RNA-seq Analysis Reveals Mode of Splicing Inefficiency: Coordinated Intron Retention

Jodi Lee[1], Carmelle Catamura[2], Dennis Mulligan[2], Angela Brooks[2], Melissa Jurica[3]
[1]Department of Chemistry and Biochemistry, University of California Santa Cruz, Santa Cruz, CA, USA
[2]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, USA
[3]Department of Molecular, Cell and Developmental Biology, University of California Santa Cruz, Santa Cruz, CA, USA

Acknowledgements:
I would like to thank Dr. Melissa Jurica for mentoring me on this project and the Jurica Lab for generating the dataset. Additionally, thank you to Carmelle Catamura, Dennis Mulligan, and Dr. Angela Brooks for performing upstream computational analysis and providing computational tool MESA. Finally, I would like to thank my family and friends for supporting me throughout my academic and scientific journey. It has been a long, yet rewarding journey!

Abstract:
Recently, an anti-tumor drug targeting a key component of the spliceosome has been identified: spliceostatin A (SSA). However, the mechanism of splicing interruption and regulation on gene expression due to SSA has not been characterized. To understand how SSA can inhibit cancer cell growth, RNA-seq data provides a rich resource for the detection of alternative splicing events. In this approach, an algorithm was built to test whether splicing inhibitors lead to a new mode of splicing inefficiency characterized by coordinated intron retention (IR) events across an entire transcript. Transcriptome wide analysis was performed to gain a deeper insight into full detection of IR events among all transcripts (rather than known IR events). Indeed, cells treated with SSA expressed a greater number of transcripts showing splicing inefficiency in contrast to non-treated cells. Unexpectedly, SSA RNA-seq captured a binary distinction of IR such that all introns of a transcript are retained/ coordinated or none are affected. A subset of gene transcripts involved in the innate immune response were observed to contain upregulated coordinated IR events. As a whole, results from this study offers a starting point to help guide chemotherapy development by offering insight into the effects on splicing.

Introduction:
Pre-mRNA splicing by the spliceosome is an essential step in regulating gene expression in eukaryotes. The spliceosome removes intervening sequences (introns) from gene transcripts and joins coding sequences (exons) to form mature mRNAs, which are later translated into proteins. As a mode of alternative splicing (AS), intron retention (IR) has been classified as splicing events in which introns are not removed from pre-mRNA and are retained in mature mRNAs (Figure 1). Despite being a major type of alternative splicing (AS), IR has previously been considered as 'noise'. Limitations in transcriptomic analysis failed to distinguish exon skipping and alternative 5' and 3'acceptor sites from intron retention events.

As a result, the role of intron retention (IR) in gene expression regulation had been overlooked. However, recent advances in RNA-seq analysis have discovered characteristic IR patterns. Accumulating evidence suggests IR has a role in regulating gene expression in pathways such as neuronal and germ cell differentiation and CD4+ T cell activation; IR is also associated with Alzheimer's disease and cancer[1].

Interestingly, a natural product called spliceostatin A (SSA) has been shown to inhibit the spliceosome by binding to SF3B1, a core component of the spliceosome[4]. As a result, it has been observed that the spliceosome U2snRNP base pairs with 'decoy' sequences 5' of the conventional BP, which weakens the overall stability and assembly of the spliceosome at the anchoring BP site[4]. Essential steps in 3'splice site (ss) recognition are affected causing potential interference with splicing altered protein and function[4]. Additionally, SSA has been shown to inhibit cancer growth, suggesting that interfering with the spliceosome can be beneficially effective when splicing becomes a rate-limiting step for high gene expression [5,6,7]. Thus, SSA offers a potential drug for cancer.

However, the mechanism by which the growth of cancer cells is selectively inhibited as a result of SSA

targeting the SF3B component remains unknown. Previous studies have shown that SF3B inhibitors appear to induce differential IR patterns that show splicing inefficiency[3]. Given these observations, a deeper understanding of how selectivity among specific transcripts, sequence elements, and their combinations is required to modulate differential drug sensitivity and perturb the viability of cancer cells.

Thus, transcriptomic analysis appears as an optimal approach to analyze sequence and gene transcript selectivity towards IR. However, only a small percentage of IR events have been classified by known IR sequences and from existing gene transcripts[1,3,4]. Moreover, alignment artifacts have led to staggering numbers of false positive events. Such a computational approach to accurately classify all IR events regardless of relative gene transcripts and position has not been created[1,4]. Therefore, by creating an algorithm to quantify all IR events, **I hypothesize that splicing inhibitors lead to a new mode of splicing inefficiency characterized by coordinated IR events across an entire transcript.**

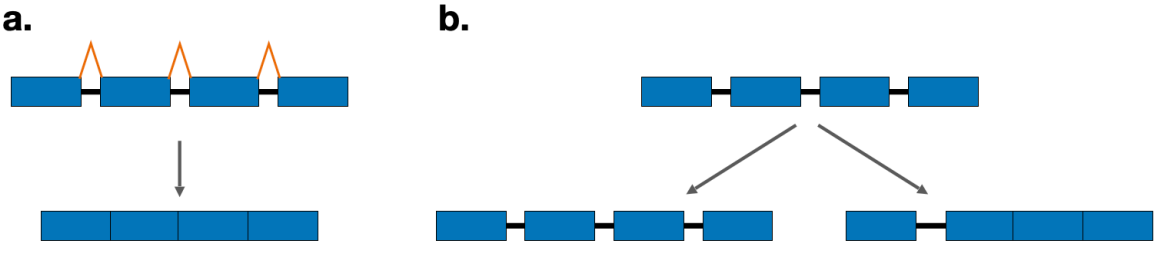

Figure 1: Alternative splicing modes.
a, Exon skipping displaying an event where all exons are retained.
b, Intron retention (IR) displaying coordinated IR event (left) and non-coordinated IR event (right)

Methods:

Data Set Description

To generate the data set, upstream processes included treating HeLa cells with DMSO, 10nM SSA, and 100nM SSA (three samples per treatment). RNA was extracted by selecting for poly-A enriched RNA-seq, capturing mRNA transcripts against rRNA, while minimizing nascent and premature RNAs.

To detect low frequency IR transcripts and avoid alignment biases, only splice junction reads by Transcript per Million (TPM) were considered as input. An optimal library size with roughly 150 million mapped short read RNA-seq at a higher sequencing-depth distinguished between splicing variations and statistically classified IR events as biologically significant.

While previous approaches have only considered reads along the exon/exon and exon/intron boundaries, computational tool Mutually Exclusive Splicing Analysis (MESA) developed by the Brooks Lab measured genome-wide IR events from short read RNA-seq[1,4]. MESA assessed all reads aligning to intron/intron and intron/exon boundaries and developed a measurement score for read distribution to describe an IR event.

Data Set Analysis

An algorithm written in Python was developed to specifically quantify coordinated IR patterns. Only multiple IR events per gene transcript were considered; Single IR events per gene transcript were separated. To capture coordinated IR with more specificity, this algorithm filtered against quantifications of IR reads falling within larger intron boundaries and alternative 5' and 3' splice-sites. To capture significant IR events, different levels of delta values demarcating the difference of median IR reads between control (DMSO) and treated (SSA) were considered. Various delta values to capture a large number of coordinated IR events were used to optimize for a delta value of below -0.2, which was used to mark significant IR events. Next, fractions measuring the number of significant IR events divided by total IR events per gene transcript were generated per gene transcript. Further analysis of IR distribution across transcripts included the calculation of the standard deviation of intron reads per gene transcript. Genes with closer standard deviation to zero demarcated uniform intron measurements per IR event compared to genes with larger standard deviations. Gene Enrichment Analysis (Gene Ontology) was performed on lists of significant coordinated IR events[10,11].

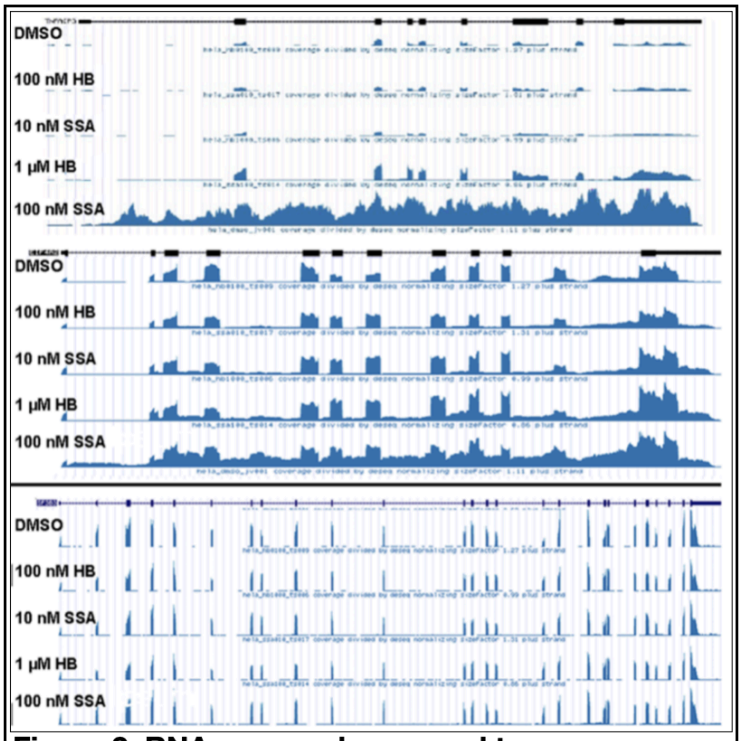

Figure 2: RNA-seq reads mapped to genome browser.
RNA-seq reads from cells treated with increasing doses of SSA and HB inhibitors mapped to two representative genes with altered splicing patterns. Most genes were unaffected (bottom row).

Results:

Algorithm for coordinated IR pattern detection

An algorithm was developed to quantify coordinated IR patterns. Cells treated with SSA expressed a greater number of transcripts showing splicing inefficiency in contrast to non-treated cells (dimethyl sulfoxide, DMSO). Among treated (SSA) samples, preliminary data captured a binary distinction of IR such that all introns of a transcript are retained/ coordinated or none are affected. In Figure 2, RNA-seq read quantification in genome browser showed treatment of HeLa cells with varying concentrations of SF3B inhibitors (herboxidiene (HB) and SSA) captured this binary distinction among transcripts.

In Figure 3, more than 100 genes were identified to contain multiple IR events greater than 10 introns, but 10% of IR events were considered significant. Significance was determined by IR events falling below a delta value outlined in Methods. In contrast, transcripts with less than 3 introns were considered highly significant with more than 90% IR events. Furthermore, genes with an upregulation of significant coordinated IR events were analyzed using gene set enrichment analysis (Gene Ontology).

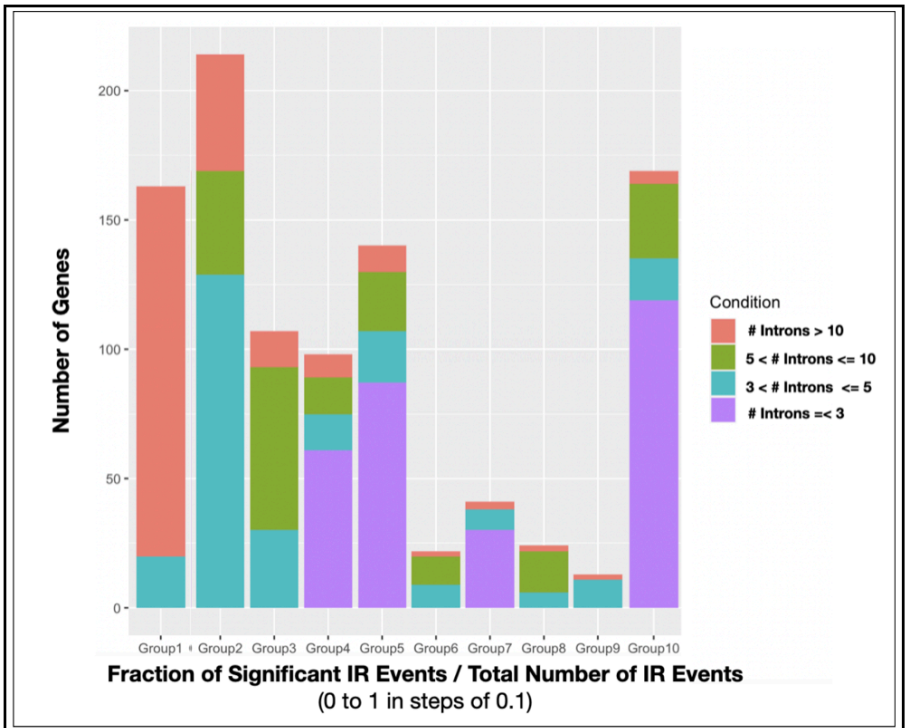

Figure 3: Distribution of intron retention across gene transcripts.
Bars show the fraction of IR events per gene with 10% increase in intron retention. Total number of IR events / gene is indicated by color.

Seen in Figure 4, an immune response gene (*Cd83*) showed coordinated IR pattern, in contrast to *Pde4d*, a phosphodiesterase that did not show coordinated IR. Unexpectedly, in Figure 5, the following genes upregulated in coordinated IR were identified with the immune response pathway (Figure 5).

Discussion:

To test our hypothesis and address the need of a systematic evaluation of splicing outcomes due to SF3B1 inhibition, an algorithm was created to capture and quantify coordinated IR events. Furthermore, this unique algorithm considered looking at all IR events across an annotated genome, as supposed to capturing only known IR events[1,4].

Coordinated events were robustly captured by computationally characterizing IR events along a transcript. This was performed by focusing on multiple IR events, omitting

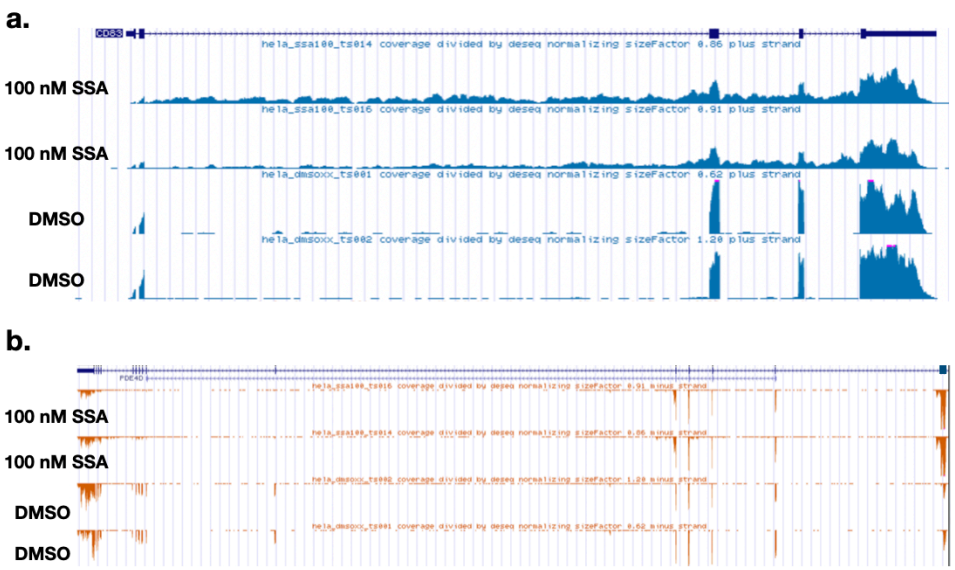

Figure 4: Binary distinction of IR shown in RNA-seq reads mapped to genome browser.
a, RNA-seq reads mapped to *Cd83* show coordinated IR in SSA treated cells.
b, RNA-seq reads mapped to *Pde4d* show no coordinated IR in SSA treated cells.

sub-introns with the same start or end sites within larger introns, and omitting IR events with alternative 5' and 3' splice-sites.

Development of this algorithm supported my hypothesis that splicing inhibitors lead to a new mode of splicing inefficiency characterized by coordinated IR events across an entire transcript. Notably, treatment of splicing inhibitor (SSA) contained a greater number of transcripts with splicing inefficiency compared to control treatment. Analysis of SSA RNA-seq is suggestive of a binary distinction of IR such that all introns of a transcript are retained/coordinated or none are affected. Unexpectedly, results of coordinated IR patterns were upregulated in a set of gene transcripts rather than all gene transcripts suggesting that constitutive splicing events are not considered as independent events. Rather, IR may be dependent on an entire transcript, as supposed to the intron and adjacent exon sequences of an intron. Furthermore, significant IR events among genes related to the innate immune response in Group 10 of Figure 3 suggested a possible response of cells against SSA (Figure 5).

Immune Response Genes
Adm
Aj271736
Ankrd49
Cd83
Cxcl2
Cxcl8
Id2
Il6
Micb
Nr4a3
Nfkb2
Nfkbia
Nfkbiz
Prdm1
Thbs1
Traf4

Figure 5: Identified genes related to innate immune response with significant coordinated IR.

Likewise, these preliminary results have also corroborated scattered reports of coordinated IR in the literature; However, these reports have not characterized nor discussed mechanism of this pattern [3,12, 13]. It is possible that the observation of coordinated IR among subsets of gene transcripts can be explained by splicing inhibitors directly impacting branch sequence recognition[13]. Another working explanation that the Jurica Lab and I propose of specificity in coordinated IR include a limited quantity of splicing factors (not affected by inhibitors) that are being competed for by transcripts in the nucleus. It is possible that due to an abundance of unspliced transcripts, this may lead to their leakage out of nucleus and triggering of genes involved in the innate immune response in the cell.

Future work include fine tuning biases introduced by sequencing and alignment artifacts upstream of this algorithm in mapping IR events. Further development of this algorithm and experimental validation will also resolve why coordinated IR events greater than 10 introns contained around 10% significance while transcripts with lower IR events were more significant (Figure 3). Possible explanations for this observation may be due to biases in capturing and sequencing against longer transcripts. Other variables to investigate include intron length, intron GC content, transcript length, and transcript expression levels.

Ultimately, results of this study is suggestive that a new mode of splicing inefficiency across entire gene transcripts can be selective. As a whole, these results offer a starting point to help guide chemotherapy development by offering insight into the effects on splicing.

References:

[1] Green, C. J. et al. (2018). MAJIQ-SPEL: web-tool to interrogate classical and complex splicing variations from RNA-Seq data. *Bioinformatics (Oxford, England)*, *34*(2), 300–302

[2] Vanichkina, D. P. et al. (2018). Challenges in defining the role of intron retention in normal biology and disease. *Seminars in cell & developmental biology*, *75*, 40–49.

[3] Vigevani, L. et al. (2017). Molecular basis of differential 3' splice site sensitivity to anti-tumor drugs targeting U2 snRNP. *Nature communications*, *8*(1), 2100.

[4] Wang, Q. et al. (2018). JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(35), E8181–E8190.

[5] Hsu, T. Y. et al. The spliceosome is a therapeutic vulnerability in MYC-driven cancer. *Nature* 525, 384–388 (2015).

[6] Dvinge, H. et al. RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* 16, 413–430 (2016).

[7] Yoshimoto, R.et al. (2021). Spliceostatin A interaction with SF3B limits U1 snRNP availability and causes premature cleavage and polyadenylation. *Cell chemical biology*, *28*(9), 1356–1365.e4.

[8] Yoshimoto, R. et al. Global analysis of pre-mRNA subcellular localization following splicing inhibition by spliceostatin A. *RNA* 23, 47–57 (2017).

[9] Bonnal, S. et al. The spliceosome as a target of novel antitumour drugs. Nat. Rev. *Drug Discov*. 11, 847–859 (2012).

[10] Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. 25(1):25-9. (2000).

[11] The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 49(D1):D325-D334. (2021).

[12] Wong, J. J. et al. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell*, *154*(3), 583–595.

[13] Ni, T. et al. Global intron retention mediated gene regulation during CD4+ T cell activation. *Nucleic acids research*, *44*(14), 6817–6829 (2016).

Note:

MESA (**M**utually **E**xclusive **S**plicing **A**nalysis) is currently in development.

GitHub - BrooksLabUCSC/mesa: A tool for detecting and quantifying alternative splicing with RNA-seq, 2022.