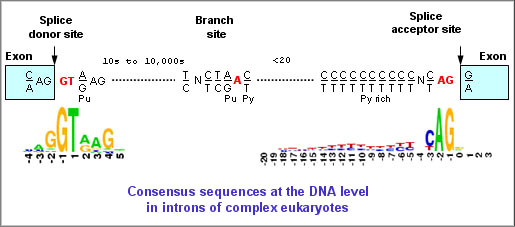
Splice Site Sequences of Interest and What to do about them:

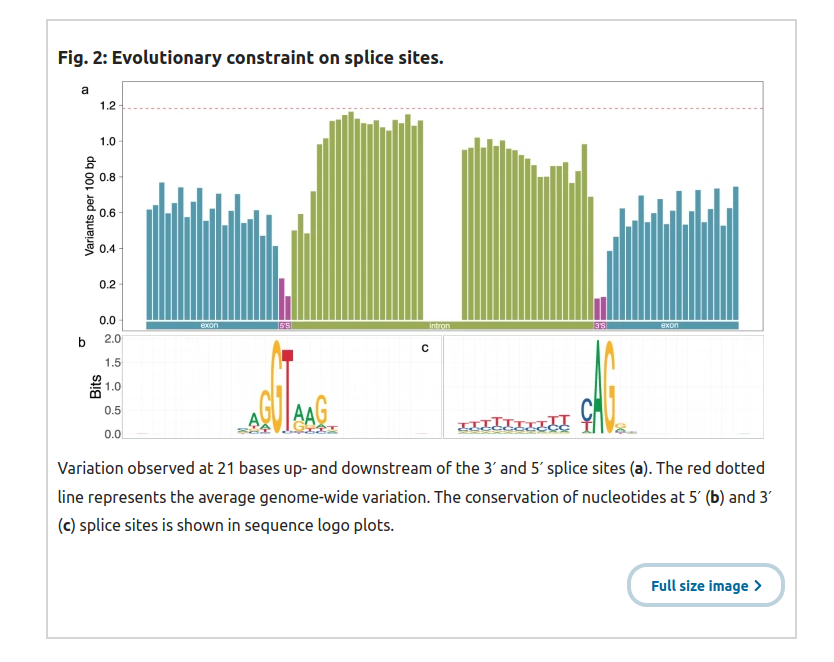
Basically, we have 4 different sequences to worry about with splicing:

* The 5’ splice site, typically GU
* The 3’ splice site, typically AG
* The branch point sequences (which can seemingly be anywhere in the intron)
* The 3’ polypyrimidine tract (mostly T’s and C’s) that is between branch point and 3’ splice site.



**Splice Sites:**

It would be interesting to recreate this plot for the 5’ and 3’ splice sites between the two conditions.



* Sequence compositioDon’t have the Zoom Workplace app installed
* n of these sites could be interesting and informative, and we could test for for differences between these two conditions quite easily.

**Branch Point Sequences and Polypyrimidine Tracts:**

* Branch Point: "The branch point sequence is a cis-acting intronic motif required for mRNA splicing. Despite their functional importance, branch point sequences are not routinely annotated"
* Polypyrimidine tracts: "The polypyrimidine tract is one of the important cis-acting sequence elements directing intron removal in pre-mRNA splicing. Progressive deletions of the polypyrimidine tract have been found to abolish correct lariat formation, spliceosome assembly and splicing."

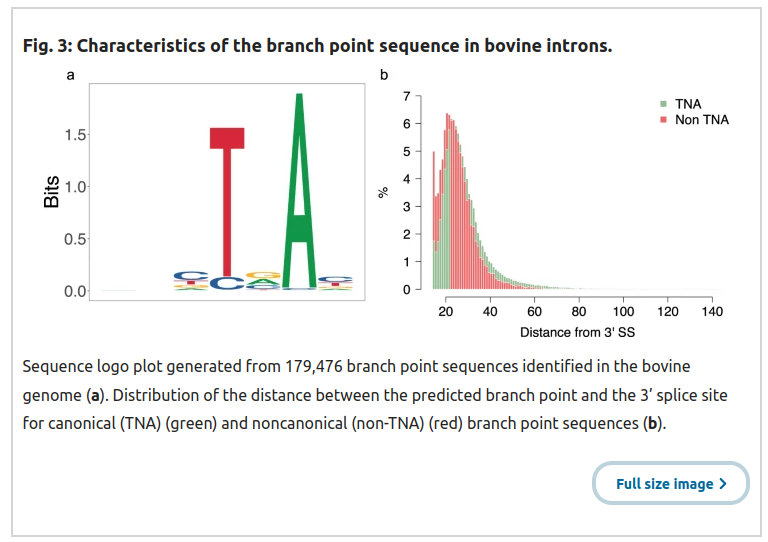
The location of these sequences is also nonrandom, despite branch points being a little more variable than the polypyrimidine tracts which as the name implies are very C and T rich. Previous papers in bovine used this method (<https://academic.oup.com/bioinformatics/article/33/20/3166/3870482>) to detect branch points and PPTs.

They created inputs using the following methods:

“First, we obtained the weighted octanucleotides from the bovine introns following the approach proposed by Zhang and colleagues[18](https://www.nature.com/articles/s42003-021-02725-7" \l "ref-CR18). Briefly, we extracted 14 bp long intronic sequences from the polypyrimidine tract region (3–16 bp upstream of the 3’ splice site) and the background region (187–200 bp upstream of the 3’ splice site) of 146,992 introns that were longer than 300 bp. The octanucleotide frequency was calculated in both regions separately and scored as proposed by Zhang and colleagues[18](https://www.nature.com/articles/s42003-021-02725-7" \l "ref-CR18).”

“Next, using these weighted octanucleotides and the position weight matrix (PWM) of predicted human branch point sequences (downloaded from <https://github.com/zhqingit/BPP>), we predicted the branch point sequence in bovine introns that were at least 20 bp long. We considered 179,476 unique intronic sequences (up to 250 bp upstream of 3’ splice site) from protein-coding transcripts. Within each intron, we retained the branch point sequence with the highest score.”

Using these methods, we can identify BP and PPT regions, creating graphs like this and also performing motif analysis using homor, at least for the shorter PP sequences.



If they are looking for broader patterns explaining these differential splicing, I think this is the clearest way to present this information.

References:

1. https://www.nature.com/articles/s42003-021-02725-7

2. https://academic.oup.com/bioinformatics/article/33/20/3166/3870482

3. https://github.com/zhqingit/BPP

ACTUAL METHOD WE ARE USING:

1.) Extract 200 bp window upstream of the splice site acceptor for each intron… in the entire genome... The first 30 bp upstream of the acceptor will be treated as the PPT while the remaining 170 bp upstream will be represent the background. (Bed file → Fasta)

2.) Extract all possibly octanucleotides and calculate frequency. (Fasta → octanucleotide table)

Frequencies were incremented by one to avoid zero division.

3.) Extract 250 bp windows upstream of significantly DE transcripts, run BPP on these sequences to predict branch point sequences.

4.) Run homer on the the predicted branch point sequences and the PPT sequences identified for the significantly DE transcripts. We kept BP prediction with highest score per intron (same as Bovine paper).

Basically, the idea here is that we are identifying differences in the sequences relevant to splicing because differences in these sequences result in different binding affinities for the spliceosome. Clearly, enrichment of certain sequences between the treatments in these splicing related regions indicates some sort of differential preference from the spliceosome that is likely the induced by the treatment condiction.

######Interestingly enough, “PB.11982.13”, the most significant DE transcript, is a transcript coding for the UNC13D protein in which certain mutations have been shown to impair splicing, or at least alter it. These results beg the question of whether differential expression of this UNC13D isoform is impacting splicing in a way that is mediate through altering the spliceosome structure so that it preferentially prefers certain BP and PPT sequences. This may just be correlation and not causation though, pretty much the problem with all RNAseq.