SQANTI3 filter report

Total Genes: 656

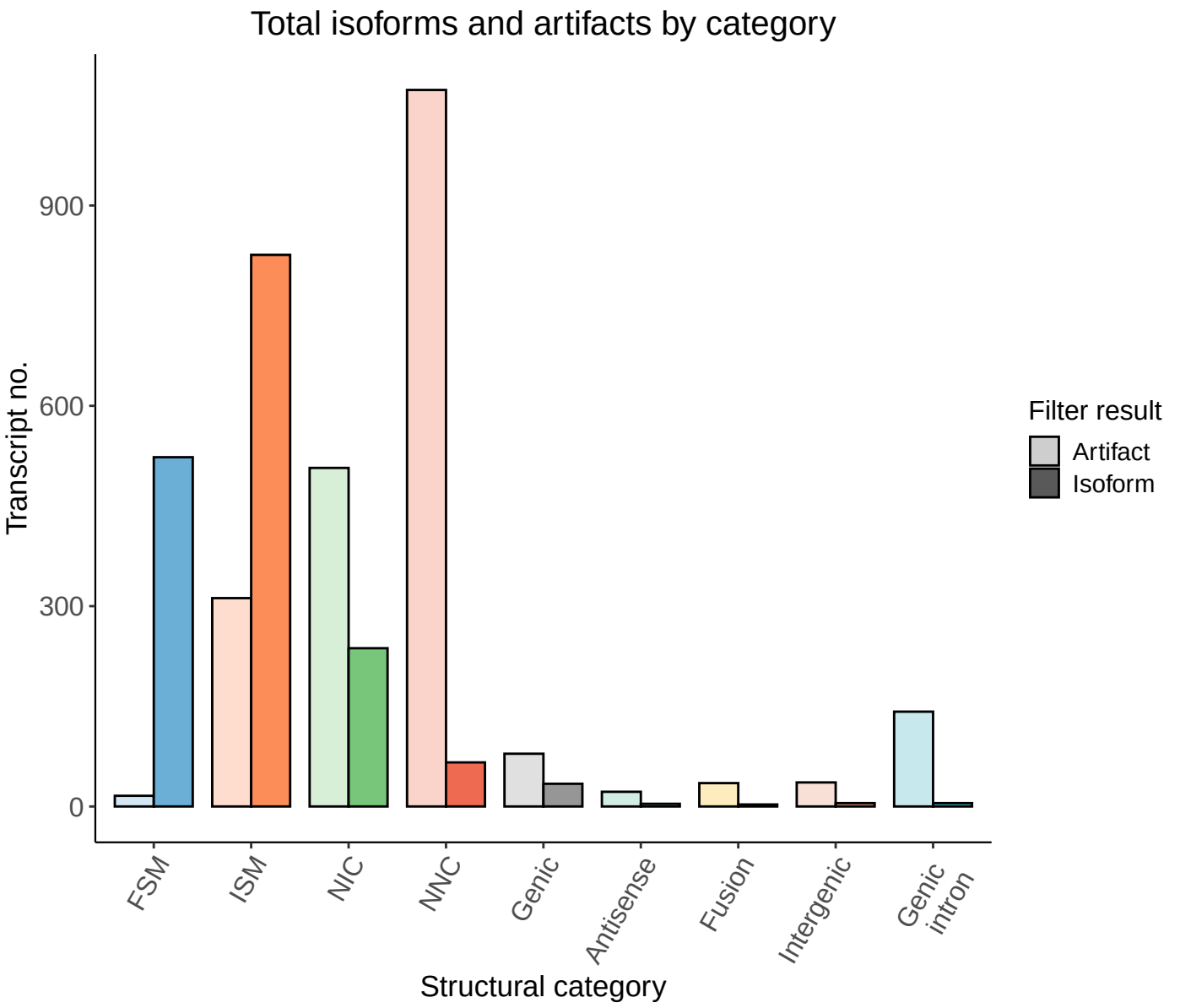
Total Transcripts: 3925

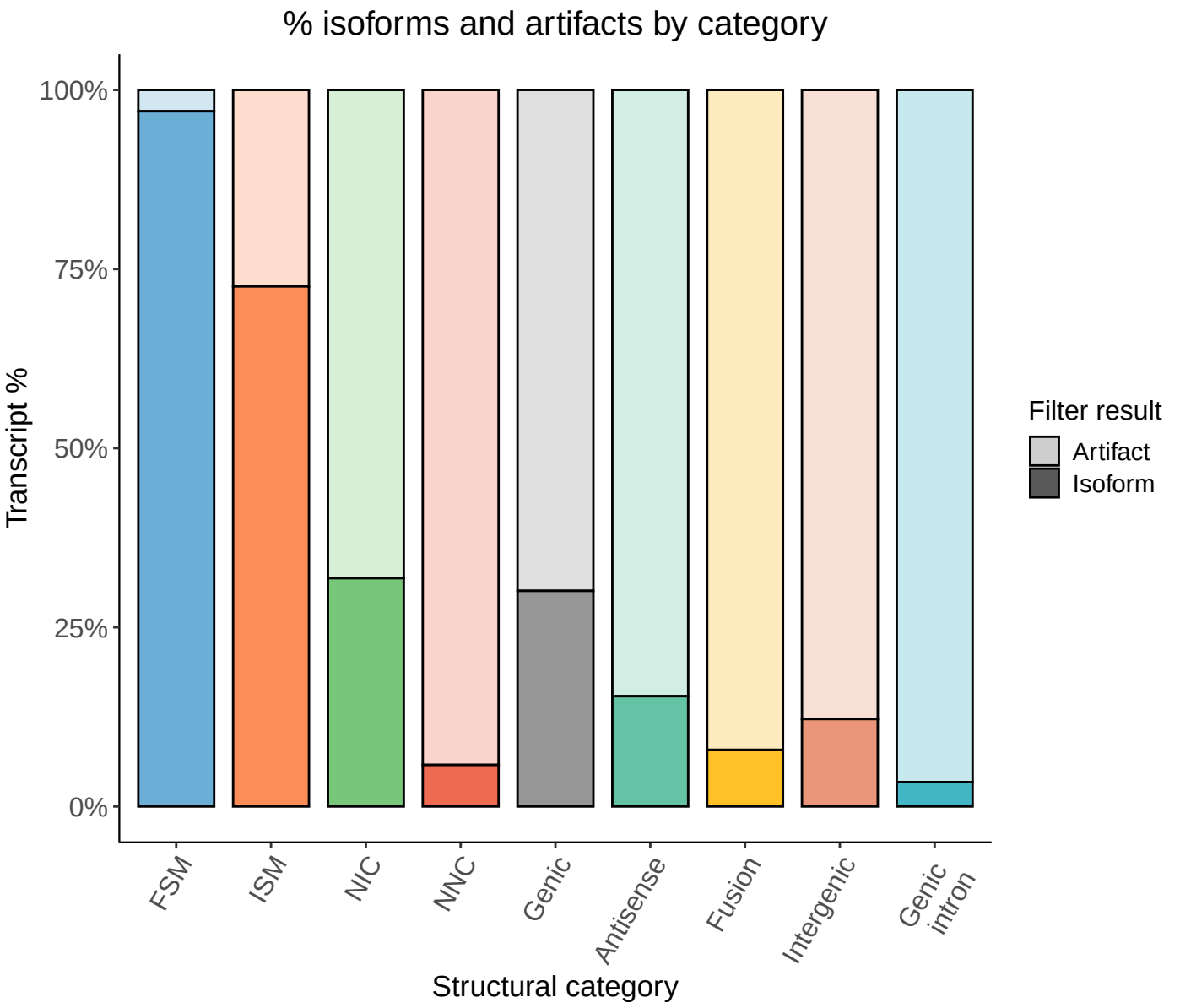
− Isoforms: 1703 (43%)

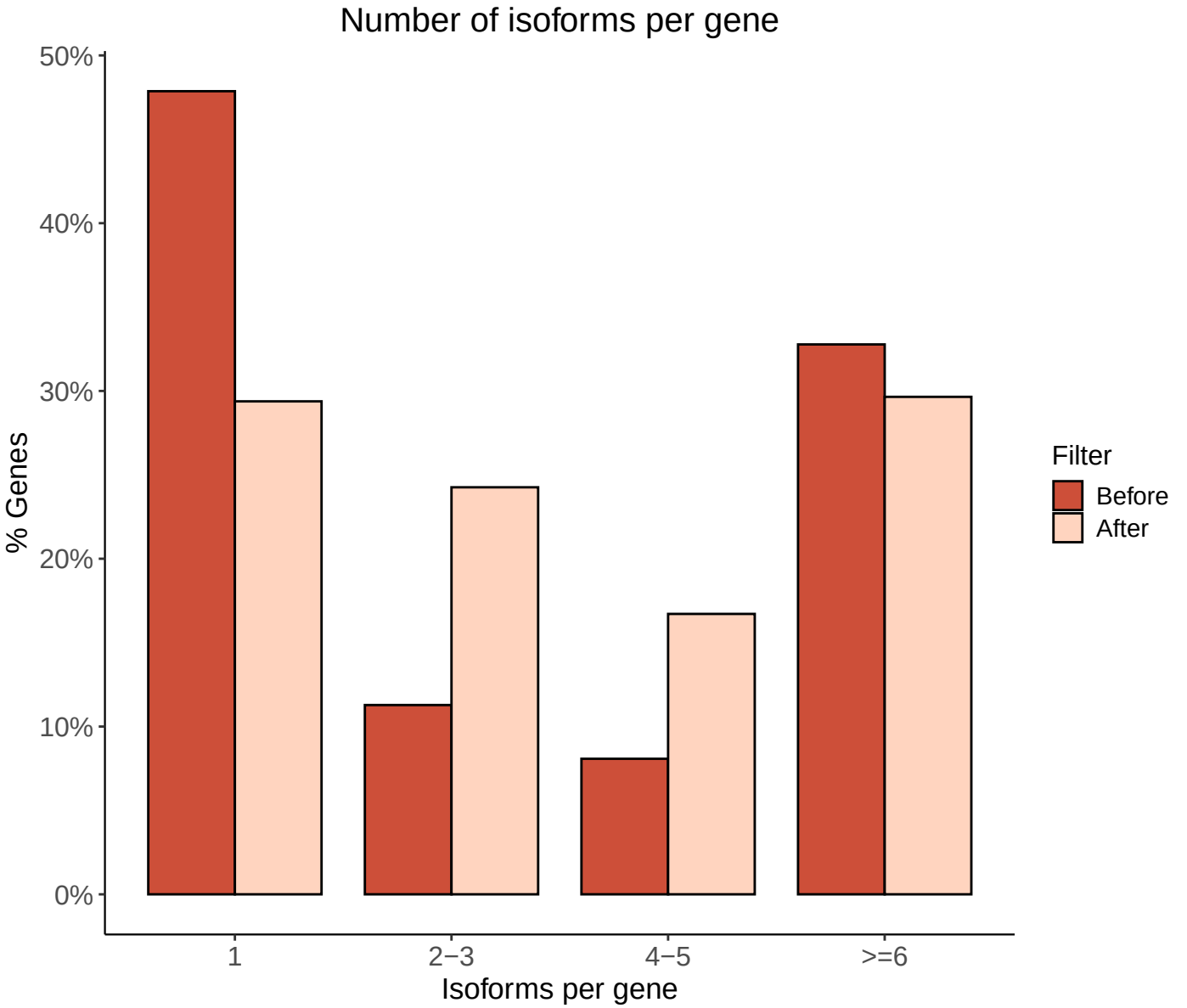
− Artifacts: 2222 (57%)

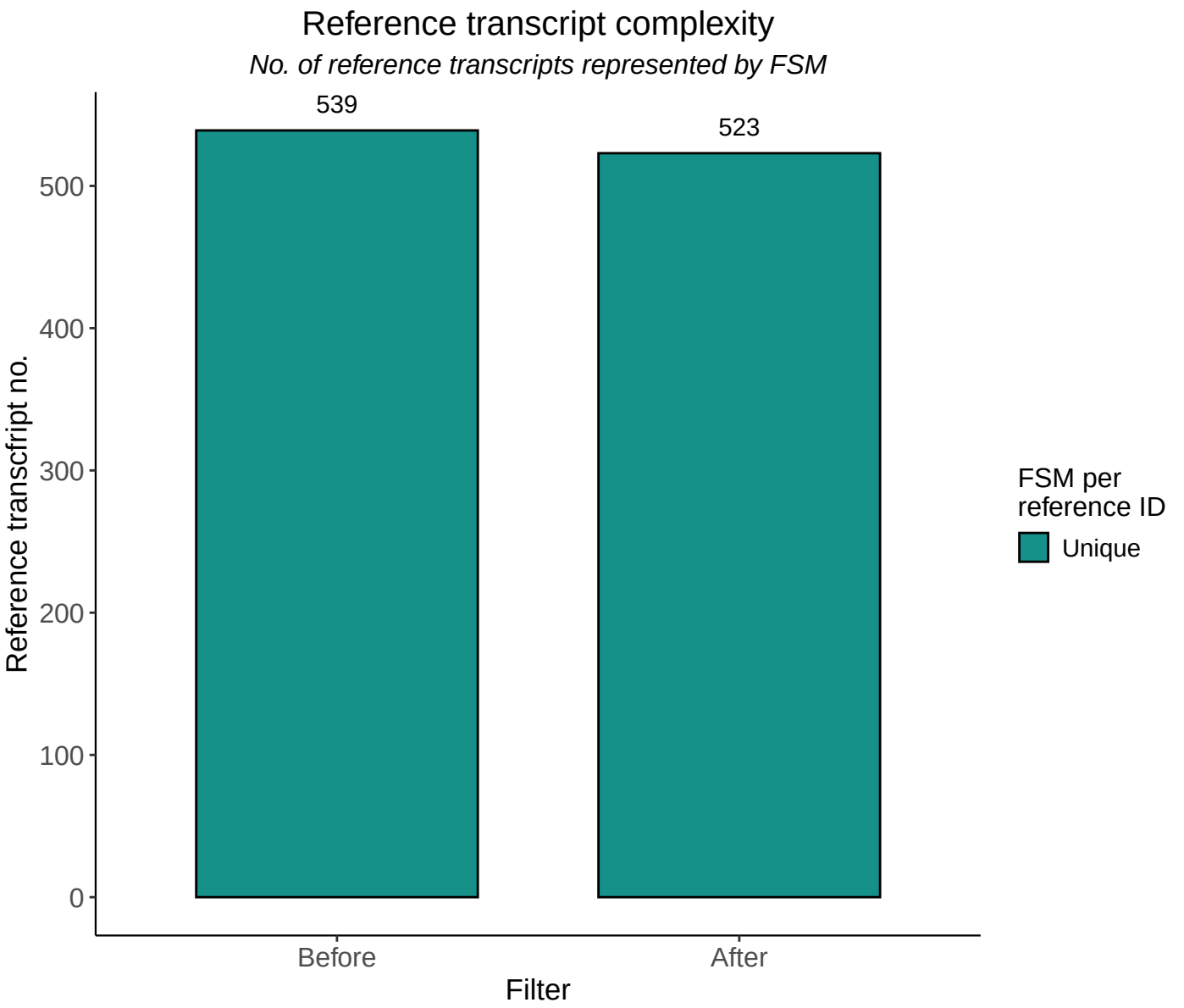
Gene category	Gene no.	No. of genes with artifacts only
Annotated	443	86
Novel	213	199

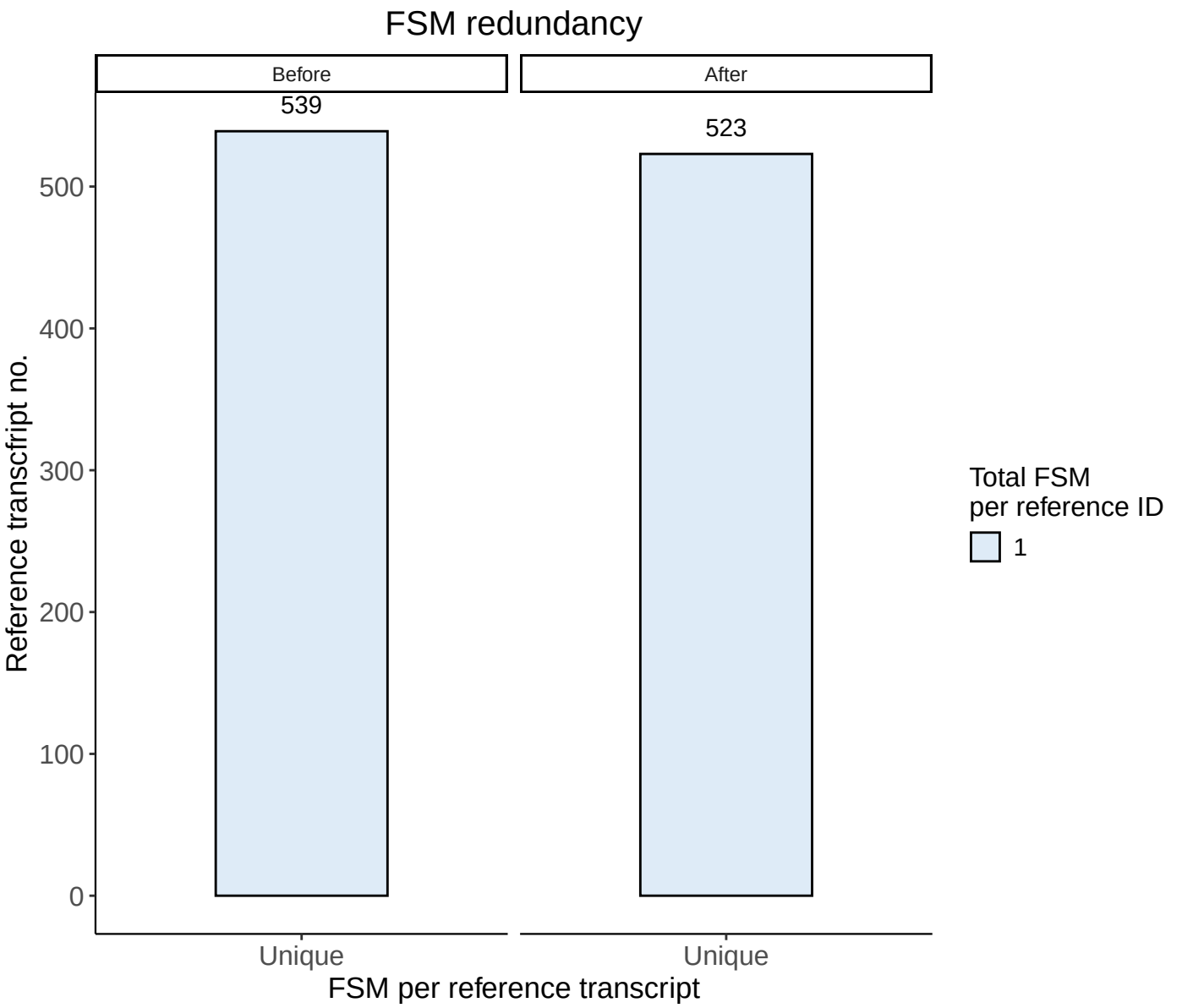
Structural category	Artifact no.	Isoform no.
FSM	16	523
ISM	312	826
NIC	507	237
NNC	1073	66
Genic	79	34
Antisense	22	4
Fusion	35	3
Intergenic	36	5
Genic intron	142	5

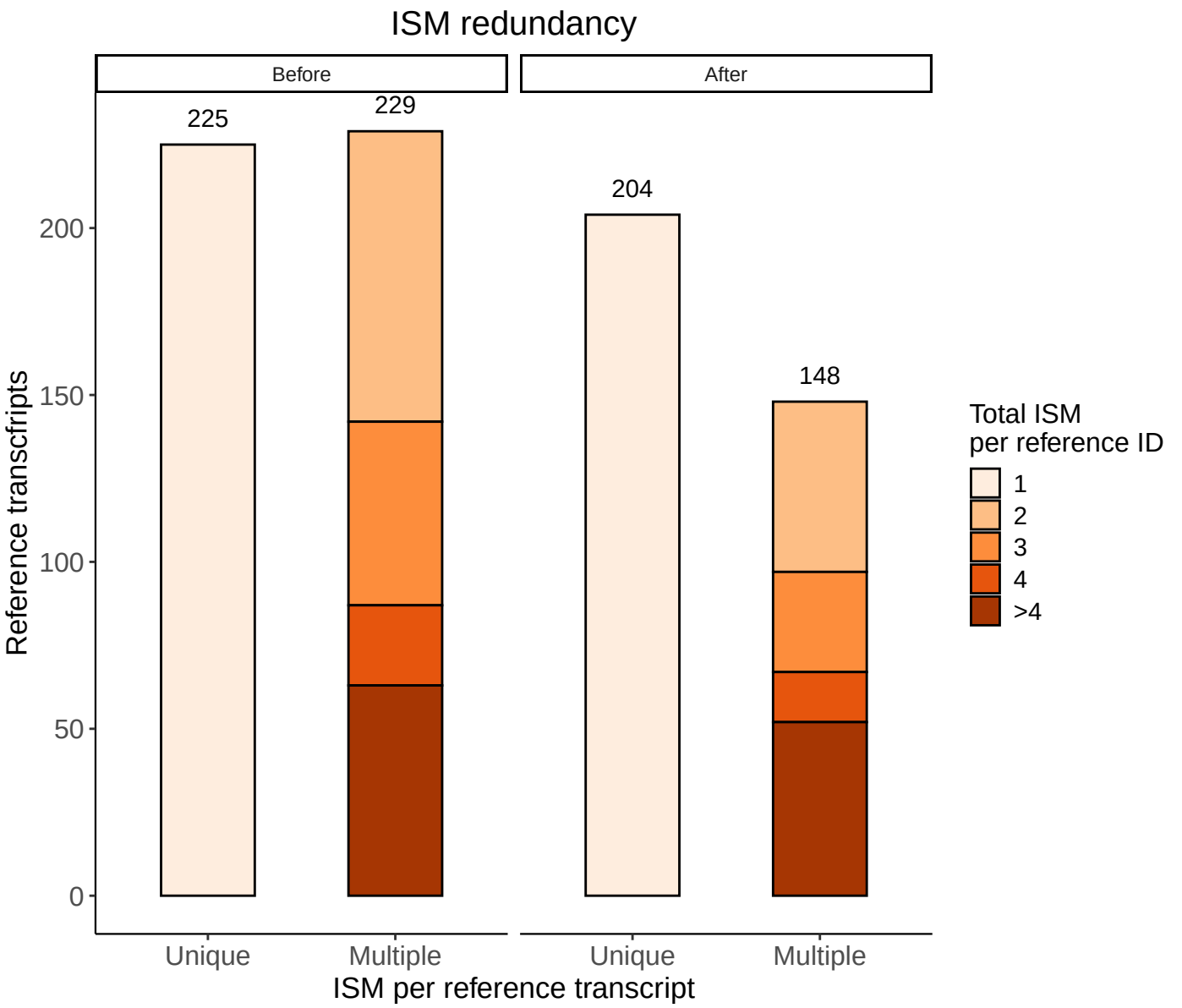

Total isoforms and artifacts by category
Transcript no.
900
600
300
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

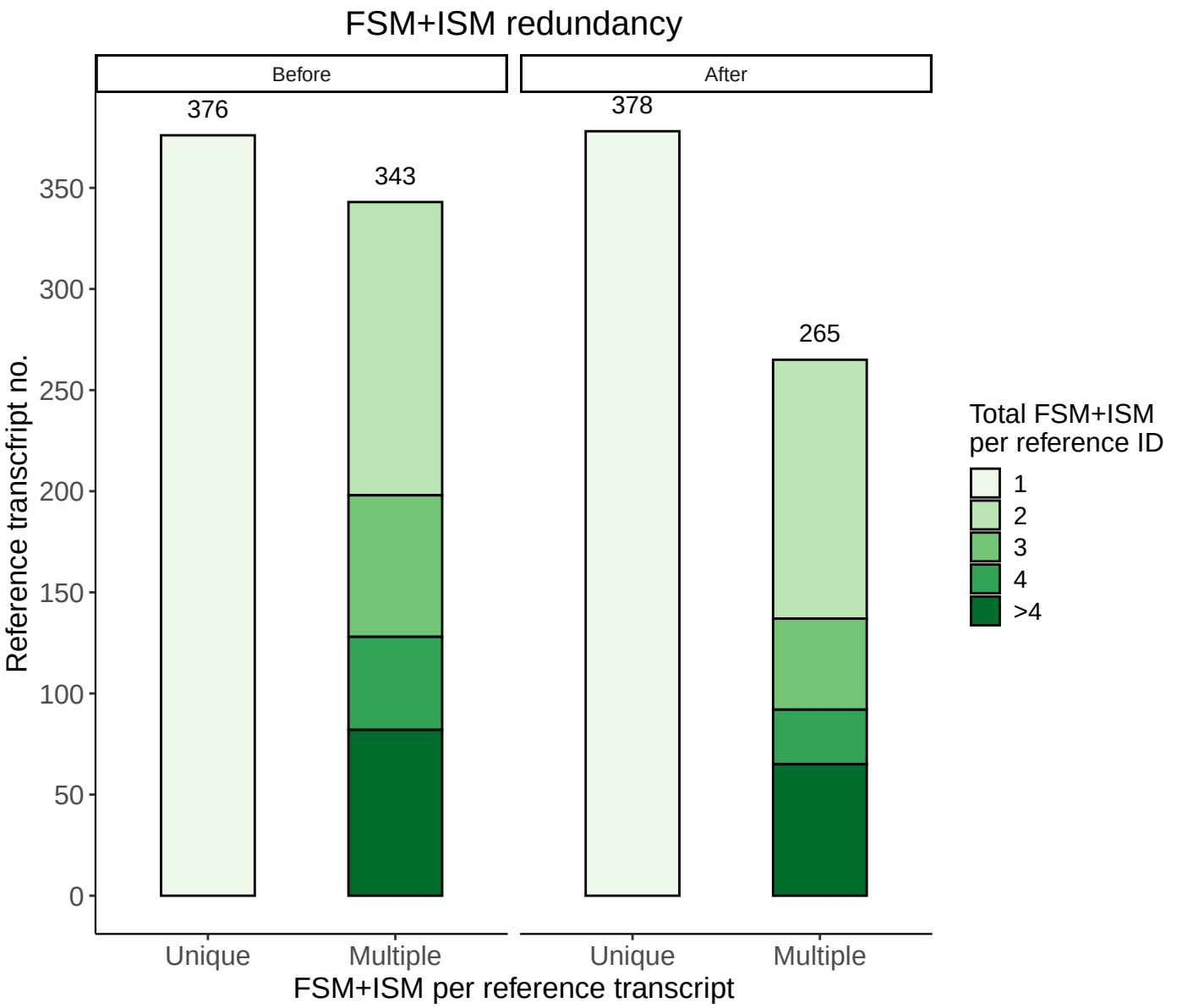

% isoforms and artifacts by category
100%
75%
50%
25%
0%
Transcript %
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform


Number of isoforms per gene
50%
40%
30%
20%
10%
0%
% Genes
1
2−3
4−5
>=6
Isoforms per gene
Filter
Before
After


Reference transcript complexity
No. of reference transcripts represented by FSM
539
523
500
400
300
200
100
0
Reference transcfript no.
Before
After
Filter
FSM per reference ID
Unique


FSM redundancy
Before
After
539
523
500
400
300
200
100
0
Reference transcfript no.
Unique
Unique
FSM per reference transcript
Total FSM
per reference ID
1


ISM redundancy
Before
After
225
229
204
148
200
150
100
50
0
Reference transcfripts
Unique
Multiple
Unique
Multiple
ISM per reference transcript
Total ISM
per reference ID
1
2
3
4
>4


FSM+ISM redundancy
Before
After
376
343
378
265
Reference transcfript no.
0
50
100
150
200
250
300
350
Unique
Multiple
Unique
Multiple
FSM+ISM per reference transcript
Total FSM+ISM per reference ID
1
2
3
4
>4

ML classifier performance report

Classification model performance on test set

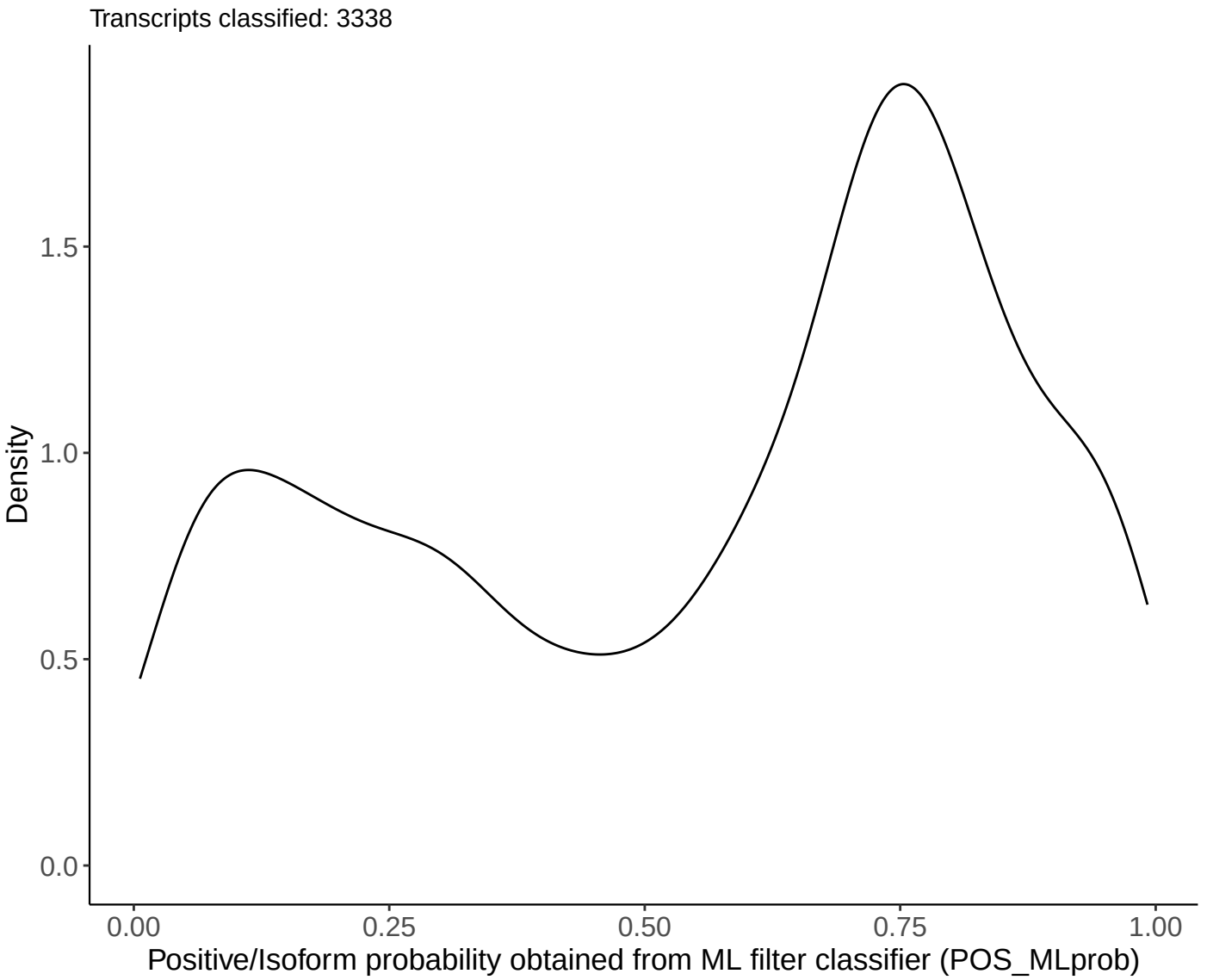
Performance metrics

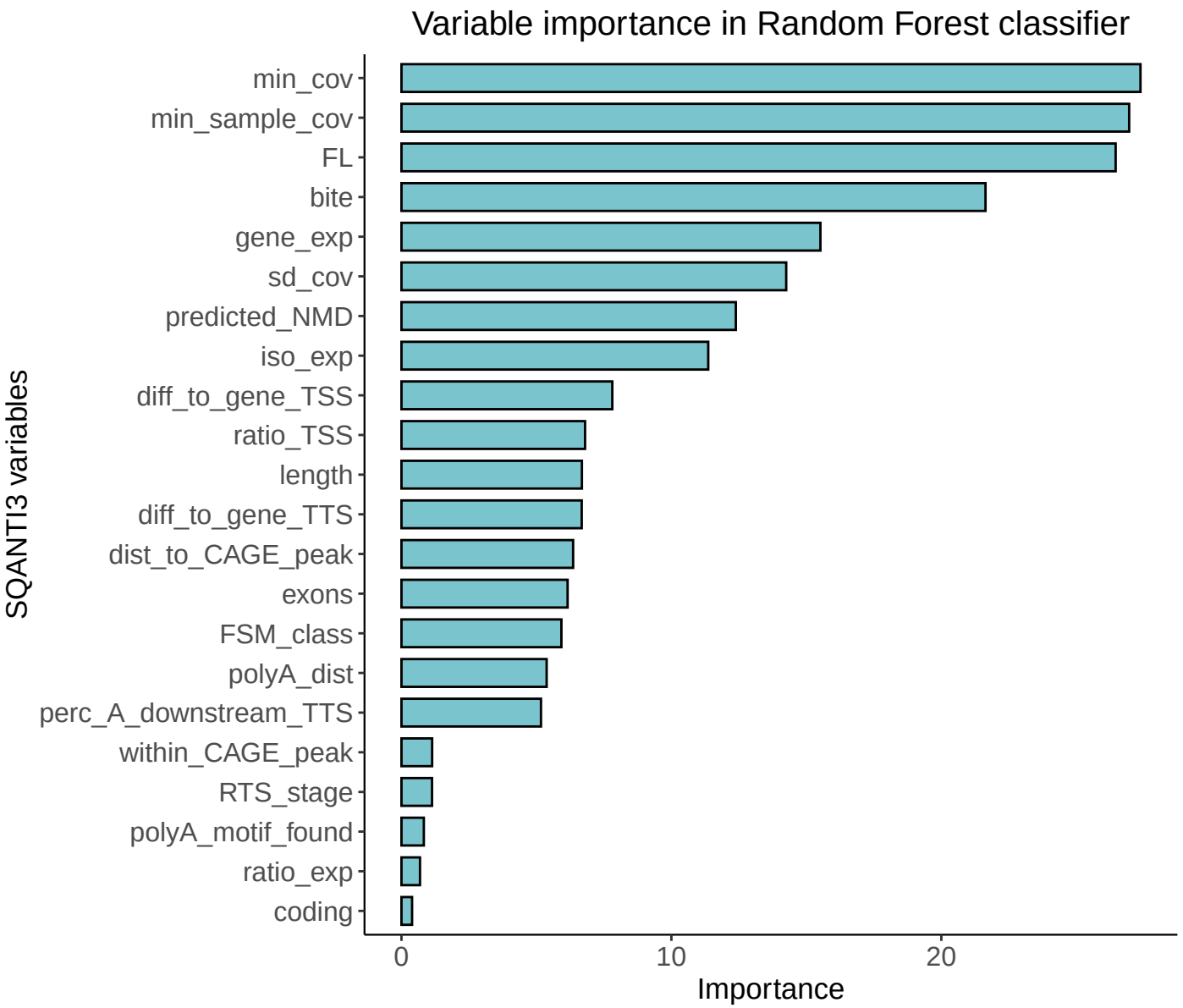
Metric	Value
Accuracy	0.912
Kappa	0.825
AccuracyLower	0.845
AccuracyUpper	0.957
AccuracyNull	0.5
McnemarPValue	0.0269
Sensitivity	0.842
Specificity	0.982
Pos Pred Value	0.980
Neg Pred Value	0.862
Precision	0.980
Recall	0.842
F1	0.906
Prevalence	0.5
Detection Rate	0.421
Detection Prevalence	0.430
Balanced Accuracy	0.912

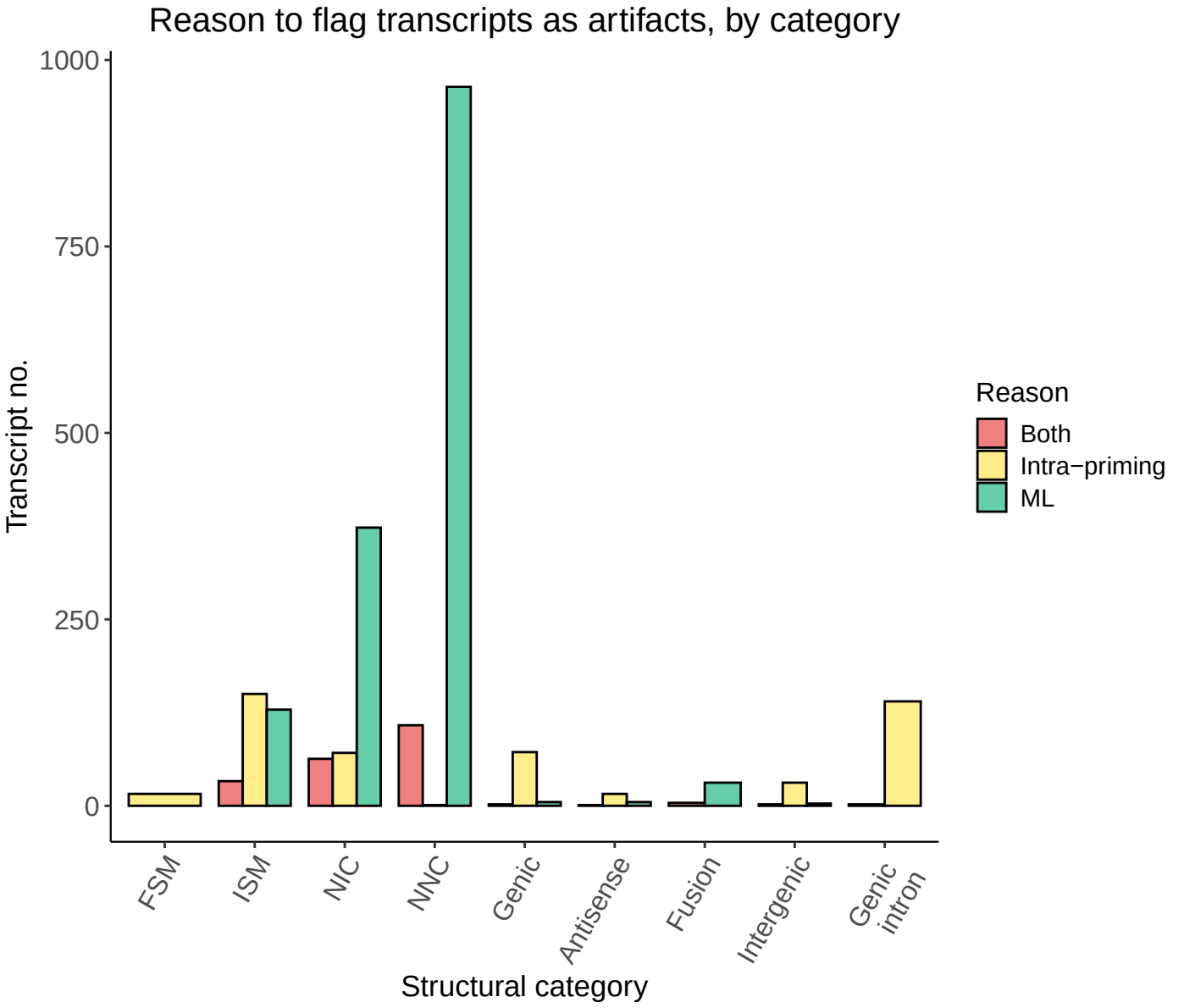
AccuracyPValue	3.617604e−21

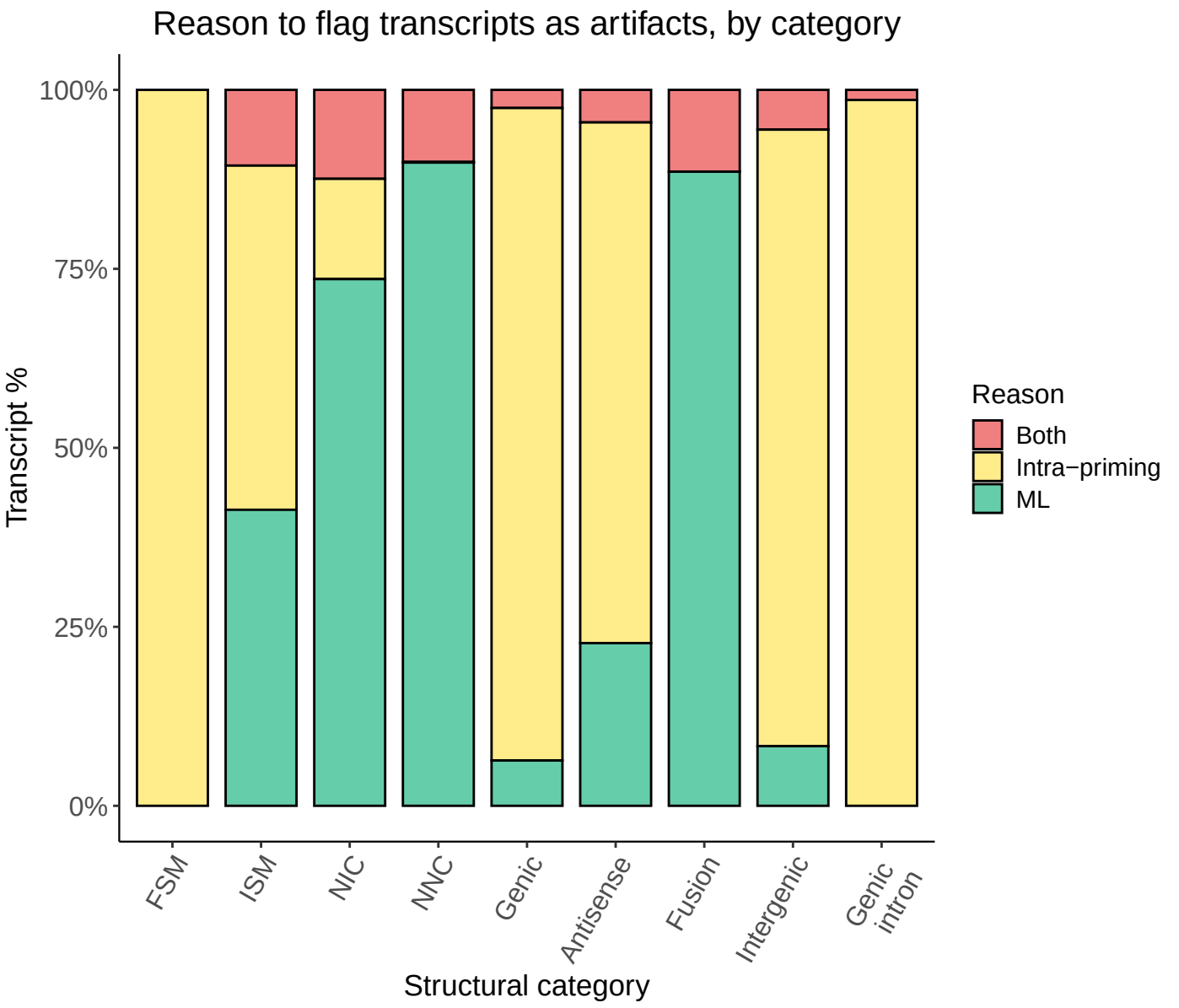
Confusion matrix

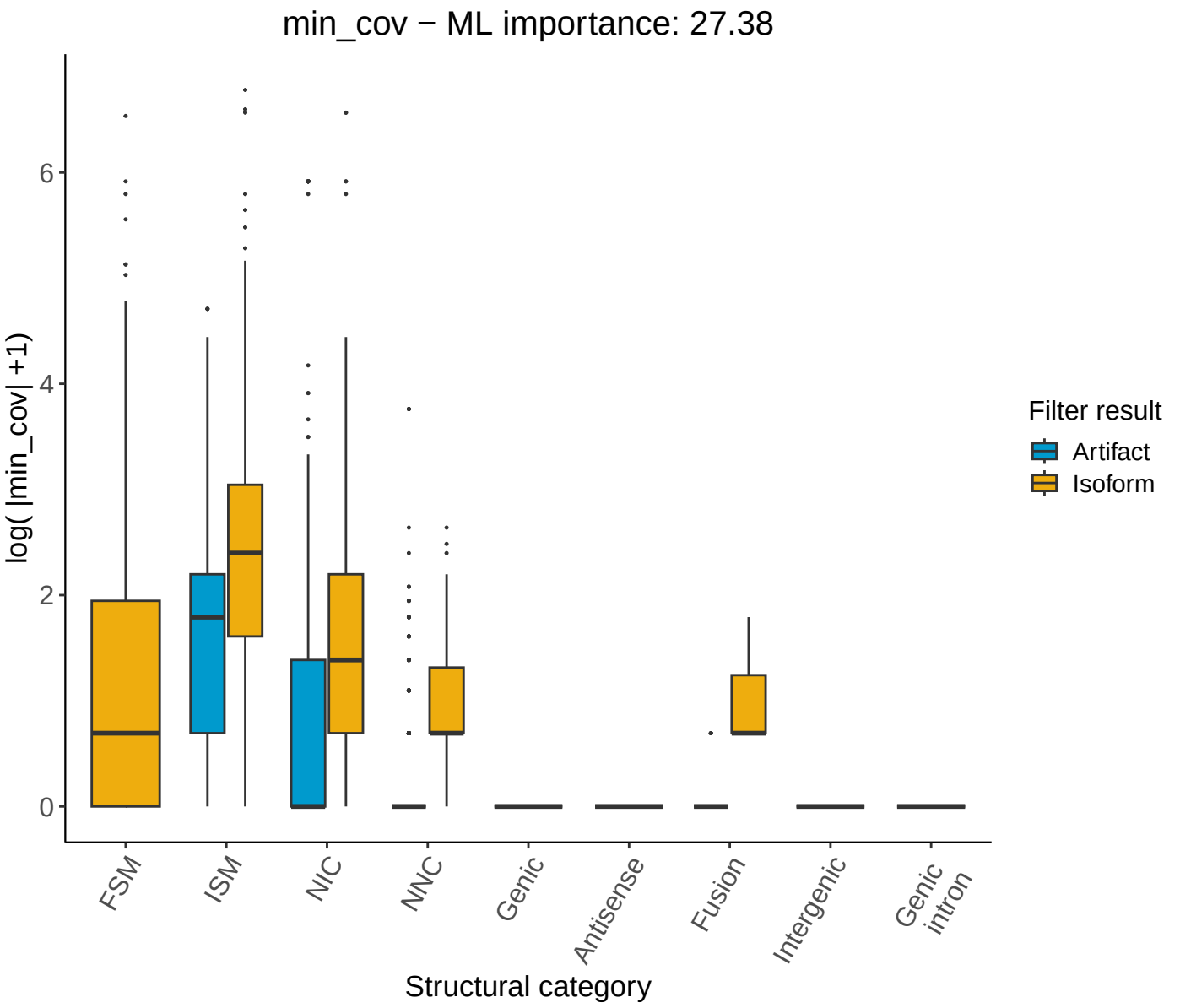
Prediction	Reference	Freq
POS	POS	48
NEG	POS	9
POS	NEG	1
NEG	NEG	56

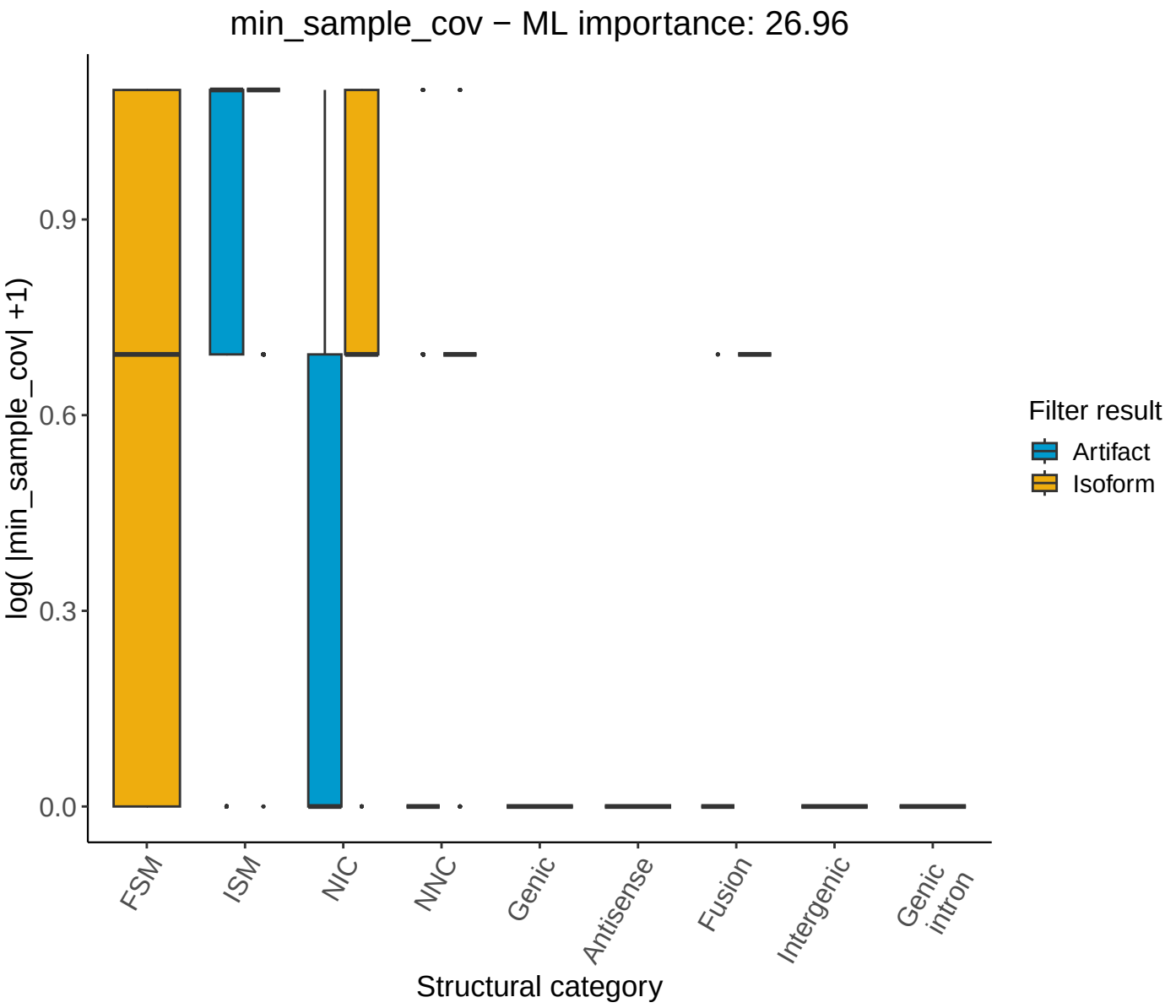

Transcripts classified: 3338
Density
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
Positive/Isoform probability obtained from ML filter classifier (POS_MLprob)

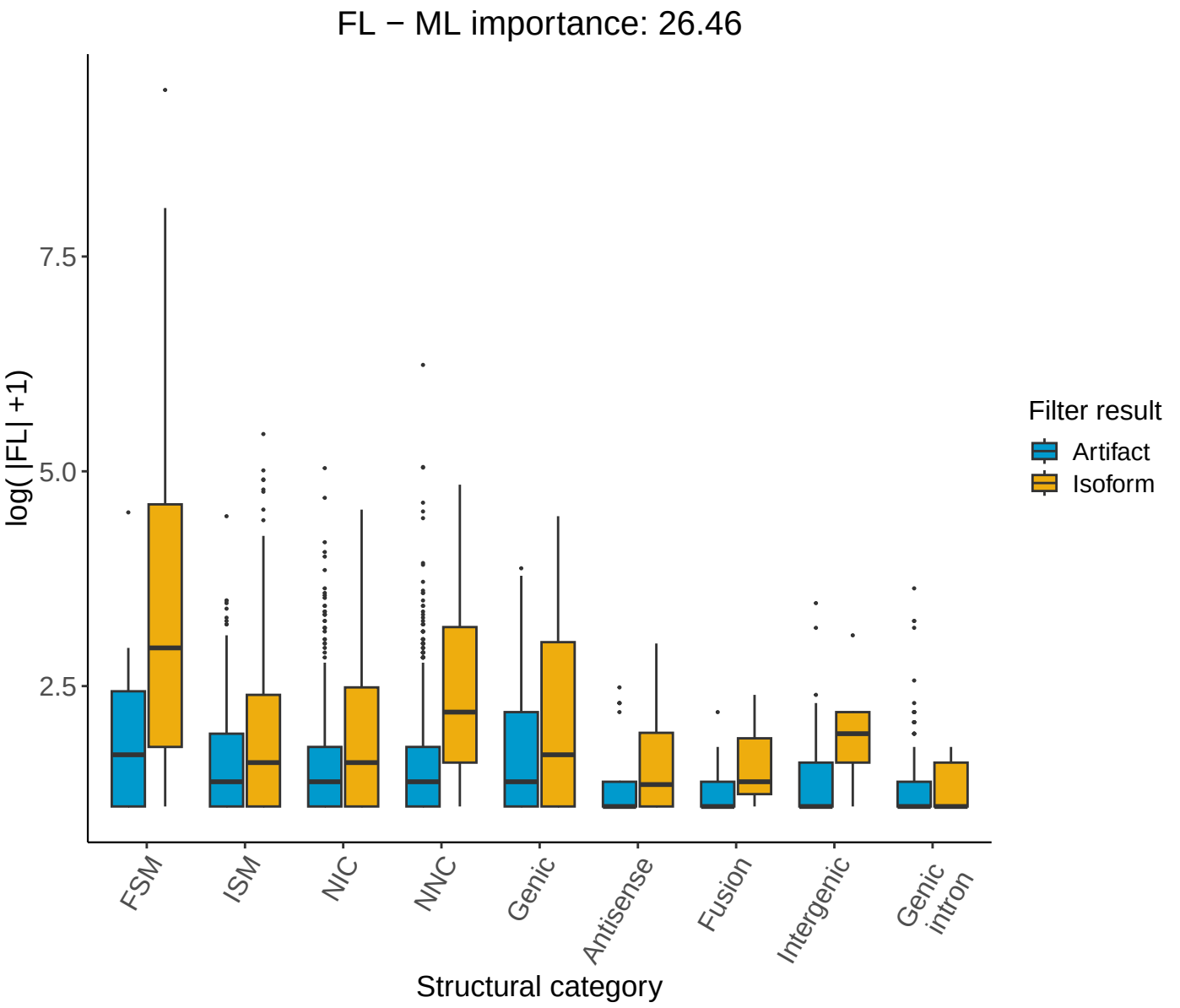

Variable importance in Random Forest classifier
SQANTI3 variables
min_cov
min_sample_cov
FL
bite
gene_exp
sd_cov
predicted_NMD
iso_exp
diff_to_gene_TSS
ratio_TSS
length
diff_to_gene_TTS
dist_to_CAGE_peak
exons
FSM_class
polyA_dist
perc_A_downstream_TTS
within_CAGE_peak
RTS_stage
polyA_motif_found
ratio_exp
coding
0
10
20
Importance

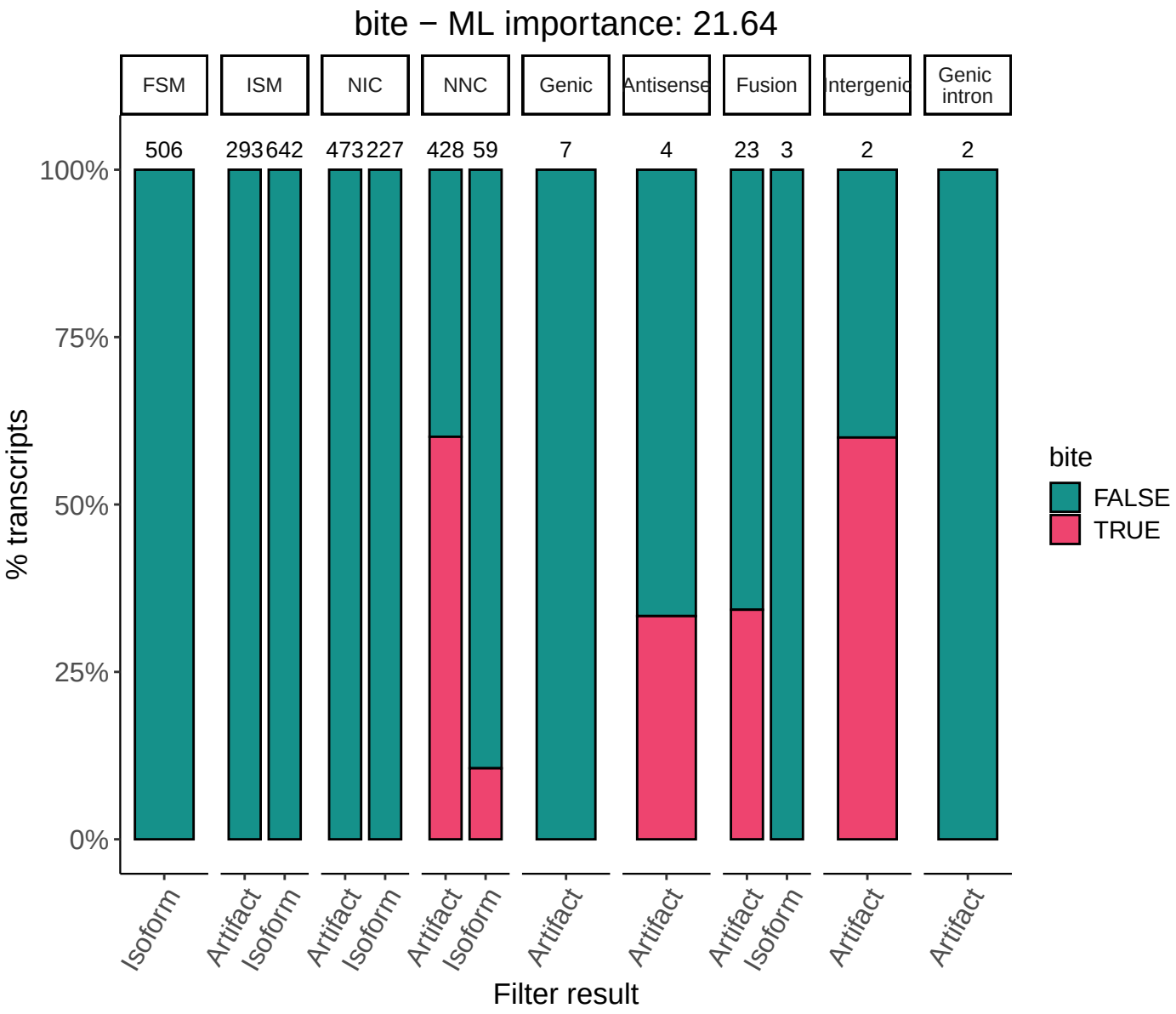

Reason to flag transcripts as artifacts, by category
Transcript no.
1000
750
500
250
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Reason
Both
Intra−priming
ML

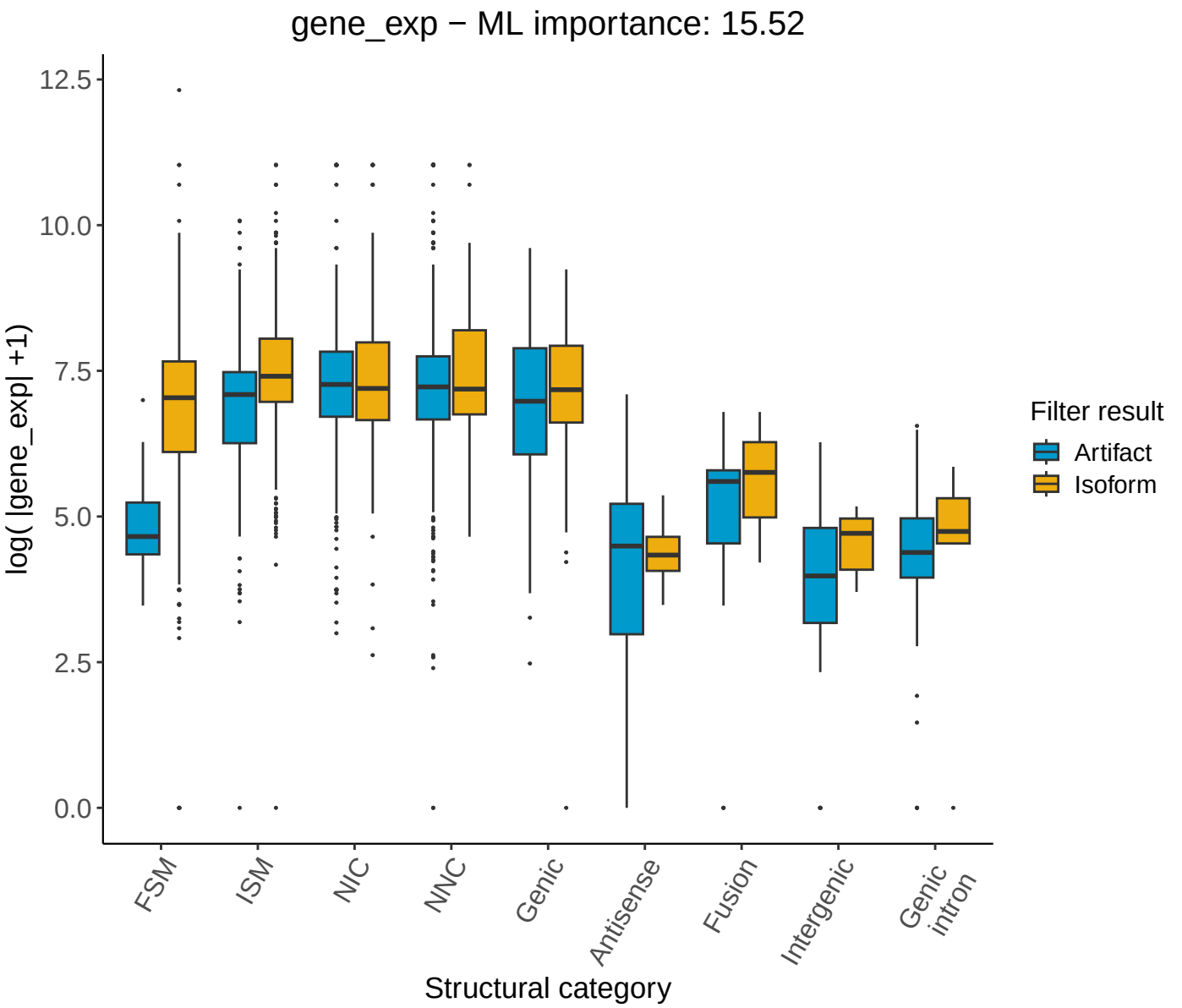

Reason to flag transcripts as artifacts, by category
100%
75%
50%
25%
0%
Transcript %
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Reason
Both
Intra−priming
ML

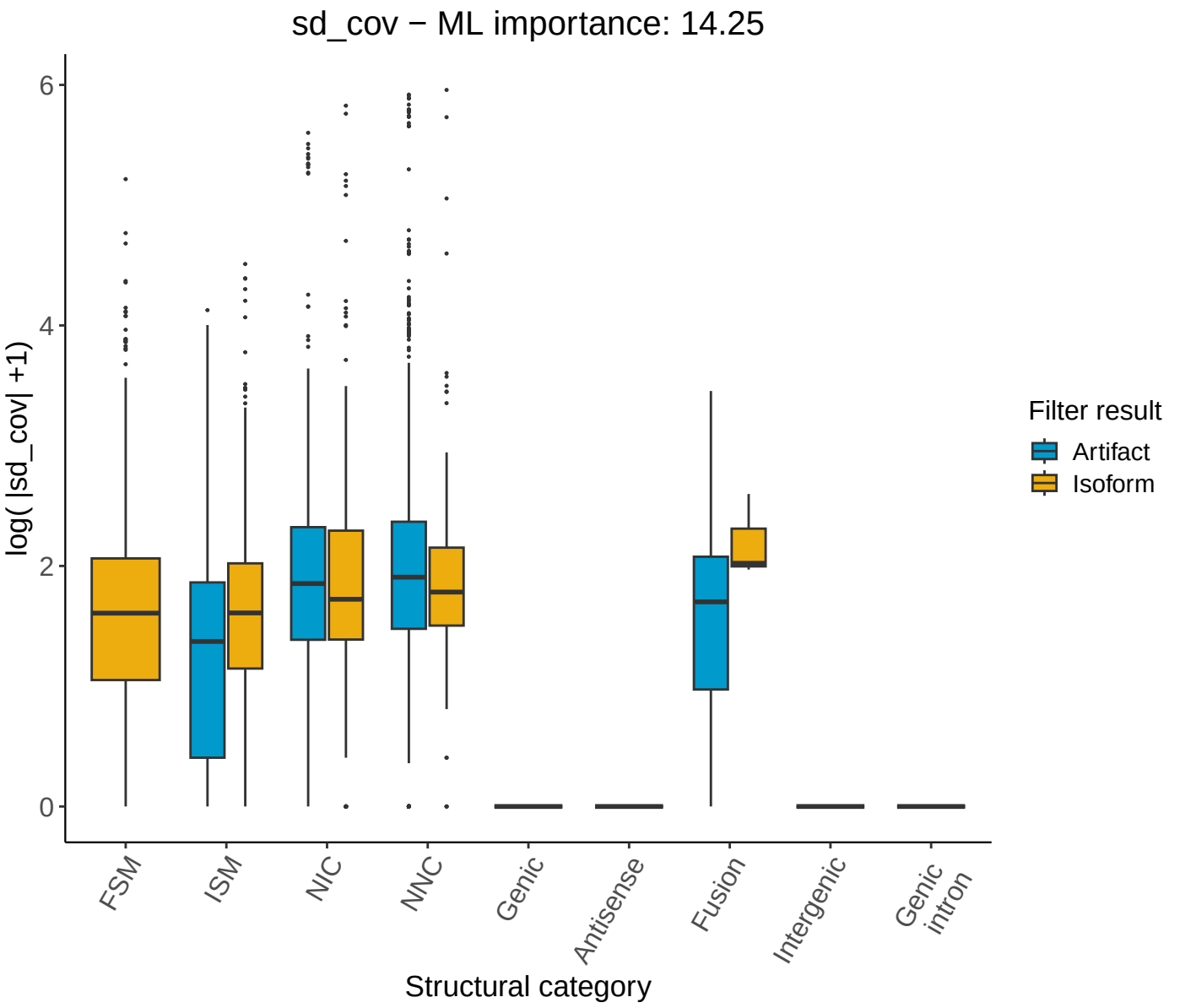

min_cov − ML importance: 27.38
log(|min_cov| +1)
6
4
2
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

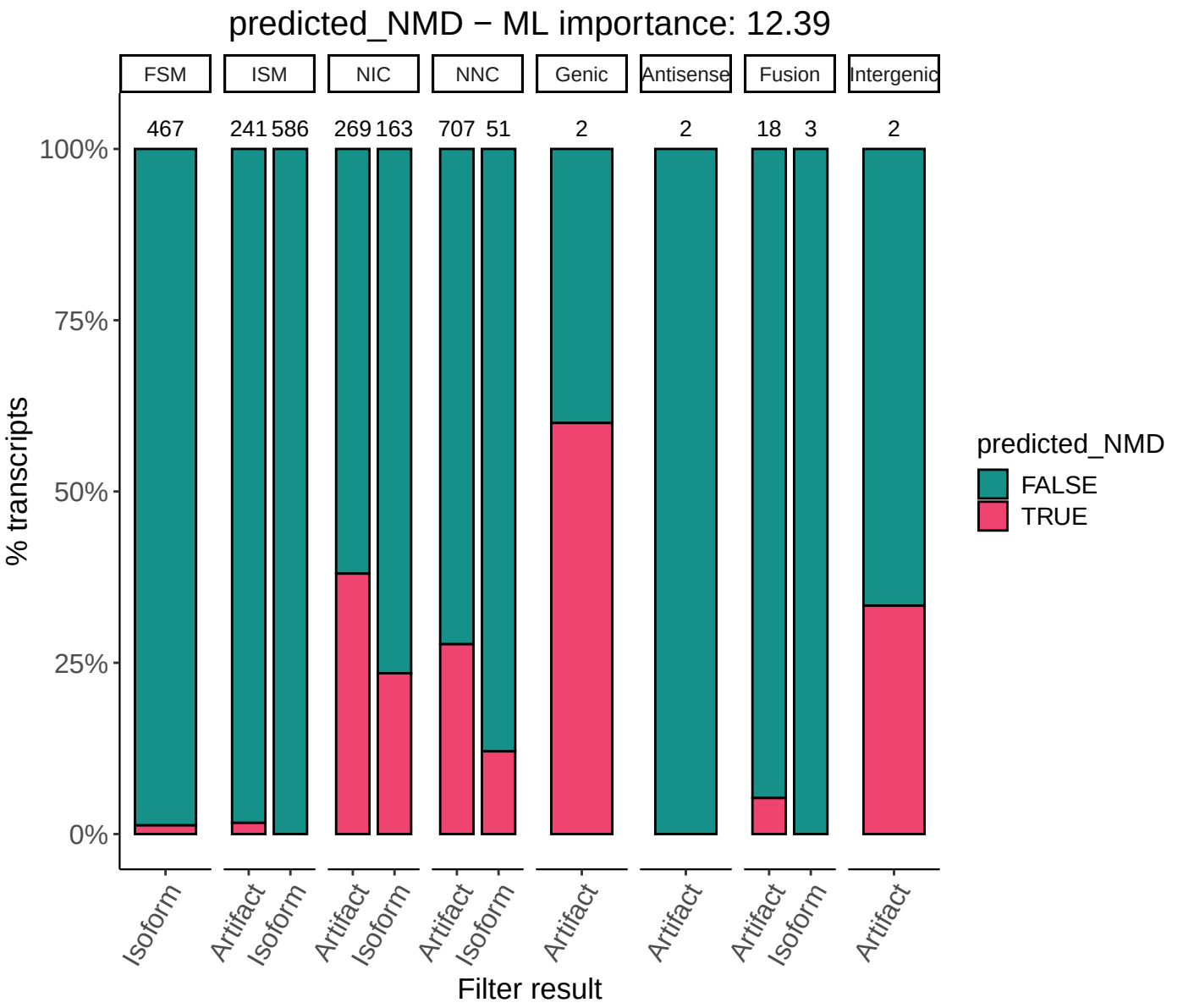

min_sample_cov − ML importance: 26.96
log(|min_sample_cov| +1)
0.0
0.3
0.6
0.9
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

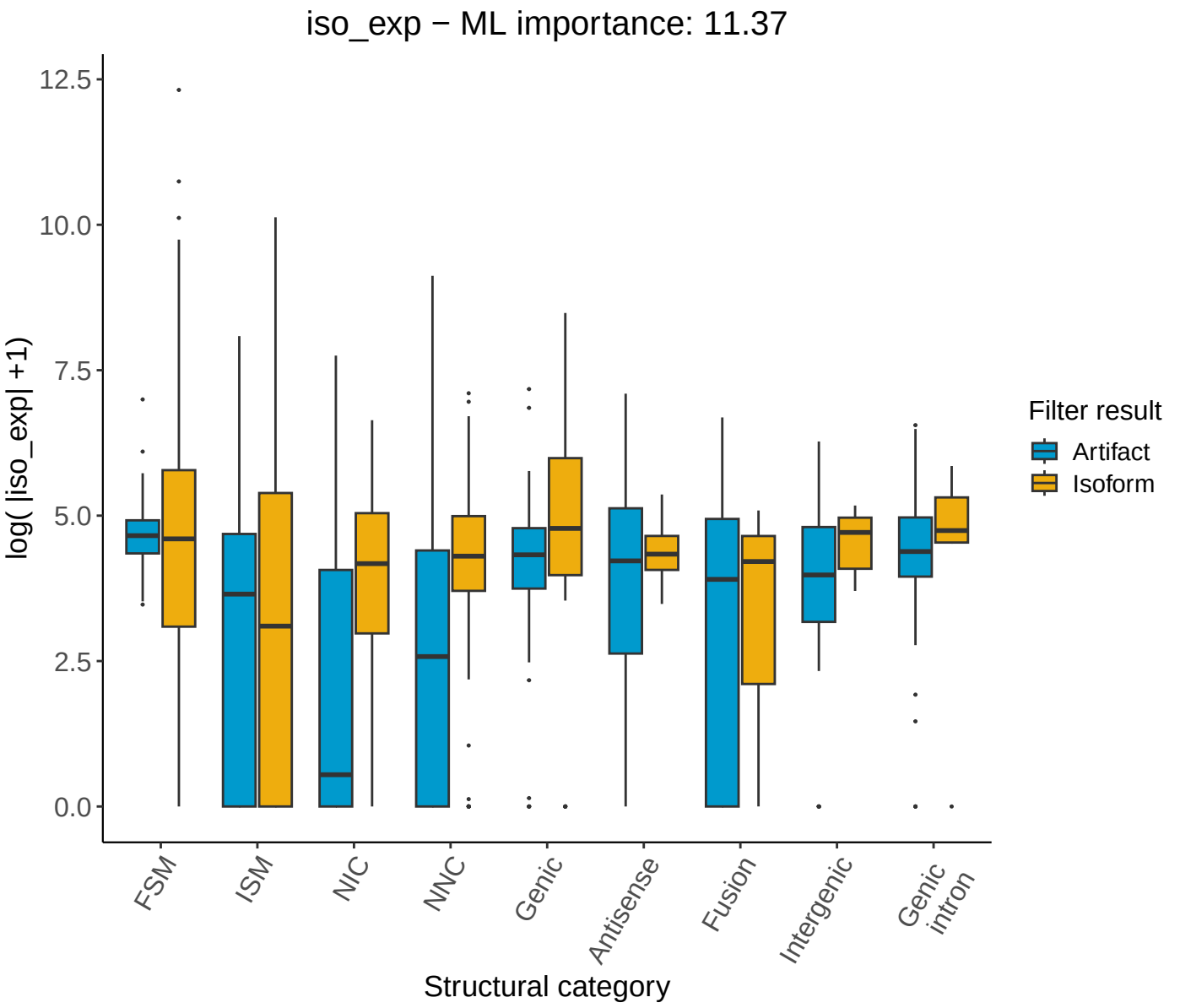

FL − ML importance: 26.46
log(|FL| +1)
7.5
5.0
2.5
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

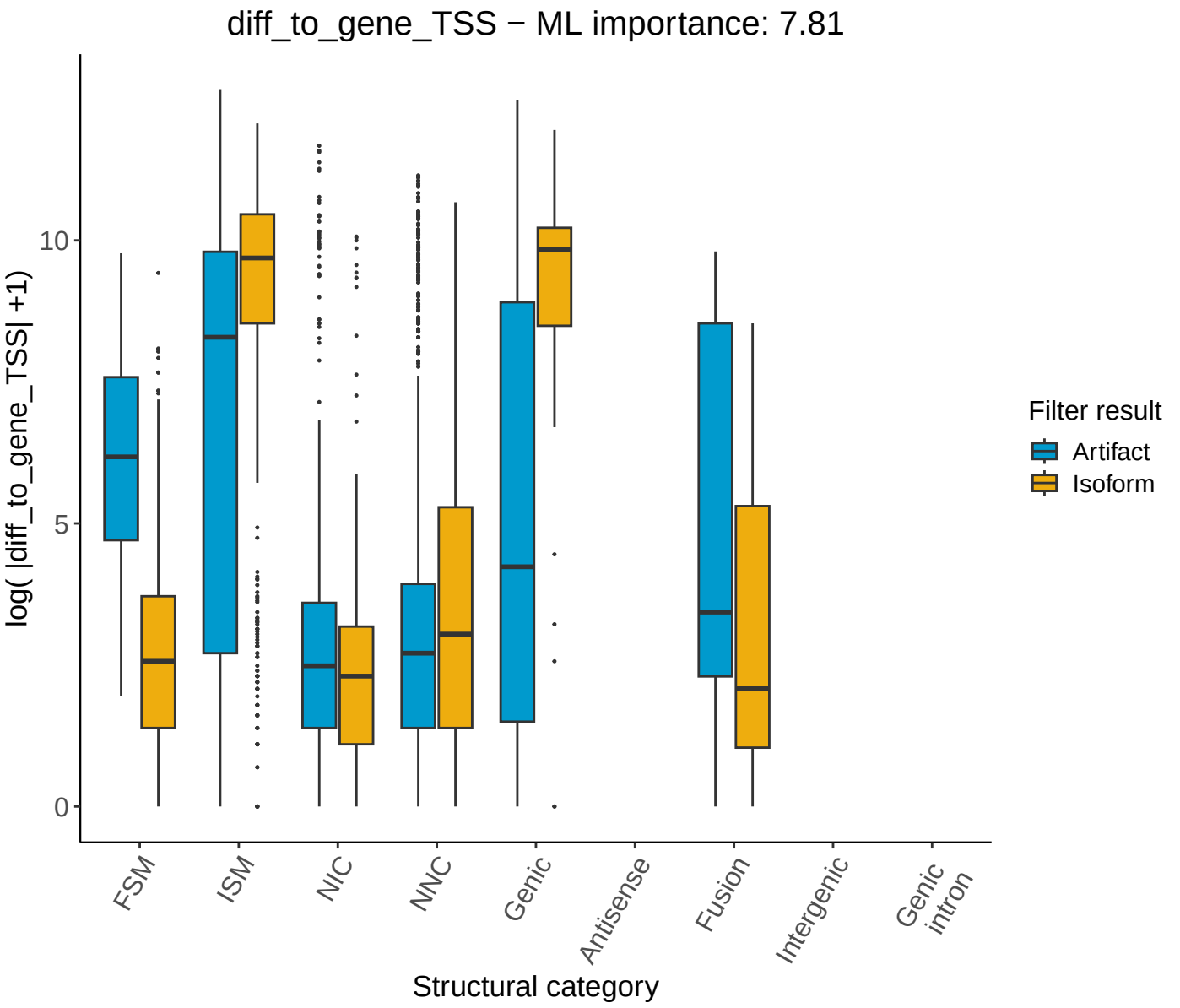

bite − ML importance: 21.64
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
506
293
642
473
227
428
59
7
4
23
3
2
2
100%
75%
50%
25%
0%
% transcripts
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Artifact
Artifact
Isoform
Artifact
Artifact
Filter result
bite
FALSE
TRUE

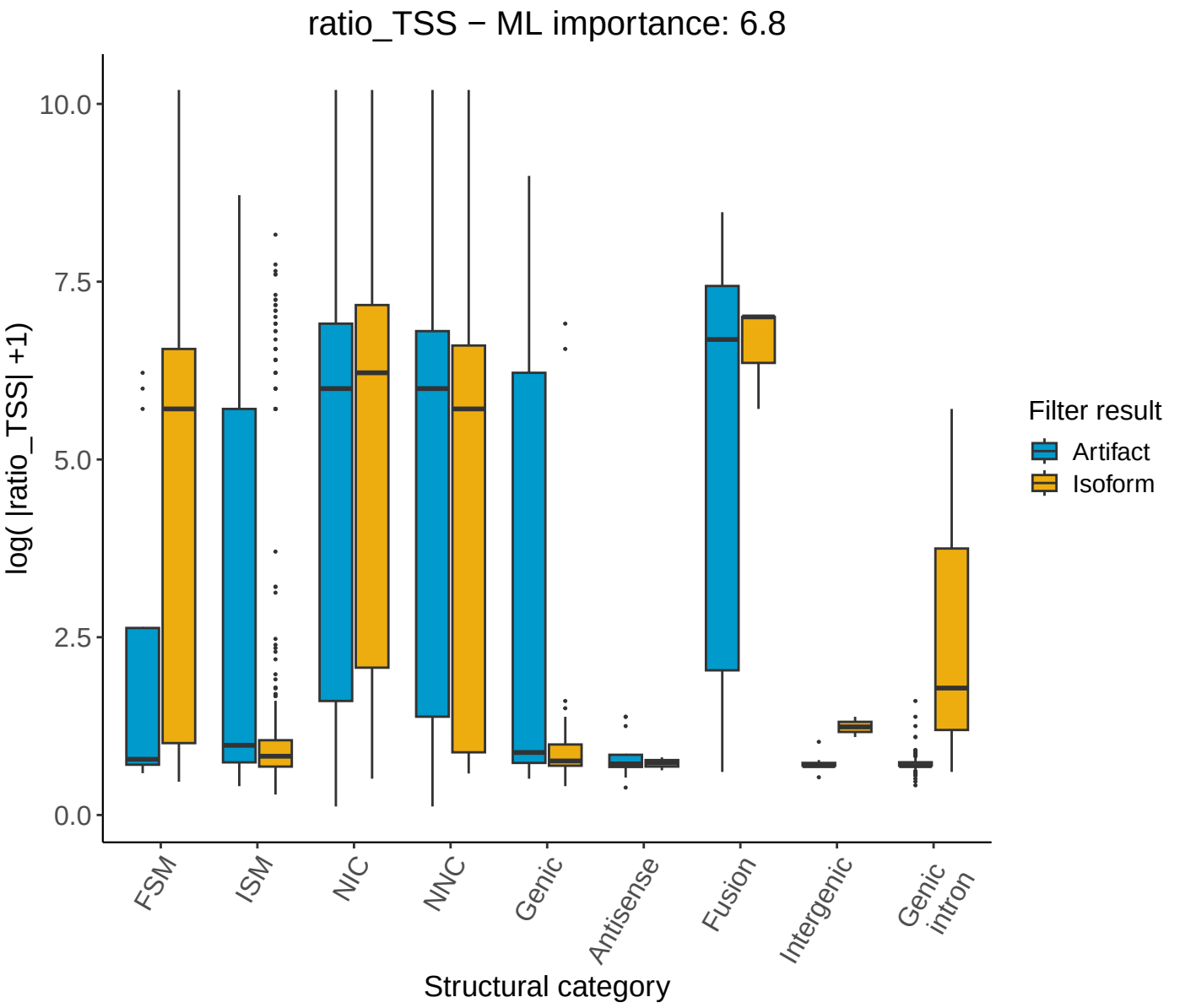

gene_exp − ML importance: 15.52
log(|gene_exp| +1)
12.5
10.0
7.5
5.0
2.5
0.0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

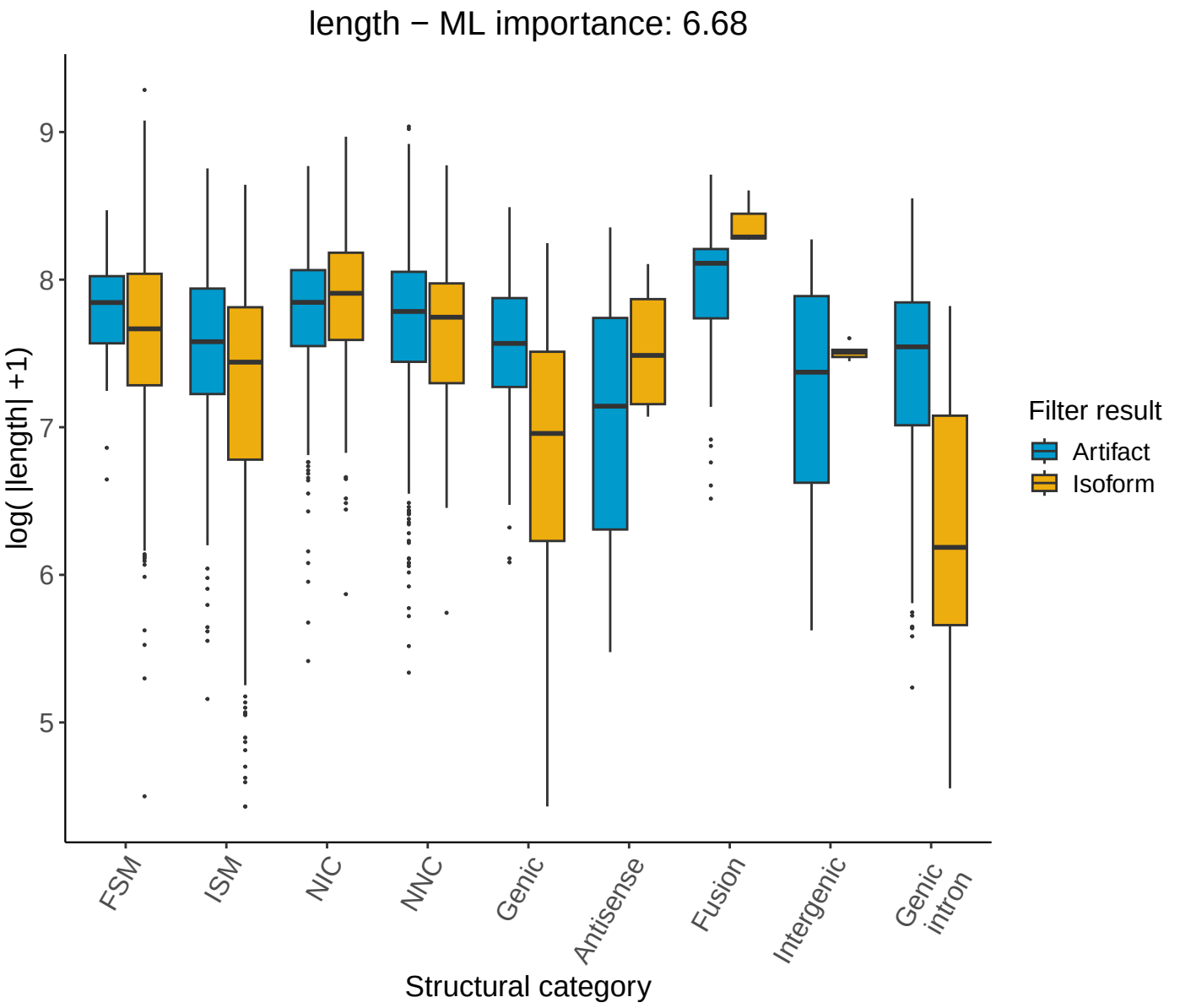

sd_cov − ML importance: 14.25
log(|sd_cov| +1)
6
4
2
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

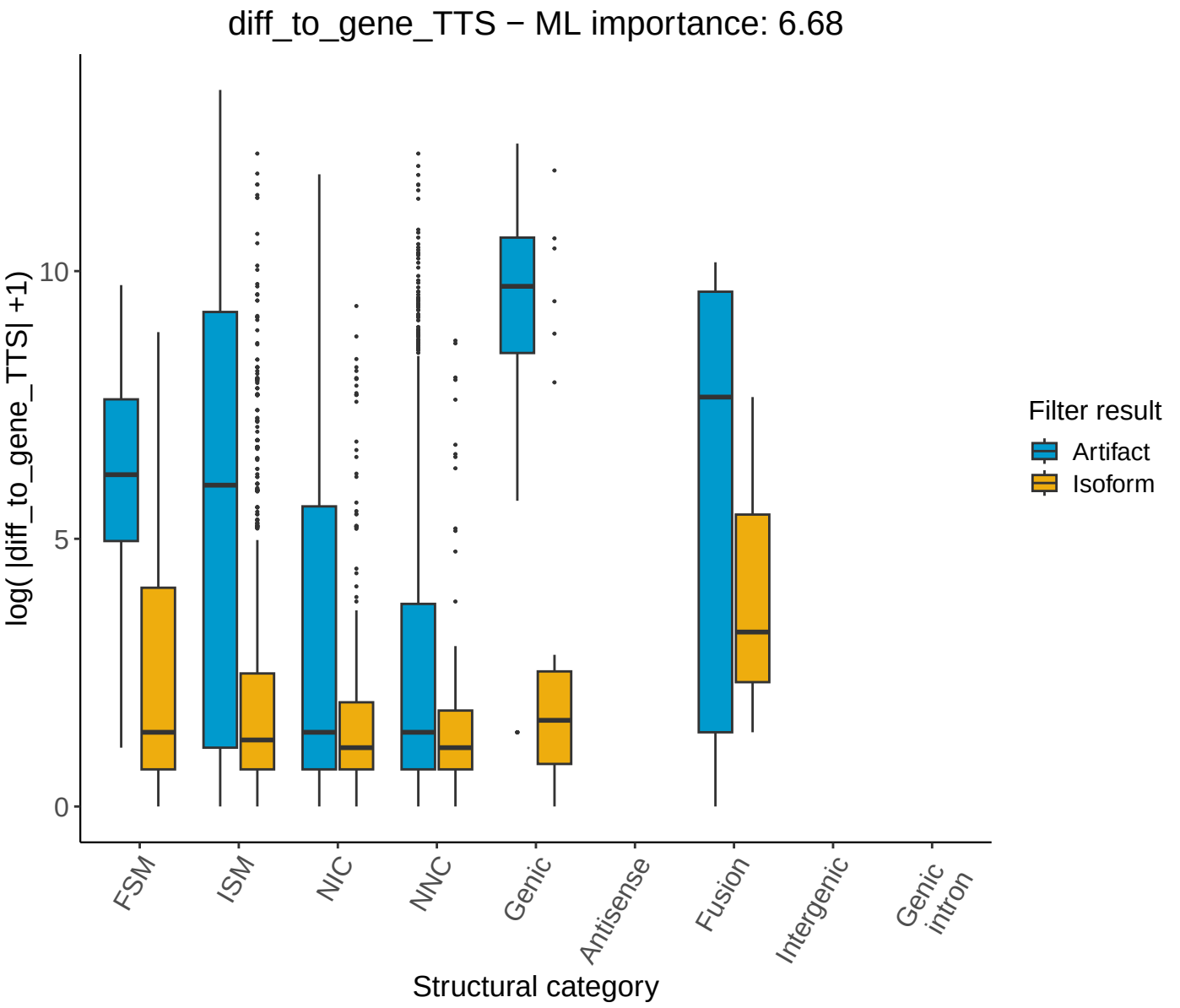

predicted_NMD − ML importance: 12.39
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
467
241
586
269
163
707
51
2
2
18
3
2
100%
75%
50%
25%
0%
% transcripts
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Artifact
Artifact
Isoform
Artifact
Filter result
predicted_NMD
FALSE
TRUE

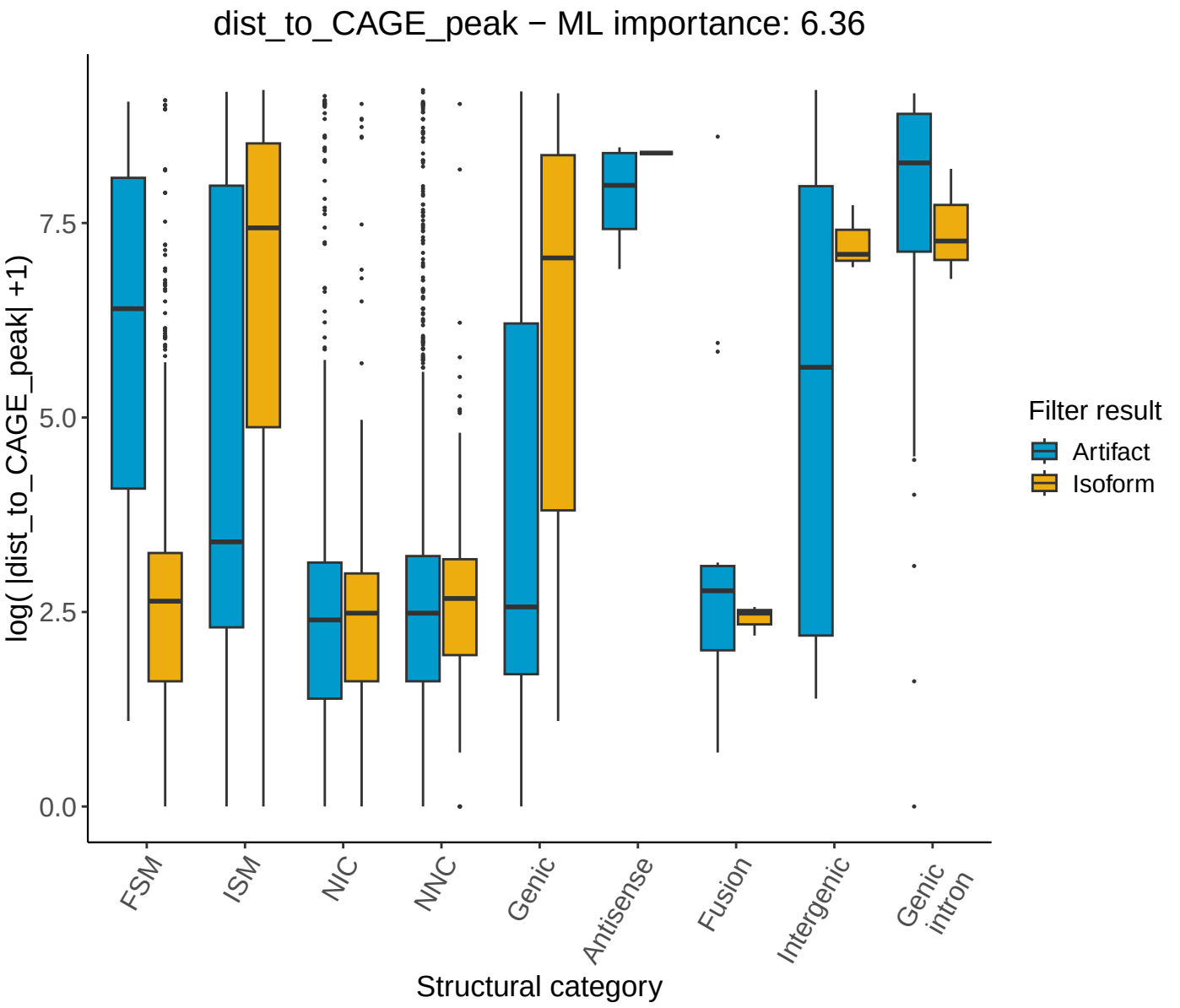

iso_exp − ML importance: 11.37
log(|iso_exp| +1)
12.5
10.0
7.5
5.0
2.5
0.0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

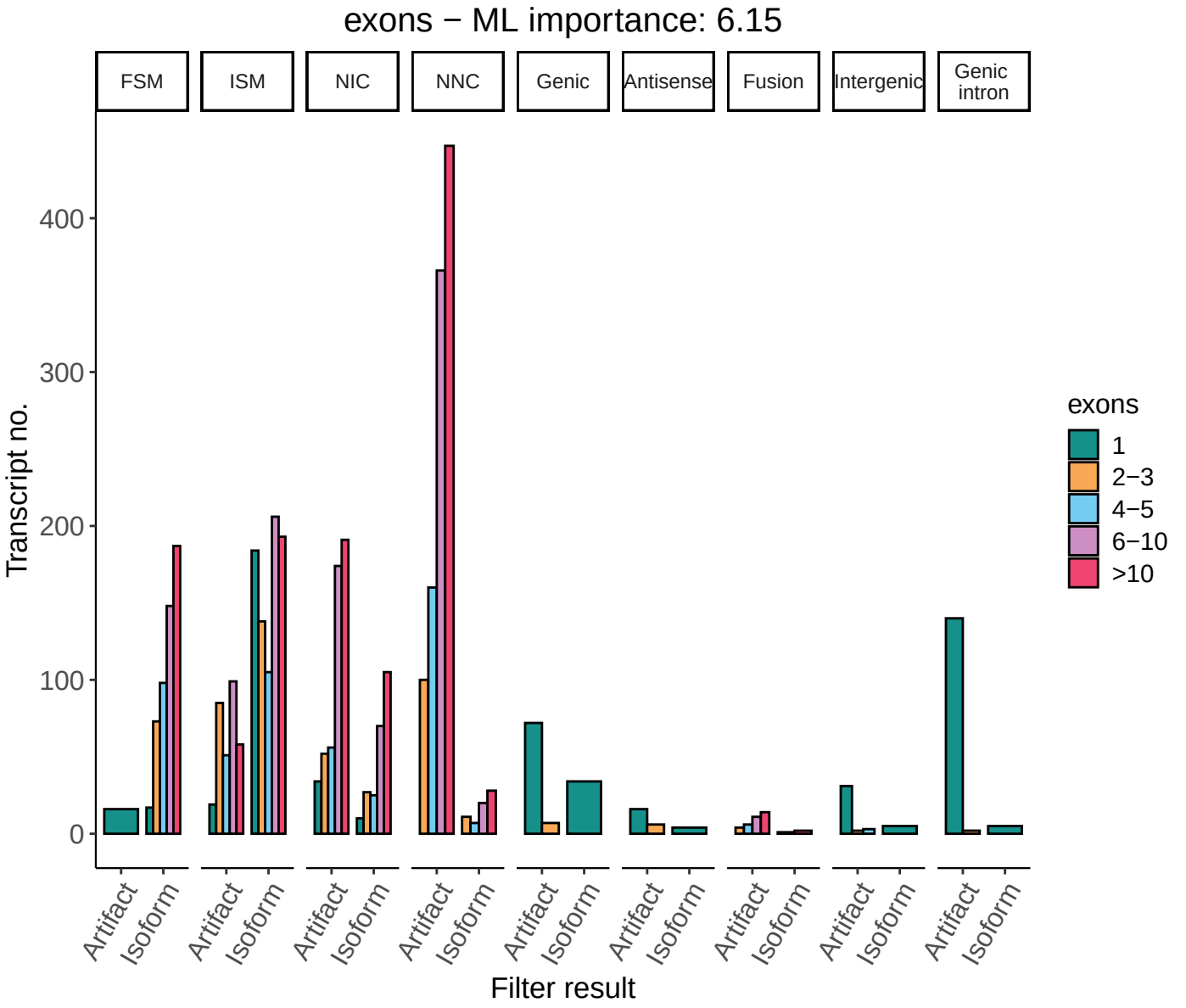

diff_to_gene_TSS − ML importance: 7.81
log(|diff_to_gene_TSS| +1)
10
5
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

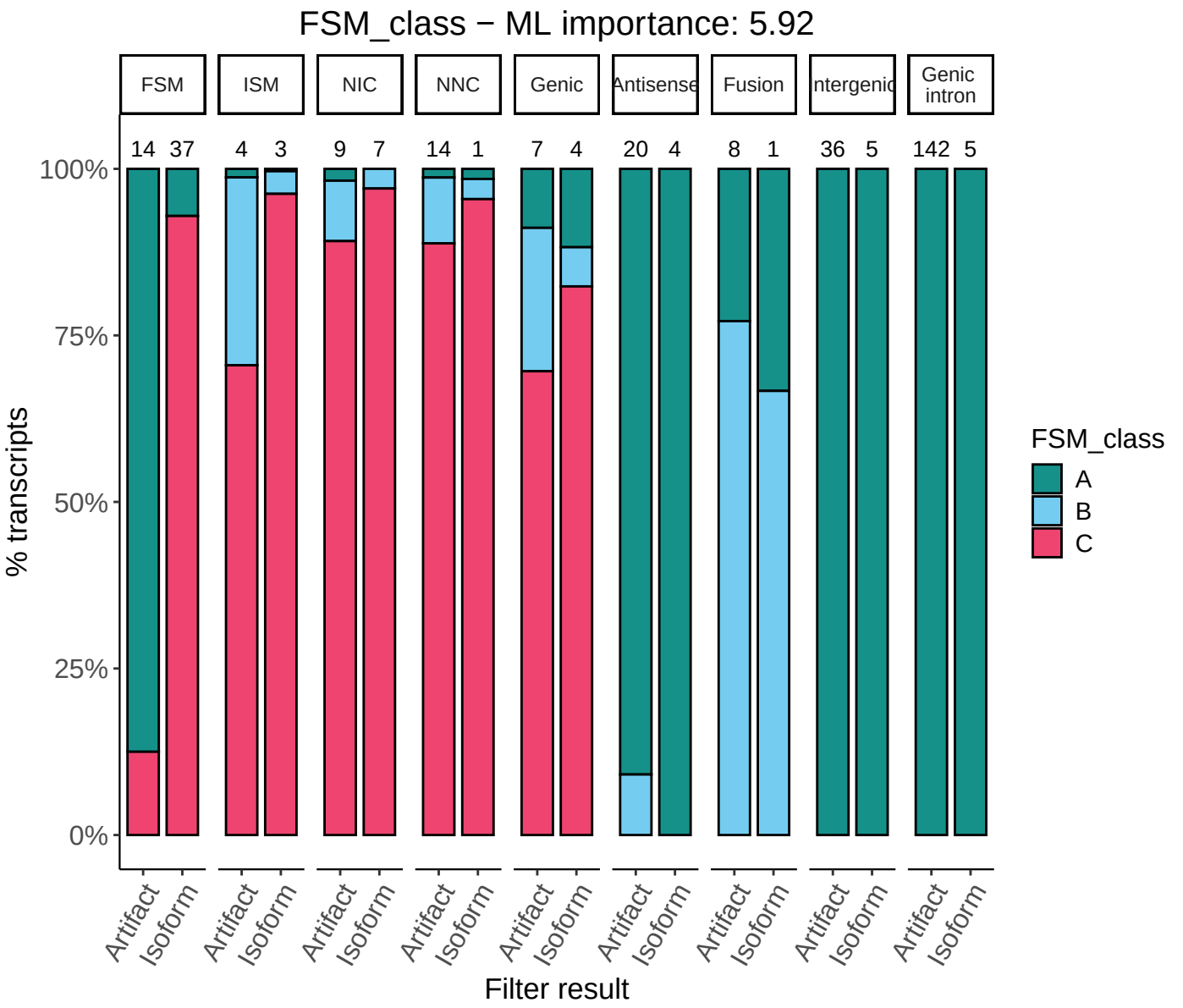

ratio_TSS − ML importance: 6.8
log(|ratio_TSS| +1)
10.0
7.5
5.0
2.5
0.0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

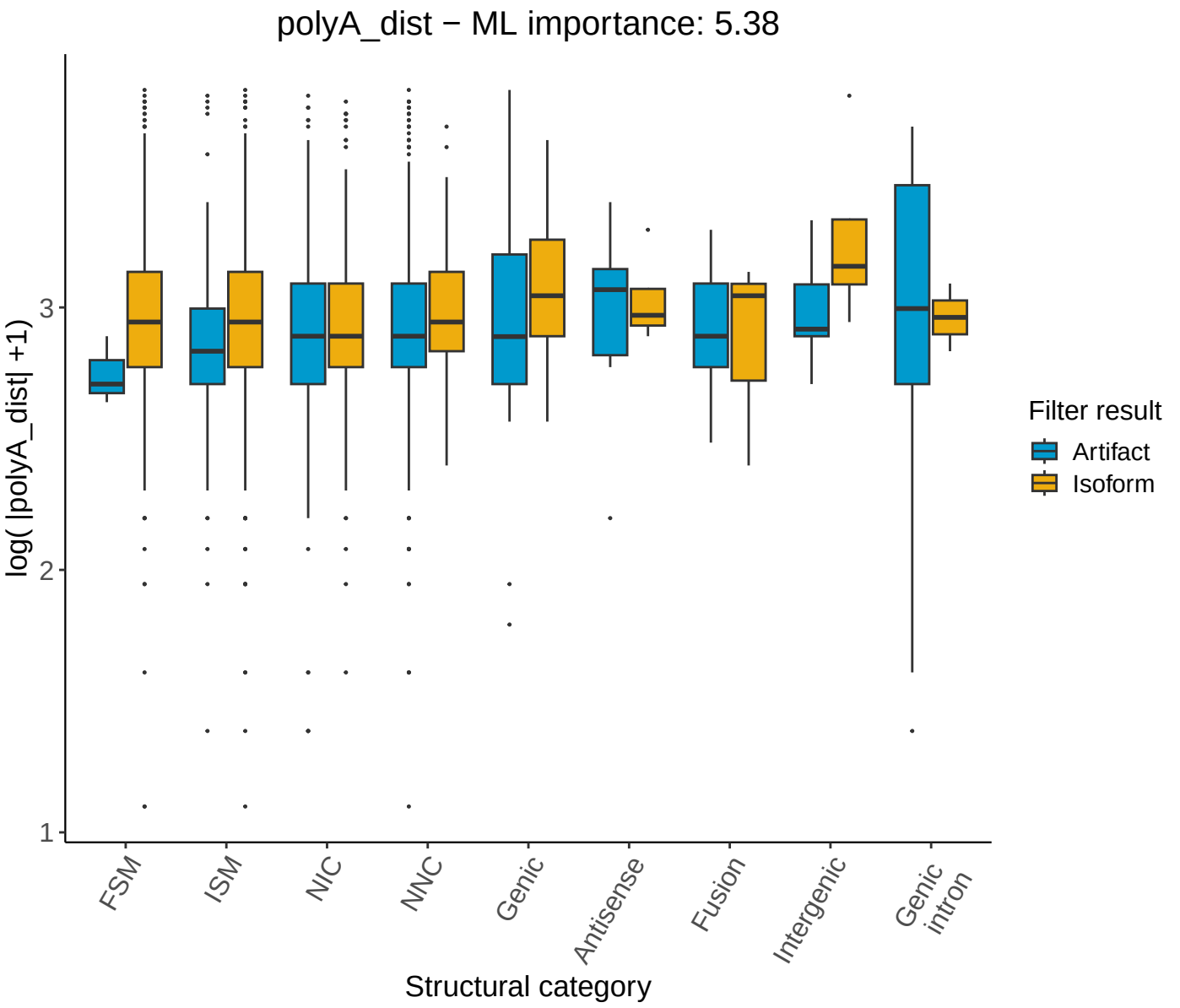

length − ML importance: 6.68
log(|length| +1)
9
8
7
6
5
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

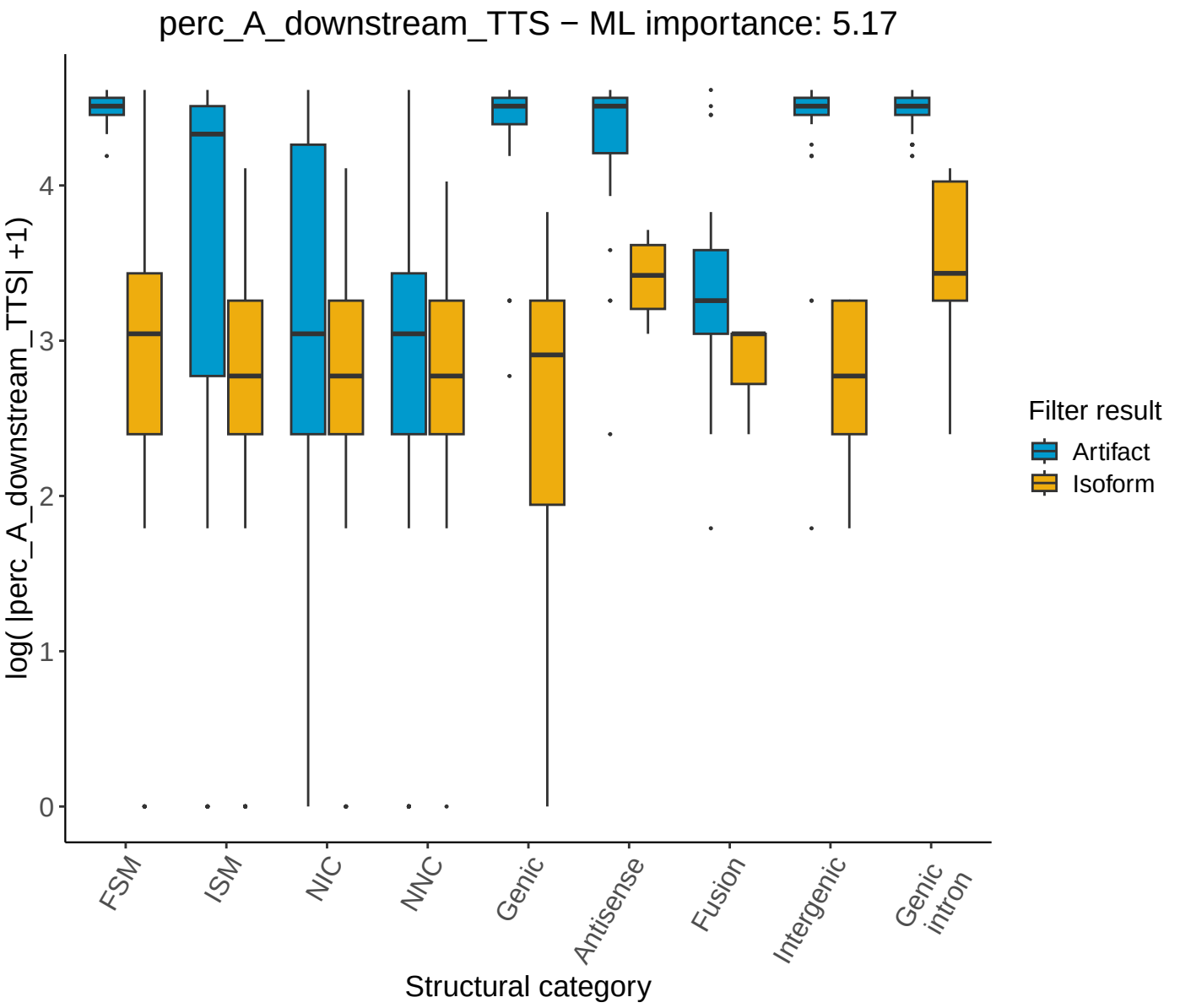

diff_to_gene_TTS − ML importance: 6.68
log(|diff_to_gene_TTS| +1)
10
5
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

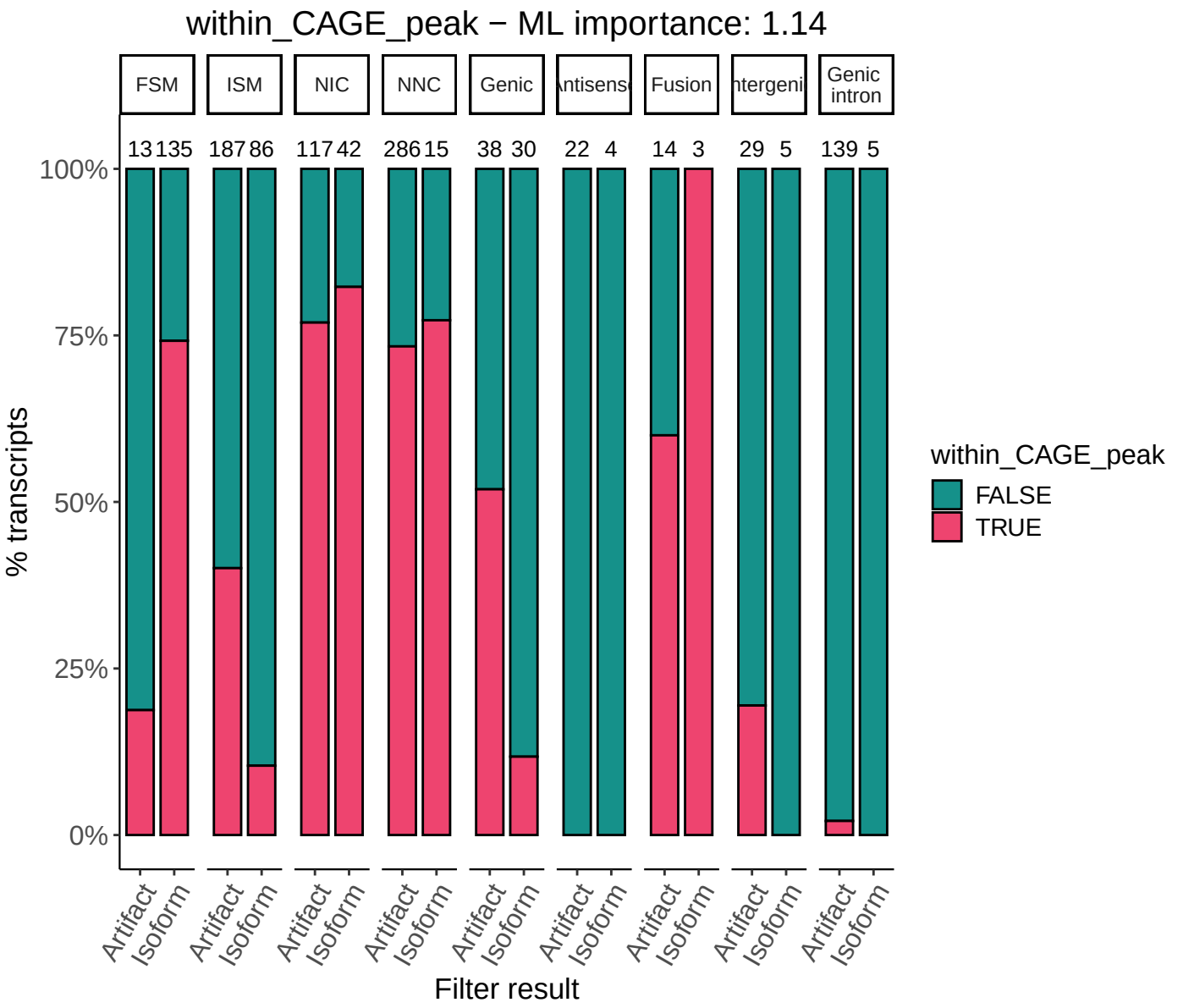

dist_to_CAGE_peak − ML importance: 6.36
log(|dist_to_CAGE_peak| +1)
7.5
5.0
2.5
0.0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

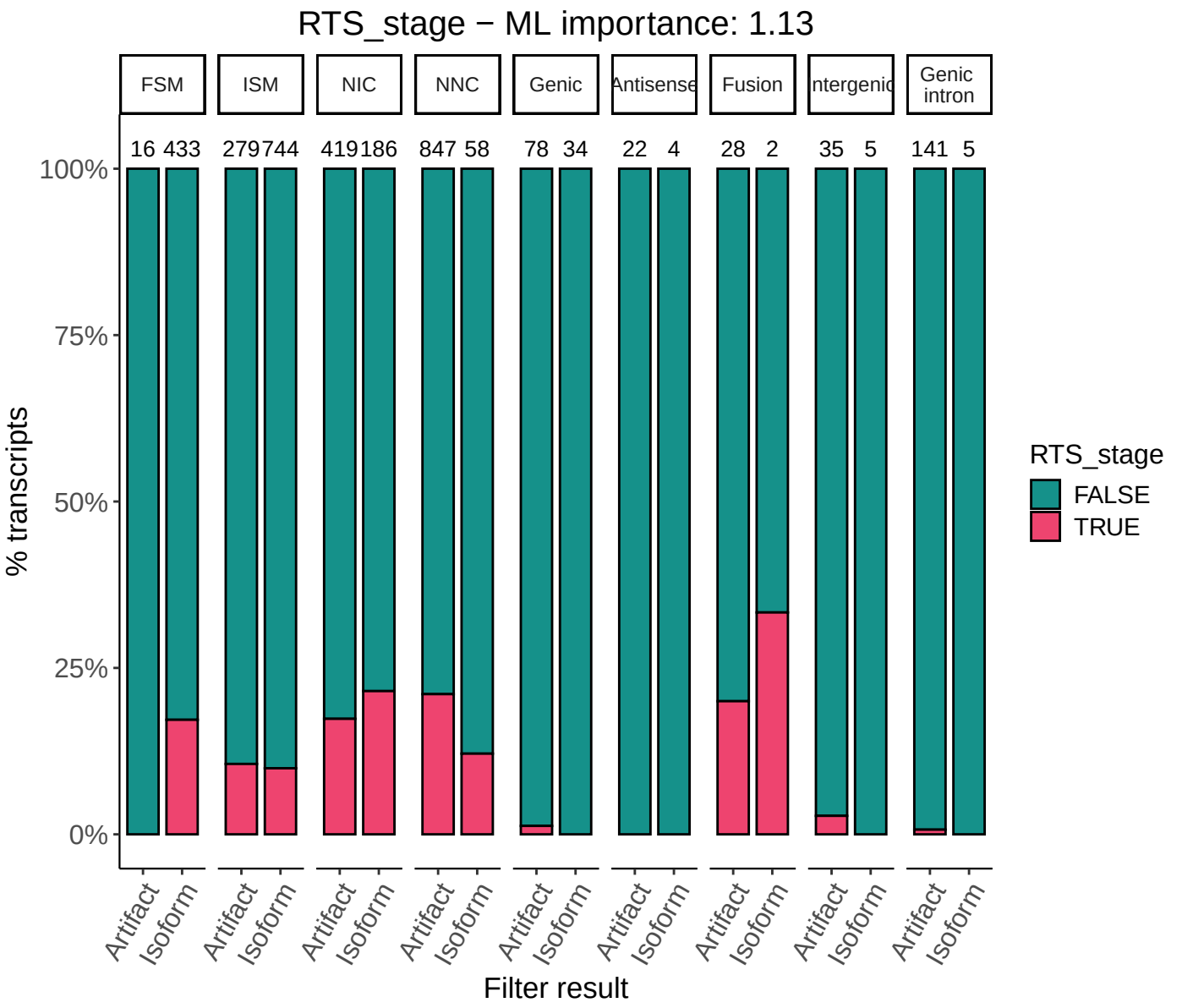

exons − ML importance: 6.15
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Transcript no.
0
100
200
300
400
Artifact
Isoform
Filter result
exons
1
2−3
4−5
6−10
>10

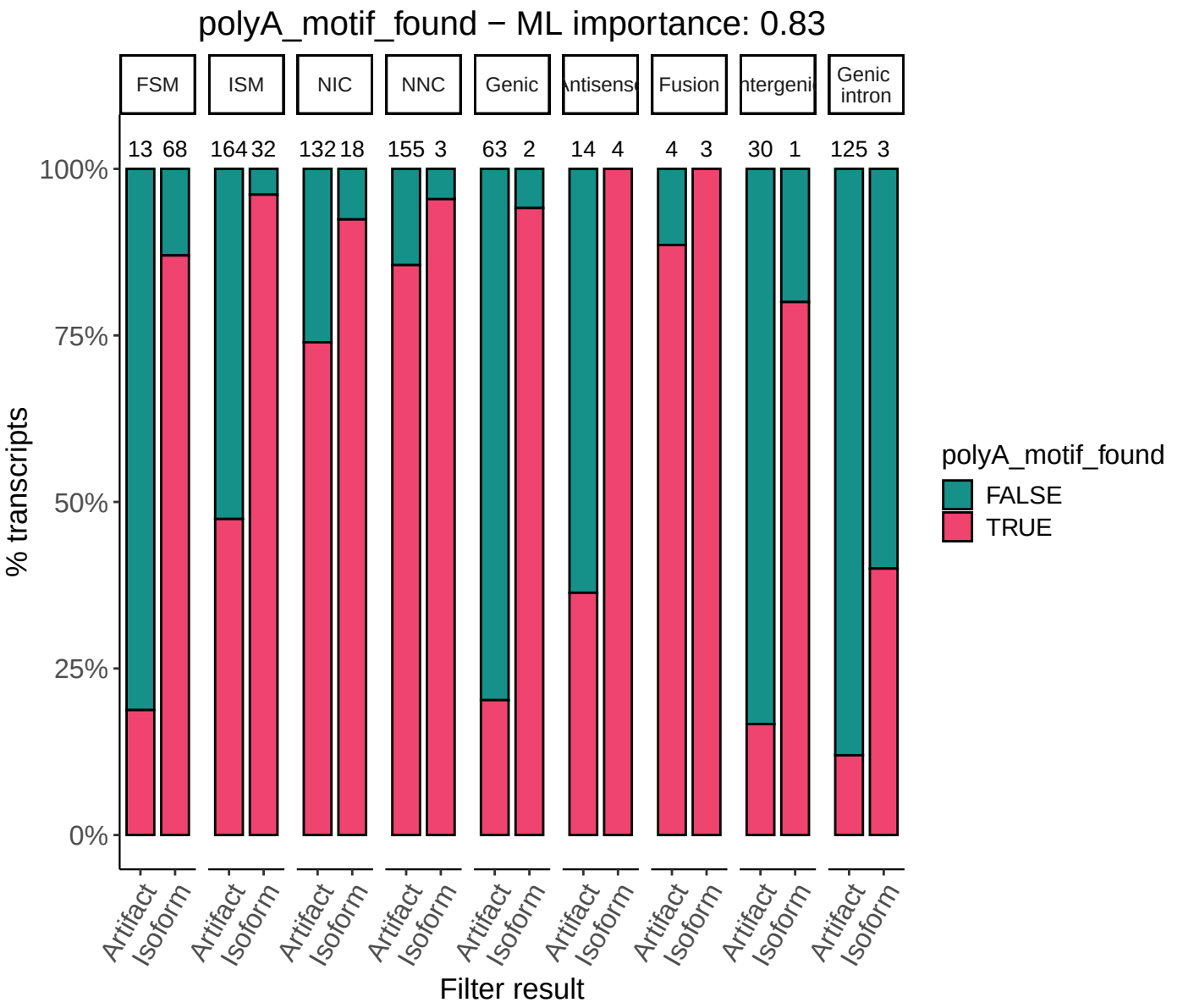

FSM_class − ML importance: 5.92
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
ntergenic
Genic intron
14 37
4 3
9 7
14 1
7 4
20 4
8 1
36 5
142 5
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
FSM_class
A
B
C

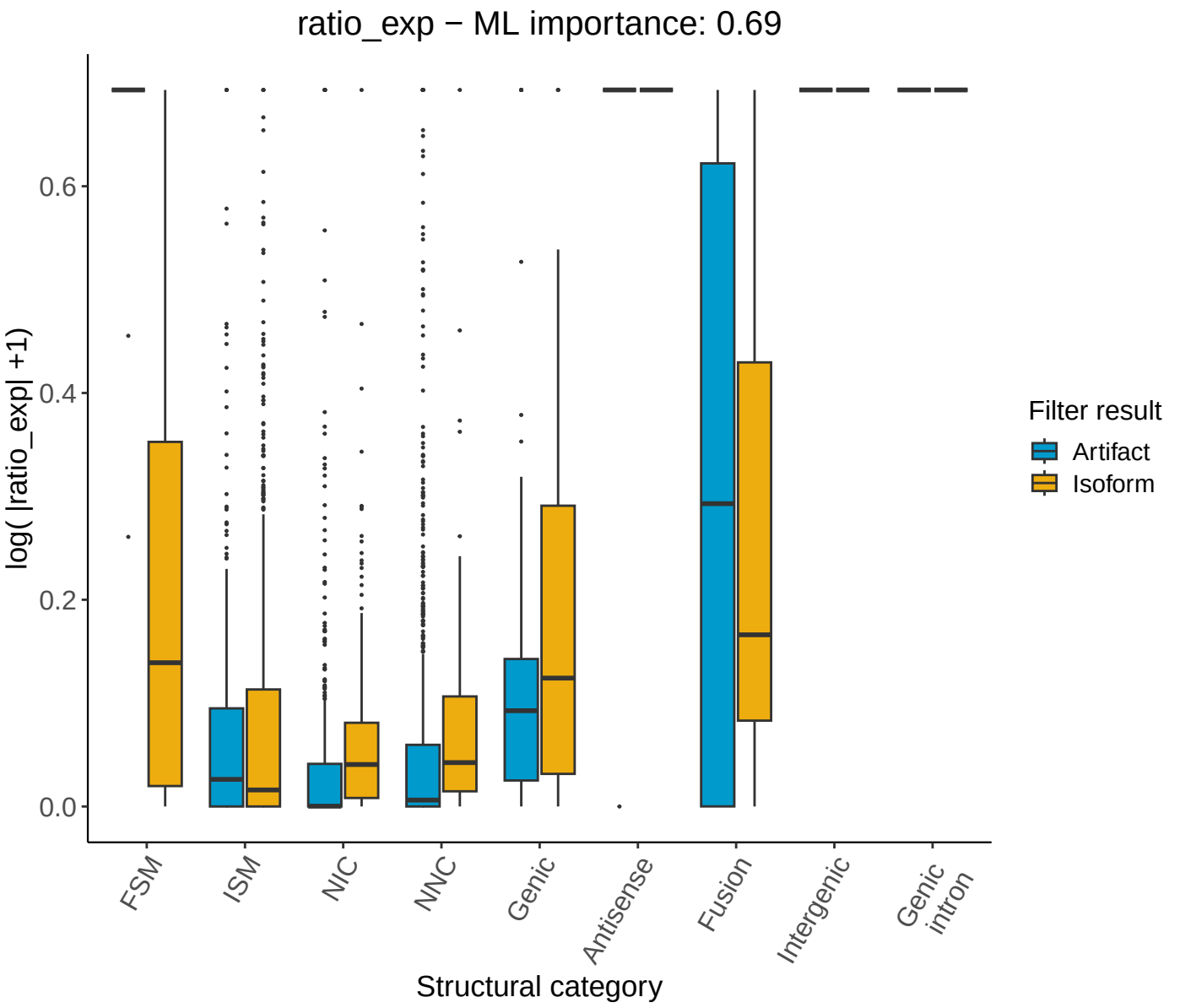

polyA_dist − ML importance: 5.38
log(|polyA_dist| +1)
3
2
1
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

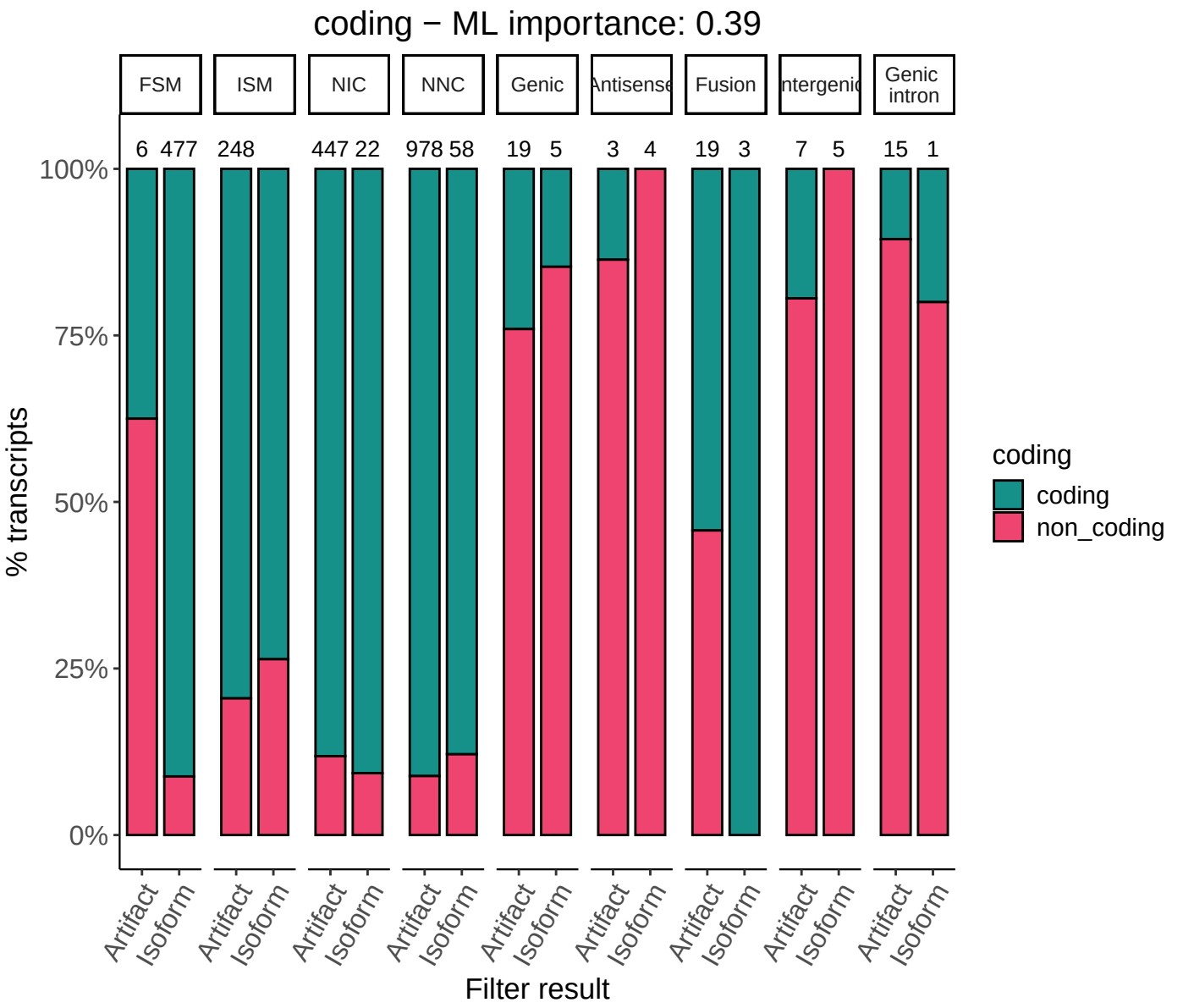

perc_A_downstream_TTS − ML importance: 5.17
log(|perc_A_downstream_TTS| +1)
0
1
2
3
4
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform


within_CAGE_peak − ML importance: 1.14
FSM
ISM
NIC
NNC
Genic
ntisens
Fusion
ntergeni
Genic intron
13 135
187 86
117 42
286 15
38 30
22 4
14 3
29 5
139 5
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
within_CAGE_peak
FALSE
TRUE

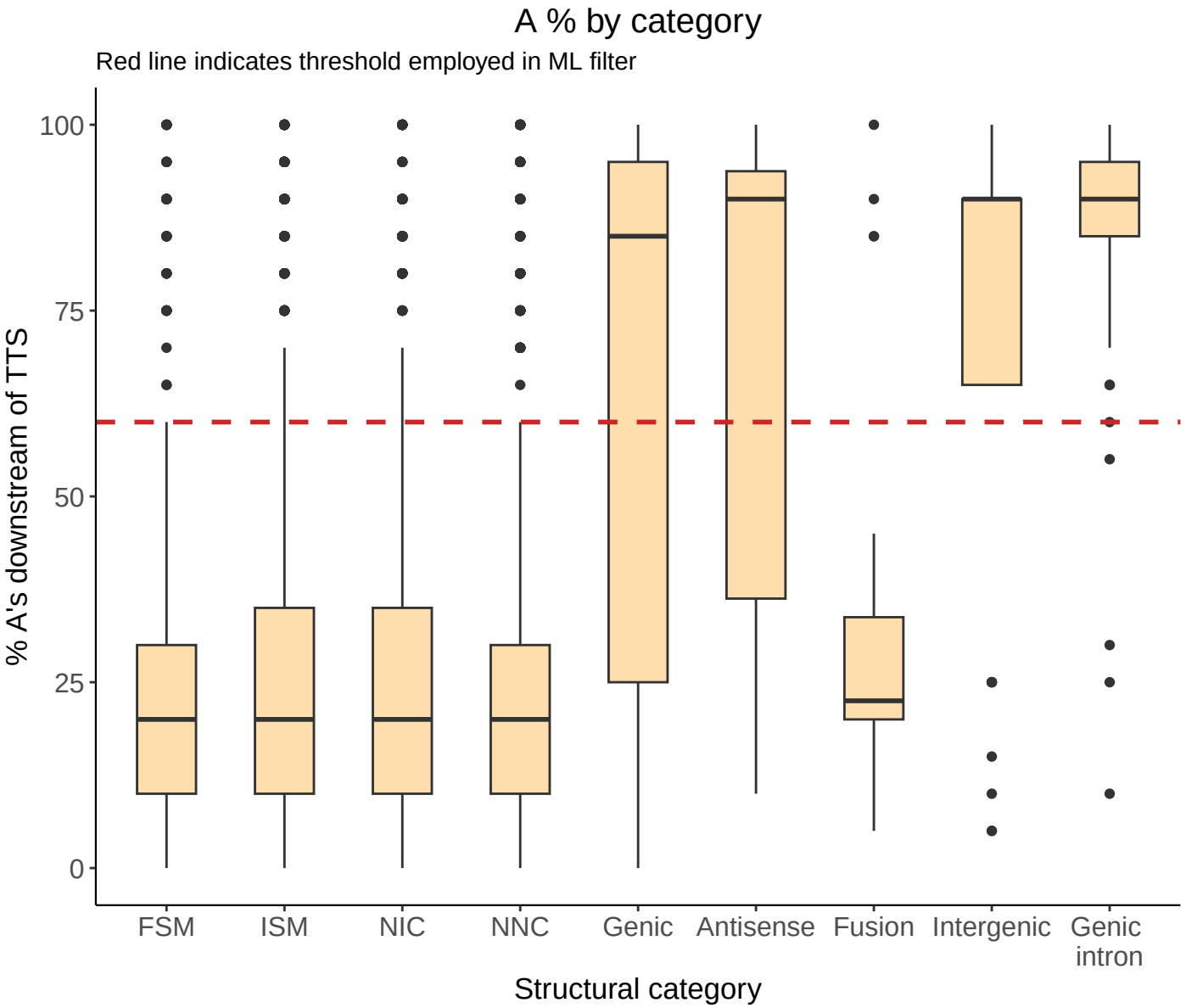

RTS_stage − ML importance: 1.13
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
ntergenic
Genic intron
16 433
279 744
419 186
847 58
78 34
22 4
28 2
35 5
141 5
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
RTS_stage
FALSE
TRUE

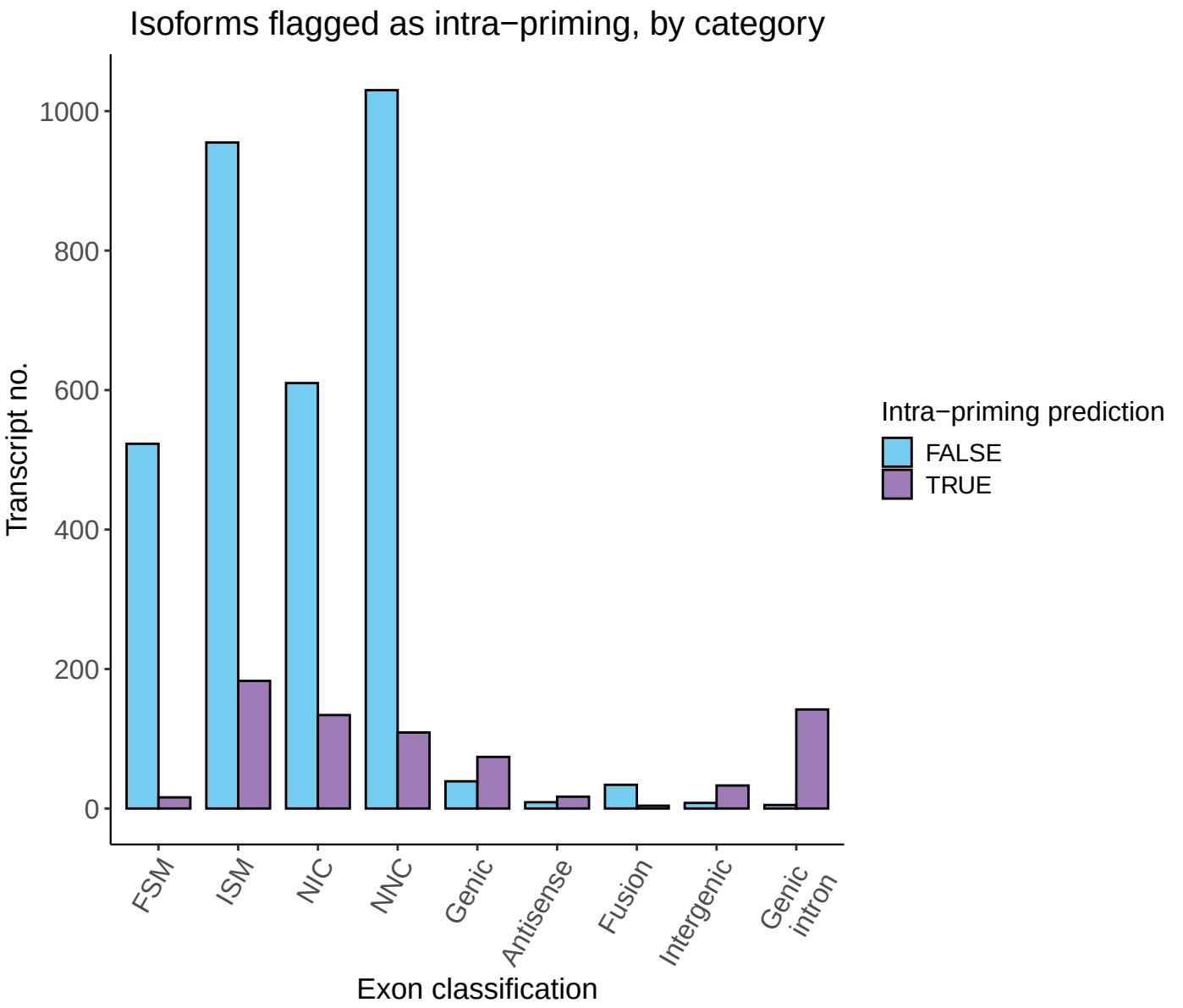

polyA_motif_found − ML importance: 0.83
FSM
ISM
NIC
NNC
Genic
ntisens
Fusion
ntergeni
Genic intron
13 68
164 32
132 18
155 3
63 2
14 4
4 3
30 1
125 3
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
polyA_motif_found
FALSE
TRUE

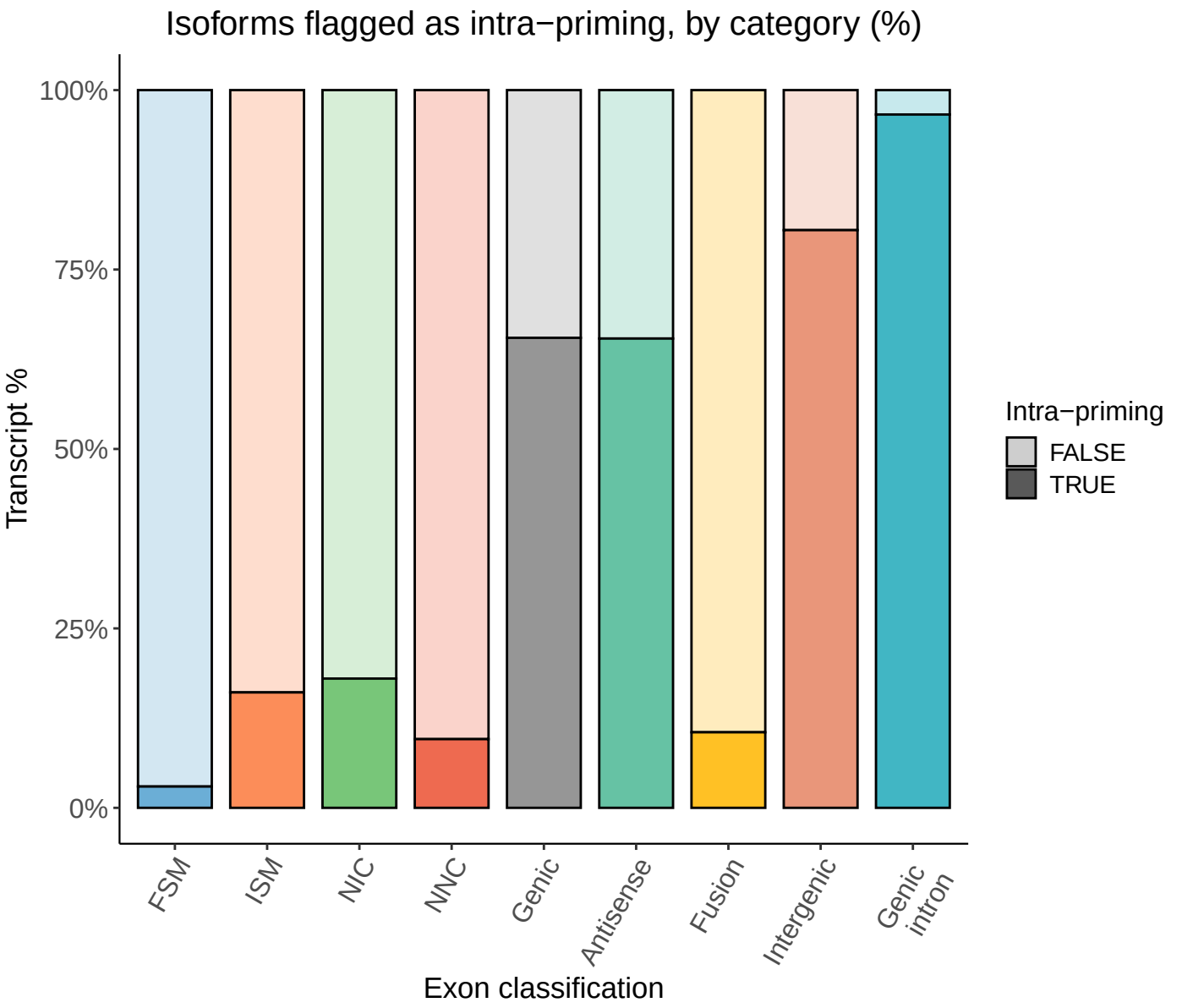

ratio_exp − ML importance: 0.69
log(|ratio_exp| +1)
0.0
0.2
0.4
0.6
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

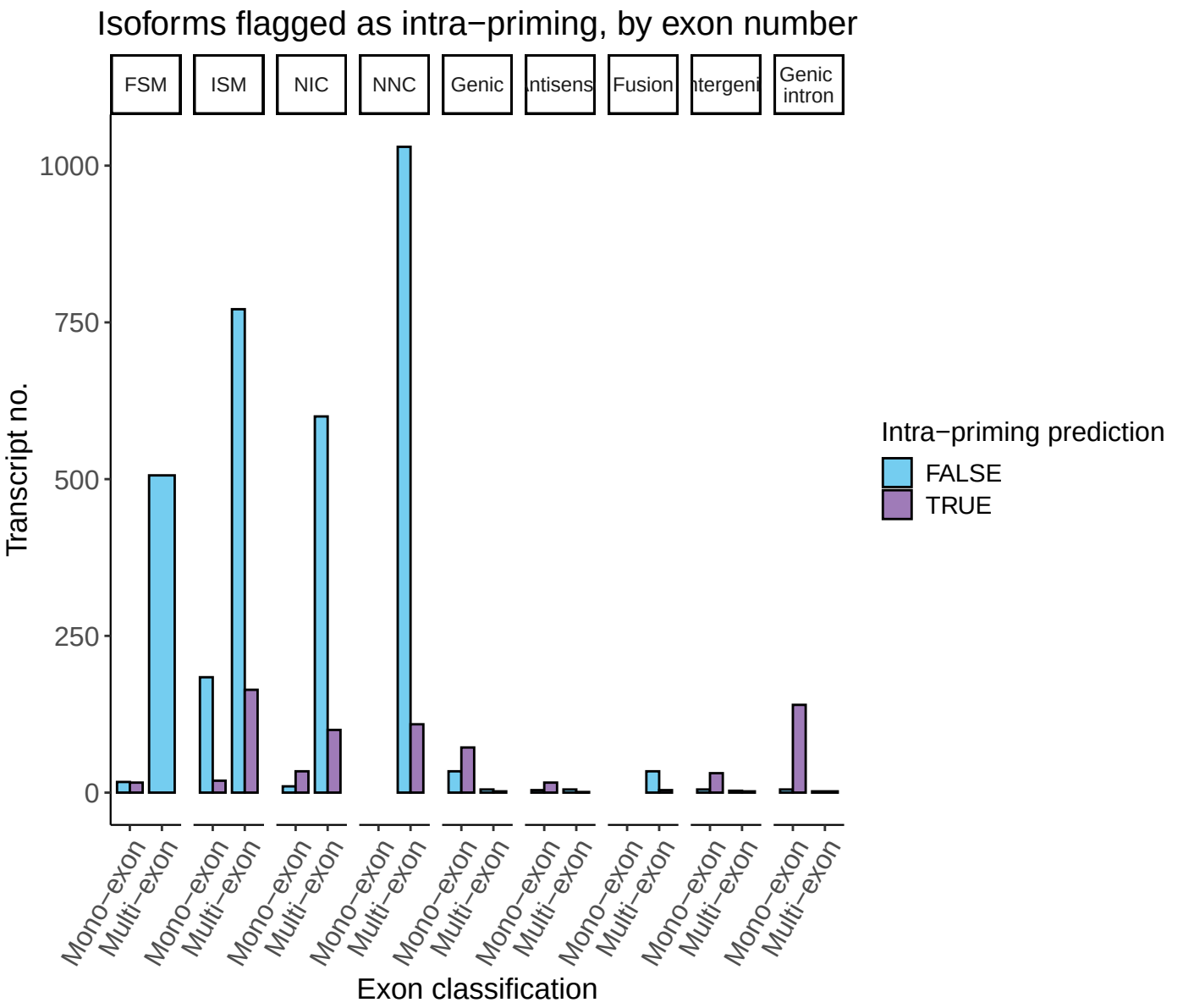

coding − ML importance: 0.39
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
6 477
248
447 22
978 58
19 5
3 4
19 3
7 5
15 1
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
coding
coding
non_coding

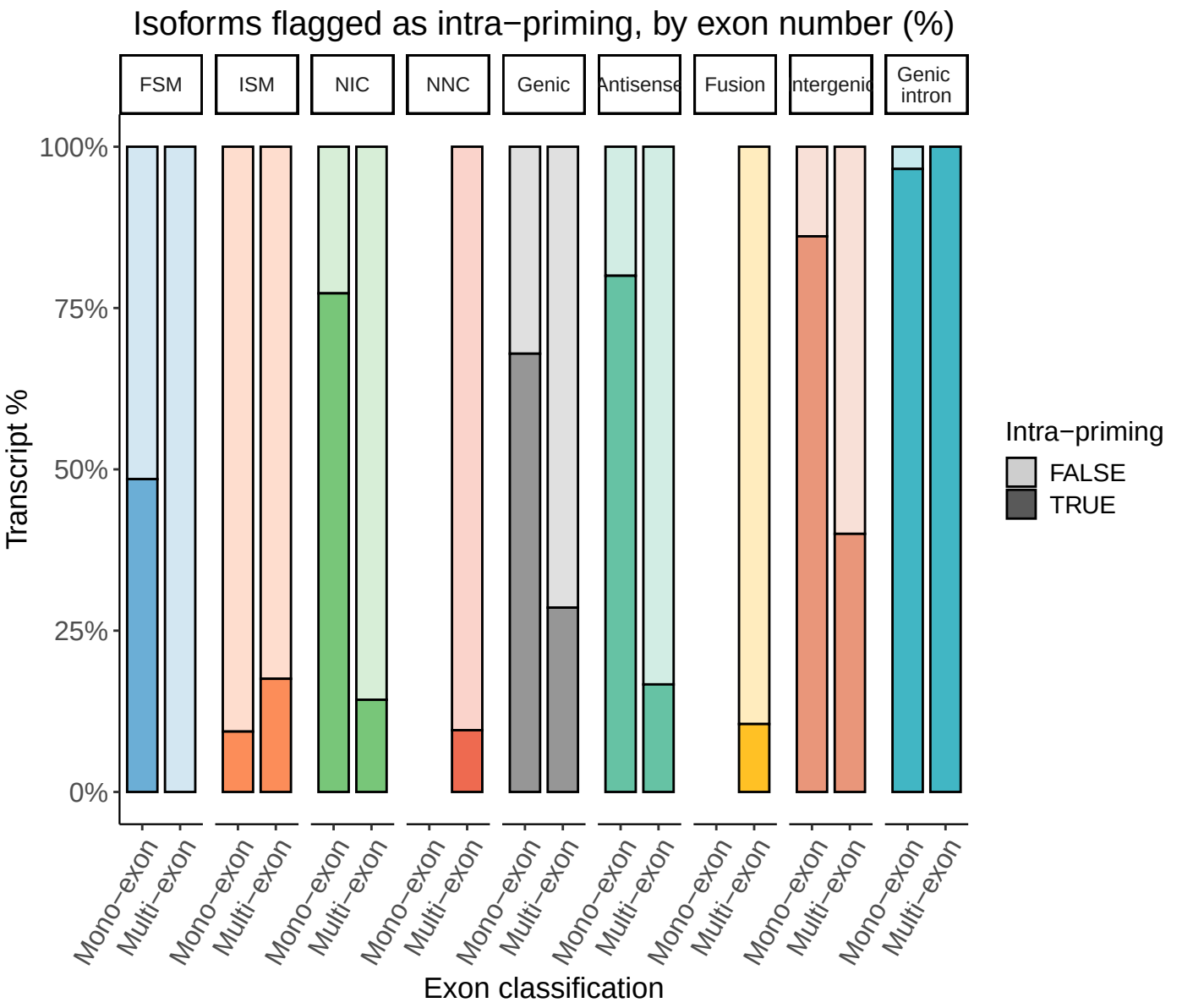
Intra−primming filter report


A % by category
Red line indicates threshold employed in ML filter
% A's downstream of TTS
100
75
50
25
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category


Isoforms flagged as intra−priming, by category
Transcript no.
1000
800
600
400
200
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Exon classification
Intra−priming prediction
FALSE
TRUE


Isoforms flagged as intra−priming, by category (%)
100%
75%
50%
25%
0%
Transcript %
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Exon classification
Intra−priming
FALSE
TRUE


Isoforms flagged as intra−priming, by exon number
FSM
ISM
NIC
NNC
Genic
ntisens
Fusion
ntergeni
Genic intron
1000
750
500
250
0
Transcript no.
Mono−exon
Multi−exon
Exon classification
Intra−priming prediction
FALSE
TRUE


Isoforms flagged as intra−priming, by exon number (%)
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
ntergenic
Genic intron
100%
75%
50%
25%
0%
Transcript %
Mono−exon
Multi−exon
Exon classification
Intra−priming
FALSE
TRUE