Assessment of long read transcript annotation

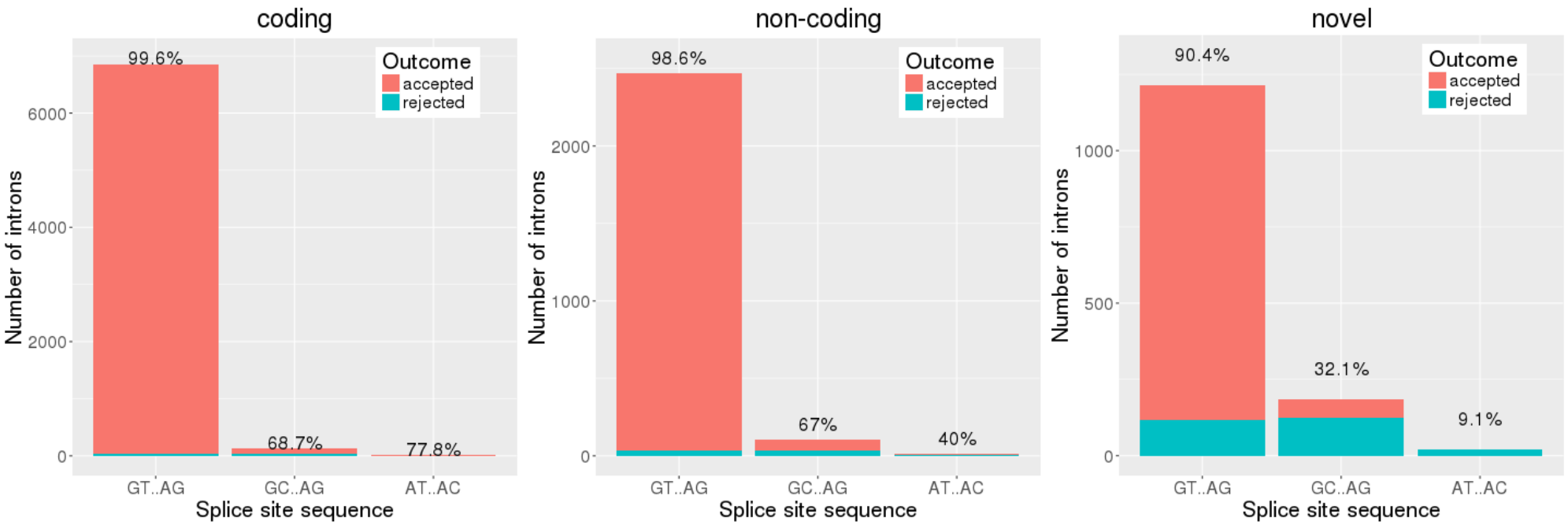
The quality of a set of ~2000 transcripts derived from long read transcriptomics was manually assessed by gene annotators, who classified them into *accepted* or *rejected*. Individual introns were also classified in the same way.

Outcome of the assessment

	Introns	Transcripts
Accepted	10601	1592
Rejected	409	392
Total	11010	1984
Rejection rate	3.7%	19.7%

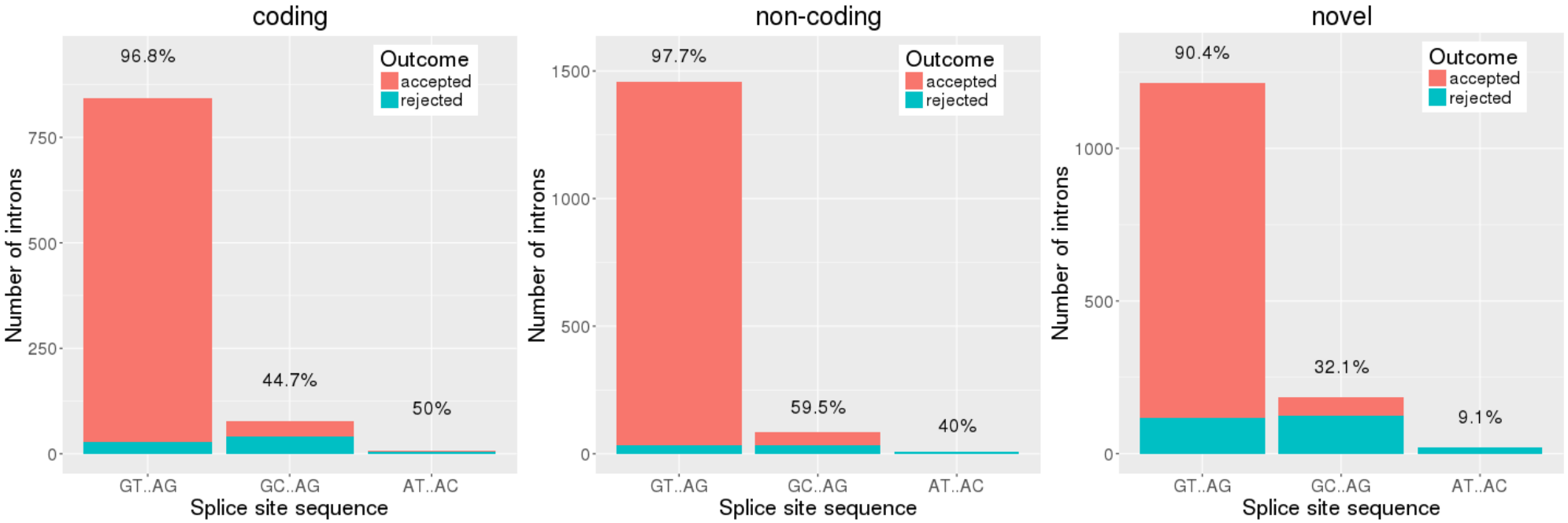
A small number of features were identified that could help predict the transcript or intron quality. The following slides show an exploratory analysis of the significance of those features.

Outcome by splice site sequence



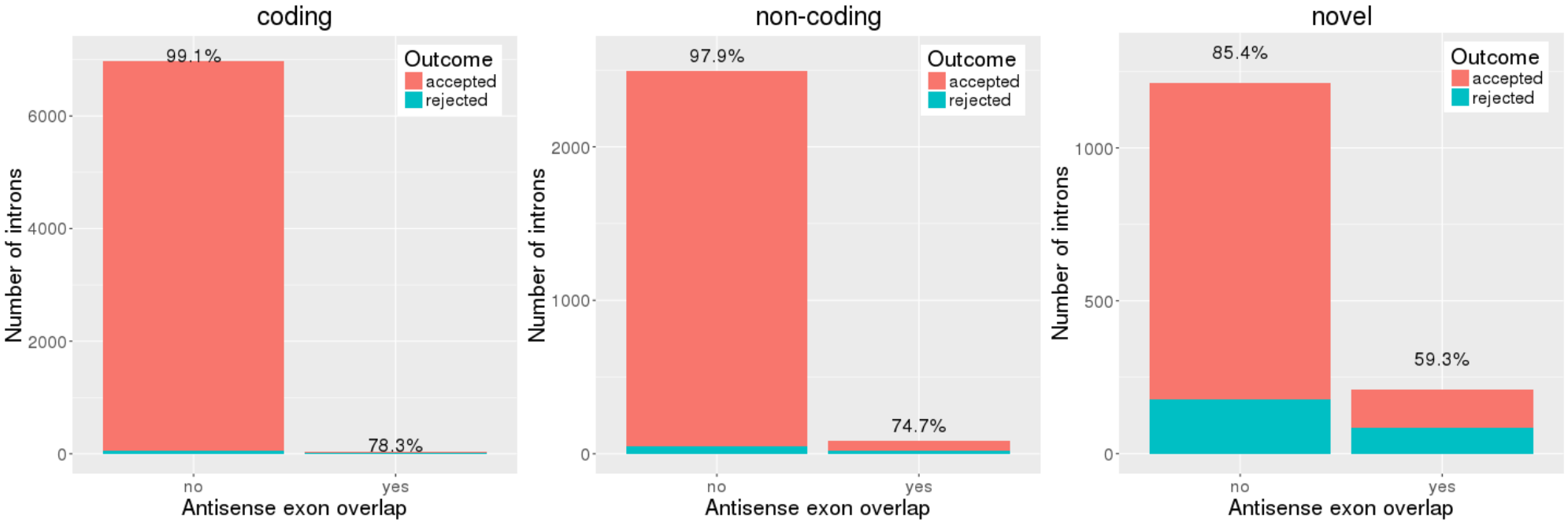
Acceptance rate of splice junctions for the three canonical splice site sequences. Results are broken down by gene category (protein-coding, non-coding and novel). Splice junctions with GC-AG and AT-AC splice sites were more likely to be rejected, especially in novel genes.

Outcome by splice site sequence (novel introns only)



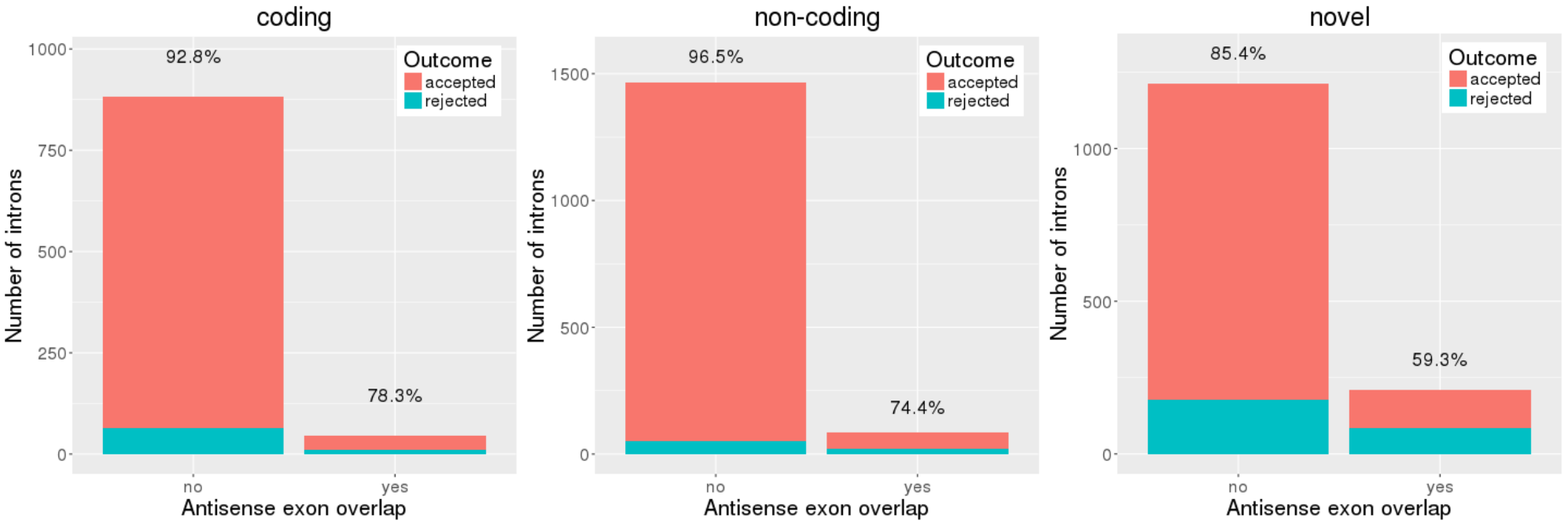
Same as in the previous slide but analysing splice junctions that were not present in the GENCODE human annotation.

Outcome by antisense overlap



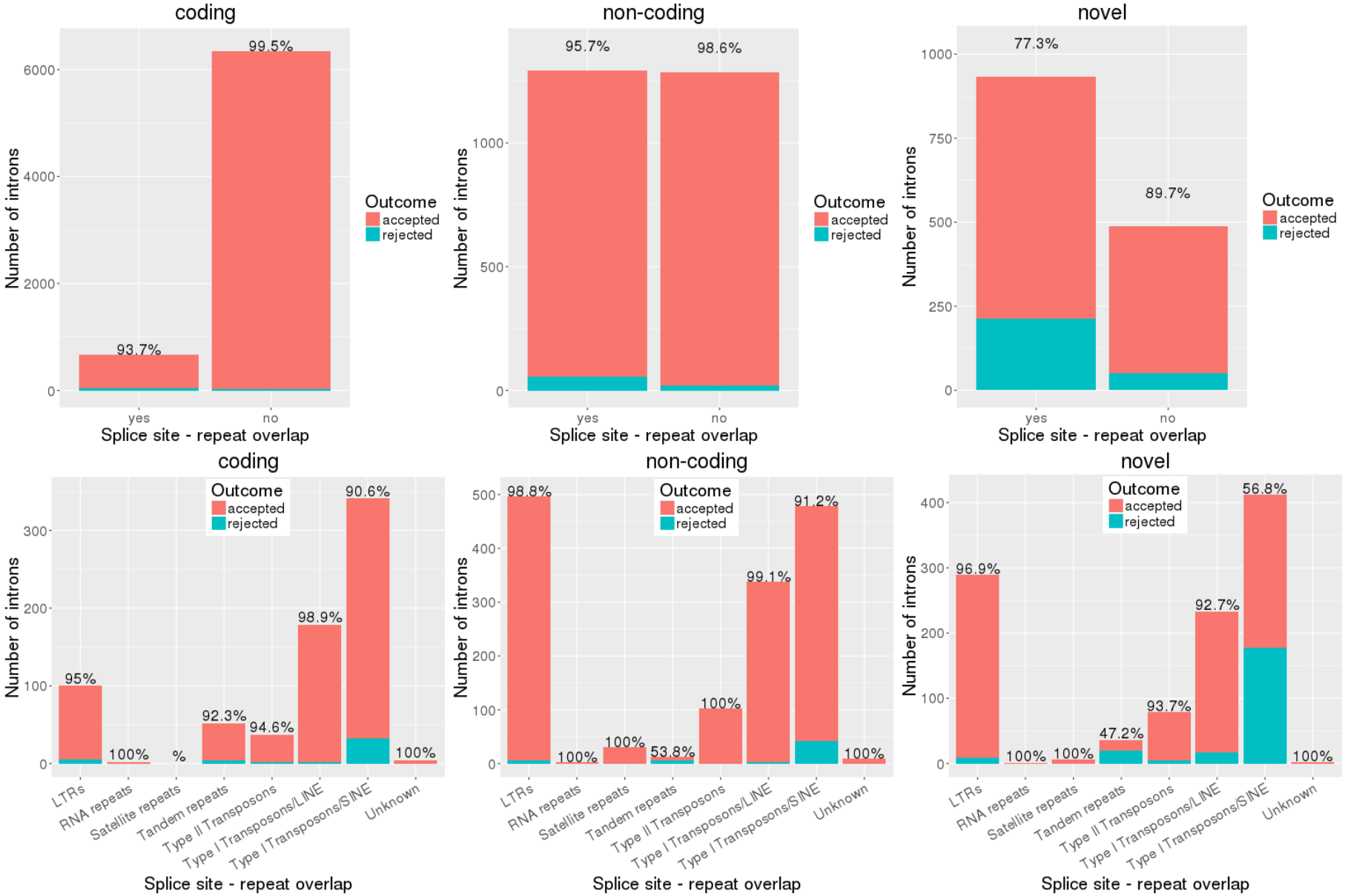
Acceptance rate of splice junctions that overlap an annotated exon on the opposite strand. Results are broken down by gene category (protein-coding, non-coding and novel). Splice junctions with antisense exons were somewhat more likely to be rejected, especially in novel genes.

Outcome by antisense overlap (novel introns only)



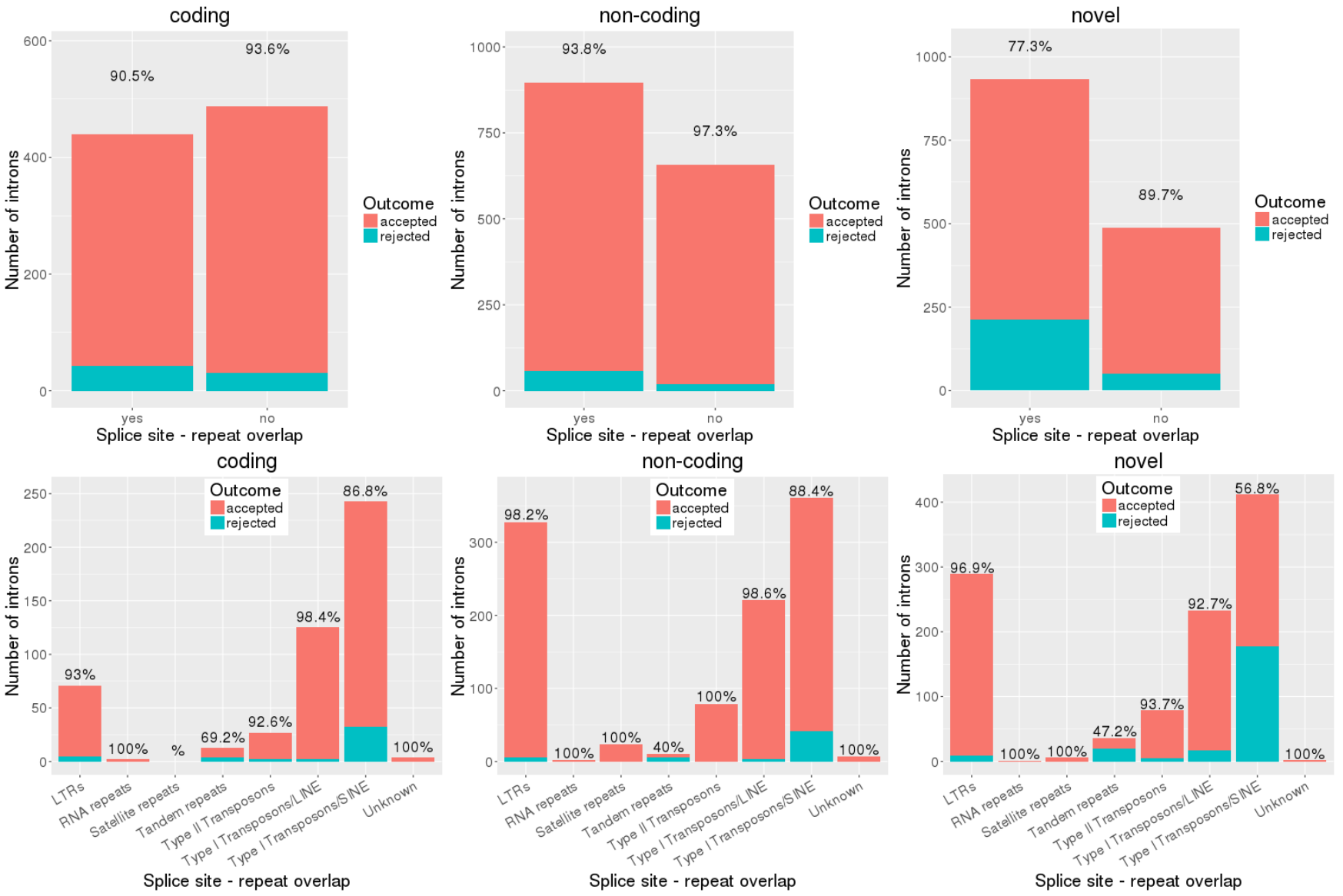
Same as in the previous slide but analysing splice junctions that were not present in the GENCODE human annotation.

Outcome by repeat overlap



Acceptance rate of splice junctions that overlap a sequence repeat features. Results are broken down by gene category (protein-coding, non-coding and novel). Splice junctions overlapping type I transposons or tandem repeats were more likely to be rejected, more significantly in novel genes.

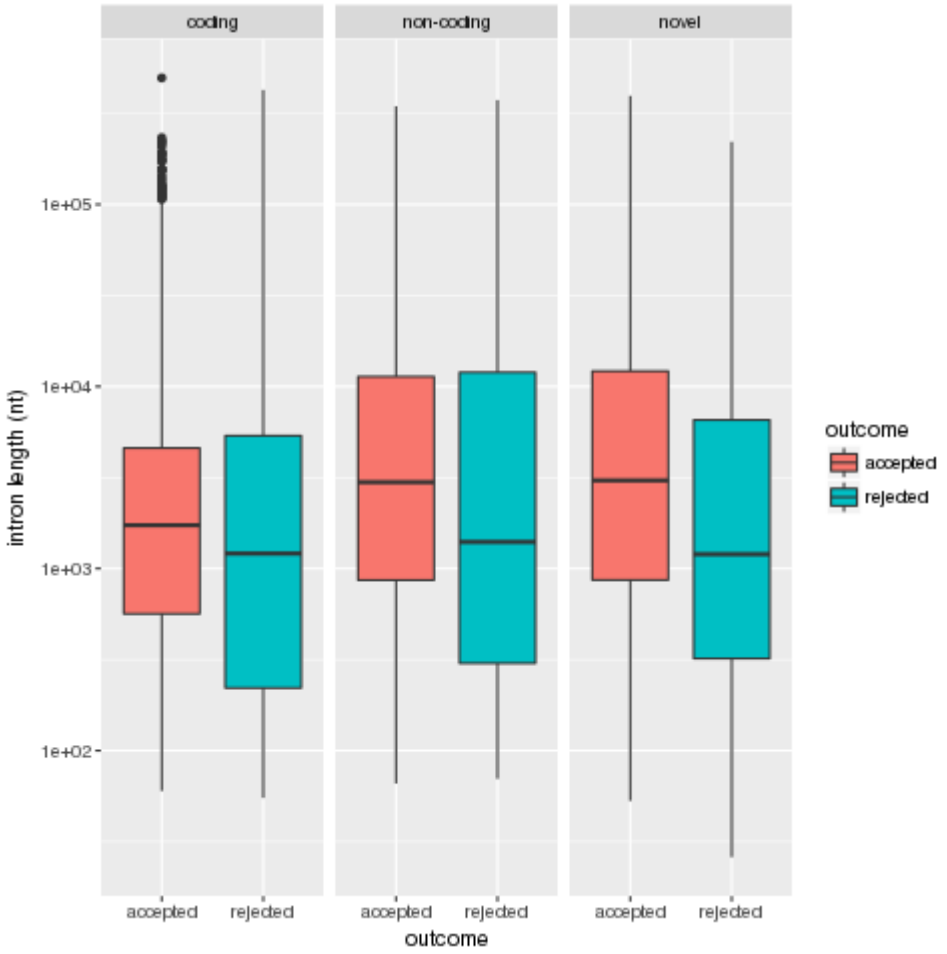
Outcome by repeat overlap (novel introns only)



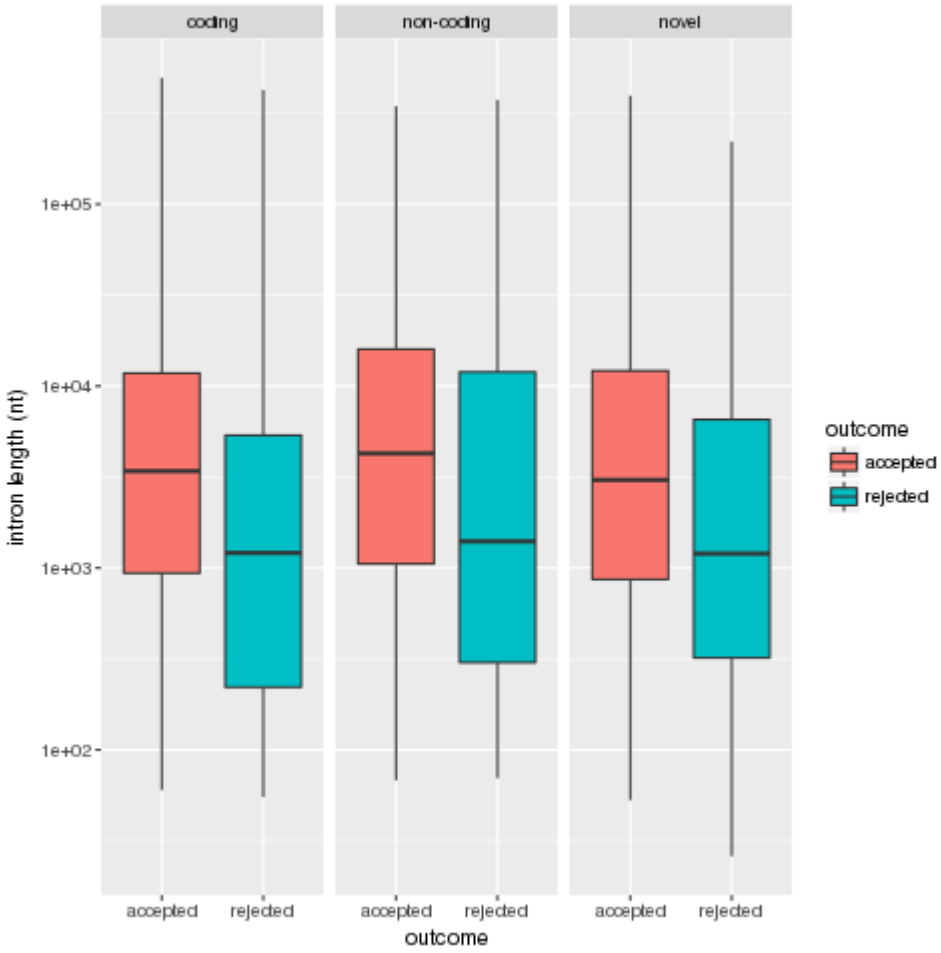
Same as in the previous slide but analysing splice junctions that were not present in the GENCODE human annotation.

Outcome by intron length

All introns



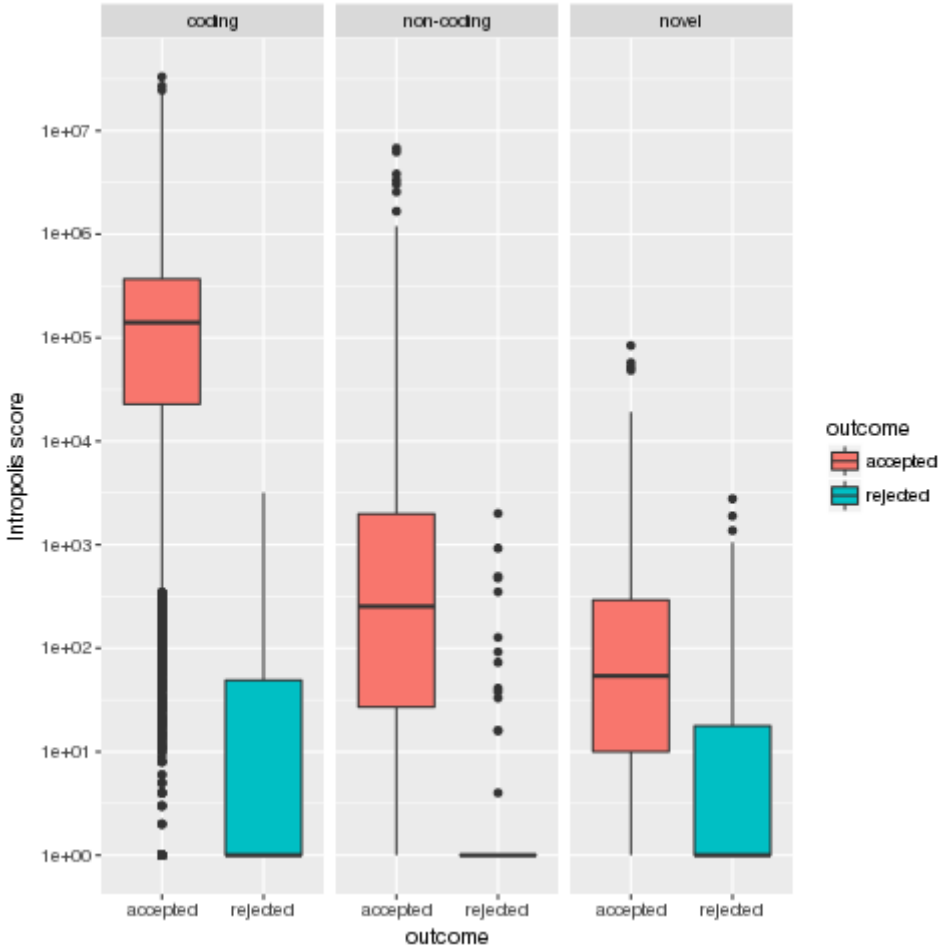
Novel introns



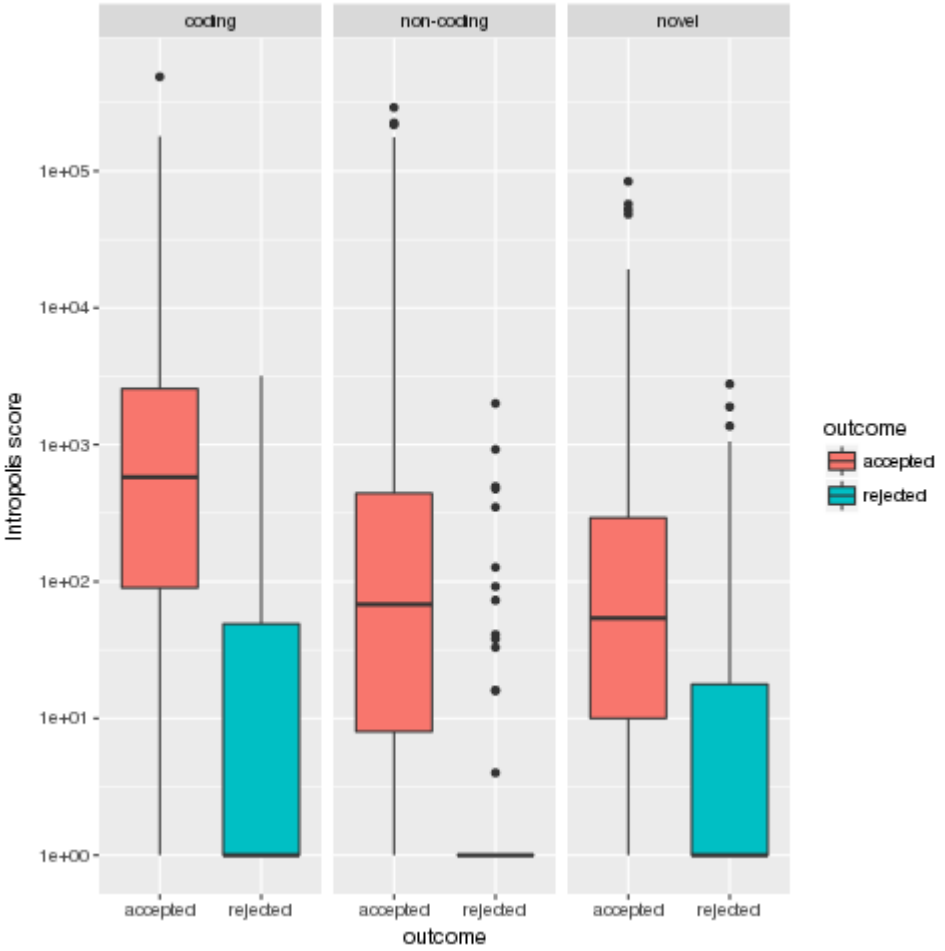
Intron length distributions for introns accepted and rejected in three gene categories.

Outcome by intron score

All introns



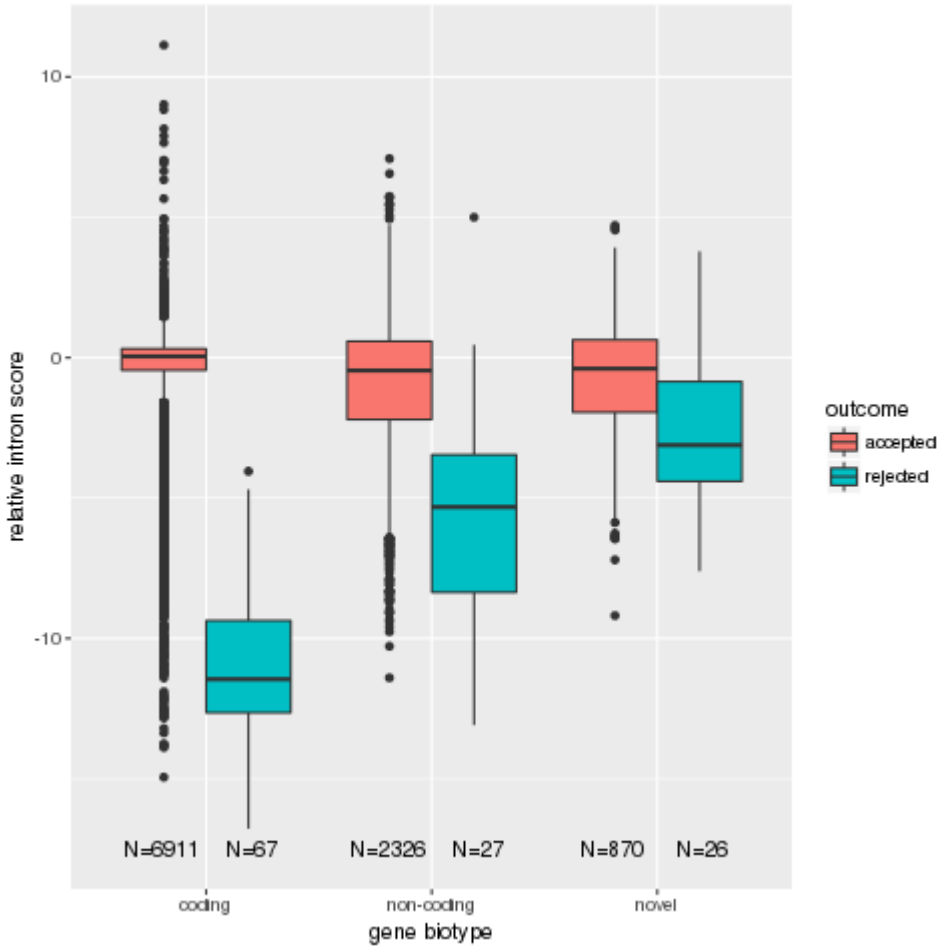
Novel introns



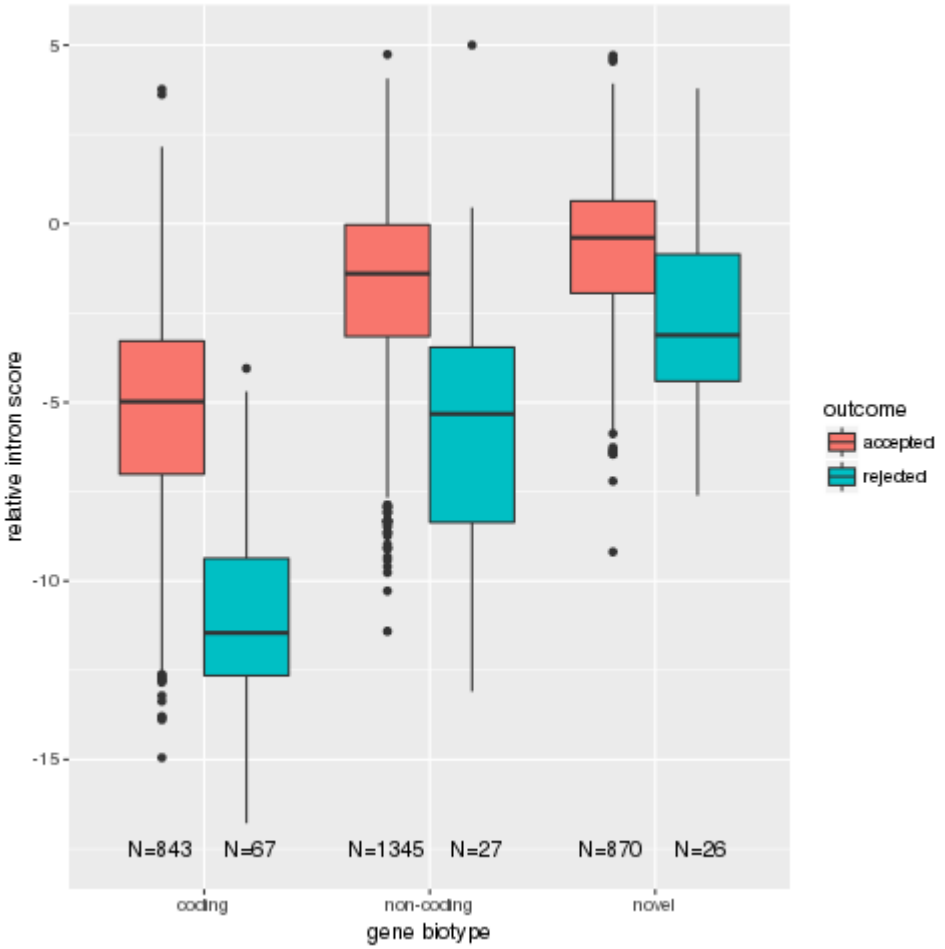
Intron score distributions for introns accepted and rejected in three gene categories. The intron scores represent the number of RNA-seq short reads that support each splice junction, using data across multiple experiments that was compiled in an earlier version of the Recount3 resource.

Relative intron support = log(intron score / average score of other introns)

All introns



Novel introns



Relative intron support distributions for introns accepted and rejected in three gene categories. The relative intron support measures the relative score of a splice junction in comparison to the scores of other splice junctions in the same transcript. Splice junctions with very low relative intron support are more likely to be false.

Features and thresholds

Following the exploratory analysis, a set of filtering steps based on those features were tested on the transcript and intron dataset.

A transcript with at least one intron meeting any of these conditions would be rejected:

- splice site sequence is not GT-AG
- splice site overlaps an exon on the opposite strand
- splice site overlaps a type I transposon or a tandem repeat
- intron length is smaller than 50 nt
- intron has no RNA-seq short read support
- relative intron support is smaller than -7

Outcome prediction using the chosen features

The transcript and intron quality predicted using the filtering steps described in the previous slide was compared with the outcome of the manual assessment by annotators (slide 1).

	Introns	Transcripts[a]
Accepted	9134	851
Rejected	1876	1133
Rejection rate	17.0%	57.1%
TP	9133	820
FP	1	65[b]
TN	408	327
FN	1468	772
Sensitivity	86.1%	51.5%
Specificity	99.8%	83.4%

a) a transcript is rejected if at least one of its introns is rejected

b) mostly already in annotation, wrong locus or extension of existing transcript

	TP	FP	TN	FN	FPR	sensitivity	specificity
Splice site seq	10448	177	232	153	0.433	0.986	0.567
Antisense	10377	292	117	224	0.714	0.979	0.286
Repeats	9812	129	280	789	0.315	0.926	0.685
Intron length	10601	405	4	0	0.99	1	0.01
Intron score	10066	158	251	535	0.386	0.95	0.614
Relative intron score	10312	341	68	289	0.834	0.973	0.166
All	9133	1	408	1468	0.002	0.862	0.998