AQR Plays a Key Role in Splicing Fidelity in Higher Organisms

Meng Li (✉ limeng49631@aliyun.com)

CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences

Research article

Keywords: AQR, cryptic 3' splice site, cryptic 5' splice site, cryptic exon, splicing fidelity

Posted Date: March 31st, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-20141/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

The structures of the spliceosome have been resolved in different complexes, but the functions of spliceosome components in splicing fidelity are largely unclear. A few proteins, such as *SLU7* and EJC, are known to have roles in splicing fidelity.

Results

In this work, through a systematic analysis of the cryptic splice sites in humans, mice and fruit flies using shRNA-seq and siRNA-seq datasets, I ranked the importance of different proteins in splicing fidelity. The important proteins include IBC, EJC, *SLU7*, *Prp22* and other splicing-related factors, especially *AQR*, whose knockdown induced the highest number of cryptic splice sites in K562, N2A, and CGR8, suggesting its central role in splicing fidelity in higher organisms. The metaprofiles of the cryptic splice sites for *AQR* and *Prp17* further indicate that cryptic 3' splice sites can induce cryptic exons, which then contribute to cryptic 5' splice site generation.

Conclusions

I analyzed siRNA-seq/shRNA-seq from a new angle. I found that *AQR* knockdown induced the highest number of cryptic 5' and 3' splice sites in three cell lines and that RBP knockdown mainly induced cryptic 3' splice sites, which then induced cryptic 5' splice sites.

Background

Extensive data on the spliceosome structure have been reported due to the development of Cryo-EM. Different stages of the spliceosome complex have been resolved in both yeast and humans [1, 2]. Although the interaction of spliceosome components and their structures in different stages have been elucidated, their detailed functions in splicing regulation are unclear. The functional studies have lagged largely behind the structural analyses [3].

One major function of the spliceosome is ensuring splicing fidelity (or splicing proofreading) [4], and cryptic splice site usage directly reflects the fidelity of splicing. Both mutations in the cis-elements and mutation/knockdown of trans-factors can induce cryptic splice sites. Several studies have shown the roles of some RBPs in splice site recognition, e.g., *SLU7* participates in the second step of the splicing reaction, and knockdown of *SLU7* induces cryptic 3' splice sites [5]. Another example is the EJC; two studies recently showed that EJC's components protect the exons, and their knockdown-induced cryptic 3' and 5' splice sites occur mainly in exon regions [6, 7]. Although several roles of RBPs in splicing fidelity, such as those of *PRP22* and the abovementioned EJC, are well known, the functional roles of most RBPs in splice site recognition are unclear. Thus, a systematic study on cryptic splice sites is needed to determine the role of different RBPs in splicing fidelity.

Here, I analyzed siRNA-seq and shRNA-seq in humans, mice and fruit flies using thousands of NGS datasets. The results showed that several RBPs, including IBC (*ISY1*, *CRNKL1*, *ZNF830*, *SYF1*, *AQR*), EJC (*EIF4A3*, *MAGOH*, *RBM8A*, *CASC3*), poly-U (poly-T) binding proteins (*hnRNP C*, *U2AF2* and *PUF60*), *SLU7* and *Prp22*, play an important role in splicing fidelity. In addition, through meta-analyses of cryptic 5' splice sites, cryptic 3' splice sites, and cryptic exons, I found that RBP knockdown mainly induced cryptic 3' splice sites and cryptic 3' splice sites can induce cryptic exons, which then induce cryptic 5' splice sites.

Results

AQR knockdown induced the highest number of cryptic splice sites

Five cell lines from three organisms were used here to analyze cryptic splice sites. The five cell lines are K562, HepG2, S2, N2A, and CGR8. The shRNA-seq/siRNA-seq data for K562, HepG2 and the fruit fly S2 cell lines were directly obtained from the ENCODE project. The mouse N2A and CGR8 cell lines contain ~1,400 RBP/TF siRNA-seq datasets, which are pooled sequencing data, and 52 cassette splicing events were sequenced [8]. The general workflow is shown in Figure 1A. For each shRNA-seq/siRNA-seq dataset, I used STAR aligner [9] to obtain all the junctions (both new junctions and annotated junctions). Then, these junctions were converted into 5' splice sites and 3' splice sites for each shRNA-seq/siRNA-seq. For exclusion of the known junctions and noisy background junctions, these collected 5' splice sites and 3' splice sites were then filtered by 5' splice sites and 3' splice sites from gene annotation and all the control RNA-seq datasets (Figure 1A). The same workflow was performed for the other cell lines.

Among the 241 RBPs' shRNA-seq datasets in K562, knockdown of *AQR* induced the highest number of cryptic splice sites (Figure 1B). In addition to that of *AQR*, knockdown of several other RBPs induced a relatively higher number of cryptic 3' splice sites and cryptic 5' splice sites: EJC components, SF3A and SF3B complex components, the spliceosome core protein *RRPF8*, the poly-U binding proteins *hnRNP C*, *U2AF2* and *PUF60*, and the m6A reader *FTO* (Additional file 1). Similar results were obtained using the HepG2 cell line (Figure 1C and Additional file 2).

The mouse dataset contains ~1,400 siRNA-seq datasets and almost all spliceosome components. Among the IBC components (*AQR*, *ISY1*, *SYF1*, *ZNF830*, and *CRNKL1*), three of them are the top genes in the datasets when considering cryptic 3' splice sites (Figure 1D). Other top genes included *SLU7* and *DHX8/PRP22* (Additional file 3). For cryptic 5' splice sites, *AQR* was still the top gene (Figure 1E). The 2nd highest was *Med26*, which belongs to the mediator complex (Additional file 3). One study reported that mediator and polymerase II can influence cryptic 5' splice site recognition in yeast [10]. The above results were consistent between the two cell lines (N2A and CGR8, please see Additional file 3).

Previous studies on the ENCODE fruit fly siRNA-seq dataset showed that spliceosome component knockdown caused a smaller PSI change than knockdown of alternative splicing regulation RBPs [11], while the above results showed that cryptic splice sites are mainly caused by knockdown of spliceosome components. Thus, I further tested whether spliceosome component knockdown induced the highest

number of cryptic splice sites in this fly siRNA-seq dataset, which may indicate that spliceosome components are more related to splicing fidelity than alternative splicing regulation. The results showed that RBPs whose knockdown induced the highest number of cryptic sites are spliceosome components in fruit flies, as shown in Figure 1F and Additional file 4. *Pea/DHX8/Prp22* knockdown induced the highest numbers of cryptic 3' splice sites (Additional file 4), and this protein knockdown also induced one of the highest numbers of cryptic 3' splice sites in the mouse dataset, as shown above.

Sequence and motif properties of the cryptic splice sites

Next, I analyzed the properties of these cryptic sites. The number of cryptic 3' splice sites was greater than the number of cryptic 5' splice sites for all RBPs in the K562 and HepG2 cells (Figure 2A). This result suggests that 3' splice sites undergo more regulation than 5' splice sites by the RBPs studied here. I calculated the maxEntScore [12] for three different types of splice sites: all AG/GT splice sites in introns, AG/GT splice sites that are used when proteins are knocked down, and canonical intron splice sites. The results indicated that sequence preference has a strong effect on cryptic splice site usage (Figure 2B&C). I then performed motif analysis for all the RBPs' cryptic 5' splice sites and cryptic 3' splice sites for K562, and all the motifs were similar to AGGTAAG. The corresponding sequence in U1 snRNA is UCCAUUC, so all the motifs around the cryptic 5' splice sites are complementary to the U1 snRNA (Figure 2D&E). The complementation between the cryptic 5' splice site motif and U1 snRNA indicated that 5' splice sites are mostly regulated by U1 snRNA. For cryptic 3' splice sites, only the motifs of *hnRNP C* contained an obvious poly-U. This result is consistent with one previous study on *hnRNP C* competing with *U2AF2* to repressing the cryptic 3' splice sites in the poly-U region of ALU [13]. Poly-U is bound by *U2AF2* to initiate spliceosome assembly, while the lack of poly-U in the cryptic 3' splice site motifs suggests that splicing fidelity regulation by these RBPs mainly occurs after *U2AF2* participates in the spliceosome. A consistent motif result was observed using the fruit fly dataset. The cryptic splice sites induced by knockdown of most RBPs are strongly enriched near the canonical 3' or 5' splice sites (Additional file 5). The result of cryptic sites' positions near canonical sites suggest they usually don't disrupt exon definition and this agreed with that RBPs mainly regulate splice sites after *U2AF2* participated in spliceosome when near canonical splice sites.

Some of the RBPs used above may not be directly involved in splicing fidelity, and they may act as a regulation layer for other RBPs. To find the potential similar functions and potential regulation between the RBPs in splicing fidelity, I overlapped the cryptic junctions between the shRNA-seq dataset of every two RBPs and performed a Poisson test. As shown in Additional file 1, junctions overlapping within EJC's components showed very significant p-values, and junctions overlapping between *U2AF1* and *U2AF2* also showed very significant p-values. Similar results were obtained using the HepG2 cell line (Additional file 2). Overlapping junctions were also determined for the N2A and CGR8 cell lines to test if cryptic junctions were stable between these cell lines. This result showed that junctions overlapping within the components of IBC had the highest ranks (Additional file 3).

Analyzing the cryptic splice sites in different regions

The regulation mechanism of splicing is different in deep introns and near canonical splice sites [14]. Thus I further classify the cryptic splice sites into different regions. I analyzed cryptic splice sites in ALU elements, cryptic splice sites in exons, cryptic splice sites near canonical splice sites (< 200bp) and cryptic splice sites that are in deep introns (distance from canonical splice sites > 200bp) respectively. Knockdown of *HnRNP C* induced the highest number of cryptic 3' splice sites in ALU elements (Additional file 1&2). Knockdown of *RAVER1* and *SRSF3* induced the highest number of cryptic 3' splice sites in deep introns in HepG2 cells (Additional file 2). When considering cryptic splice site motifs in deep introns, most RBPs' cryptic 3' splice site motifs contained more obvious poly-U, and cryptic 5' splice site motifs are more complementary with U1 snRNA (Additional file 5). In total, this analysis suggests that in deep introns, cryptic 3' splice sites need poly-U to initiate spliceosome assemble. While around canonical splice sites, cryptic 3' splice sites shared the poly-U of the canonical splice sites.

Metaprofiling of cryptic exons and cryptic splice sites suggests that cryptic 5' splice sites are induced by cryptic exons

Exons can be accurately identified by sequencing reads that span two or more junctions [6] (Figure 3A). I used this method to evaluate all shRNA-seq datasets and found that *AQR* knockdown still induced the highest number of cryptic exons (Figure 3B). To determine why *AQR* knockdown simultaneously induced the highest number of cryptic 5' splice sites, cryptic 3' splice sites, and cryptic exons, I generated the metaprofiles of the three and aligned them in Figure 3C. The cryptic exon 5' positions are highly consistent with the cryptic 5' splice sites. This result suggests that cryptic 5' splice sites are induced by cryptic exons. Note that cryptic exons are much harder to detect than cryptic splice sites since one read must contain at least two junctions to detect a cryptic exon, and the read length constraints must be considered (100 bp for ENCODE shRNA-seq). The similarity between the motifs of cryptic 5' splice sites and cryptic exons' 5' splice sites further indicates that they are produced by the same mechanism (Figure 3C).

To further test this, I generated the above profiles for *PRP17* in HepG2 cells, whose knockdown (by CRISPR) induced one of the highest numbers of cryptic 3' splice sites in HepG2 cells (Additional file 2). A similar profile was observed compared with that of *AQR* in K562 cells (Figure 3D). The motifs of the cryptic exons' 5' splice sites after knockdown of *AQR* and *PRP17* are highly complementary to the U1 snRNA recognition motif.

Discussion

Higher organisms contain much longer introns than yeast, and how the spliceosome recognizes the correct splice sites among the much greater number of cryptic splice sites is unknown. In this work, through a systematic analysis of cryptic splice sites in three organisms, I found that several proteins play a role in this process.

Splicing fidelity and *AQR*

Among the RBPs whose knockdown induced the cryptic splice sites mentioned above, some molecules' mechanisms are relative clear. EJC and EJC peripheral bind the exon region and prevent the exon from re-splicing [6, 7], *SLU7* fixes upstream exon in appropriate position in spliceosome catalyzation [5], *SF3B1* binds and regulates the branch point position [15], *hnRNP C* competes with *U2AF2* to regulate 3' splice sites [13], and *PRP22* proofreads splicing by competing with exon ligation [4]. *PRP16* and *RBM17* have also been reported to influence splicing fidelity [4, 16], but their knockdown didn't induce significant number of cryptic splice sites. Some others molecules, especially IBC, whose knockdown induced the highest number of cryptic splice sites in mice and humans, have unclear mechanisms. In an extreme example of *RPL11*, almost all the AG in this intron near canonical 3' splice site were used as cryptic 3' splice sites after knockdown of *AQR* (Figure 4A). IBC joins the spliceosome during B^{act} complex [17], so it regulates splicing fidelity after B complex formed. Previous studies found that *AQR* knockdown can disrupt snoRNA metabolism[18]. I further found that snoRNA host introns naturally contain a much greater number of branch points than other introns (Figure 4B). This result suggests that *AQR* may be related to branch site position regulation. In addition, *AQR* in IBC is one of the splicing factors (the only RNA helicase missing) that is missing in *S. cerevisiae*'s spliceosome but exists in *S. pombe*'s spliceosome (Additional file 5), suggesting its potential role in the much larger number of introns in *S. pombe* (~5,000 introns) than *S. cerevisiae* (~250 introns).

IBC and EJC play an important role in splicing fidelity and both are peripheral to the spliceosome

EJC components' cryptic 5' splice sites and cryptic 3' splice sites are usually enriched in exon regions [6, 7], also as shown in Additional file 5, and their knockdown shortens the exons. Cryptic splice sites induced by knockdown of *AQR* are enriched in intron regions (Additional file 5), and their knockdown shortens the introns. Both IBC and EJC are in the periphery of the spliceosome (Figure 4C). This result suggests that the periphery of the spliceosome is the same or more important than the interior of the spliceosome in splicing fidelity in mice and humans.

***AQR* knockdown induced cryptic 3' splice sites →induced cryptic exons →induced cryptic 5' splice sites**

AQR knockdown induced the highest number of cryptic exons through cryptic 3' splice sites, as shown in Figure 4D. When cryptic 3' splice sites are relatively far from canonical 3' splice sites, the new cryptic downstream exons will be too long, while the median length of the human exon is 136 bp. A longer exon may cause the cryptic 5' splice sites in the new exon regions to be recognized and induce cryptic exons. Overall, cryptic 3' splice sites induced cryptic exons, which then induced cryptic 5' splice sites. These results can also explain why all cryptic 5' splice site motifs are complementary to U1 snRNA because splicing factors mainly regulate 3' splice sites and cryptic 5' splice sites are induced by cryptic 3' splice sites.

Limitations

This work examines the importance of different RBPs in splicing fidelity. Although the results are stable between different organisms or different cell lines, the regulatory mechanism of most proteins analyzed

here is unclear. Thus, future works should elucidate the detailed mechanisms of different proteins in splicing fidelity and the relationship between 3' splice sites and exon definition.

Conclusions

Reads that support cryptic splice sites usually account no more than 0.1% of totally mapped reads in each siRNA-seq/shRNA-seq. Instead of differential gene expression and differential alternative splicing, I analyzed cryptic splice sites in three higher organisms. I found that *AQR*, which belongs to IBC, plays a key role in splicing fidelity. In addition, I found that knockdown of RBPs mainly induced cryptic 3' splice sites, which then induced cryptic exons and cryptic 5' splice sites. The important RBPs found here are stable between cell lines or organisms.

Methods

Datasets and annotations used

shRNA-seq for K562 and HepG2 cells was downloaded from the ENCODE project (<https://www.encodeproject.org/>) [19], and raw fastq files were used here. All the control shRNA-seq, control CRISPR-seq, and normal RNA-seq data were merged to further filter the cryptic splice sites. Mouse data were downloaded from NCBI SRA (SRP073198), and the siNT, nontargeting siRNA and mock controls were merged together to filter the cryptic splice sites. The fruit fly siRNA-seq dataset was downloaded from the modENCODE project [20]. Because only two control RNA-seq datasets exist in the siRNA-seq dataset of fruit flies, all the normal RNA-seq data in fruit flies available in ENCODE were used as controls to filter cryptic splice sites independent of the cell line. GENCODE hg19 annotation was used for human. GENCODE mm10 was used for mouse, and Dm6 from ENSEMBL [21] was used for fruit fly. The positions of ALU elements are downloaded from UCSC [22] and only kept ALU elements with length < 350 and > 250.

STAR and cryptic sites

The 2-pass mode was used for the STAR aligner to find cryptic junctions, and an overhang of at least 8 base pairs was set in the STAR to avoid spurious junctions. The STAR aligner can output 'SJ' files that contain all the junctions detected, nucleotides of the anchor sites and read count information, and the cryptic splice sites can be obtained by simply processing these files. At least two unique mapped reads were required to define a cryptic splice site for shRNA-seq and siRNA-seq in the human and fruit fly datasets. The mouse data were from a low coverage dataset due to pooled sequencing, and only 1 unique mapped read was required to define a cryptic site. For all three organisms, at least 1 unique mapped read was required to define a splice site for the control RNA-seq in the three organisms. For each siRNA-seq/shRNA-seq, the splice sites of all the controls were merged together to filter noisy background splice sites. I filtered the 3' and 5' splice sites by their canonical nucleotides, I mainly concentrated on the

GT-AG type when performing downstream analysis, and over 99% of the cryptic splice sites are the GT-AG type.

Cryptic exons

After getting the aligned BAM files, cryptic exons were called from a custom Java program. The 'jl' tag in the STAR aligner contained the junction sites for each read. The Java program detected reads that contain two or more junctions ('jl' tag length >3) and output the junctions; the regions between these junction sites are either introns or exons. At least 1 unique read was required to define a cryptic exon, and 1 unique read was required to define exons in the control. The cryptic exons in each shRNA-seq dataset were obtained by subtracting the exons detected in the annotation and all the control RNA-seqs. The detailed analysis workflow is similar to the cryptic splice site analysis, as shown in Figure 1A; the only difference is the splice sites were replaced by exons.

Metaprofiling of cryptic splice sites

Metaprofiles were based on R and Bioconductor. Unique introns in GENCODE hg19 with lengths > 500 were used here to generate the metaprofiles. The branch point positions were obtained from one previous study [23].

Overlapping junctions and IDR

The cryptic junctions are junctions not in the control or the annotation and with at least two reads for support. The workflow to determine the cryptic junctions is similar to Figure 1A, but spliced sites were replaced with junctions. Fisher's exact test will be biased when the number of junctions is very large. Therefore, the Poisson test was used to test whether the overlapping cryptic junctions between the two proteins were significant. The detailed formula is as follows:

$$\text{Null hypothesis: } A\&B \sim \text{Poisson}\left(A \times \frac{B}{\text{total cryptic junctions}}\right)$$
$$\text{p-value} = P\left(A\&B > A \times \frac{B}{\text{total cryptic junctions}}\right)$$

where A is the number of cryptic junctions after knockdown of protein A and B is the number of cryptic junctions after knockdown of protein B. A&B is the number of same overlapping cryptic junctions between knockdown of protein A and B. Total cryptic junctions is the number of all unique cryptic junctions in this cell line.

IDR was calculated for the mouse data to find the most stable proteins that influence splicing fidelity in the two cell lines [24]. The R package called 'idr' performed this analysis.

Abbreviations

ALU: A kind of short repeat element

EJC: Exon junction complex

IBC: Intron binding complex

IDR: Irreproducible discovery rate

NGS: Next-generation sequencing

PSI: Percent spliced in

RBP: RNA binding protein

Declarations

Ethics approval and consent to participate

No ethics approval was required for the study.

Consent for publication

Not applicable.

Acknowledgments

The author is grateful to Dr. Zefeng Wang for his key suggestions and support for this work. The author also thanks Yun Yang for his insightful discussion and other Wang lab members for their helpful discussion. The author is grateful to the producer of the datasets for sharing their published data freely online, thus making this analysis possible. I acknowledge the ENCODE Consortium and the ENCODE production laboratories, which generated the RNA-seq datasets I analyzed here.

Competing interests

The author declares no competing interest.

Funding

None.

Availability of data and materials

All the major datasets or results generated or analyzed during this study are included in this published article. Source code and other intermediate result are available in [25].

Authors' contributions

Meng Li designed the study, implemented the analysis, and wrote the manuscript.

References

1. Kastner B, Will CL, Stark H, Lührmann R: **Structural Insights into Nuclear pre-mRNA Splicing in Higher Eukaryotes.** *Cold Spring Harb Perspect Biol* 2019.
2. Shi Y: **Mechanistic insights into precursor messenger RNA splicing by the spliceosome.** *Nat Rev Mol Cell Biol* 2017, **18**(11):655-670.
3. Burke JE, Longhurst AD, Merkurjev D, Sales-Lee J, Rao B, Moresco JJ, Yates JR, Li JJ, Madhani HD: **Spliceosome Profiling Visualizes Operations of a Dynamic RNP at Nucleotide Resolution.** *Cell* 2018, **173**(4):1014-1030.e1017.
4. Semlow DR, Staley JP: **Staying on message: ensuring fidelity in pre-mRNA splicing.** *Trends Biochem Sci* 2012, **37**(7):263-273.
5. Chua K, Reed R: **The RNA splicing factor hSlu7 is required for correct 3' splice-site choice.** *Nature* 1999, **402**(6758):207-210.
6. Blazquez L, Emmett W, Faraway R, Pineda JMB, Bajew S, Gohr A, Haberman N, Sibley CR, Bradley RK, Irimia M *et al*: **Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing.** *Mol Cell* 2018, **72**(3):496-509.e499.
7. Boehm V, Britto-Borges T, Steckelberg AL, Singh KK, Gerbracht JV, Gueney E, Blazquez L, Altmüller J, Dieterich C, Gehring NH: **Exon Junction Complexes Suppress Spurious Splice Sites to Safeguard Transcriptome Integrity.** *Mol Cell* 2018, **72**(3):482-495.e487.
8. Han H, Braunschweig U, Gonatopoulos-Pournatzis T, Weatheritt RJ, Hirsch CL, Ha KCH, Radovani E, Nabeel-Shah S, Sterne-Weiler T, Wang J *et al*: **Multilayered Control of Alternative Splicing Regulatory Networks by Transcription Factors.** *Mol Cell* 2017, **65**(3):539-553.e537.
9. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29**(1):15-21.
10. Albulescu LO, Sabet N, Gudipati M, Stepankiw N, Bergman ZJ, Huffaker TC, Pleiss JA: **A quantitative, high-throughput reverse genetic screen reveals novel connections between Pre-mRNA splicing and 5' and 3' end transcript determinants.** *PLoS Genet* 2012, **8**(3):e1002530.
11. Brooks AN, Duff MO, May G, Yang L, Bolisetty M, Landolin J, Wan K, Sandler J, Booth BW, Celniker SE *et al*: **Regulation of alternative splicing in Drosophila by 56 RNA binding proteins.** *Genome Res* 2015, **25**(11):1771-1780.
12. Yeo G, Burge CB: **Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.** *J Comput Biol* 2004, **11**(2-3):377-394.
13. Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stevant I, Reyes A, Anders S, Luscombe NM, Ule J: **Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements.** *Cell* 2013, **152**(3):453-466.
14. Anna A, Monika G: **Splicing mutations in human genetic disorders: examples, detection, and confirmation.** *J Appl Genet* 2018, **59**(3):253-268.

15. Darman RB, Seiler M, Agrawal AA, Lim KH, Peng S, Aird D, Bailey SL, Bhavsar EB, Chan B, Colla S *et al*: **Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point.** *Cell Rep* 2015, **13**(5):1033-1045.
16. Lallena MJ, Chalmers KJ, Llamazares S, Lamond AI, Valcarcel J: **Splicing regulation at the second catalytic step by Sex-lethal involves 3' splice site recognition by SPF45.** *Cell* 2002, **109**(3):285-296.
17. Zhang X, Yan C, Zhan X, Li L, Lei J, Shi Y: **Structure of the human activated spliceosome in three conformational states.** *Cell Res* 2018, **28**(3):307-322.
18. Hirose T, Ideue T, Nagai M, Hagiwara M, Shu MD, Steitz JA: **A spliceosomal intron binding protein, IBP160, links position-dependent assembly of intron-encoded box C/D snoRNP to pre-mRNA splicing.** *Mol Cell* 2006, **23**(5):673-684.
19. Sundararaman B, Zhan L, Blue SM, Stanton R, Elkins K, Olson S, Wei X, Van Nostrand EL, Pratt GA, Huelga SC *et al*: **Resources for the Comprehensive Discovery of Functional RNA Elements.** *Mol Cell* 2016, **61**(6):903-913.
20. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF *et al*: **Identification of functional elements and regulatory circuits by Drosophila modENCODE.** *Science* 2010, **330**(6012):1787-1797.
21. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG *et al*: **Ensembl 2018.** *Nucleic Acids Res* 2018, **46**(D1):D754-D761.
22. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool.** *Nucleic Acids Res* 2004, **32**(Database issue):D493-496.
23. Pineda JMB, Bradley RK: **Most human introns are recognized via multiple and tissue-specific branchpoints.** *Genes Dev* 2018, **32**(7-8):577-591.
24. Qunhua L, B. BJ, Haiyan H, J. BP: **Measuring reproducibility of high-throughput experiments.** *Ann Appl Stat* 2011, **5**(3):1752-1779.
25. Li M: **Cryptic splice site code.** <https://github.com/limeng12/abs>. In., vol. Accessed 1 March 2020; 2020.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AdditionalFiles.zip](#)