Intron detection notes

Anh Vu Le

Comparison of splice site classification models (on *Thega1* phylum)

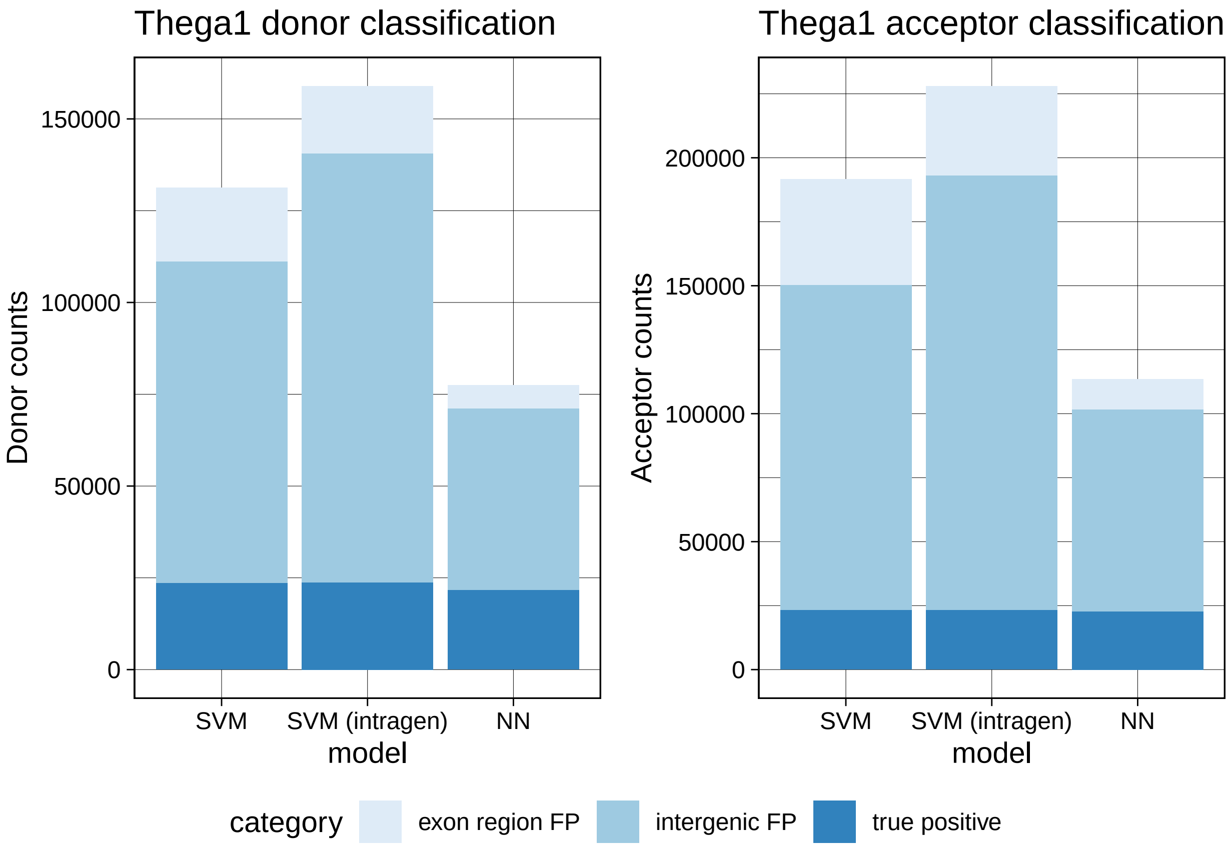
This section will be devoted to comparing splice site classification models. In total 3 models were developed - two SVM models and NN model (technically there 6 of them as donors/acceptor site require each its own predictor). The SVM models differ in data, which they were trained on. First one (let's call it "general SVM") was given negative splice site candidates sampled from the entire genome with no distinction, whether they come from an intergenic or an intragenic (i.e. exon) region. Second SVM was given only negative candidates from within exons ("intragenic SVM"). NN models were also trained ignoring intergenic sequences. The motivation behind this decision comes from the assumption, that the accuracy in intergenic regions is irrelevant for the task as hand, which is basically about recognizing intron sequences from exon sequences.

In general validation, the models ended up toe to toe in sense of both precision and recall. The tests however do not capture the real nature of the task, which is intron cutting for the purposes of enhancing BLAST search results. Good results here are measured in a slightly more complex manner, as we will see shortly.

Intron cutting is a classification task. And as in all such tasks, we measure how we successful we are at solving such a task by means of metrics such as precision and recall. In this case however, we need to adjust (or rather extend) the notion of these metrics, especially precision. In intron classiciation and cutting, not all false positives "hurt" the same - in fact, some false positives are even beneficial. This is because we deal with two types of false positives introns here - those which are in intergenic regions and those that are inside exons. The former do not concern us at all. If we cut into the intergenic region, it is only for good, as BLAST will recieve smaller data at the cost of nothing. Intra-exon cuts are on the other hand the "real" false positives, as they will lower the quality of BLAST hits.

One way to compare the two models is therefore to look at the composition of examples, to which the models say they are introns (or splice sites). How many of them are true positives? And if they are false positives, how many of them will actually cause damage? If we compare SVM and NN splice site models in this fashion, we will reveal, that they are not quite the same, as they may seem after the first round testing [1].

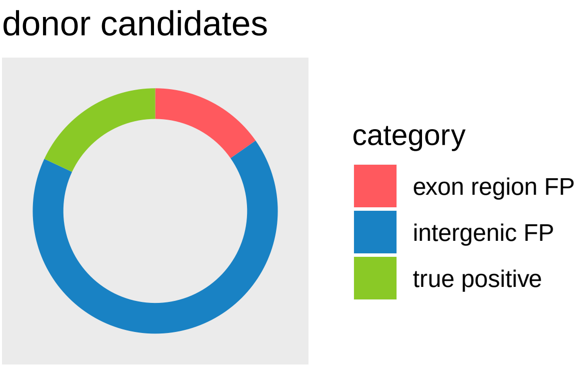
[1] Numbers taken from the output of *intragen/intragen_splice_sites_fp.py* script

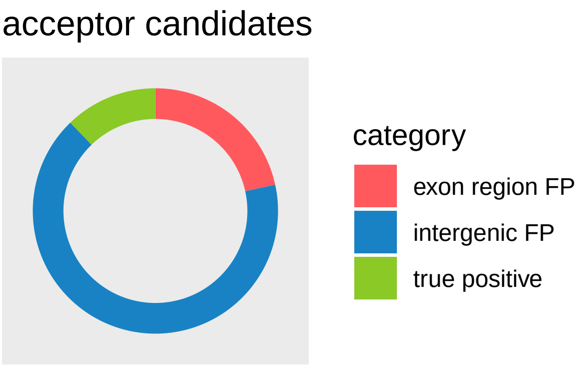


The chart captures several interesting facts about the models - first of all, all models recognized very similar number of true donors and acceptors. This might indicate, that there is set of introns which cannot be detected by neither of the methods. We will verify this hypothesis later on. Second, NN models produce substanitally less false positives, than both SVM models, which carries forward to intragenic false positives, the harmful type of false positives. We can also see, that the intragenic SVM, even though producing more raw FP, than the general SVM, has actually better results in exon regions. Afterall, the model was built for such purpose.

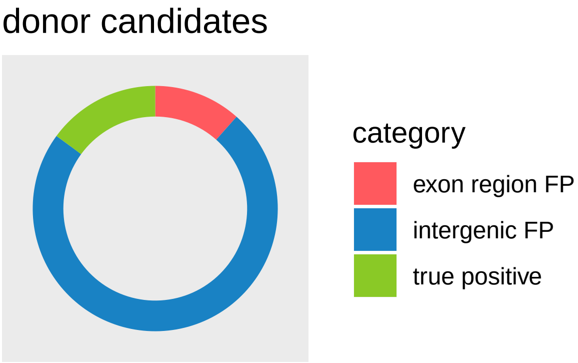
Let's visualize solely the prportions of the types of classified true candidates. We will be particularly interested in ratios between correctly classified donors/acceptors (green) and falsely classified candidates, that lie inside an exon (red).

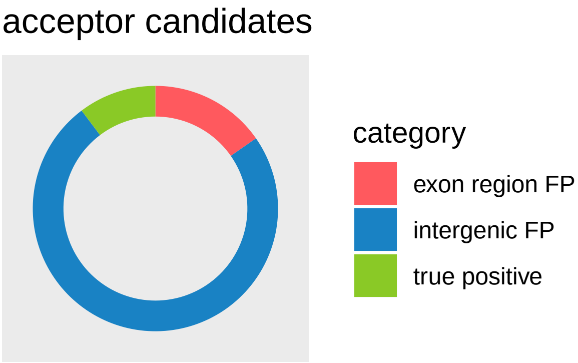
SVM models



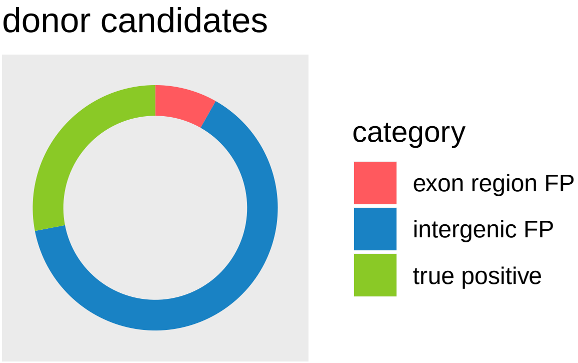


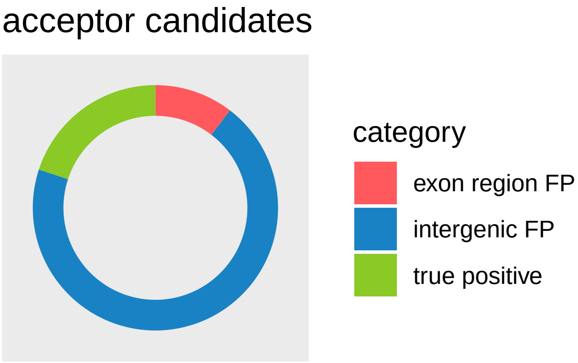
SVM intragen models





NN models



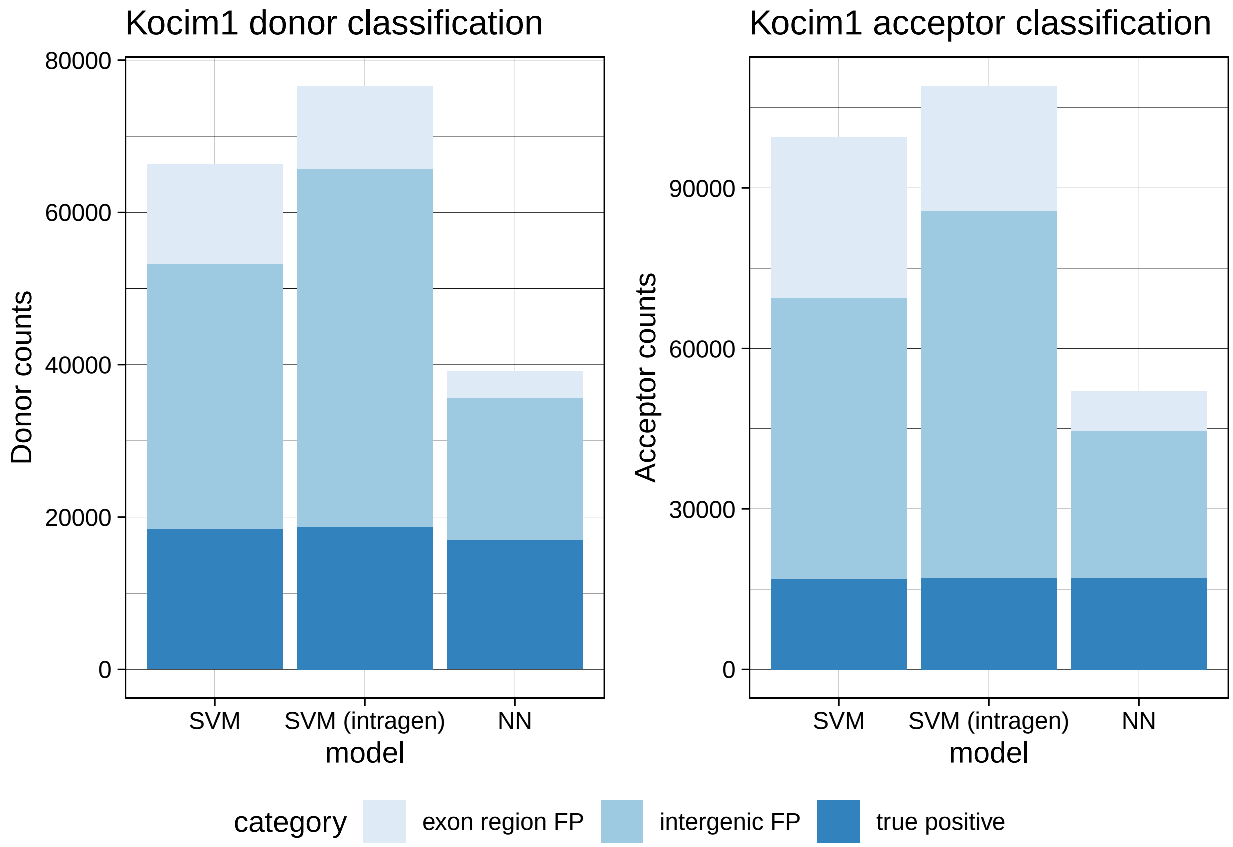


Looking at the three pairs of graphs and the proportions of green-red regions we can clearly see, that NN models have significantly better ratio between correctly classified splice site candidates and intra-exon false positives. Note, that the green regions capture in all cases very similar number of examples. In case of SVM models, they appear to produce one false intragenic splice site for each correctly detected one. This ratio seems unfavorable and this is why a subsequent intron model is needed to improve those odds. Next, let's look at the the same analysis, this time fo *Kocim1*.

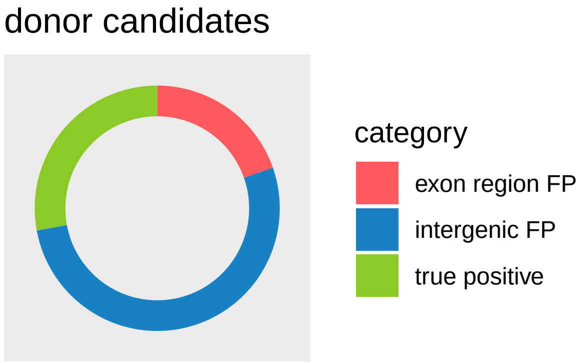
Comparison of splice site classification models (on *Kocim1* phylum)

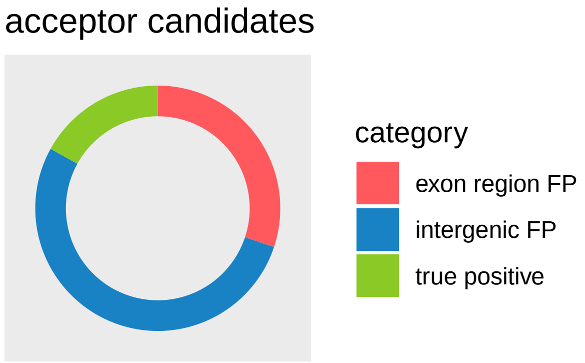
The graphs tell a similar story as in the *Thega1* case. Models find almost the same number of correct splice sites, NN however produce less false positives, both intergenic and - more imporantly - intragenic. Intragenic SVM model produces more FP, than the general SVM, but slightly less intragenic FP [2].

[2] Numbers taken from the output of *intragen/intragen_splice_sites_fp.py* script

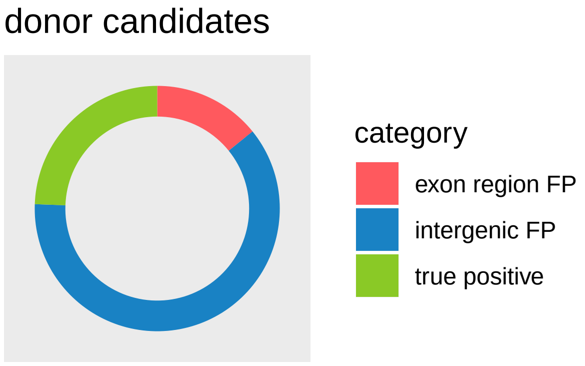

Kocim1 donor classification
Kocim1 acceptor classification
Donor counts
80000
60000
40000
20000
0
Acceptor counts
90000
60000
30000
0
SVM
SVM (intragen)
NN
SVM
SVM (intragen)
NN
model
model
category
exon region FP
intergenic FP
true positive

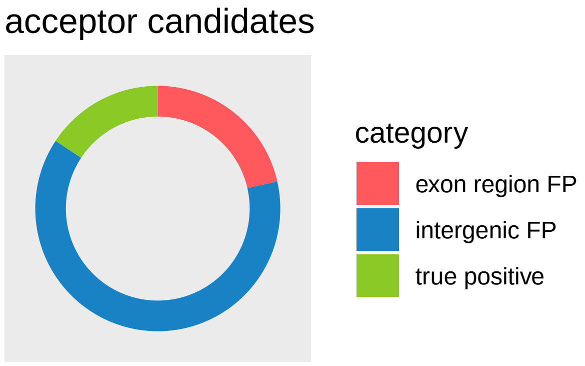
SVM models


donor candidates
category
exon region FP
intergenic FP
true positive

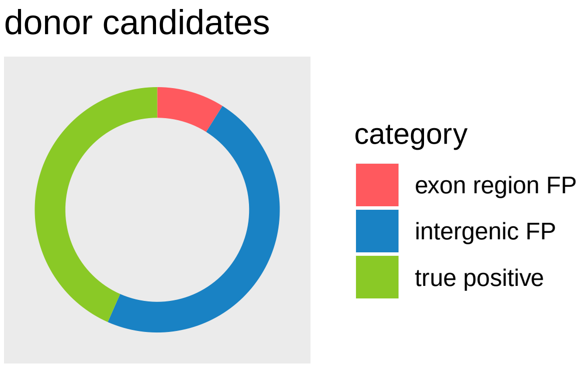

acceptor candidates
category
exon region FP
intergenic FP
true positive

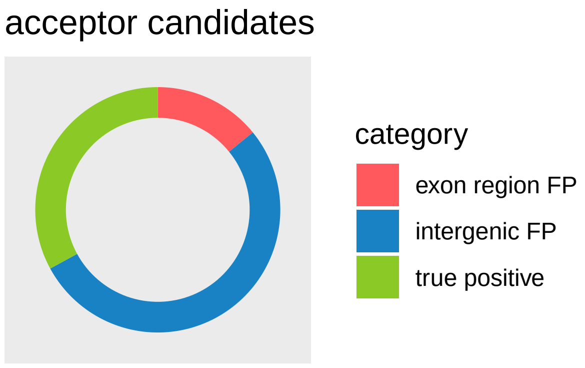
SVM intragen models





NN models



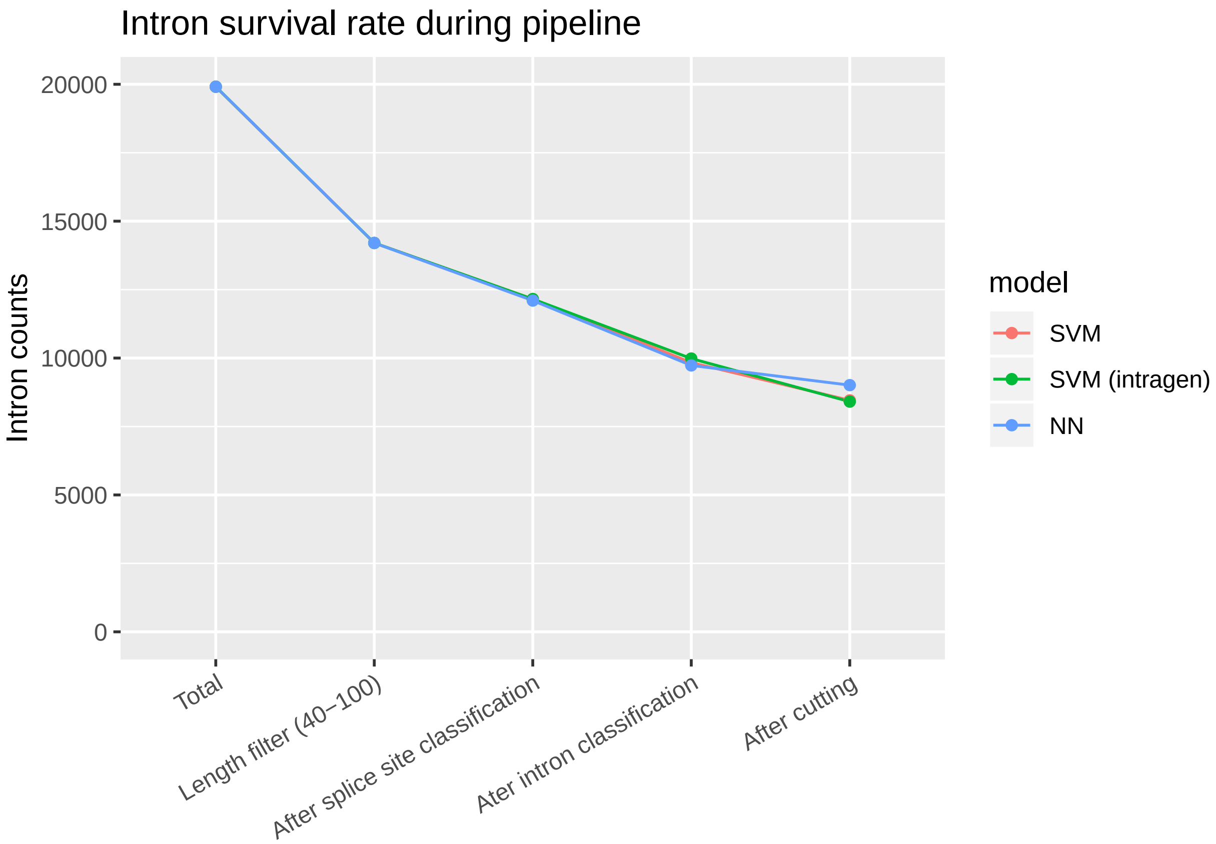


Follow-up intron classification (*Kocim1*)

Next step after splice site classification is pairing positively classified donors/acceptor to form intron candidates. Intron model is then trained and candidates classified. Positively classfied introns are then cut out of a scaffold.

It is prominent, that data introm model gets is determined by what splice site models produce. It is therefore reasonable, that each donor/acceptor pair of models should have a dedicated intron model. At this stage however, our main focus are still splice site models and how they influence the intron one. Therefore, for benchmarking purposes, we use an intron model trained with general SVM data. Even though this will give the general SVM models a slight edge, the comparison should still be possible. As we will see later, this is thanks to the fact there is a considerable overlap in the splice site models outputs.

As mentioned bevore, intron cutting and purging is done in steps. Each steps looses a portion of introns due to some filtration or non-perfect recall. Let's visualize, how many introns are lost in each step. The first stage is the number of introns we start with - all introns on the positive strand. Next, we will loose those introns, which are not in the length range of 40-100. Undetected donors and acceptor will also cause some introns to be lost. Intron classification is an apparent source of losses and some are lost during cutting as well. This final decrease is due to incorrect decisions when dealing with candidate overlaps - two (or more) positively classified introns overlap and only one can be cut. And it may happen, that the wrong candidate is chosen over the correct one.

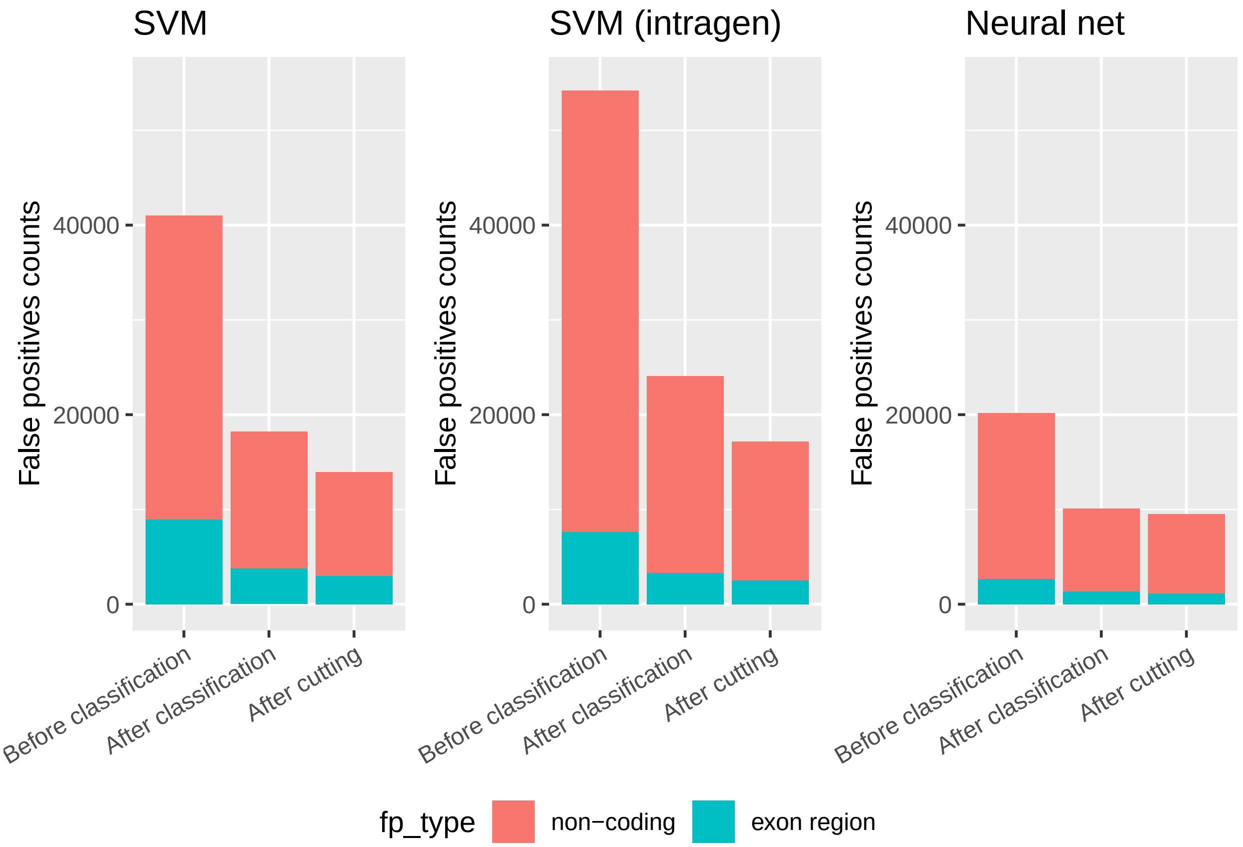


We saw earlier, that the recall of splice site models is more or less the same. The area, where the models differed the most, are numbers of false positives and their locations (inter vs. intra exons). Now, let's see, how this translates into intron classification. The IG SVM has by far the most number of false positives. It however becomes better than the general SVM, if we take into account only intragenic false positives.

Vastly better than both is the NN model, having only 55% and 42% of other two's FPs respectively. The ratio is even better for intra-genic false positives - NN makes only 38% and 46% of "serious" mistakes other two models do, which is a huge reduction.

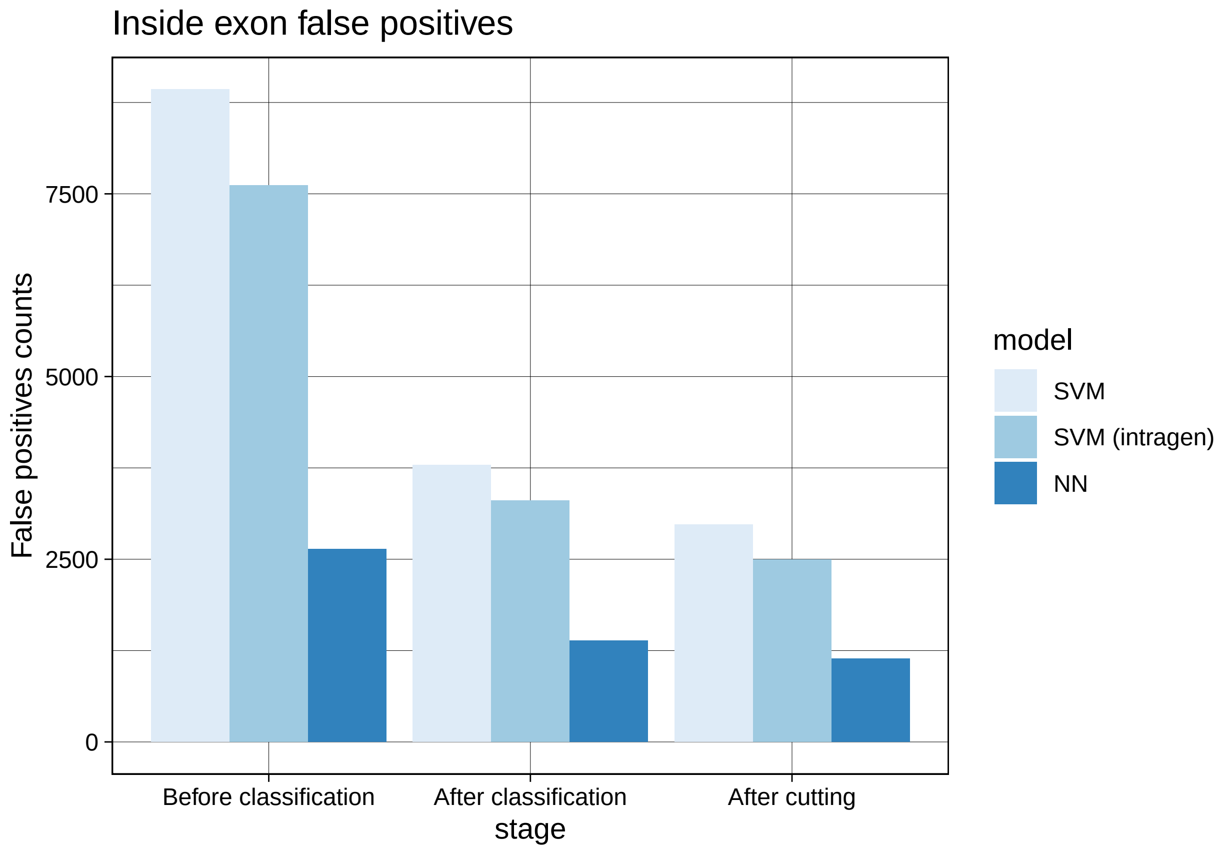
In absolute numbers, using SVM for splice site classification we were able to correctly cut out 8456 introns at the cost of "damaging" 2975 exons with wrong cuts (i.e. 2.84 correct cuts for each damaged exon). In case of IG SVM the ratio was 3.37 (8409 correct, 2496 damaged exons). For NN, the ratio was stunning 7.88 with 9011 correctly cut introns and only 1143 damaged exons.

Precision metric, that treats non-exon false positives cuts as true positives (as they don't matter), gives for the three models values 86.71%, 90.24% and 93.47%. Recall, is comparable.



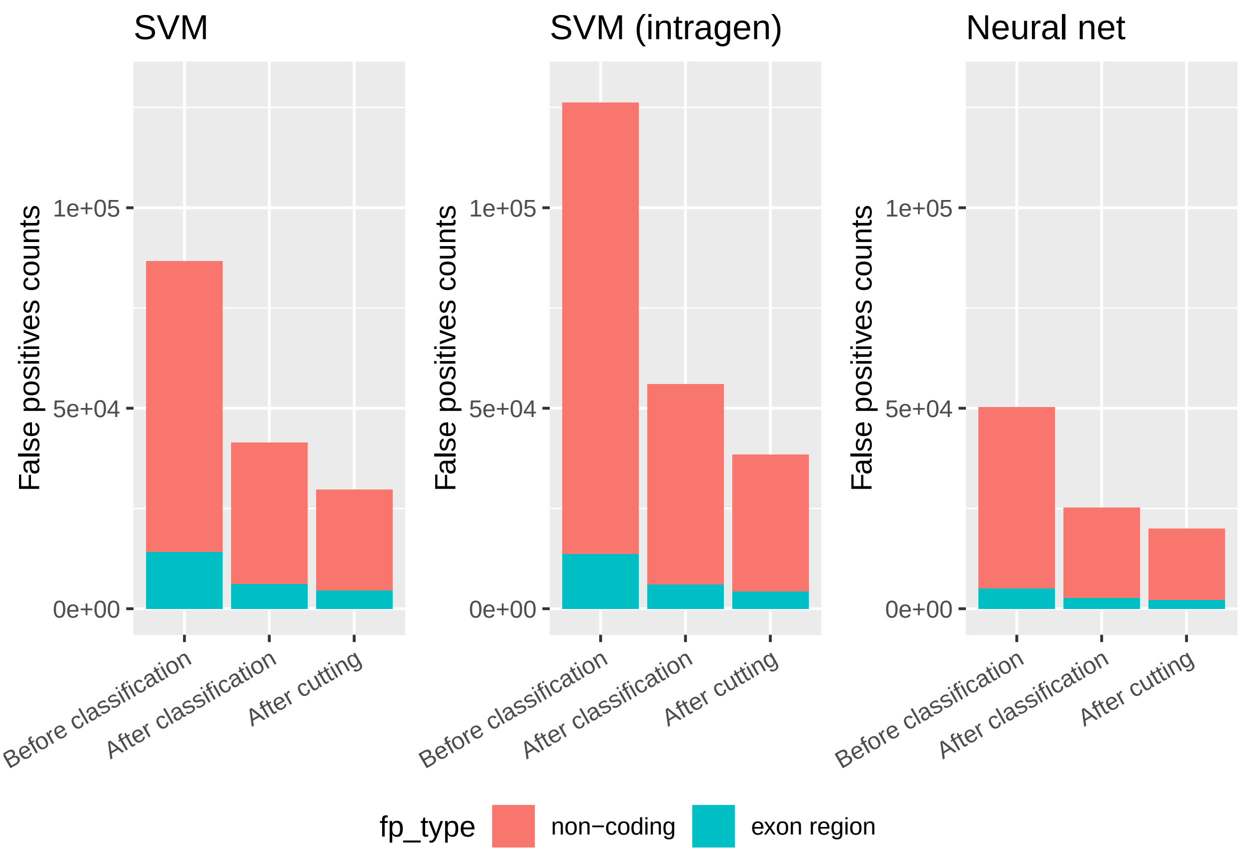
In the setting, where the cutting only prepares data for BLAST, high number of false positives might even be beneficial. Considering the number of intra-genic FP (the only adversary FPs) is in both types of SVMs similar, it can be beneficial to employ the IG version as it will shorten the input for the BLAST search. Still, the difference is marginal (after intron classification) and cannot make up for considerably better performance of NN.

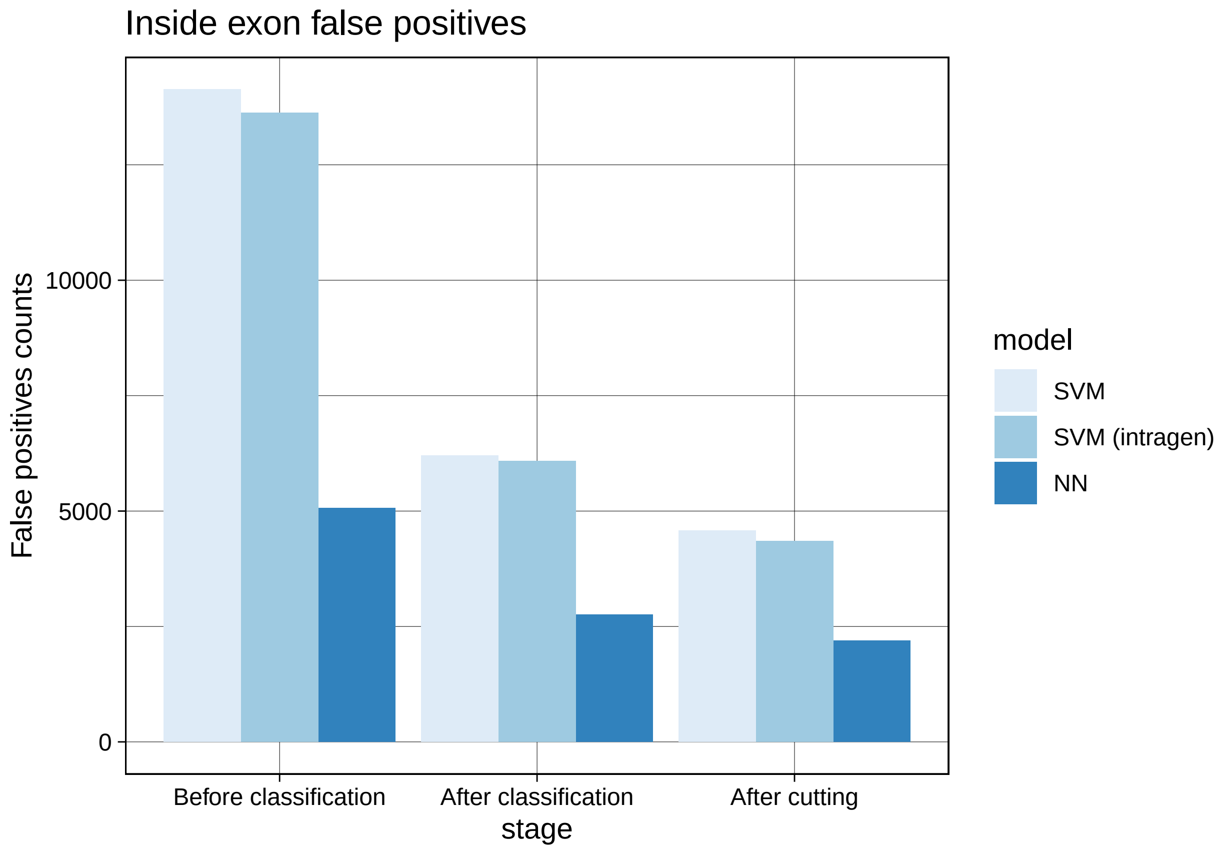
Zooming in and considering the intragenic FP reveals slight advantages of IG SVM. NN models are however still vastly superior to both of them.



Follow-up intron classification (*Thega1*)

Similar results as in splice site classification.


SVM
SVM (intragen)
Neural net
False positives counts
1e+05
5e+04
0e+00
Before classification
After classification
After cutting
fp_type
non−coding
exon region


Inside exon false positives
False positives counts
10000
5000
0
Before classification
After classification
After cutting
stage
model
SVM
SVM (intragen)
NN

In absolute numbers, using SVM for splice site classification we were able to correctly cut out 15335 introns at the cost of "damaging" 4577 exons with wrong cuts (i.e. 3.35 correct cuts for each damaged exon). In case of IG SVM the ratio was 3.48 (15168 correct, 4353 damaged exons). For NN, the ratio was stunning 6.81 with 14955 correctly cut introns and only 2195 damaged exons.

Precision metric, that treats non-exon false positives as true positives (as they don't matter), gives for the three models values 89.83%, 91.88% and 93.72%. Recall, is comparable.

Misc

SVM and NN repressent two distinct approaches in splice site classification. In previous section we could witness, that recalls of the models is more or less the same. Naturally, it raises a question, whether only the number of true positives is similar, or whether also the set of true positives themselves is similar. This question can further extended to false positives - do the models make similar mistakes? Analyzing classification outputs for donors, we obtain following numbers:

True positives shared between models: 16954. The intersection makes 90.41% of SVM-intragen model and 99.58% of NN model results False positives shared between models: 19645. The intersection makes 33.93% of SVM-intragen model false positives and 88.61% of NN model false positives Intragenic false positives shared between models: 2837. The intersection makes 26.11% of SVM-intragen model false positives and 81.08% of NN model false positives [3]

Several conclusions can be drawn. First of all - the models, despite different in approach, discover very similar set of true introns (over 90% is shared). The fact 99% of NN true donors is shared, basically means, that all splice sites discovered by NN models are discovered by IG SVM model as well. This has a notable caveat - considering non-perfect recall of both models (87% to be precise), there exists a considerable set of introns, which neither model is able to detect.

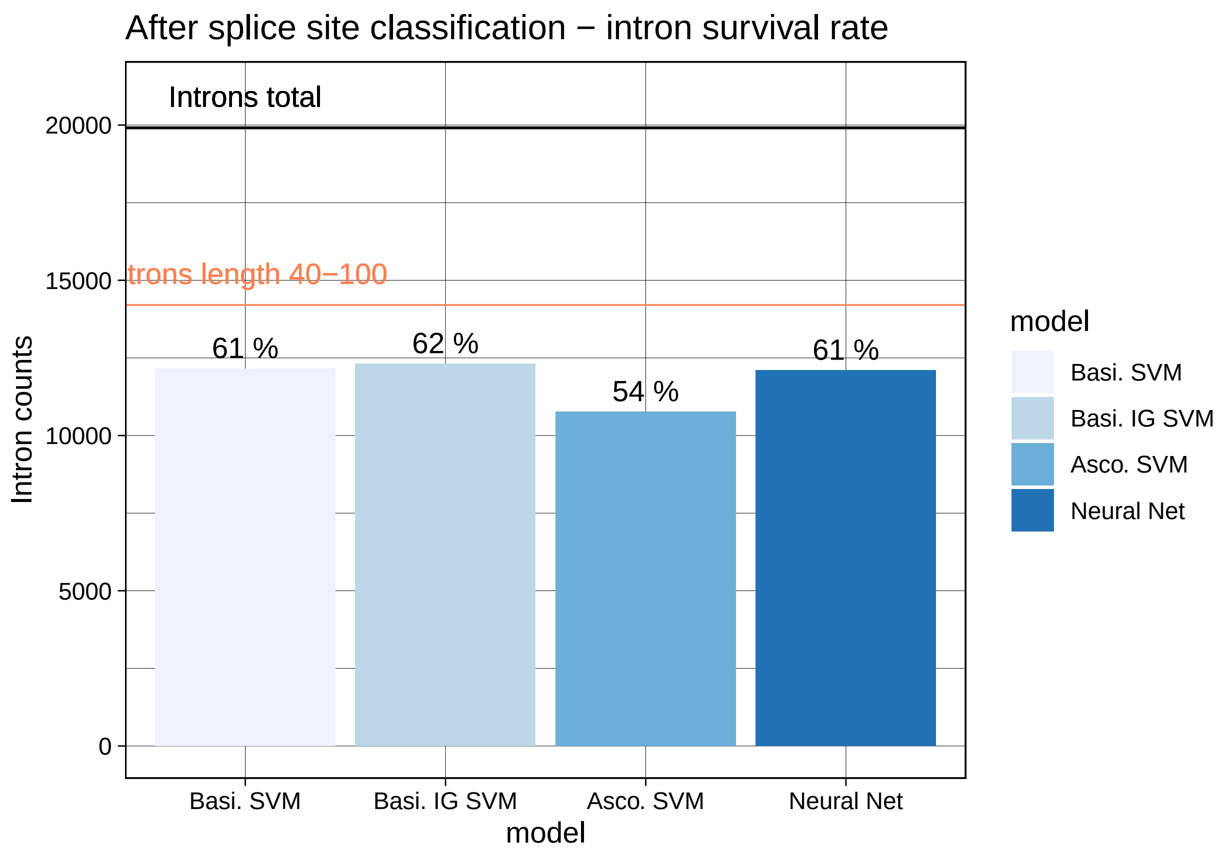
In case of false positives, it can be also seen, that intersection of false positives take up almost all NN false positives, meaning SVM makes almost all mistakes NN does. The 33% intersection on the other hand indicates, that SVM makes the same mistakes as NN and a large number of additional ones. This applies also for intra-genetic false positives. The caveat is again, that there is a sizable number of false positives, on which both models make mistakes.

The proportions are very similar also for *Thega1* division.

Intron model recall

This chart shows how many introns we would be able to detect, if we had an intron model with perfect recall. I.e. we see here, how many true introns survived the splice site classification step (and pairing of GT-AG candidates). The orange line represents the relative 100% recall level, as we only pair GT-AG candidates in a range of 40-100nt. Longer or shorter introns are therefore not even included into the classification dataset, not speaking of being detected.

[3] Output of the *intragen/compare-splice-site-classification.py* script with parameters (model1 = 'SVM-intragen', model2 = 'NN') for *Kocim1* donors


After splice site classification − intron survival rate
Introns total
trons length 40−100
61 %
62 %
54 %
61 %
Intron counts
20000
15000
10000
5000
0
Basi. SVM
Basi. IG SVM
Asco. SVM
Neural Net
model
model
Basi. SVM
Basi. IG SVM
Asco. SVM
Neural Net