

mechanism for carrying out large introns. The recursive 3'SS samples have higher memberships of topic 3, which is not observed in canonical 3'SS. Interpretation of topics indicated that AG and GA dinucleotides are enriched in topic 3, consistent with the high fold changes of tetramers with AG or GA between recursive and canonical 3'SS. Positional fractions of AG and GA dinucleotides showed more AG or GA at about 26~37bp upstream of recursive 3'SS than canonical 3'SS (Fig3.5E). To evaluate whether the AG and GA signals are from the branch points, which usually appear at the same position, we calculated the positional fraction of branch site consensus (CTNA). Surprisingly, compared with all kinds of canonical 3'SS, the recursive 3'SS has lower fractions of the branch site consensus across the region of 50 to 5bp upstream of 3'SS, suggesting that branch sites cannot explain the AG and GA features (Fig3.5F).

Sequences near 3'SS are vital for recruiting splicing factors, and U2AF is one of the essential factors and interacts with 3'SS and the polypyrimidine tract (Shao et al., 2014; Wu & Fu, 2015). A thermodynamic U2AF binding model was built using RNA bind-n-seq sequencing result and motif discovery tool proBound (Rube et al., 2022). The primary binding mode of U2AF reflects the polypyrimidine tract feature, assembled by multiple continuous Ts and a few Cs (Fig3.5G). The 3'SS consensus was not observed from the primary binding mode, which could be related to the fact that U2AF is involved in splicing in an AG-independent manner when strong pyrimidine tracts are present (Shao et al., 2014). We calculated the relative binding affinity of U2AF using sequences from upstream 32 to upstream 2bp of all types of 3'SS. Strikingly, the affinity value distribution suggested that recursive 3'SS have

relatively higher U2AF binding affinities (Fig3.5H). Further, we applied kernel density calculations to compare pairs of canonical and recursive 3'SS from RS introns. Both RS3 introns and nested introns contain a substantial amount of examples where recursive 3'SS has stronger U2AF affinity than the canonical site (Fig3.5I).

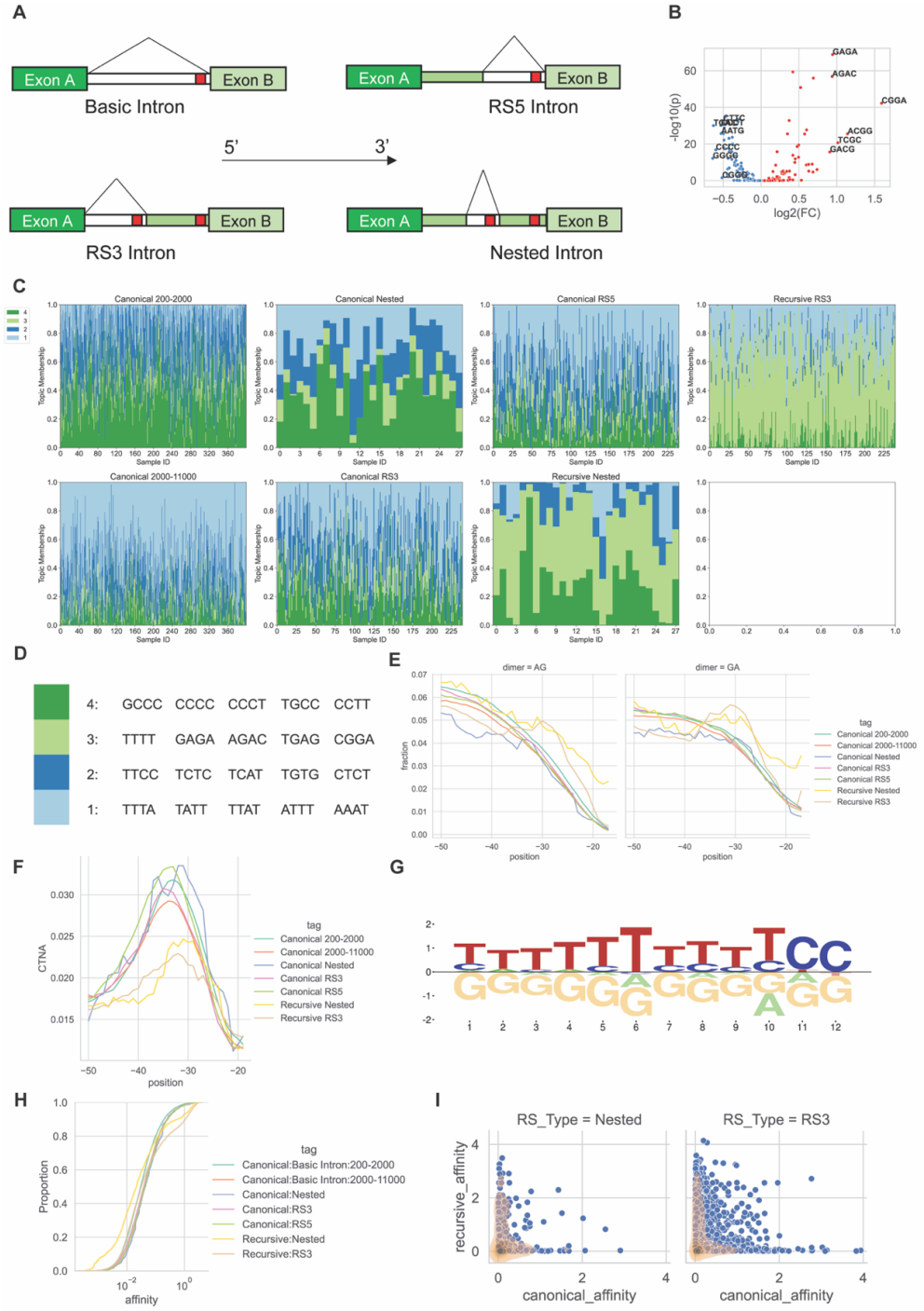


Figure 3.5 Sequence feature and U2AF binding affinities of canonical and recursive 3'SS. (A) Diagram of the sequences selected for 3'SS analysis. The region from upstream 50bp to upstream 5bp of canonical 3'SS of basic introns, canonical 3'SS of RS introns, and recursive 3'SS are obtained. (B) The volcano plot of tetramer expression in RS and non-RS introns. The red dots indicate tetramers expressed more (positive log(FC)) in recursive 3'SS, while blue dots are tetramers with more occurrences in canonical 3'SS. (C) The structure plots of seven groups of 3'SS. LDA was fitted by tetramer occurrences and 4 topics were generated. (D) The driving k-mers of (C). (E) The line plots of positional fractions of AG and GA across sequences. (F) The positional fractions of the branch point consensus CTNA across sequences. (G) The primary U2AF binding mode calculated by proBound using RBNS data. (H) The cumulative distribution of U2AF binding affinities of the region from upstream 32bp to 2bp of the 3'SS. (I) The scatter plots and kernel density plots of U2AF binding affinity pairs from RS introns with recursive 3'SS.

Potential RS regulator proteins are obtained by sequence feature analysis and RBP-binding profiles.

To identify potential trans-acting factors of recursive splicing, we investigated genome-wide RBP-binding profiles from the ENCODE project eCLIP and RBNS assays, as well as U2AF RBNS results from Larson lab (The ENCODE Project Consortium, 2012; Van Nostrand, Freese, et al., 2020; Van Nostrand, Pratt, et al., 2020) . The RBNS provides the enrichment of all possible short k-mers by comparing the observations from the RBP samples of different concentrations and the control sample (Lambert et al., 2014). Since the topics fitted by the mixture models are matrices of probabilities of k-mer observations, we can transform the k-mer topics into RBP topics to represent the enrichment expectations. By calculating the Kullback-Leibler divergence of the new topic matrices, we obtained the 'driving proteins' from the upstream exon, 5'SS, and 3'SS analysis separately (Dey et al., 2017). Here we use 'enrichment prospects' to refer to the sum of the product of k-mer frequencies from sequences and k-mer R-values from RBNS to compare the relative binding strength to a protein between two groups of sequences.

To calculate the enrichment prospects for different types of sequences, we made a new subset of data where each sequence type group has an equal number of samples, and each sequence is 50bp long. Unlike the sequence feature evaluation, the 3'SS and 5'SS were included in the analysis. Consistent with our observation of recursive 3'SS sequence feature modeling, the mixture models of tetramer counts in the region from 50bp upstream of 3'SS to the 3'SS also revealed a topic of AG-rich driving k-mers (Fig3.6A, Fig3.6B). The top 10 driving proteins include RBFOX3, ESRP1, RBM5,

and SFPQ of different concentrations (Fig3.6C). Previous studies have shown that SFPQ regulates the splicing of long introns (Stagsted et al., 2021). We used a new sequence set to evaluate the single-sequence enrichment prospects of the top 5 proteins, which showed that the recursive 3'SS may have tighter binding than the canonical 3'SS (Fig3.6D). We also calculated the enrichment prospects across positions by grouping sequences by intron types and counting the tetramers at each position. The line plots showed that the binding strength of all 5 top proteins at canonical 3'SS sequences drops as approaching the 3'SS and reaches the minimum at 10bp upstream of 3'SS, which is followed by a sharp increase. However, the recursive 3'SS from both nested introns and RS3 introns reach the minimum with relatively higher values at 20bp upstream of 3'SS and then start to increase (Fig3.6E).

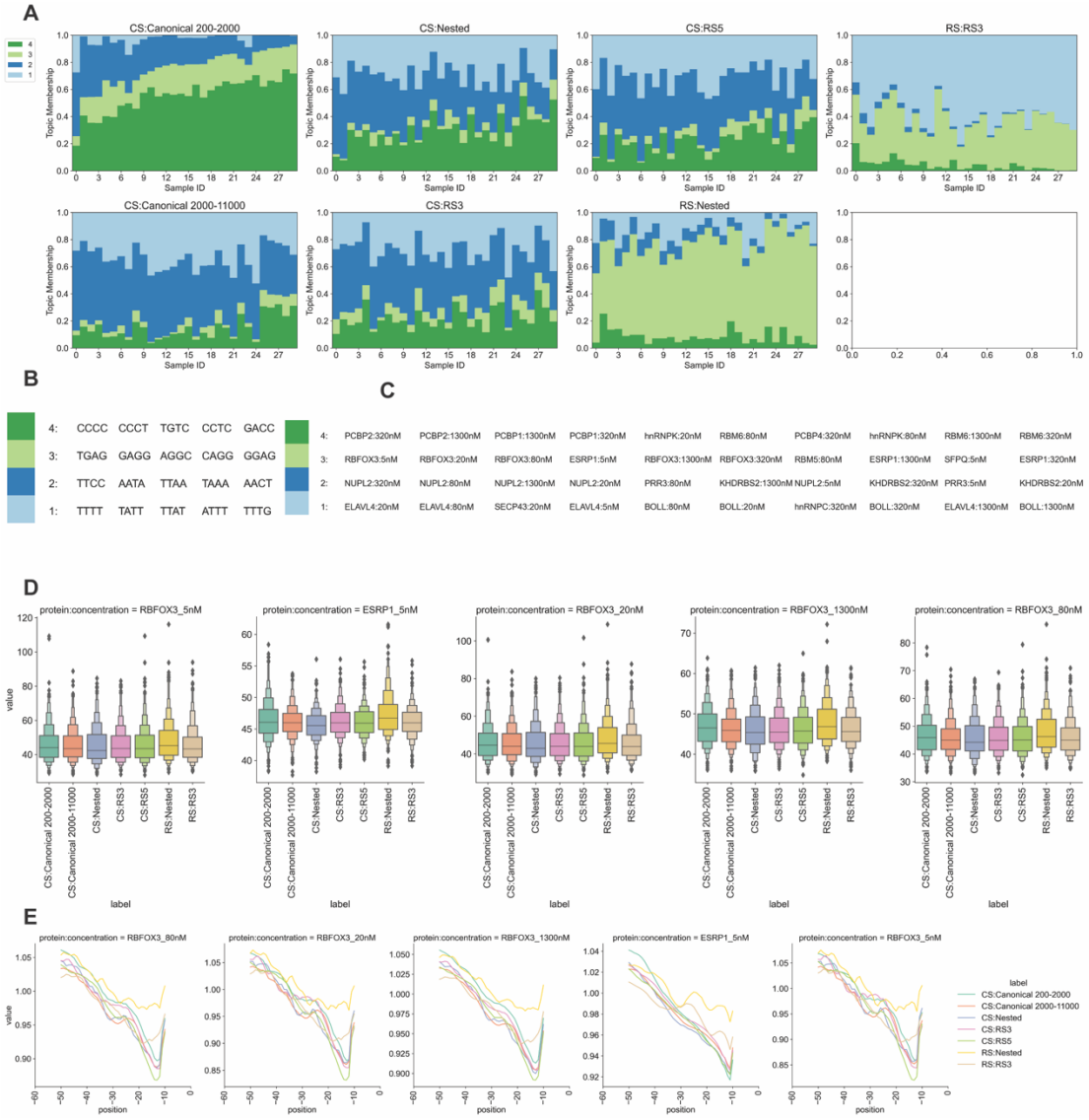


Figure 3.6 Potential recursive 3'SS regulatory RBPs by non-positional 3'SS sequence analysis. (A) Structure plot of the sequences from upstream 50bp to 3'SS. (B) Top 5 driving k-mers in (A). (C) Top 5 driving proteins in (A) and (B). (D) Distribution of the protein enrichment prospect values. Values were calculated for single sequences that were not used in model fitting. (E) Topic 3 protein enrichment prospect across positions in the region between upstream 50 to the 3'SS.

To find other RBPs that have position-specific binding preferences, we applied LDA to model the tetramer counts by positions (Fig3.7A). Kullback-Leibler divergence revealed two topics with different distributions between recursive and canonical 3'SS. One of them is a topic with TTTT or triple Ts with an A, and is represented by driving proteins hnRNPC, hnRNPCL and ELVAL (Fig3.7B, Fig3.7C). Plotting the

binding strengths by enrichment prospect calculation indicated that the recursive 3'SS can be bound tighter than the canonical 3'SS, and binding is stronger when approaching 3'SS (Fig3.7D). On the other hand, the canonical 3'SS was more enriched in a topic of TCs, and the difference grows more significant as approaching the 3'SS. The driving proteins from the topic cover PCBPs and RBM6 (Fig3.7B, Fig3.7C). The two topics also indicate that the pyrimidine tracts of the recursive and canonical 3'SS may differ, which is consistent with our observation that recursive 3'SS have higher U2AF binding affinities (Fig3.7E).

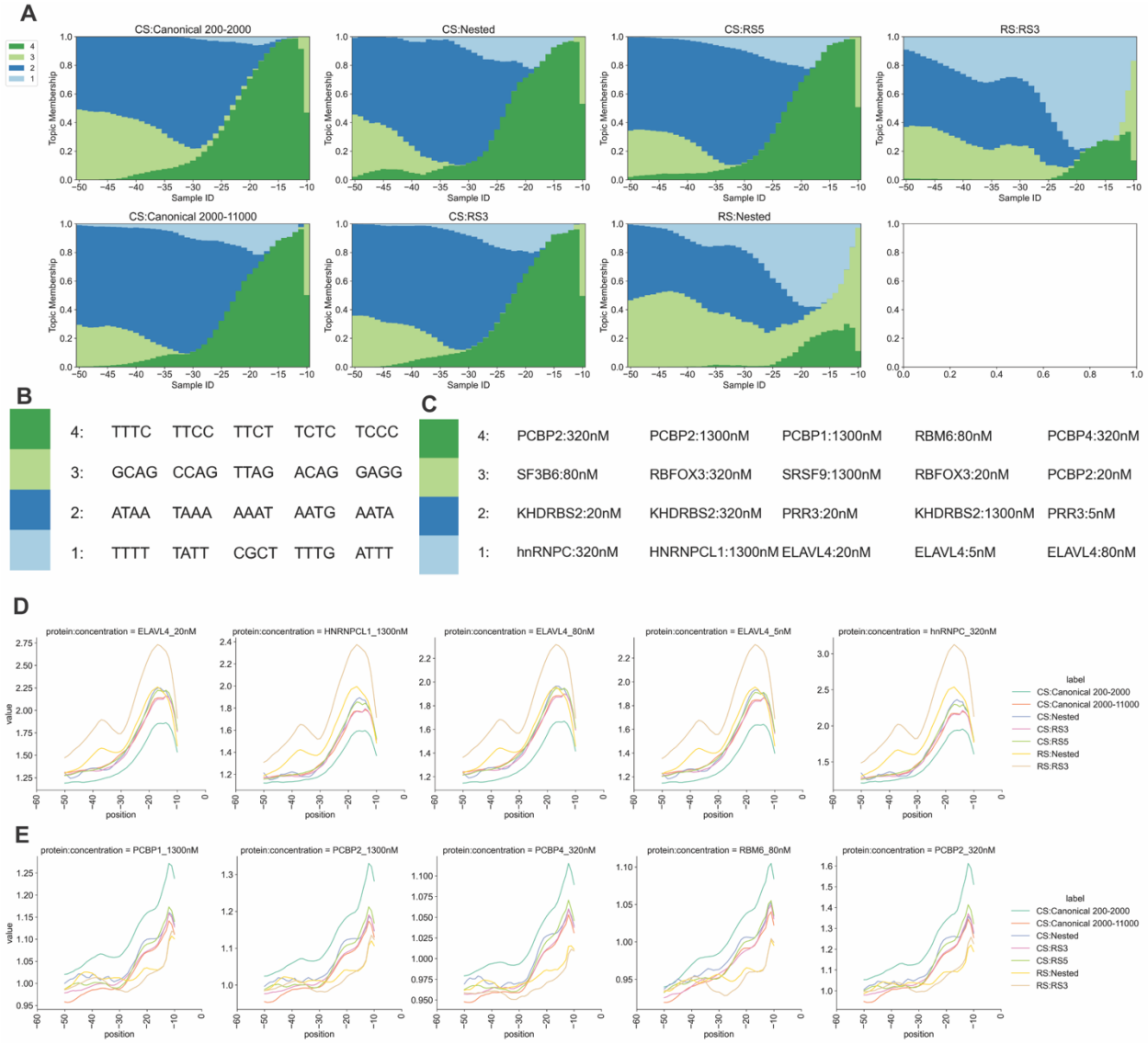


Figure 3.7 Potential recursive 3'SS regulatory RBPs by positional 3'SS sequence analysis. (A) Structure plot of sequences in the region between upstream 50bp to the 3'SS. Each position was modeled by the k-mer counts of all sequences in the group with a sliding window of 4bp. (B) Top 5 driving k-mers of (A). (C) Top 5 driving proteins of (A) and (B). (D) Topic 4 protein enrichment prospect values across positions. (E) Topic 1 protein enrichment prospect values across positions.

We applied the same method to 5'SS and upstream exon sequences. We found potential position-dependent driving proteins in the 5'SS region, such as ELVAL4 and NOVA1, and non-positional driving proteins in the upstream exon region, including RBM4B and ZC3H10 (Fig3.8-Fig3.10). We repeated the pipeline for different subsets of sequences using different k-mer sizes and topic numbers and

summarized 80 candidate proteins for our following experiments to test the function of RBPs in regulating recursive splicing.

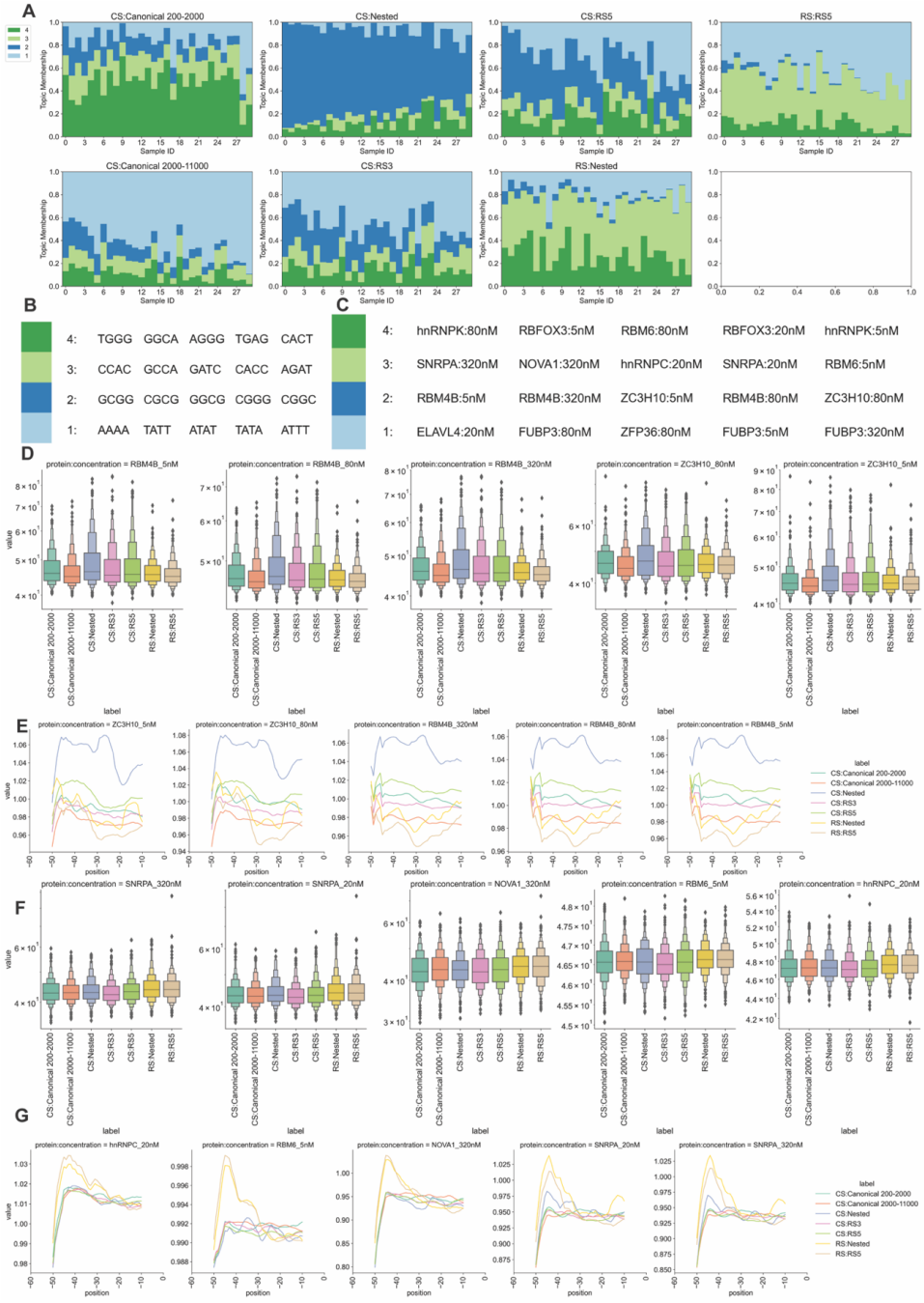


Figure 3.8 Potential recursive 5'SS regulatory RBPs by non-positional 5'SS sequence analysis. (A) Structure plot of sequences in the region between 5'SS and downstream 50bp. (B) Top 5 driving k-mers in (A). (C) Top 5 driving proteins in (A) and (B). (D) Distribution of topic 2 protein enrichment prospects by single sequences. Sequences were not used in the model fitting process. (E) Topic 2 protein enrichment prospect values across positions. (F) Distribution of topic 3 protein enrichment prospects by single sequences. (G) Topic 3 protein enrichment prospect values across positions.

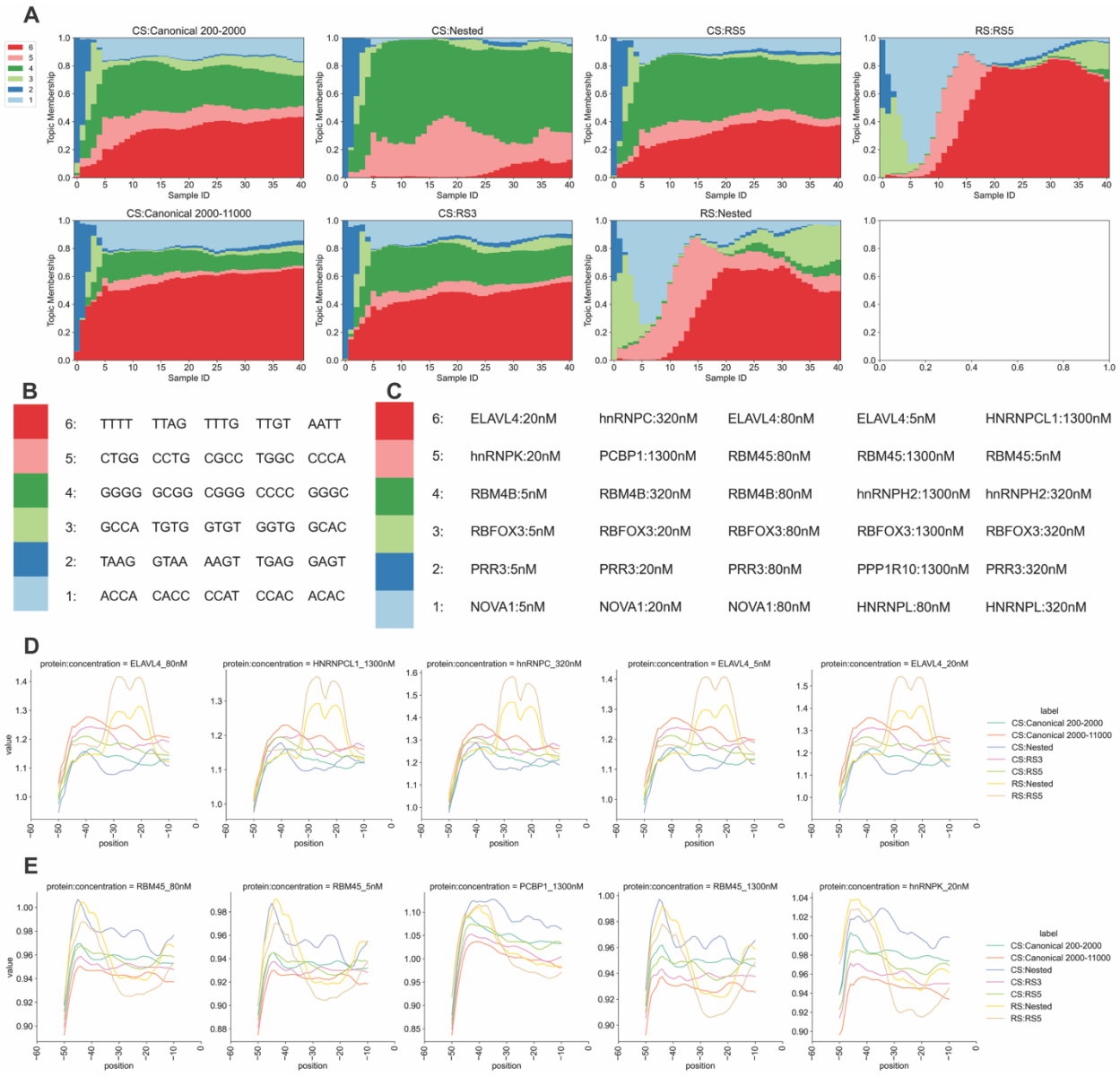


Figure 3.8 Potential recursive 5'SS regulatory RBPs by positional 5'SS sequence analysis. (A) Structure plot of sequences in the region between 5'SS and downstream 50bp. Each position was modeled with k-mer counts from all the sequences at this position in the group with a sliding window of 4bp. (B) Top 5 driving k-mers in (A). (C) Top 5 driving proteins in (A) and (B). (D) Topic 6 protein enrichment prospect values across positions. (E) Topic 5 protein enrichment prospect values across positions.

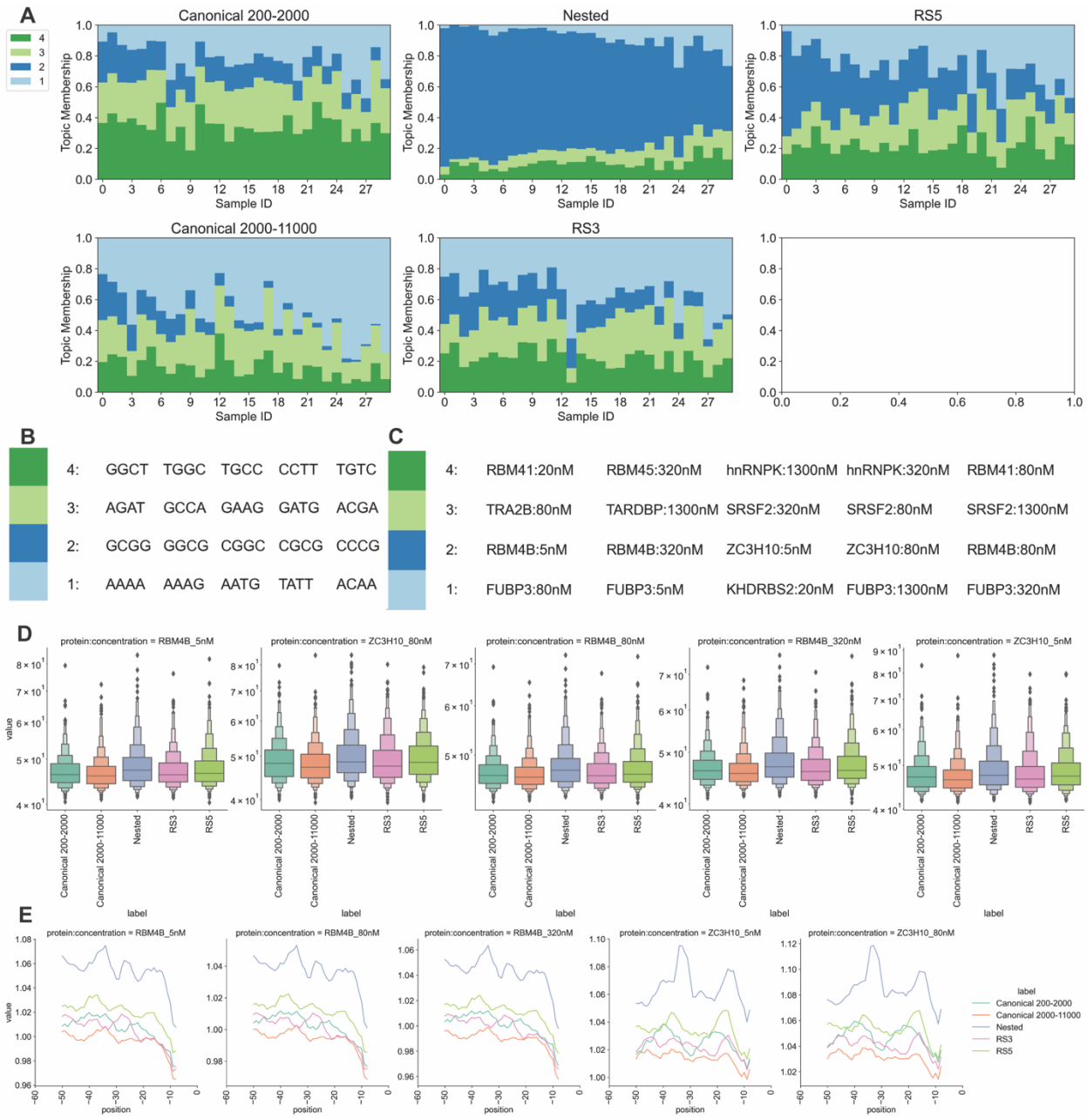


Figure 3.9 Potential recursive splicing regulatory proteins by non-positional upstream exon sequence analysis. (A) Structure plot of sequences in the region of 50bp before the annotated 5'SS. (B) Top 5 driving k-mers in (A). (C) Top 5 driving proteins in (A) and (B). (D) Distribution of topic 2 protein enrichment prospects calculated for single sequences. The sequences were not used in the model fitting process. (E) Topic 2 protein enrichment prospects across positions.

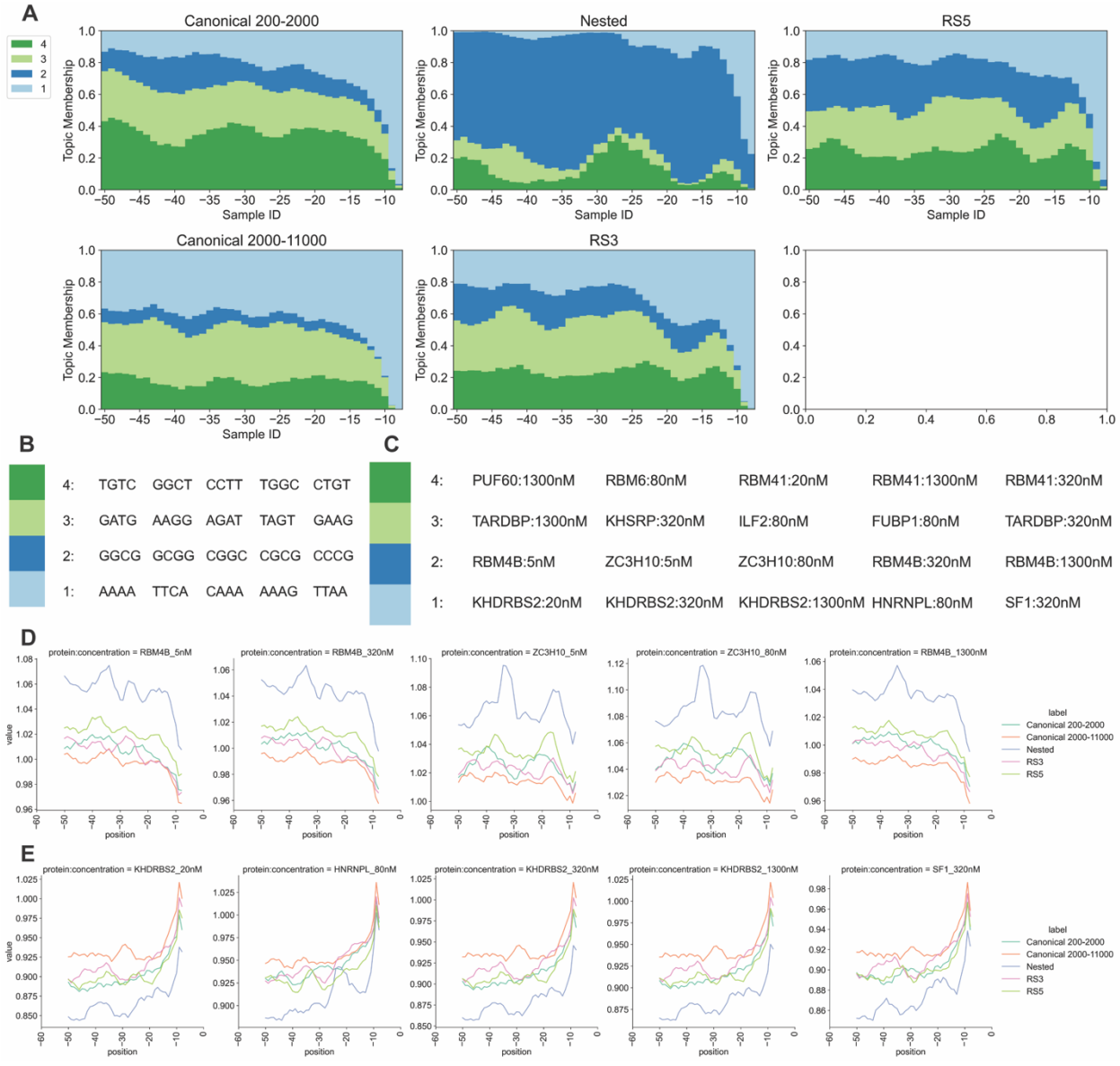


Figure 3.10 Potential recursive splicing regulatory RBPs by positional upstream exon sequence analysis. (A) Structure plot of sequences in the region of 50bp upstream of annotated 5'SS. Each position was modeled by the sum of k-mer counts of all sequences in the group with a sliding window of 4bp. (B) Top 5 driving k-mers in (A). (C) Top 5 driving proteins in (A) and (B). (D) Topic 2 protein enrichment prospect values across positions. (E) Topic 1 protein enrichment prospect values across positions.

Discussion

Identifying cis-regulatory sequences in RNA has long been a challenge. Here, we have used a suite of next-generation approaches aimed at recognizing the sequence features associated with recursive splicing. By modeling tetramer occurrences in the

sequences flanking the recursive and canonical splicing sites, we can compare the sequence features represented by topic distribution by LDA and recognize the potential cis-acting recursive splicing regulators. The experimental input sequences used in our analysis were carefully prepared using various filtering criteria to ensure they were not from sequencing errors, annotated alternative splicing events, or transposable elements. The mixture model fitted with the sequences can reveal sequence features, sequence subgroups, and feature interpretations concurrently. Our findings suggest that first, recursive splicing tends to occur in the first introns. Second, the first exons upstream of RS first introns are enriched with CG-rich subsequences. Third, these exons associated with RS have lower CpG methylation. Third, when the first intron of a transcript is recursively spliced, the downstream introns from the same transcript are more likely to be recursively spliced. Finally, in RS introns, the recursive 3'SS tend to have higher U2AF binding affinities than the corresponding canonical 3'SS. Based on the cis-acting features we observed, we generated a list of potential trans-acting RBPs that may function as recursive splicing factors based on RBP-sequence binding assays. These analyses lay out a blueprint for interrogating the underlying biochemical mechanism of recursive splicing.

Enrichment of CG k-mers in the first exons be associated with recursive splicing in both the first introns and downstream introns.

The mixture models and the enrichment analysis of the tetramer abundance in the first exon regions posit that recursively spliced introns are more enriched in CG-rich subsequences in their upstream exon region than basic introns which are removed by the spliceosome in a single step. By calculating the average exon CpG methylation,

we confirmed that the RS first exons are less methylated than the non-RS first exons. Previous studies have shown that changes in DNA methylation can alter exon inclusions by regulating spliceosome assembly and splicing factor activities (Lev Maor et al., 2015b; Maunakea et al., 2013; Yearim et al., 2015; Young et al., 2005). It has been shown that increasing DNA methylation of an entire gene resulted in a higher inclusion rate of alternative exons (Yearim et al., 2015). Combined with the findings that recursive splicing usually requires RS-exons, one explanation is that the low methylation rate in the upstream exon region can help suppress the inclusion of the RS-exons, which promotes the splicing efficiency and maintains splicing fidelity at the same time.

Alternatively, DNA methylation is mutagenic, and methylated CpG sites can serve as mutation hot spots in cancer and other diseases (Hanson & Liebl, 2022). Evidence showed that methylated cytosines have a mutation rate 20,000 times that of unmethylated cytosines (Danchin et al., 2019). Besides happening in somatic mutations and participating in DNA replication, such mutations caused by methylation can also occur in the germline and be inherited. Thus, it is possible that the low rate of DNA methylation in the upstream exons protects the CG-rich feature from mutations, preventing the recursive splicing regulation function from being lost through mutations. One potential regulation function could be the inhibition of EJC formation at the upstream exon, which marks the completion of splicing and prevents the usage of adjacent splicing sites (Blazquez et al., 2018; Boehm et al., 2018).

We also discovered that transcripts with their first introns recursively spliced have a higher fraction of downstream introns with recursive splicing evidence. The first

exons of transcripts with downstream recursive splicing events are also enriched in CG-rich features and less methylated. Thus, we hypothesize that the usage of recursive splicing sites is determined at the transcription initiation stage. Other aspects of mRNA homeostasis also are determined by co-transcriptional events which occur during transcription initiation such as RNA decay , localization, and translation (Bregman et al., 2011; Trcek et al., 2011; Zid & O'Shea, 2014).

U2AF affinity regulates 3'SS usage.

RBP-binding modeling tools have enabled the in vivo binding prediction through investigating high-throughput data by in vitro assays. We built U2AF binding modes by applying proBound methods to U2AF RBNS data and confirmed that the major binding mode resembled the pyrimidine tract, which is known to be the binding site of U2AF2, with concentration dependence (Rube et al., 2022). Sequence U2AF binding affinity evaluation showed that the recursive 3'SS tends to have higher relative affinity than the paired canonical 3'SS in the same intron. It is well known that U2AF binding is proportional to the alternative exon inclusion, and the 3'SS of alternative exons are relatively weaker than the competing exons (Shao et al., 2014). The higher U2AF affinity values of the recursive sites suggest that they can bind stronger to U2AF than the canonical 3'SS, resulting in the splicing junctions that overlap a pair of canonical 5'SS and recursive 3'SS.

Evidence shows that splicing can occur co-transcriptionally, with the spliceosome physically close to Pol II in vivo (Görnemann et al., 2005; Herzel et al., 2017; Kotovic et al., 2003; Shenasa & Bentley, 2023). In this model, the spliceosome may take the first possible 3'SS signal that is transcribed to carry out an intron sequence,

even though the downstream sequences are still absent. From our affinity evaluations, there are recursive 3'SS that do not have higher U2AF affinity than the canonical sites but still have decent affinity values, which should be able to recruit U2AF before the downstream canonical sites are made.

Future Directions

Our first exon sequence feature analysis indicated that from transcripts with downstream intron recursive splicing events, the first exons have more CG-rich features and lower CpG DNA methylation profiles. We will confirm recursive splicing in two situations. First, we will select transcripts where first introns have no recursive splicing evidence but high fractions of downstream introns are recursively spliced. Since nascent RNA-Seq may not capture all the recursive splicing events due to splicing speed and sequence coverage, we will try to confirm if recursive splicing can be found in the first introns with targeted primer extension. Conversely, we will select transcripts where the first introns are recursively spliced and enriched in CG-rich features, but none of the downstream introns have recursive splicing events. We will then test whether recursive splicing can be discovered in these downstream introns.

This study also predicted a list of candidate proteins with potential different binding activities between sequences flanking introns with or without recursive splicing sites. Dr. Jee Min Kim from Larson lab will perform arrayed CRISPR screen with an image-based assay to investigate the RBP knockdown effect upon recursive splicing. While waiting for the assays, we will also be improving the performance of our recursive splicing prediction models.

Limitations of the study

The nascent RNA-Seq and lariat sequencing provide a solution to splicing intermediate detection, as well as challenges to the investigation of recursive splicing mechanisms analysis. First, due to the various speeds of splicing and the possibility that splicing can happen within seconds, many sequencing events can not be captured (Shenasa & Bentley, 2023; Wan et al., 2021). Thus, the background intron sequences that were labeled as basic introns contain an undetermined amount of false negatives, which could be the reason that I observed several constitutive intron sequences with relatively higher GC content and low CpG methylation, even though there were no annotated alternative splicing records.

This study suggests that the CG-rich subsequences enrichment and low CpG methylation profiles of the upstream exons may induce recursive splicing. And the relatively higher U2AF binding affinity at the recursive 3'SS may recruit splicing factors to carry out the recursive splicing junctions. However, from both analyses, there are a substantial number of sequences that do not follow the patterns predicted, suggesting there may be other recursive splicing regulations that remain to be investigated.

I introduced the usage of mixture models to investigate differences between introns with different types of splicing sites. The topics computed by LDA reflect the Dirichlet distribution of k-mer selections, and further calculations are required to form more reliable functional motifs. Also, other models could also be used to identify the recursive splicing signals, such as Markov models proposed for the

transcription and splicing dynamics (Wan et al., 2021) with states to reflect a splicing decision or not.

Methods

Data preparation for RS and non-RS intron sequences

To lower the probability of RS events from sequencing errors or transposable elements, we applied filtering criteria to make lists of candidate sequences. The nascent RNA splicing junctions and the lariats records are from (Wan et al., 2021). The reference intron and exon sequences are downloaded from the UCSC genome browser (Kent et al., 2002).

1. Background intron sequences must be constitutive introns. From the hg19 reference intron lists, we remove introns that are involved in any annotated alternative splicing events or introns that have smaller annotated introns within their coordinates (Church et al., 2011). The intron list was downloaded from the UCSC genome browser (Kent et al., 2002). The annotated alternative splicing events were generated using SUPPA2 (Trincado et al., 2018).

2. Background intron sequence must be expressed in the sample. We used only intron sequences that were discovered as annotated splicing junctions by STAR with at least 2 nascent RNA-Seq unique spanning reads (Dobin et al., 2013).

3. Unannotated splicing junctions must not have a significant blast hit against the Alu sequence (Price et al., 2004). We found that one Alu element has the

potential recursive splicing feature AGGT and we used blast to evaluate if any splicing junctions (from 68bp upstream of the 5'SS to 217bp downstream of the 5'SS) were Alu elements. Only sequences without any blast score were used for further analysis (Altschul et al., 1990).

4. Unannotated splicing junctions must not be from highly repetitive regions. Regions with long repeats are error-prone in sequencing. Thus, we only use unannotated junctions with less than 80% of the sequences from repetitive regions.

5. Unannotated splicing junctions must be located in the annotated intron regions. We compared the coordinates of unannotated splicing junctions with our list of reference introns that cannot be recursively spliced and cannot have smaller introns inside.

6. Unannotated splicing junctions must have enough reads covering the junctions. We only used unannotated splicing junctions with at least 2 unique spanning reads as potential recursive splicing junctions.

7. Unannotated splicing junctions must not be introns from annotated alternative splicing events.

8. Recursive splicing junctions must be at least 75bp away from the canonical splicing sites. Based on the recursive splicing model, after the observed splicing junctions are removed, the spliceosome will carry out the rest of the intron sequences, which requires the intron to be long enough. We only used

unannotated splicing junctions that can leave at least 75bp introns to be carried out next for further analysis.

9. Recursive splicing junctions in nascent RNA-Seq must not be present in regular RNA-Seq of the same cell line (with >=2 unique spanning reads).

10. The upstream exon sequences and downstream exon sequences must be at least 50bp long. We evaluated the sequence features of the upstream and downstream exon sequences of annotated and unannotated splicing junctions. If the exon sequences are too small, we may not obtain enough k-mer composition information for model fitting.

11. Potential recursive splicing events from lariat sequencing should have the junction sizes much smaller than annotated introns and meet all criteria applied to the nascent RNA-Seq. We also used the lariat sequencing profile to identify potential recursive splicing events. Besides the filtering criteria of nascent RNA-Seq, we also made sure that the lariats sequences must be at least 75bp away from both annotated 5'SS and 3'SS.

Creating mixture models and sequence topic distribution analysis

We used the LDA method developed by the Sci-Kit learn Python package (Pedregosa et al., 2011). The hyper-parameters of the model were selected based on the performance of the classifiers to distinguish between RS and basic introns and the biological meanings from the interpretable topics. The driving k-mer from each topic was calculated by the distinctiveness values followed by the Kullback-Leibler divergence (Dey et al., 2017). The structure

plot visualization was developed using Matplotlib Python package (Hunter, 2007). Matrices were calculated with Numpy package (Harris et al., 2020).

RS predictions

We fitted random forest classifier to predict recursive splicing in the first introns and in the downstream introns using features including topic distributions and exon CpG methylation status (Pedregosa et al., 2011).

K-mer enrichment and multi-testing corrections

We calculated the log fold change of the k-mer fractions from different groups of introns to evaluate the k-mer enrichment. Since the tetramer counts involved some overlapping subsequences and thus cannot be viewed as purely independent, we applied multi-testing corrections provided by the Sci-Py Python package to correct the p-values (Virtanen et al., 2020).

Protein candidate selections

We modeled sequences flanking different types of 3'SS and 5'SS by LDA. Matrix multiplication of the topic-kmer matrix and the kmer-protein R-value matrix generates the topic-protein matrix (The ENCODE Project Consortium, 2012). We used Kullback-Leibler divergence for the topics to get the top 'driving protein' list.

Chapter 4: U2AF Binding Affinity in RNA Sequences

Introduction

RNA-binding proteins (RBPs) are an essential group of proteins that regulate gene expression. RBPs bind to RNA sequences by directly recognizing sequence signals or through aids from other proteins, altering the fate of the RNA sequences and their products (Dominguez et al., 2018). The investigation of RBP-sequence interactions requires various binding assays, including in vivo assays such as multiple types of crosslinking and immunoprecipitation (CLIP), and in vitro assays such as RNA bind-and-seq (RBNS), followed by high-throughput sequencing and computationally binding site searching (Lambert et al., 2014; Van Nostrand, Pratt, et al., 2020). Moreover, RBP knockdown followed by RNA-Seq reflects the gene expression and splicing changes in response to the focused RBP removal. Immunofluorescence techniques provide a solution to RBP localizations (Van Nostrand, Freese, et al., 2020).

The Encyclopedia of DNA Elements (ENCODE) project phase III applied the above methods to introduce a dataset that resolved the RBP-sequence interactions of various post-transcriptional regulations (The ENCODE Project Consortium, 2012). The in vitro binding site motifs calculated from RBNS data are highly overlapped with eCLIP motifs, and always present at eCLIP peaks, confirming that the in vivo binding can be specifically controlled by in vitro RBP affinities. Upon RBP knockdown, the genes that are significantly upregulated or downregulated overlap with genes enriched in the RBP binding motifs, suggesting the RNA-protein interactions regulate the fate

of the genes. The integration between RBP abundance and alternative splicing also revealed the association between RBP and splicing regulations (Van Nostrand, Freese, et al., 2020).

As one of the fundamental splicing factors, the heterodimer U2AF has been well-studied to reveal the splicing mechanisms and regulations (Shao et al., 2014; Soares et al., 2006; Wu & Fu, 2015). During the RNA splicing process that is catalyzed by the major spliceosome, at the complex E stage, U2AF1 binds to the 3'SS, and U2AF2 binds to the polypyrimidine tract. Studies also found cases where U2AF is not required as the canonical splicing mechanism, such as the AG-independent splicing manner when strong pyrimidine tract recruits U2AF2 and splicing without U2AF when the level of SR proteins is very high. Both U2AF1 and U2AF2 can regulate splicing. Studies suggest that the concentration of U2AF and the location of U2AF binding can affect splicing (Lim et al., 2011). U2AF can also be recruited to the splicing sites through interactions with other proteins, which makes it also a target for other splicing regulators (Fu & Ares, 2014).

U2AF mutations have been observed as driver mutations in hematopoietic disorders, and U2AF1 S34F is particularly associated with poor prognosis in patients with myelodysplastic syndrome (Esfahani et al., 2019; Seiler et al., 2018). Besides altering the splicing site selection, U2AF1 S34F mutation also promotes cancer progression as a translation regulator (Akef et al., 2020; Palangat et al., 2019). Studies revealed that U2AF binds to mRNA in the cytoplasm and can function noncanonically as a translation repressor. S34F mutation can inhibit the function and increase gene expressions that cause oncogenic phenotypes (Palangat et al., 2019).

In this chapter, I presented three data analysis projects that aim to resolve U2AF-RNA binding activities. The first project is focused on calculating U2AF binding models. The predicted binding models were generated by applying an online RBP analyzing tool proBound to our U2AF RBNS data (Rube et al., 2022). The RBNS experiments and k-mer enrichment calculations are conducted by Dr. Benjamin T. Donovan from Larson lab at the National Cancer Institute. With the binding models, I can calculate the relative U2AF binding affinity for any sequence and confirmed the binding of U2AF is concentration-dependent.

The second project is evaluating the U2AF binding affinities of U2AF1 binding sites revealed by photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (Danan et al., 2016). This is part of a project conducted by Dr. Gloria R. Garcia from Larson lab at the National Cancer Institute, aiming to reveal the function of U2AF1 in mitochondria and the effect of S34F mutation upon translation. By Immunoprecipitation-Mass Spectrometry (IP-MS), Dr. Garcia found that U2AF1 interacts with proteins involved in translation regulations. The high-throughput sequencing after PAR-CLIP revealed that U2AF1 binds to mitochondrial-associated mRNAs and single-exon genes. After analyzing the U2AF binding affinity distributions, we observed that in various genomic regions, the highest abundance of U2AF1 binding was found in coding regions, while among all types of binding sites, the locations near 3'SS have the highest binding affinities.

Led by Dr. Benjamin T. Donovan, the third project aims to discover how splicing factors DDX42 regulate U2AF binding. After the search for U2AF interaction partners through mass spectrometry by Dr. Garcia, it was confirmed that DDX42 is

the most enriched factor. Dr. Donovan performed DDX42 knockdown, which led to more stable U2AF binding, and conducted RNA-Seq for KD samples and control samples. In this project, I built a pipeline with open-source software to analyze the RNA-Seq data to discover the changes in splicing and gene expression. Though the numbers of alternative splicing events from all samples are similar, we observed that the selection of alternative 3'SS is associated with the binding affinity ratio between the site pairs, which can be altered by DDX42 knockdown.

Results

The primary binding mode of U2AF resembles the polypyrimidine tract motif with concentration-dependent binding activities.

Previous studies have shown that the U2AF2 RNA recognition motifs (RRMs) bind to 9bp of the RNA sequences (Glasser et al., 2022); I decided to model the U2AF binding sites of 12 nucleotides and expected to capture motifs that cover the known signals. The primary binding modes from both wildtype U2AF and U2AF with U2AF1 S34F mutation represent the polypyrimidine tract, with high scores for Ts at the first ten positions and relatively higher C scores at the last two nucleotides. The second and third binding modes from S34F and the third binding mode from wildtype U2AF contain GAGA, which has AG but does not follow the 3'SS consensus CAG. The second binding mode of wildtype U2AF contains CAG, but the score of the A site is much less significant than the G signal (Fig4.1A, Fig4.1B). The relative affinity distribution of the RBNS sequences showed that for both primary binding modes, the U2AF samples have higher affinity values than the control samples and

the 20nM samples have the highest affinities, followed by 80nM and 320nM samples. The concentration-dependent manner was not observed for other motifs, while the second binding model of wild type indicates that the U2AF samples have slightly higher affinity than the control. Thus, I focused on the primary binding modes for the next steps (Fig4.1C).

The deep RNA-Seq of RBNS generated over 100,000 reads for each sample, and I randomly selected 50,000 reads to build the models. To confirm the generalization of the calculated motifs, I made two new subsamples and applied the primary binding modes to them. The calculated binding affinities revealed that the motifs I made could adapt to the new datasets of U2AF RBNS (Fig4.1D).

The difference in U2AF binding between canonical and recursive 3'SS has been observed, and the level of U2AF can also affect the splicing outcome. I calculated the relative binding affinity with the primary binding modes for annotated constitutive introns and introns that are involved in alternative splicing (Fig4.1E). Compared with regular introns, the alternative 3'SS and 3'SS from annotated intron retention events have relatively lower U2AF affinities. Interestingly, it is observed that the upstream 3'SS of possible skipped exons have similar affinities with canonical introns, while the downstream 3'SS have stronger U2AF affinities. This observation is consistent with previous findings that the 3'SS of the alternative exons are relatively weaker than flanking exons (Shao et al., 2014).

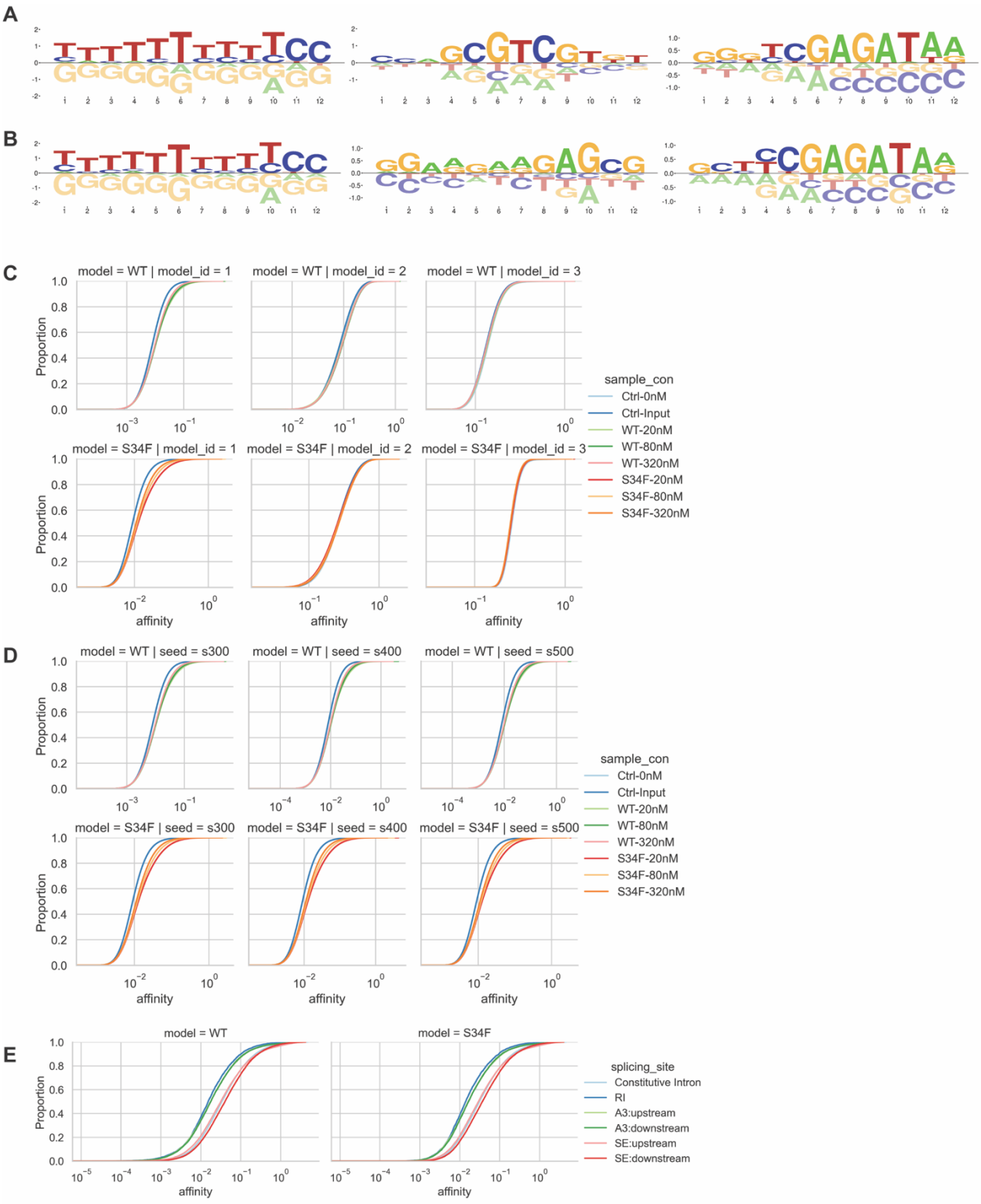


Figure 4.1 Binding modes of wildtype U2AF and U2AF with U2AF1 S34F mutation. (A) Three binding modes of wildtype U2AF. The primary binding mode resembles the polypyrimidine tract feature. (B) Three binding modes of U2AF with S34F mutation. (C) The cumulative distributions of relative binding affinities calculated using RBNS sequences and 6 modes. (D) The cumulative distributions using different random subsets of RBNS sequences. All subsets show concentration-dependent binding. (E) The cumulative distribution of sequences that range from upstream 32 to 2bp of annotated 3'SS in different types of alternative splicing events and canonical introns.

U2AF clustered at coding regions of low-affinity binding sites at mRNA sequences.

Transcriptome-wide U2AF-RNA interaction sites were captured in the nucleus, chromatin, and cytoplasm by Dr. Garcia using PAR-CLIP. Since the size of the signals varied, and the affinity comparison could only be applied to sequences of the same length, I selected records of at least 12bp long and scanned every 12-mer to pick the highest value as the binding site's affinity. The wildtype samples and the U2AF1 S34F mutation samples from all three locations show the highest U2AF affinities at sites flanking annotated 3'SS, which is consistent with the known function of U2AF1 in RNA splicing (Fig4.2A). The signals from the cytoplasm and nucleus have overall higher affinity values than signals in chromatin. The wildtype samples also have higher affinities than the samples of U2AF1 S34F, indicating that the S34F mutation alters the splicing site selection, and the sites being used are less sensitive to the wildtype U2AF (Fig4.2B).

A substantial amount of U2AF signals were mapped to CDS and UTR regions, suggesting that U2AF may have other gene regulatory functions other than splicing. To evaluate how the binding activities changed at different positions along coding sequences, we made metagene plots using 20bp of UTRs and whole CDS. Because the CDS sizes vary, we scaled the coding sequences from 50bp to the end of the sequences. The metagene of signal counts shows that the CDS region has more abundant U2AF binding targets than the UTRs (Fig4.2E). However, the metagene of U2AF affinity values calculated using the signal sequences indicates that the CDS region is relatively less sensitive to U2AF than UTR5 or UTR3, and the target

sequences from wildtype samples have higher affinities than the S34F signals (Fig4.2F). The difference between wildtype and S34F sequence affinities is consistent with the result from the RBNS sequence affinity calculations that the sequences bound by U2AF with S34F mutation are less sensitive to the wildtype U2AF, while the fact that CDS regions have lower U2AF affinity but more binding signals suggests that U2AF may be recruited to the coding sequences by a different mechanism than the primary binding mode.

The pull-down proteins from U2AF1 IP-Mass spectrometry include 12 out of 13 subunits of the EIF3 complex, which has been proven a mitochondrial and transcription regulator (A. S. Y. Lee et al., 2015; Lin et al., 2020). Moreover, overlaps between the PAR-CLIP signals of U2AF1 and EIF3 in mitochondria-associated mRNAs have been observed. To test if U2AF and EIF3 have the same binding targets, I built EIF3D binding model with the EIF3D RBNS sequencing data from ENCODE and the proBound software (Rube et al., 2022; The ENCODE Project Consortium, 2012). Both the motifs from proBound and RBNS contain GAGA, which also exists in the U2AF secondary binding modes (Fig4.2C). I observed that the EIF3D RBNS sequences have higher affinity values than the control samples, but the binding strength is not concentration-dependent (Fig4.2D). The metagene plot of the EIF3D binding affinity indicates that the CDS region has similar affinity values as UTRs, and there is a smooth decreasing trend from the end of the 5' UTR region to the beginning of the 3' UTR (Fig4.2G).

I compared U2AF binding affinity from sequences in U2AF and EIF3 PAR-CLIP signals. The RBP FMR1 was used as a negative control (Ascano et al., 2012). The

affinity values from FMR1 targets are relatively stable. In the region of CDS and 5' UTRs, EIF3 targets have higher U2AF binding affinities, while in the intron, 3' UTRs, and 3'SS regions, the targes from U2AF PAR-CLIP data can be bound by U2AF tighter (Fig4.2H). On the contrary, the EIF3D affinity distribution shows that both wildtype and S34F U2AF PAR-CLIP targets have higher EIF3D affinities in CDS and 5' UTR regions than 3'SS, 3' UTRs, and introns, suggesting that U2AF binding at the low-affinity CDS sites may be affected by EIF3 (Fig4.2I).

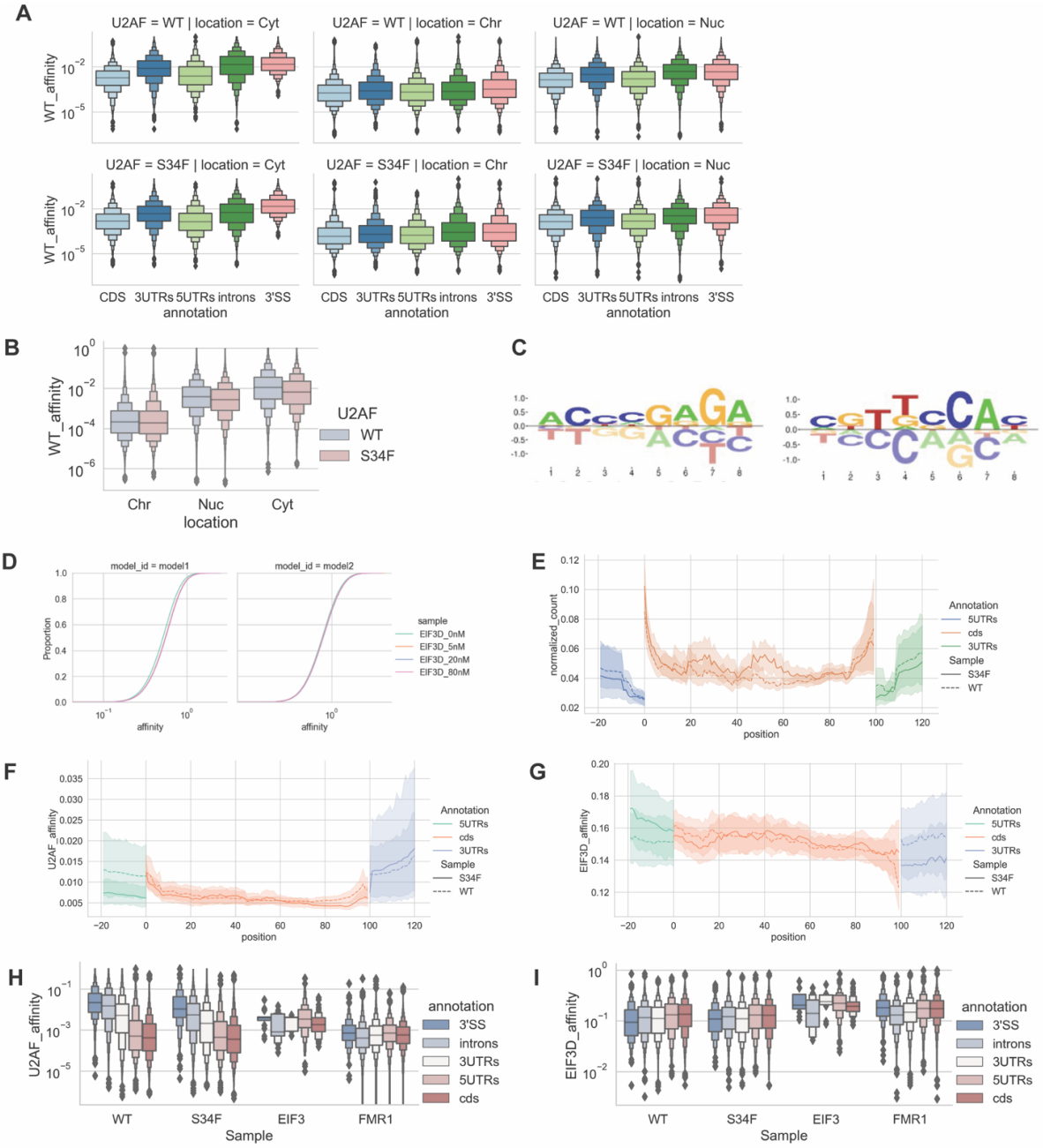


Figure 4.2 U2AF binding affinities of PAR-CLIP target sequences. (A) The distributions of U2AF binding affinities of wildtype and mutant U2AF PAR-CLIP target sequences. Sequences were annotated with regions on RNA sequences and were collected from samples from cytoplasm, chromatin, and nucleus. The affinity values were calculated using wildtype U2AF primary binding mode of 12-mer (Fig4.1A). (B) The distribution of all the target sequences from PAR-CLIP. The S34F target sequences are less sensitive to the wildtype U2AF. (C) The primary and secondary binding modes of EIF3 calculated by proBound. (D) The cumulative distributions of EIF3D RBNS sequences calculated with EIF3D binding modes. (E) The metagene plot of U2AF PAR-CLIP target counts on mRNA sequences. The region from 50bp to 100bp of the CDS represents scaled CDS from 50bp to the end of coding sequences. (F) The metagene plot of U2AF binding affinities of the U2AF PAR-CLIP targets on mRNA sequences. (G) The metagene plot of EIF3D binding affinities of the U2AF PAR-CLIP targets on mRNA sequences. (H) The distribution of U2AF affinities of targets from different PAR-CLIP datasets. Target sequences were annotated in different regions on RNA sequences. (I) The distribution of EIF3D affinities of targets from different PAR-CLIP datasets.

3’ splice site selection is affected by U2AF affinity and may be regulated by DDX42 indirectly.

I built a pipeline to analyze alternative splicing, and gene expression changes from RNA-Seq of DDX42 KD and control samples (Fig4.3A). Differential expression showed that 10 out of 13 transcripts of DDX42 have decreased expression in DDX42 KD samples (Fig4.3B). The control samples and DDX42 KD samples have similar numbers of annotated and unannotated splicing junctions (Fig4.3C). All the samples also have similar numbers of alternative splicing events (Fig4.4D).

I found over 24,000 alternative 3'SS (A3) events from control and knockdown samples. Over 80% of the involved 3'SS have U2AF affinity less than 0.1 (Fig4.3E, Fig4.3F). Thus, we investigated the effect of U2AF upon 3'SS selection among low-affinity sites by calculating the relationship between the percent spliced-in (PSI) of the alternative exon and the U2AF affinity ratio between 3'SS. We binned the natural log of the affinity ratios to explore the relationship between the affinity ratios and the 3'SS selections (Fig4.3G). The linear relationship between the median of the binned values and the PSI of alternative exon sequences follows a sigmoid-like function where the PSI of the alternative exon sequences increases as the ratio between upstream and downstream 3'SS grows. This observation suggests that though most 3'SS are weak U2AF binding sites, the 3'SS usage is biased towards the stronger U2AF binding sites. Studies also revealed that the inclusion of exons from potential skipped exon (SE) events is affected by the competition between two 3'SS (Shao et al., 2014). I removed SEs where the alternative exons can provide potential recursive splicing sites to avoid interference of inaccurate PSI calculations by incomplete

multistep splicing. SE also shows a sigmoid-like relationship between the U2AF affinity ratios and the inclusion rate of the alternative exon (Fig4.3H-4.3J). Though there are unique 3'SS that were only used by control or knockdown samples, both sample types show the same relationship between splice site selection and U2AF affinity ratio.

To evaluate if DDX42 knockdown affected the biological functions, I performed Gene Ontology (GO) enrichment analysis using differentially expressed genes (Thomas et al., 2022). The top molecular function GO of the genes that are highly expressed in control samples are polyU binding and polypyrimidine tract binding, which is consistent with the U2AF functions, indicating that DDX42 may regulate U2AF binding activities indirectly. I also observed that genes that are relatively highly expressed in DDX42 KD samples are involved in positive regulations of RNA bindings (Fig4.3K).

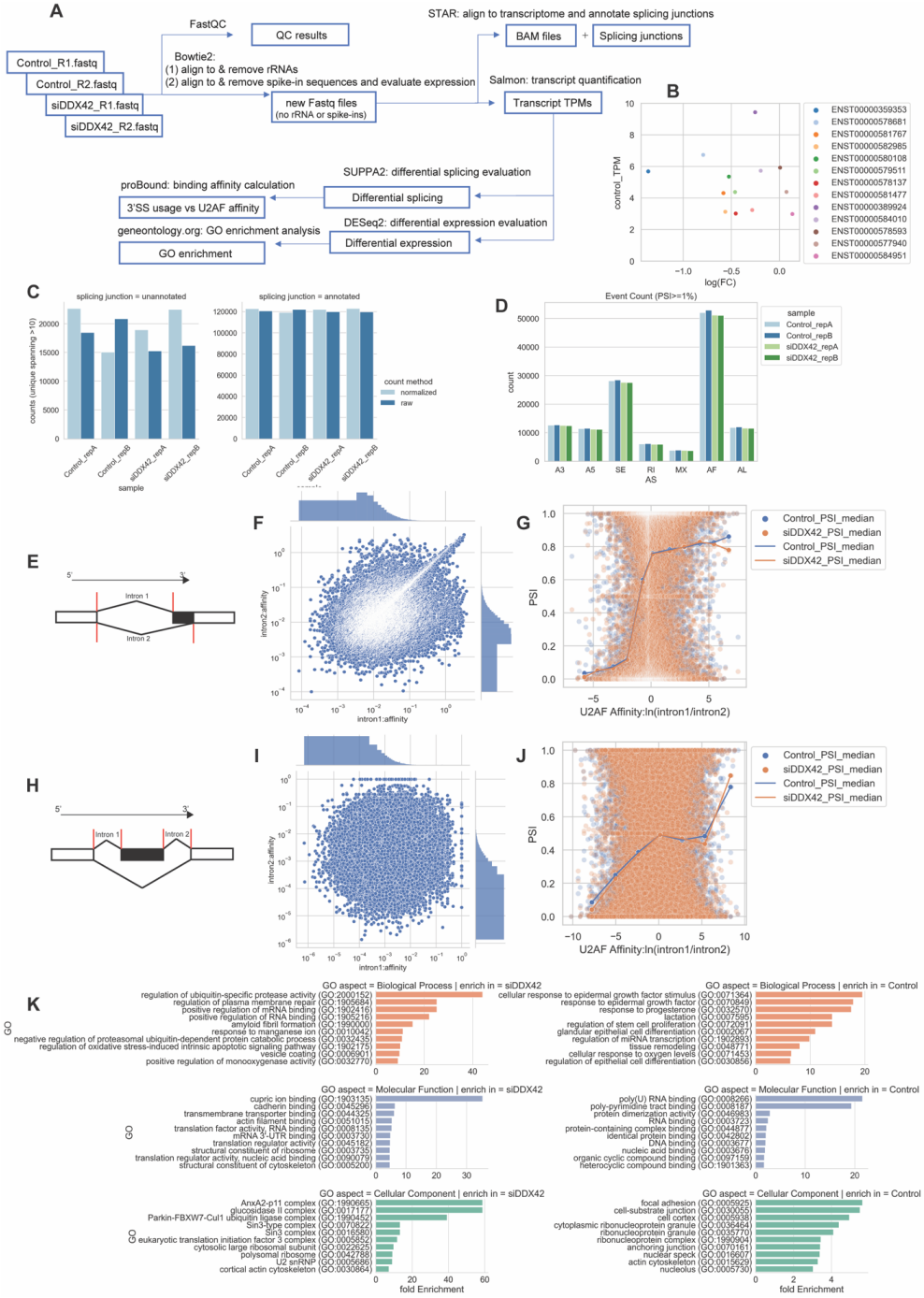


Figure 4.3 Differential splicing and gene expression analysis of DDX42 KD samples. (A) Diagram of the pipeline of the RNA-Seq analysis. (B) The scatterplot of the log fold change of transcript per million reads (TPM) and the average TPM in control samples of DDX42 isoforms. (C) The count plots of annotated and unannotated splicing junctions in the samples. (D) The count plot of different types of alternative splicing events in the samples from alternative 3'SS events. (E) Diagram of 3'SS selection in alternative 3'SS events. (F) The scatterplot of the upstream and downstream 3'SS U2AF affinities in all AS3. (G) The scatterplot of the log ratio of alternative 3'SS and the alternative exon sequence percent spliced in (PSI) values. The line plot represents the median values from binned data. (H) Diagram of the 3'SS selection in skipped exon events. (I) The scatterplot of the upstream and downstream 3'SS U2AF affinities in all SE. (J) The scatterplot of the log ratio of two 3'SS and the PSI of alternative exons from skipped exon events. (K) Gene Ontology enrichment analysis using genes that are up- and down-regulated in DDX42 KD samples. Only the top 10 GO were shown.

Discussion

By establishing the binding modes of U2AF, I observed that the in vitro U2AF-RNA binding activities are concentration-dependent and primarily determined by U2AF2. The binding modes of U2AF allowed me to measure the U2AF relative binding affinities of annotated canonical 3'SS and 3'SS involved in alternative splicing events. The relatively higher affinities in downstream 3'SS in the skipped exon events are consistent with previous studies that the 3'SS of alternative exons are weaker than those of the flanking exons, confirming that the binding model I used reflects the U2AF functions (Shao et al., 2014; Wu & Fu, 2015).

U2AF specifically binds to coding regions of mitochondria-associated mRNA sequences with low U2AF affinity but high EIF3D affinity sites.

The PAR-CLIP cluster analysis posits that U2AF binds to distinct sites on mRNAs and is enriched in CDS regions, even though the targets in CDS have the lowest U2AF binding affinities among different annotated regions (Fig4.2H). However, the CDS targets have the highest EIF3D affinity values among the regions, including CDS, UTRs, and introns. The IP-MS implies that EIF3 is one of the vital interaction partners of U2AF. Studies have shown that EIF3 subunit knockdown affects mitochondrial functions, which was also observed in samples with U2AF S34F mutations (Lin et al., 2020). Through RBP in vitro binding modeling, we confirmed that EIF3D and U2AF have distinct binding models. Thus, it is possible that the mechanisms of U2AF binding at the CDS involve U2AF-EIF3 interactions at non-U2AF primary binding sites.

3’ splice site selection is biased towards more robust U2AF binding sites.

The calculation of alternative 3'SS U2AF binding affinities showed that the alternative exon inclusion rate relates to the ratio between two potential 3'SS, showing that the 3'SS usage is biased towards the stronger U2AF binding sites. On the other hand, the horizontal asymptote phases of the sigmoid-like relationship suggest that changes in the binding affinity ratios do not always affect the splice site selection as much as expected. Splicing site choice is affected by the abundance of various splicing factors and their interactions. Thus, the insensitivity of 3'SS selection to U2AF affinity changes may be the consequence of regulation by other factors. Further efforts are needed to complete the 3'SS selection model.

DDX42 can regulate splicing site selection in various ways. Previous studies found that the interactions between DDX42 and SF3B1 regulate alternative splicing site selections, and mutations disrupting the interactions are associated with altered splicing site usage and cancers (Yang et al., 2023; Zhao et al., 2022). Here I present that DDX42 may indirectly regulate splicing by promoting the expression of genes that recognize the pyrimidine tract. I observed that genes that are not U2AF but down-regulated in DDX42 KD samples have molecular functions involving polyU and polypyrimidine tract binding. Further analysis is required to evaluate the effect of those genes on RNA splicing.

Limitations of the study

Evaluating the RBP-sequence binding activities by the models calculated with RBNS data has a few limitations. First, I could not find a U2AF binding mode with both pyrimidine tract and AG features. Though the AG-independent splicing manner has

been observed under strong pyrimidine signals, studies also found that U2AF1 can affect the usage of 3'SS (Shao et al., 2014). Thus, the AG-independent primary binding mode of U2AF can not reflect all the features of U2AF binding activities. Second, the only EIF3 RBNS result I could find from the public database was the EIF3D sequencing data, which is only one of 13 components of the EIF3 complex. Further experiments are required to confirm the major components and their binding preferences in the interactions with U2AF.

Third, in the DDX42 knockdown RNA-Seq analysis, the changes of 3'SS selection between control and experimental samples are very subtle. Further efforts are required to confirm the significance of the observations.

Methods

U2AF and EIF3D binding mode identification and relative affinity calculations

The binding models were generated using proBound online application (Rube et al., 2022). The data used for modeling includes U2AF RBNS sequencing data by Dr. Donovan at Larson lab and EIF3D RBNS sequencing data from ENCODE (The ENCODE Project Consortium, 2012). I randomly selected 50,000 reads from each sample using seqtk (https://github.com/lh3/seqtk).

The relative binding affinity of each sequence I calculated using proBound software based on the calculated binding models. For canonical and alternative 3'SS sequences, I used 30bp sequences from upstream 32bp to 2bp of annotated 3'SS. For PAR-CLIP target sequences, because comparing affinity values of sequences of different sizes is not ideal, I scanned all the 12-mer of target sequences and selected the highest values.

RNA-Seq analysis pipeline

To remove rRNAs from sequencing reads, I aligned the fastq files against rRNA reference sequences by Bowtie2 (Langmead & Salzberg, 2012). New fastq files were generated by samtools without rRNA reads (Li et al., 2009). The RNA-Seq samples were added spike-in sequences for quantification calibration (Jiang et al., 2011). I aligned the fastq files against the reference spike-in sequences by Bowtie2 and calculated the normalized read counts. New fastq files were generated by samtools without spike-in reads. I then aligned the new fastq files against hg38 transcriptome by STAR to evaluate the splicing junctions (Dobin et al., 2013).

Transcriptome quantification was carried out by Salmon, which provided the input data for differential splicing by SUPPA2 and differential expression by DESeq2 (Love et al., 2014; Trincado et al., 2018). At the end, I performed Gene Ontology enrichment analysis using the GO consortium (Thomas et al., 2022).

Chapter 5: Conclusions and Future Directions

Conclusions

Signals in nucleotide sequences serve as recognition sites for RNA-binding proteins, promoting context-dependent control of gene expression. Interruptions in signal recognition alter the fate of gene expression products, causing dysregulation and disease. Accurate signal recognition can contribute to sequence function prediction, which will assist in understanding biological processes, disease development, and potential drug target identification. In this dissertation, I presented three research chapters about developing applications of mixture models to identify sequence functional motifs, investigating cis-acting and trans-acting factors that can regulate recursive splicing site usage in nascent RNAs, and modeling U2AF binding activities.

Mixture models can identify functional sequence motifs.

In Chapter 2, I presented the applications of mixture models in solving sequence feature identification problems under three different conditions. I confirmed that LDA has the potential to identify functional motifs, such as the polypyrimidine tract and the branch point. Also, LDA can recognize subtypes of sequences, such as long and short introns, and reading frames and species of coding sequences. The observations suggest that LDA can find motifs that only a subset of sequences possess and may be applied to sequence function annotations.

Non-coding sequences from the human genome have coding signals.

I identified non-coding sequences with CDS reading frame features by evaluating sequence features from 5' UTRs and long non-coding RNAs with the coding sequence model. This finding suggests that LDA has the potential for gene annotation.

Recursive splicing is associated with the sequence features and methylation status of the upstream exons.

In Chapter 3, I described the identification of sequence features associated with recursive splicing events. Compared with exons upstream of regular introns, the exon sequences upstream of introns containing recursive splicing sites are enriched in CG-nucleotides withing specific contexts. Analyzing whole-genome bisufite sequencing profiles from public databases revealed that RS-related upstream exons are less methylated. There are two potantial explanations that require further investigations: (1) DNA methylation rate alters the downstream splicing site strength, and (2) the CG-rich feature recruits some RBPs and prevents the formation of EJC, and the low methylation protects the feature from mutations.

Recursive 3’ splice sites provide stronger U2AF binding sites.

I made a primary U2AF binding model using RBNS RNA-Seq and proBound. By calculating sequence U2AF binding affinities, I found that in the introns with recursive 3'SS, the recursive sites can bind U2AF tighter than the downstream canonical sites. Thus, during co-transcriptional splicing, the recursive 3'SS can provide a qualified splicing site before the canonical 3'SS is made.

The 3’ splice site selection between alternative sites is biased towards stronger U2AF binding sites.

I calculated 3'SS selections from both AS3 and SE events in RNA-Seq differential splicing results. The relationship between the log ratio of 3'SS pairs and the PSI values of alternative exons follows a sigmoid-like function, where the 3'SS selection is biased towards splicing sites of higher binding affinities. However, the mechanism of the horizontal asymptote phases of the relationship where the difference between splicing sites does not affect the 3'SS selection remains to be investigated.

Future Directions

To apply LDA to more alternative splicing events and sequences flanking cryptic splicing sites

Splicing site selection is affected by the coordination of various splicing factors, which requires different RBP binding site signals. I will apply mixture models to different subtypes of intron and exon sequences based on annotated AS events and differential splicing changes upon mutations to identify unknown sequence subtypes.

To evaluate the effect of upstream exon features and effect of RBPs on recursive splicing regulation

My calculations show that the upstream exons with CG-rich features are more likely to have RS events in the downstream introns. Dr. Donovan will conduct experiments to evaluate the effect of RS site usage by the upstream exon sequences.

I also generated a list of RBPs that may regulate RS. Dr. Kim will investigate the changes of RS upon RBP knockdowns.

To investigate the 3' splice site selection changes due to DDX42 knockdown

Though significant changes in differential splicing were not detected in the RNA-seq between control and DDX42 knockdown samples, I did observe unique unannotated splicing site usage in KD samples or control samples. I will analyze sequence features, U2AF affinity values of the unique splicing sites, and sequences involved in unique AS events. I will also perform GO annotation of the unique transcripts in the sequences.

Bibliography

Akef, A., McGraw, K., Cappell, S. D., & Larson, D. R. (2020). Ribosome biogenesis is a downstream effector of the oncogenic U2AF1-S34F mutation. *PLOS Biology*, *18*(11), e3000920. https://doi.org/10.1371/journal.pbio.3000920

Akhtar, W., & Veenstra, G. J. C. (2011). TBP-related factors: A paradigm of diversity in transcription initiation. *Cell & Bioscience*, *1*(1), 23. https://doi.org/10.1186/2045-3701-1-23

Alló, M., Buggiano, V., Fededa, J. P., Petrillo, E., Schor, I., De La Mata, M., Agirre, E., Plass, M., Eyras, E., Elela, S. A., Klinck, R., Chabot, B., & Kornblihtt, A. R. (2009). Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nature Structural & Molecular Biology*, *16*(7), 717–724. https://doi.org/10.1038/nsmb.1620

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Ambrosini, G., Groux, R., & Bucher, P. (2018). PWMScan: A fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics*, *34*(14), 2483–2484. https://doi.org/10.1093/bioinformatics/bty127

Anastasiadi, D., Esteve-Codina, A., & Piferrer, F. (2018). Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. *Epigenetics & Chromatin*, *11*(1), 37. https://doi.org/10.1186/s13072-018-0205-1

Ascano, M., Mukherjee, N., Bandaru, P., Miller, J. B., Nusbaum, J. D., Corcoran, D. L., Langlois, C., Munschauer, M., Dewell, S., Hafner, M., Williams, Z., Ohler, U., & Tuschl, T. (2012). FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature*, *492*(7429), 382–386. https://doi.org/10.1038/nature11737

Bailey, T. L. (2021). STREME: Accurate and versatile sequence motif discovery. *Bioinformatics*, *37*(18), 2834–2840. https://doi.org/10.1093/bioinformatics/btab203

Bailey, T. L., Williams, N., Misleh, C., & Li, W. W. (2006). MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, *34*(Web Server), W369–W373. https://doi.org/10.1093/nar/gkl198

Baralle, F. E., & Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, *18*(7), 437–451. https://doi.org/10.1038/nrm.2017.27

Begg, B. E., Jens, M., Wang, P. Y., Minor, C. M., & Burge, C. B. (2020). Concentration-dependent splicing is enabled by Rbfox motifs of intermediate affinity. *Nature Structural & Molecular Biology*. https://doi.org/10.1038/s41594-020-0475-8

Blazquez, L., Emmett, W., Faraway, R., Pineda, J. M. B., Bajew, S., Gohr, A., Haberman, N., Sibley, C. R., Bradley, R. K., Irimia, M., & Ule, J. (2018). Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing. *Molecular Cell*, *72*(3), 496-509.e9. https://doi.org/10.1016/j.molcel.2018.09.033

Blei, D. N., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 30.

Boehm, V., Britto-Borges, T., Steckelberg, A.-L., Singh, K. K., Gerbracht, J. V., Gueney, E., Blazquez, L., Altmüller, J., Dieterich, C., & Gehring, N. H. (2018). Exon Junction Complexes Suppress Spurious Splice Sites to Safeguard Transcriptome Integrity. *Molecular Cell*, *72*(3), 482-495.e7. https://doi.org/10.1016/j.molcel.2018.08.030

Bregman, A., Avraham-Kelbert, M., Barkai, O., Duek, L., Guterman, A., & Choder, M. (2011). Promoter Elements Regulate Cytoplasmic mRNA Decay. *Cell*, *147*(7), 1473–1483. https://doi.org/10.1016/j.cell.2011.12.005

Brooks, A. N., Duff, M. O., May, G., Yang, L., Bolisetty, M., Landolin, J., Wan, K., Sandler, J., Booth, B. W., Celniker, S. E., Graveley, B. R., & Brenner, S. E. (2015). Regulation of alternative splicing in *Drosophila* by 56 RNA binding proteins. *Genome Research*, *25*(11), 1771–1780. https://doi.org/10.1101/gr.192518.115

Burnette, J. M., Miyamoto-Sato, E., Schaub, M. A., Conklin, J., & Lopez, A. J. (2005). Subdivision of Large Introns in Drosophila by Recursive Splicing at Nonexonic Elements. *Genetics*, *170*(2), 661–674. https://doi.org/10.1534/genetics.104.039701

Busch, A., & Hertel, K. J. (2012). Evolution of SR protein and hnRNP splicing regulatory factors. *WIREs RNA*, *3*(1), 1–12. https://doi.org/10.1002/wrna.100

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., … Hubbard, T. (2011). Modernizing Reference Genome Assemblies. *PLoS Biology*, *9*(7), e1001091. https://doi.org/10.1371/journal.pbio.1001091

Clower, C. V., Chatterjee, D., Wang, Z., Cantley, L. C., Vander Heiden, M. G., & Krainer, A. R. (2010). The alternative splicing repressors hnRNP A1/A2 and PTB influence pyruvate kinase isoform expression and cell metabolism. *Proceedings of the National Academy of Sciences*, *107*(5), 1894–1899. https://doi.org/10.1073/pnas.0914845107

Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L., & Myers, R. M. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Research*, *16*(1), 1–10. https://doi.org/10.1101/gr.4222606

Crick, F. (1970). *Central Dogma of Molecular Biology*.

Danan, C., Manickavel, S., & Hafner, M. (2016). PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites. In E. Dassi (Ed.), *Post-Transcriptional Gene Regulation* (Vol. 1358, pp. 153–173). Springer New York. https://doi.org/10.1007/978-1-4939-3067-8_10

Danchin, E., Pocheville, A., Rey, O., Pujol, B., & Blanchet, S. (2019). Epigenetically facilitated mutational assimilation: Epigenetics as a hub within the inclusive evolutionary synthesis. *Biological Reviews*, *94*(1), 259–282. https://doi.org/10.1111/brv.12453

De Conti, L., Baralle, M., & Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing: Exon and intron definition in pre-mRNA splicing. *Wiley Interdisciplinary Reviews: RNA*, *4*(1), 49–60. https://doi.org/10.1002/wrna.1140

Dey, K. K., Hsiao, C. J., & Stephens, M. (2017). Visualizing the structure of RNA-seq expression data using grade of membership models. *PLOS Genetics*, *13*(3), e1006599. https://doi.org/10.1371/journal.pgen.1006599

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., Van Nostrand, E. L., Pratt, G. A., Yeo, G. W., Graveley, B. R., & Burge, C. B. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell*, *70*(5), 854-867.e9. https://doi.org/10.1016/j.molcel.2018.05.001

Duff, M. O., Olson, S., Wei, X., Garrett, S. C., Osman, A., Bolisetty, M., Plocik, A., Celniker, S. E., & Graveley, B. R. (2015a). Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature*, *521*(7552), 376–379. https://doi.org/10.1038/nature14475

Duff, M. O., Olson, S., Wei, X., Garrett, S. C., Osman, A., Bolisetty, M., Plocik, A., Celniker, S. E., & Graveley, B. R. (2015b). Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature*, *521*(7552), 376–379. https://doi.org/10.1038/nature14475

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, *22*(10), 1315–1316. https://doi.org/10.1038/nbt1004-1315

Esfahani, M. S., Lee, L. J., Jeon, Y.-J., Flynn, R. A., Stehr, H., Hui, A. B., Ishisoko, N., Kildebeck, E., Newman, A. M., Bratman, S. V., Porteus, M. H., Chang, H. Y., Alizadeh, A. A., & Diehn, M. (2019). Functional significance of U2AF1 S34F mutations in lung adenocarcinomas. *Nature Communications*, *10*(1), 5712. https://doi.org/10.1038/s41467-019-13392-y

Fonseca, D. M., Keyghobadi, N., Malcolm, C. A., Mehmet, C., Schaffner, F., Mogi, M., Fleischer, R. C., & Wilkerson, R. C. (2004). Emerging Vectors in the *Culex pipiens* Complex. *Science*, *303*(5663), 1535–1538. https://doi.org/10.1126/science.1094247

Fu, X.-D., & Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, *15*(10), 689–701. https://doi.org/10.1038/nrg3778

Gehring, N. H., & Roignant, J.-Y. (2021). Anything but Ordinary – Emerging Splicing Mechanisms in Eukaryotic Gene Regulation. *Trends in Genetics*, *37*(4), 355–372. https://doi.org/10.1016/j.tig.2020.10.008

Gelfman, S., Cohen, N., Yearim, A., & Ast, G. (2013). DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon–intron structure. *Genome Research*, *23*(5), 789–799. https://doi.org/10.1101/gr.143503.112

Glasser, E., Maji, D., Biancon, G., Puthenpeedikakkal, A. M. K., Cavender, C. E., Tebaldi, T., Jenkins, J. L., Mathews, D. H., Halene, S., & Kielkopf, C. L. (2022). Pre-mRNA splicing factor U2AF2 recognizes distinct conformations of nucleotide variants at the center of the pre-mRNA splice site signal. *Nucleic Acids Research*, *50*(9), 5299–5312. https://doi.org/10.1093/nar/gkac287

Görnemann, J., Kotovic, K. M., Hujer, K., & Neugebauer, K. M. (2005). Cotranscriptional Spliceosome Assembly Occurs in a Stepwise Fashion and Requires the Cap Binding Complex. *Molecular Cell*, *19*(1), 53–63. https://doi.org/10.1016/j.molcel.2005.05.007

Grant, C. E., & Bailey, T. L. (2021). *XSTREME: Comprehensive motif analysis of biological sequence datasets* [Preprint]. Bioinformatics. https://doi.org/10.1101/2021.09.02.458722

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Guo, M., Lo, P. C., & Mount, S. M. (1993). Species-specific signals for the splicing of a short Drosophila intron in vitro. *Molecular and Cellular Biology*, *13*(2), 1104–1118. https://doi.org/10.1128/MCB.13.2.1104

Guo, M., & Mount, S. M. (1995). Localization of Sequences Required for Size-specific Splicing of a Small Drosophila Intron in Vitro. *Journal of Molecular Biology*, *253*(3), 426–437.

Guo, W., Chung, W.-Y., Qian, M., Pellegrini, M., & Zhang, M. Q. (2014). Characterizing the strand-specific distribution of non-CpG methylation in human pluripotent cells. *Nucleic Acids Research*, *42*(5), 3009–3016. https://doi.org/10.1093/nar/gkt1306

Hanson, H. E., & Liebl, A. L. (2022). The Mutagenic Consequences of DNA Methylation within and across Generations. *Epigenomes*, *6*(4), 33. https://doi.org/10.3390/epigenomes6040033

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hashim, F. A., Mabrouk, M. S., & Al-Atabany, W. (2019). *Review of Different Sequence Motif Finding Algorithms*. *11*(2), 19.

Hatton, A. R., Subramaniam, V., & Lopez, A. J. (1998). Generation of Alternative Ultrabithorax Isoforms and Stepwise Removal of a Large Intron by Resplicing at Exon–Exon Junctions. *Molecular Cell*, *2*(6), 787–796. https://doi.org/10.1016/S1097-2765(00)80293-2

Hayes, B. (1998). Computing Science: The Invention of the Genetic Code. *American Scientist*, *86*, 8–14.

Herzel, L., Ottoz, D. S. M., Alpert, T., & Neugebauer, K. M. (2017). Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. *Nature Reviews Molecular Cell Biology*, *18*(10), 637–650. https://doi.org/10.1038/nrm.2017.63

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., Single-cell Best Practices Consortium, Aliee, H., Ansari, M., Badia-i-Mompel, P., Büttner, M., Dann, E., Dimitrov, D., Dony, L.,

Frishberg, A., … Theis, F. J. (2023). Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, *24*(8), 550–572. https://doi.org/10.1038/s41576-023-00586-w

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jeong, S. (2017). SR Proteins: Binders, Regulators, and Connectors of RNA. *Molecules and Cells*, *40*(1), 1–9. https://doi.org/10.14348/molcells.2017.2319

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., & Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, *21*(9), 1543–1551. https://doi.org/10.1101/gr.121095.111

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006.

Konarska, M. M., Grabowski, P. J., Padgett, R. A., & Sharp, P. A. (1985). Characterization of the branch site in lariat RNAs produced by splicing of mRNA precursors. *Nature*, *313*(6003), 552–557. https://doi.org/10.1038/313552a0

Kotovic, K. M., Lockshon, D., Boric, L., & Neugebauer, K. M. (2003). Cotranscriptional Recruitment of the U1 snRNP to Intron-Containing Genes in Yeast. *Molecular and Cellular Biology*, *23*(16), 5768–5779. https://doi.org/10.1128/MCB.23.16.5768-5779.2003

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P. A., & Burge, C. B. (2014). RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell*, *54*(5), 887–900. https://doi.org/10.1016/j.molcel.2014.04.016

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Latchman, D. S. (1997). Transcription factors: An overview. *The International Journal of Biochemistry & Cell Biology*, *29*(12), 1305–1312. https://doi.org/10.1016/S1357-2725(97)00085-X

Lee, A. S. Y., Kranzusch, P. J., & Cate, J. H. D. (2015). eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. *Nature*, *522*(7554), 111–114. https://doi.org/10.1038/nature14267

Lee, T. I., & Young, R. A. (2013). Transcriptional Regulation and Its Misregulation in Disease. *Cell*, *152*(6), 1237–1251. https://doi.org/10.1016/j.cell.2013.02.014

Lee, Y., & Rio, D. C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry*, *84*(1), 291–323. https://doi.org/10.1146/annurev-biochem-060614-034316

Lev Maor, G., Yearim, A., & Ast, G. (2015a). The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, *31*(5), 274–280. https://doi.org/10.1016/j.tig.2015.03.002

Lev Maor, G., Yearim, A., & Ast, G. (2015b). The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, *31*(5), 274–280. https://doi.org/10.1016/j.tig.2015.03.002

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li-Byarlay, H., Li, Y., Stroud, H., Feng, S., Newman, T. C., Kaneda, M., Hou, K. K., Worley, K. C., Elsik, C. G., Wickline, S. A., Jacobsen, S. E., Ma, J., & Robinson, G. E. (2013). RNA interference knockdown of *DNA methyl-transferase 3* affects gene alternative splicing in the honey bee. *Proceedings of the National Academy of Sciences*, *110*(31), 12750–12755. https://doi.org/10.1073/pnas.1310735110

Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J., & Fairbrother, W. G. (2011). Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proceedings of the National Academy of Sciences*, *108*(27), 11093–11098. https://doi.org/10.1073/pnas.1101135108

Lin, Y., Li, F., Huang, L., Polte, C., Duan, H., Fang, J., Sun, L., Xing, X., Tian, G., Cheng, Y., Ignatova, Z., Yang, X., & Wolf, D. A. (2020). eIF3 Associates with 80S Ribosomes to Promote Translation Elongation, Mitochondrial Homeostasis, and Muscle Health. *Molecular Cell*, *79*(4), 575-587.e7. https://doi.org/10.1016/j.molcel.2020.06.003

Llorian, M., Schwartz, S., Clark, T. A., Hollander, D., Tan, L.-Y., Spellman, R., Gordon, A., Schweitzer, A. C., De La Grange, P., Ast, G., & Smith, C. W. J. (2010). Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nature Structural & Molecular Biology*, *17*(9), 1114–1123. https://doi.org/10.1038/nsmb.1881

Long, M., & Deutsch, M. (1999). Intron—Exon structures of eukaryotic model organisms. *Nucleic Acids Research*, *27*(15), 3219–3228. https://doi.org/10.1093/nar/27.15.3219

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Matsutani, T., Ueno, Y., Fukunaga, T., & Hamada, M. (2019). Discovering novel mutation signatures by latent Dirichlet allocation with variational Bayes inference. *Bioinformatics*, *35*(22), 4543–4552. https://doi.org/10.1093/bioinformatics/btz266

Maunakea, A. K., Chepelev, I., Cui, K., & Zhao, K. (2013). Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Research*, *23*(11), 1256–1269. https://doi.org/10.1038/cr.2013.110

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Ong, C.-T., & Corces, V. G. (2014). CTCF: An architectural protein bridging genome topology and function. *Nature Reviews Genetics*, *15*(4), 234–246. https://doi.org/10.1038/nrg3663

Pai, A. A., Paggi, J. M., Yan, P., Adelman, K., & Burge, C. B. (2018). Numerous recursive sites contribute to accuracy of splicing in long introns in flies. *PLOS Genetics*, *14*(8), e1007588. https://doi.org/10.1371/journal.pgen.1007588

Palangat, M., Anastasakis, D. G., Fei, D. L., Lindblad, K. E., Bradley, R., Hourigan, C. S., Hafner, M., & Larson, D. R. (2019). The splicing factor U2AF1 contributes to cancer progression through a noncanonical role in translation regulation. *Genes & Development*, *33*(9–10), 482–497. https://doi.org/10.1101/gad.319590.118

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

Pohl, M., Bortfeldt, R. H., Grützmann, K., & Schuster, S. (2013). Alternative splicing of mutually exclusive exons—A review. *Biosystems*, *114*(1), 31–38. https://doi.org/10.1016/j.biosystems.2013.07.003

Price, A. L., Eskin, E., & Pevzner, P. A. (2004). Whole-genome analysis of *Alu* repeat elements reveals complex evolutionary history. *Genome Research*, *14*(11), 2245–2252. https://doi.org/10.1101/gr.2693004

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Reyes, L., Kois, P., Konforti, B. B., & Konarska, M. M. (1996). The canonical GU dinucleotide at the 5' splice site is recognized by ~220 of the lJ5 snRNP within the spliceosome. *RNA*, *2*(3), 213–215.

Rittiner, J., Cumaran, M., Malhotra, S., & Kantor, B. (2022). Therapeutic modulation of gene expression in the disease state: Treatment strategies and approaches for the development of next-generation of the epigenetic drugs. *Frontiers in Bioengineering and Biotechnology*, *10*, 1035543. https://doi.org/10.3389/fbioe.2022.1035543

Rosner, F., Hinneburg, A., Roder, M., Nettling, M., & Both, A. (2014). *Evaluating topic coherence measures*.

Rube, H. T., Rastogi, C., Feng, S., Kribelbauer, J. F., Li, A., Becerra, B., Melo, L. A. N., Do, B. V., Li, X., Adam, H. H., Shah, N. H., Mann, R. S., & Bussemaker, H. J. (2022). Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. *Nature Biotechnology*. https://doi.org/10.1038/s41587-022-01307-0

Schroeder, H. W., & Cavacini, L. (2010). Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology*, *125*(2), S41–S52. https://doi.org/10.1016/j.jaci.2009.09.046

Seiler, M., Peng, S., Agrawal, A. A., Palacino, J., Teng, T., Zhu, P., Smith, P. G., Buonamici, S., Yu, L., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., Kasapi, M., Ferguson, M. L., Hutter, C. M., Sofia, H. J., Tarnuzzer, R., Wang, Z., Yang, L., … Mariamidze, A. (2018). Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell Reports*, *23*(1), 282-296.e4. https://doi.org/10.1016/j.celrep.2018.01.088

Shao, C., Yang, B., Wu, T., Huang, J., Tang, P., Zhou, Y., Zhou, J., Qiu, J., Jiang, L., Li, H., Chen, G., Sun, H., Zhang, Y., Denise, A., Zhang, D.-E., & Fu, X.-D. (2014). Mechanisms for U2AF to define 3′ splice sites and regulate alternative splicing in the human genome. *Nature Structural & Molecular Biology*, *21*(11), 997–1005. https://doi.org/10.1038/nsmb.2906

Shenasa, H., & Bentley, D. L. (2023). Pre-mRNA splicing and its cotranscriptional connections. *Trends in Genetics*, *39*(9), 672–685. https://doi.org/10.1016/j.tig.2023.04.008

Shepard, S., McCreary, M., & Fedorov, A. (2009). The Peculiarities of Large Intron Splicing in Animals. *PLoS ONE*, *4*(11), e7853. https://doi.org/10.1371/journal.pone.0007853

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., & Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, *479*(7371), 74–79. https://doi.org/10.1038/nature10442

Sibley, C. R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M. E., Hardy, J., Modic, M., Curk, T., Wilson, S. W., Plagnol, V., & Ule, J. (2015). Recursive splicing in long vertebrate genes. *Nature*, *521*(7552), 371–375. https://doi.org/10.1038/nature14466

Soares, L. M. M., Zanier, K., Mackereth, C., Sattler, M., & Valcárcel, J. (2006). Intron Removal Requires Proofreading of U2AF/3' Splice Site Recognition by DEK. *Science*, *312*(5782), 1961–1965. https://doi.org/10.1126/science.1128659

Stagsted, L. V. W., O'Leary, E. T., Ebbesen, K. K., & Hansen, T. B. (2021). The RNA-binding protein SFPQ preserves long-intron splicing and regulates circRNA biogenesis in mammals. *eLife*, *10*, e63088. https://doi.org/10.7554/eLife.63088

Starck, S. R., Tsai, J. C., Chen, K., Shodiya, M., Wang, L., Yahiro, K., Martins-Green, M., Shastri, N., & Walter, P. (2016). Translation from the 5′ untranslated region shapes the integrated stress response. *Science*, *351*(6272), aad3867. https://doi.org/10.1126/science.aad3867

Stormo, G. D. (2000). DNA binding sites: Representation and discovery. *Bioinformatics*, *16*(1), 16–23. https://doi.org/10.1093/bioinformatics/16.1.16

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, *14*, 117793221989905. https://doi.org/10.1177/1177932219899051

Taggart, A. J., Lin, C.-L., Shrestha, B., Heintzelman, C., Kim, S., & Fairbrother, W. G. (2017). Large-scale analysis of branchpoint usage across species and cell lines. *Genome Research*, *27*(4), 639–649. https://doi.org/10.1101/gr.202820.115

Takeda, S., Clemens, P. R., & Hoffman, E. P. (2021). Exon-Skipping in Duchenne Muscular Dystrophy. *Journal of Neuromuscular Diseases*, *8*(s2), S343–S358. https://doi.org/10.3233/JND-210682

Tareen, A., & Kinney, J. B. (2020). Logomaker: Beautiful sequence logos in Python. *Bioinformatics*, *36*(7), 2272–2274. https://doi.org/10.1093/bioinformatics/btz921

Tataru, C., Peras, M., Rutherford, E., Dunlap, K., Yin, X., Chrisman, B. S., DeSantis, T. Z., Wall, D. P., Iwai, S., & David, M. M. (2023). Topic modeling for multi-omic integration in the human gut microbiome and implications for Autism. *Scientific Reports*, *13*(1), 11353. https://doi.org/10.1038/s41598-023-38228-0

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. https://doi.org/10.1198/016214506000000302

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Thomas, P. D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L., & Mi, H. (2022). PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science*, *31*(1), 8–22. https://doi.org/10.1002/pro.4218

Trcek, T., Larson, D. R., Moldón, A., Query, C. C., & Singer, R. H. (2011). Single-Molecule mRNA Decay Measurements Reveal Promoter- Regulated mRNA Stability in Yeast. *Cell*, *147*(7), 1484–1497. https://doi.org/10.1016/j.cell.2011.11.051

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018). SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, *19*(1), 40. https://doi.org/10.1186/s13059-018-1417-1

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B. J., & Darnell, R. B. (2006). An RNA map predicting Nova-dependent splicing regulation. *Nature*, *444*(7119), 580–586. https://doi.org/10.1038/nature05304

Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J.-Y., Cody, N. A. L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L. P. B., Bergalet, J., Duff, M. O., Garcia, K. E., Gelboin-Burkhart, C., … Yeo, G. W. (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, *583*(7818), 711–719. https://doi.org/10.1038/s41586-020-2077-3

Van Nostrand, E. L., Pratt, G. A., Yee, B. A., Wheeler, E. C., Blue, S. M., Mueller, J., Park, S. S., Garcia, K. E., Gelboin-Burkhart, C., Nguyen, T. B., Rabano, I., Stanton, R., Sundararaman, B., Wang, R., Fu, X.-D., Graveley, B. R., & Yeo, G. W. (2020). Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biology*, *21*(1), 90. https://doi.org/10.1186/s13059-020-01982-9

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van Der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wallace, J. C., & Edmonds, M. (1983). Polyadenylylated nuclear RNA contains branches. *Proceedings of the National Academy of Sciences*, *80*(4), 950–954. https://doi.org/10.1073/pnas.80.4.950

Wan, Y., Anastasakis, D. G., Rodriguez, J., Palangat, M., Gudla, P., Zaki, G., Tandon, M., Pegoraro, G., Chow, C. C., Hafner, M., & Larson, D. R. (2021). Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. *Cell*, *184*(11), 2878-2895.e20. https://doi.org/10.1016/j.cell.2021.04.012

Wang, G.-S., & Cooper, T. A. (2007). Splicing in disease: Disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, *8*(10), 749–761. https://doi.org/10.1038/nrg2164

Wang, Z., & Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, *14*(5), 802–813. https://doi.org/10.1261/rna.876308

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wilhelm, B. T., & Landry, J.-R. (2009). RNA-Seq—Quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, *48*(3), 249–257. https://doi.org/10.1016/j.ymeth.2009.03.016

Wilkinson, M. E., Charenton, C., & Nagai, K. (2020). RNA Splicing by the Spliceosome. *Annual Review of Biochemistry*, *89*(1), 359–388. https://doi.org/10.1146/annurev-biochem-091719-064225

Will, C. L., & Luhrmann, R. (2011). Spliceosome Structure and Function. *Cold Spring Harbor Perspectives in Biology*, *3*(7), a003707–a003707. https://doi.org/10.1101/cshperspect.a003707

Wu, T., & Fu, X.-D. (2015). Genomic functions of U2AF in constitutive and regulated splicing. *RNA Biology*, *12*(5), 479–485. https://doi.org/10.1080/15476286.2015.1020272

Xia, X. (2012). Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction. *Scientifica*, *2012*, 1–15. https://doi.org/10.6064/2012/917540

Yan, C., Wan, R., & Shi, Y. (2019). Molecular Mechanisms of pre-mRNA Splicing through Structural Biology of the Spliceosome. *Cold Spring Harbor Perspectives in Biology*, *11*(1), a032409. https://doi.org/10.1101/cshperspect.a032409

Yang, F., Bian, T., Zhan, X., Chen, Z., Xing, Z., Larsen, N. A., Zhang, X., & Shi, Y. (2023). Mechanisms of the RNA helicases DDX42 and DDX46 in human U2 snRNP assembly. *Nature Communications*, *14*(1), 897. https://doi.org/10.1038/s41467-023-36489-x

Yearim, A., Gelfman, S., Shayevitch, R., Melcer, S., Glaich, O., Mallm, J.-P., Nissim-Rafinia, M., Cohen, A.-H. S., Rippe, K., Meshorer, E., & Ast, G. (2015). HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. *Cell Reports*, *10*(7), 1122–1134. https://doi.org/10.1016/j.celrep.2015.01.038

Yıldırım, B., & Vogl, C. (2023). Purifying selection against spurious splicing signals contributes to the base composition evolution of the polypyrimidine tract. *Journal of Evolutionary Biology*, *36*(9), 1295–1312. https://doi.org/10.1111/jeb.14205

Yoon, B.-J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, *10*(6), 402–415. https://doi.org/10.2174/138920209789177575

Young, J. I., Hong, E. P., Castle, J. C., Crespo-Barreto, J., Bowman, A. B., Rose, M. F., Kang, D., Richman, R., Johnson, J. M., Berget, S., & Zoghbi, H. Y. (2005). Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2. *Proceedings of the National Academy of Sciences*, *102*(49), 17551–17558. https://doi.org/10.1073/pnas.0507856102

Yun, J., Wang, T., & Xiao, G. (2014). Bayesian hidden Markov models to identify RNA–protein interaction sites in PAR-CLIP. *Biometrics*, *70*(2), 430–440. https://doi.org/10.1111/biom.12147

Zehir, A., Benayed, R., Shah, R. H., Syed, A., Middha, S., Kim, H. R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S. M., Hellmann, M. D., Barron, D. A., Schram, A. M., Hameed, M., Dogan, S., Ross, D. S., Hechtman, J. F., DeLair, D. F., Yao, J., … Berger, M. F. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature Medicine*, *23*(6), 703–713. https://doi.org/10.1038/nm.4333

Zeng, Y., Fair, B. J., Zeng, H., Krishnamohan, A., Hou, Y., Hall, J. M., Ruthenburg, A. J., Li, Y. I., & Staley, J. P. (2022). Profiling lariat intermediates reveals genetic determinants of

early and late co-transcriptional splicing. *Molecular Cell*, *82*(24), 4681-4699.e8. https://doi.org/10.1016/j.molcel.2022.11.004

Zhao, B., Li, Z., Qian, R., Liu, G., Fan, M., Liang, Z., Hu, X., & Wan, Y. (2022). Cancer-associated mutations in *SF3B1* disrupt the interaction between *SF3B1* and *DDX42*. *The Journal of Biochemistry*, *172*(2), 117–126. https://doi.org/10.1093/jb/mvac049

Zid, B. M., & O'Shea, E. K. (2014). Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast. *Nature*, *514*(7520), 117–121. https://doi.org/10.1038/nature13578