SQANTI3 filter report

Total Genes: 656

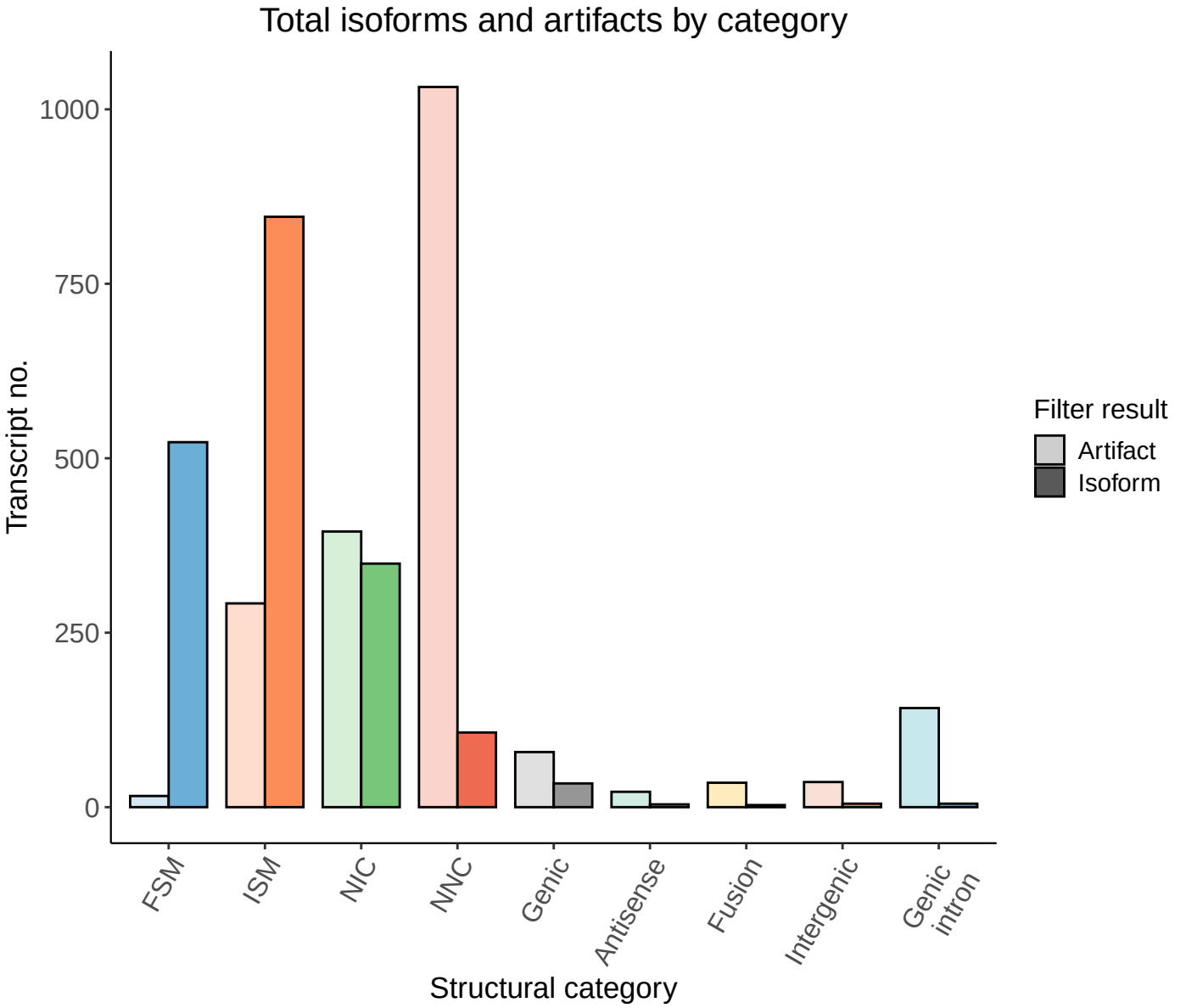
Total Transcripts: 3925

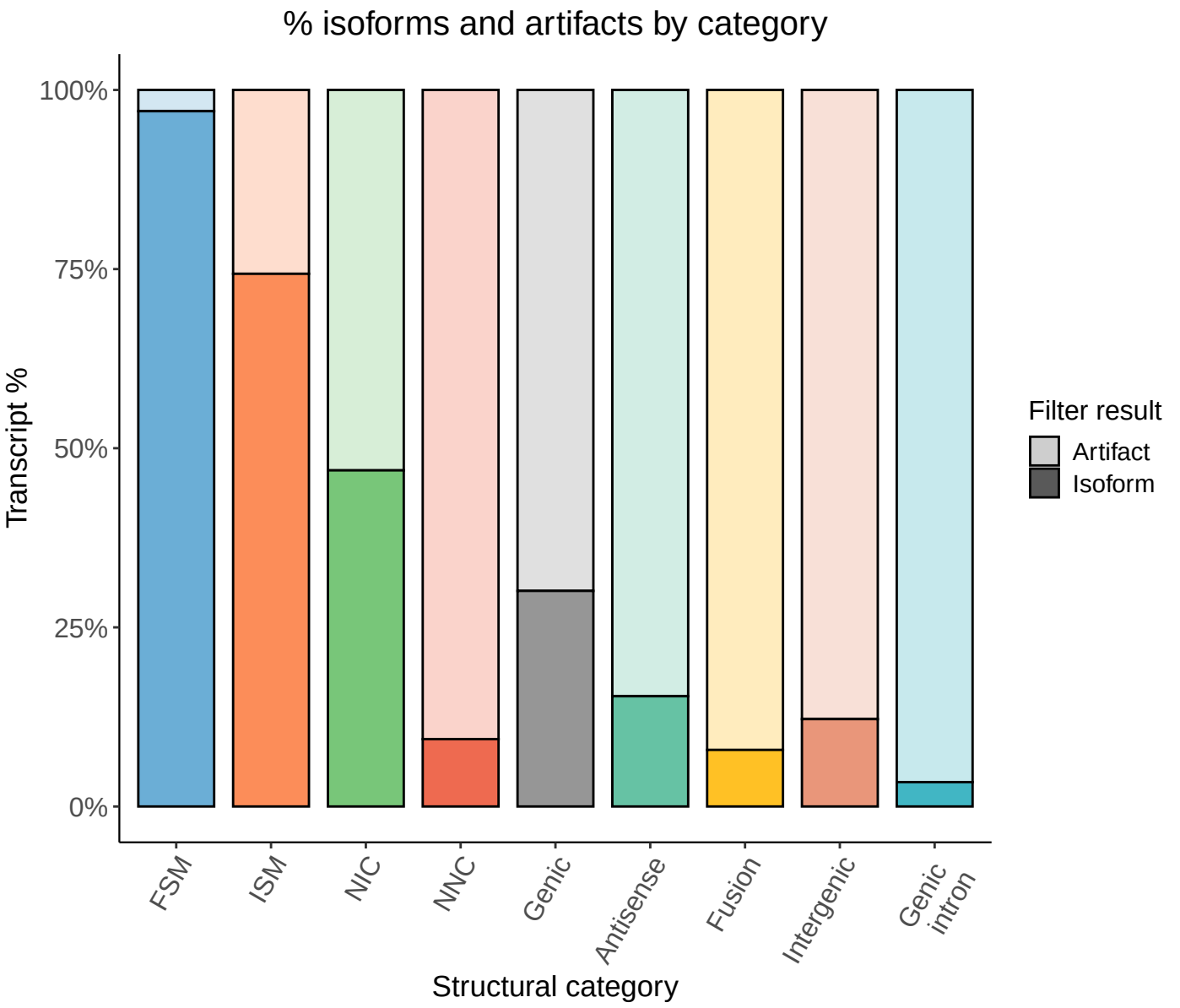
− Isoforms: 1876 (48%)

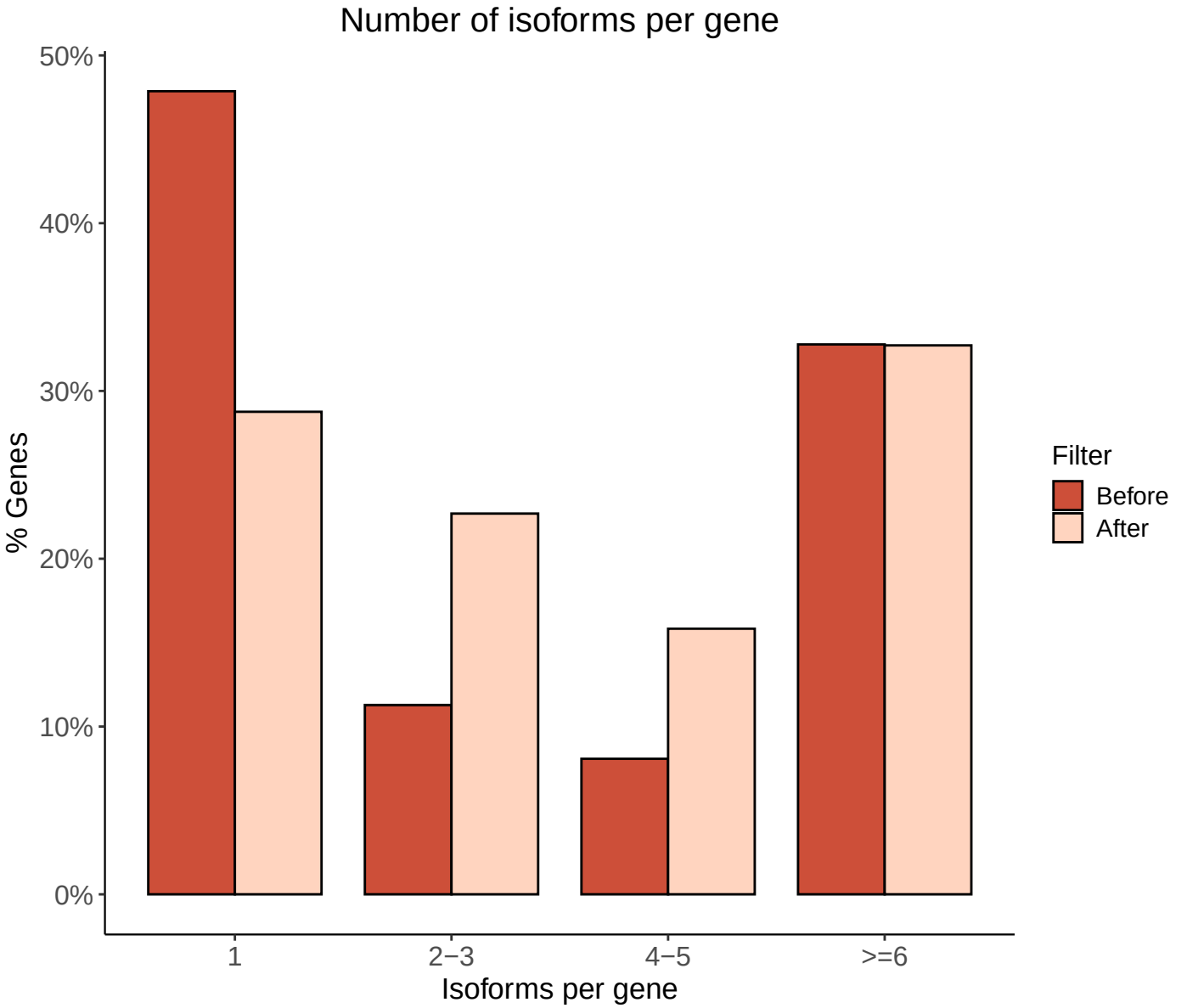
− Artifacts: 2049 (52%)

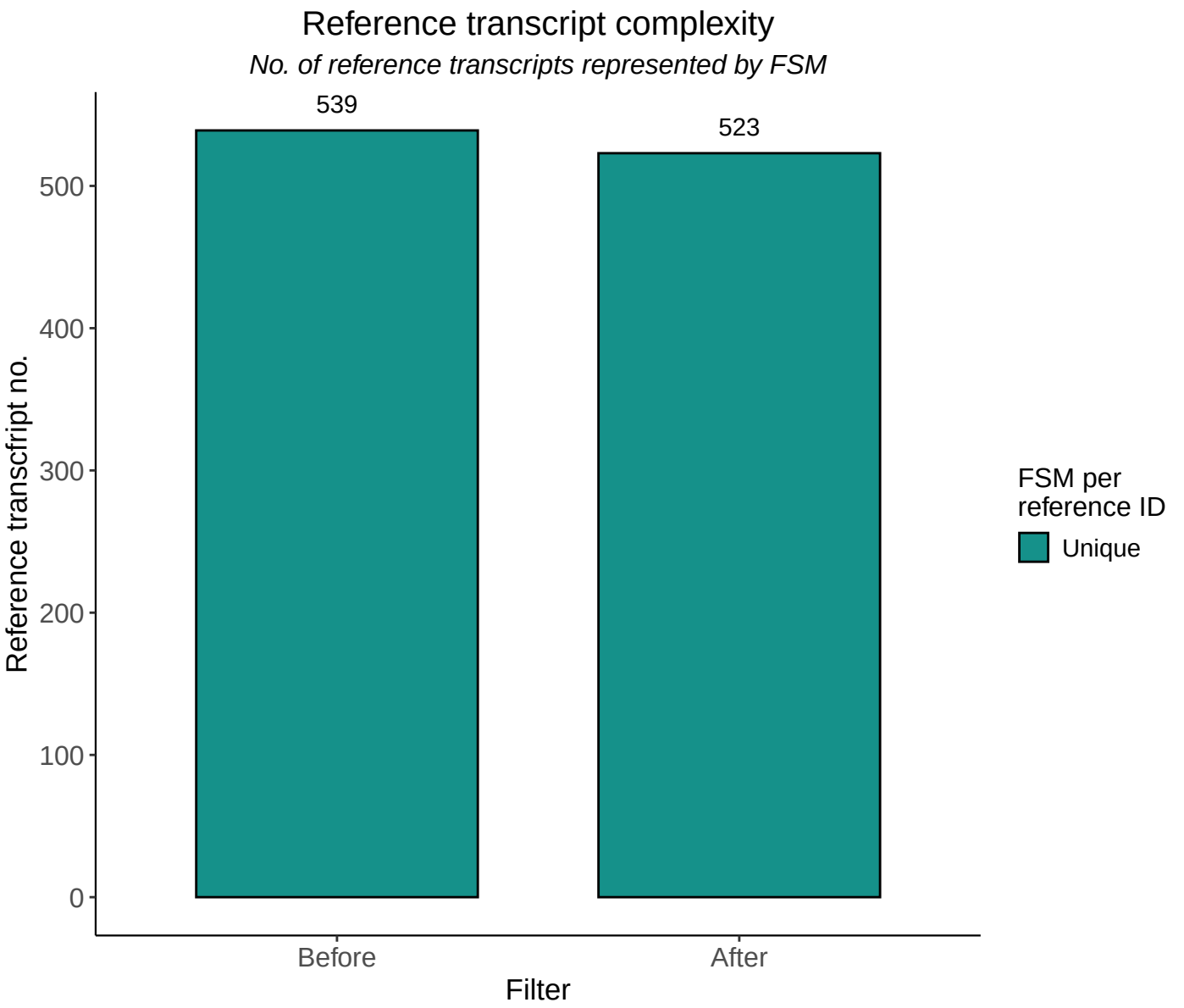
Gene category	Gene no.	No. of genes with artifacts only
Annotated	443	78
Novel	213	199

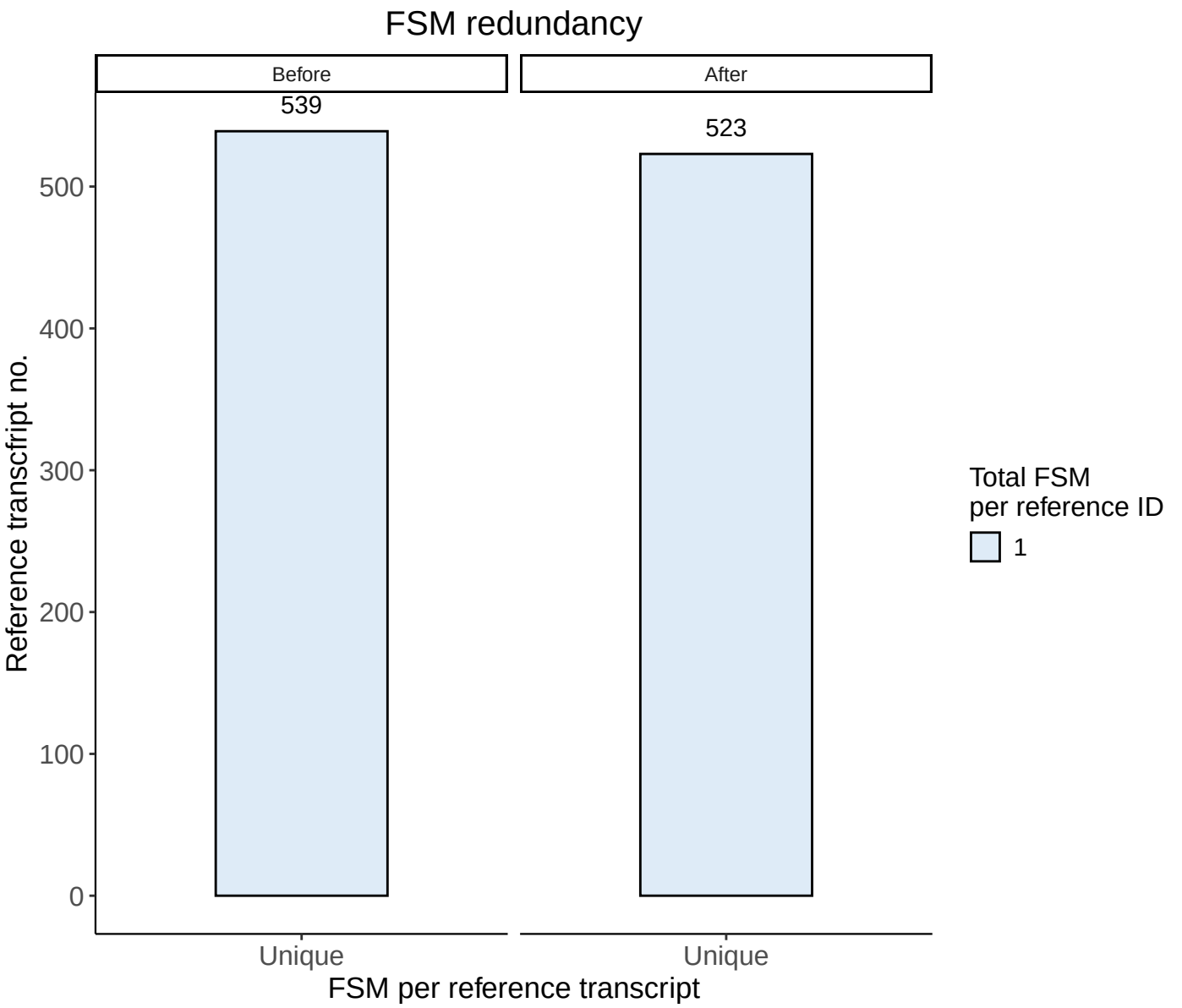
Structural category	Artifact no.	Isoform no.
FSM	16	523
ISM	292	846
NIC	395	349
NNC	1032	107
Genic	79	34
Antisense	22	4
Fusion	35	3
Intergenic	36	5
Genic intron	142	5

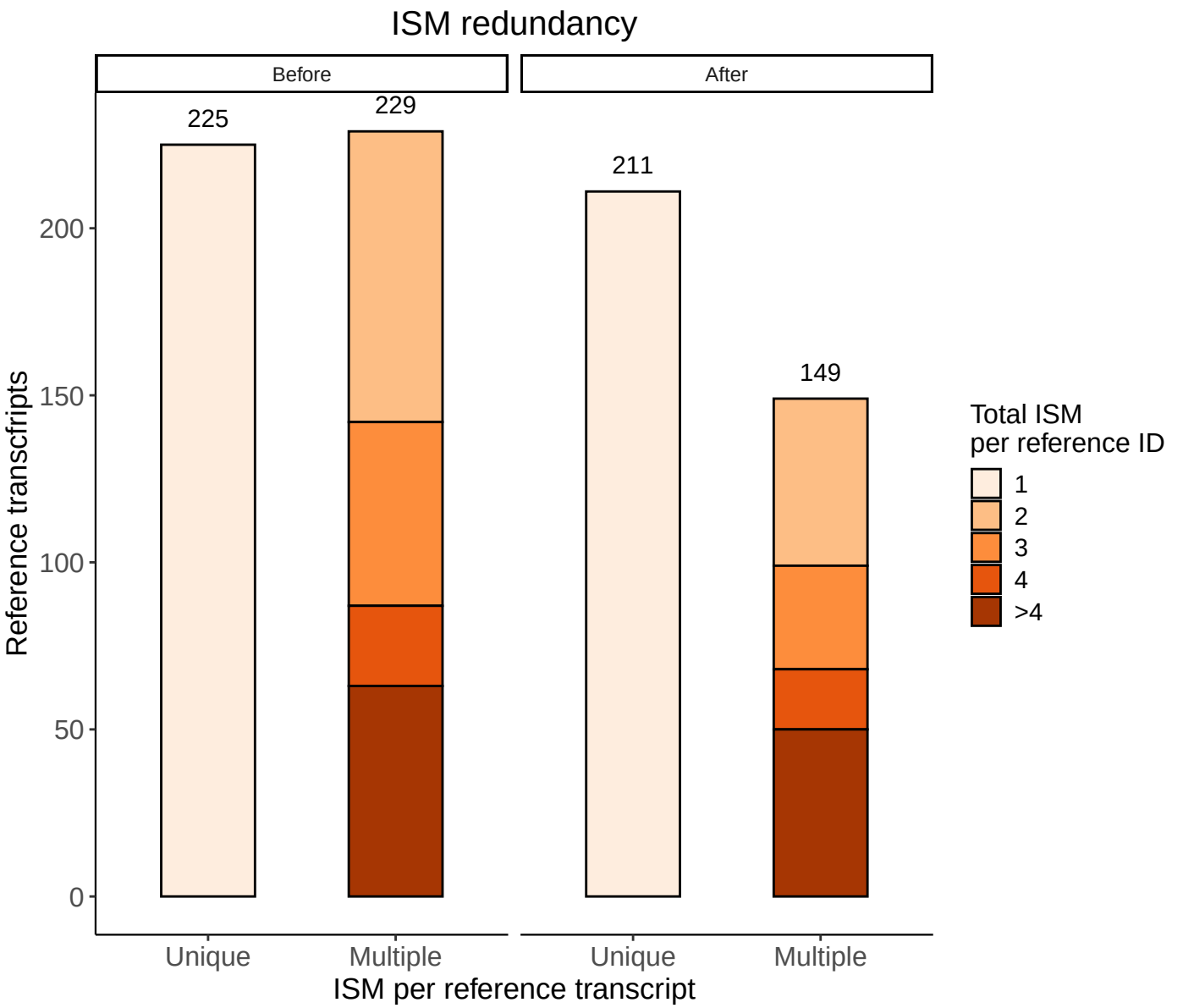

Total isoforms and artifacts by category
Transcript no.
1000
750
500
250
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

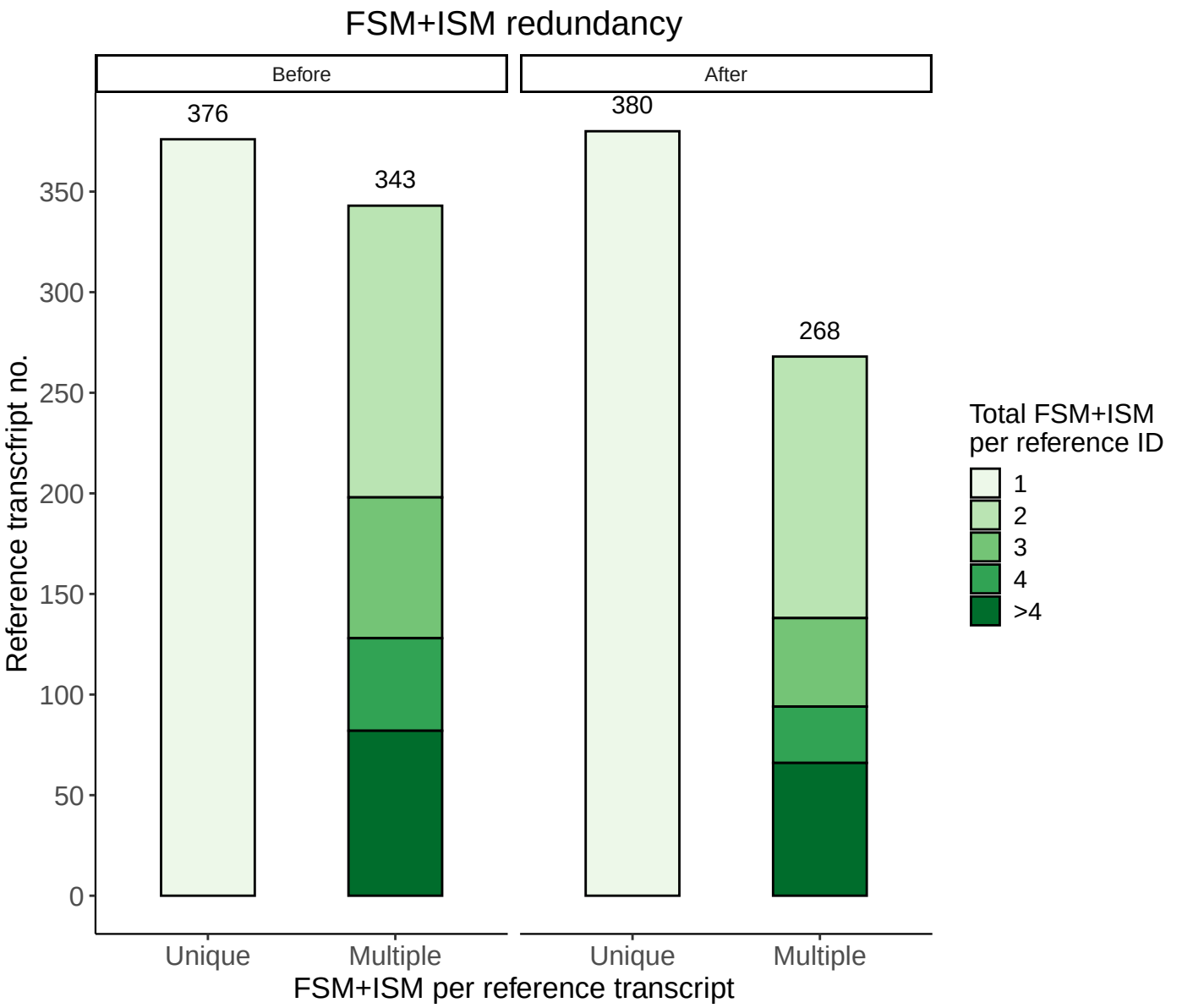

% isoforms and artifacts by category
100%
75%
50%
25%
0%
Transcript %
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform


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
211
149
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
380
268
350
300
250
200
150
100
50
0
Reference transcfript no.
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

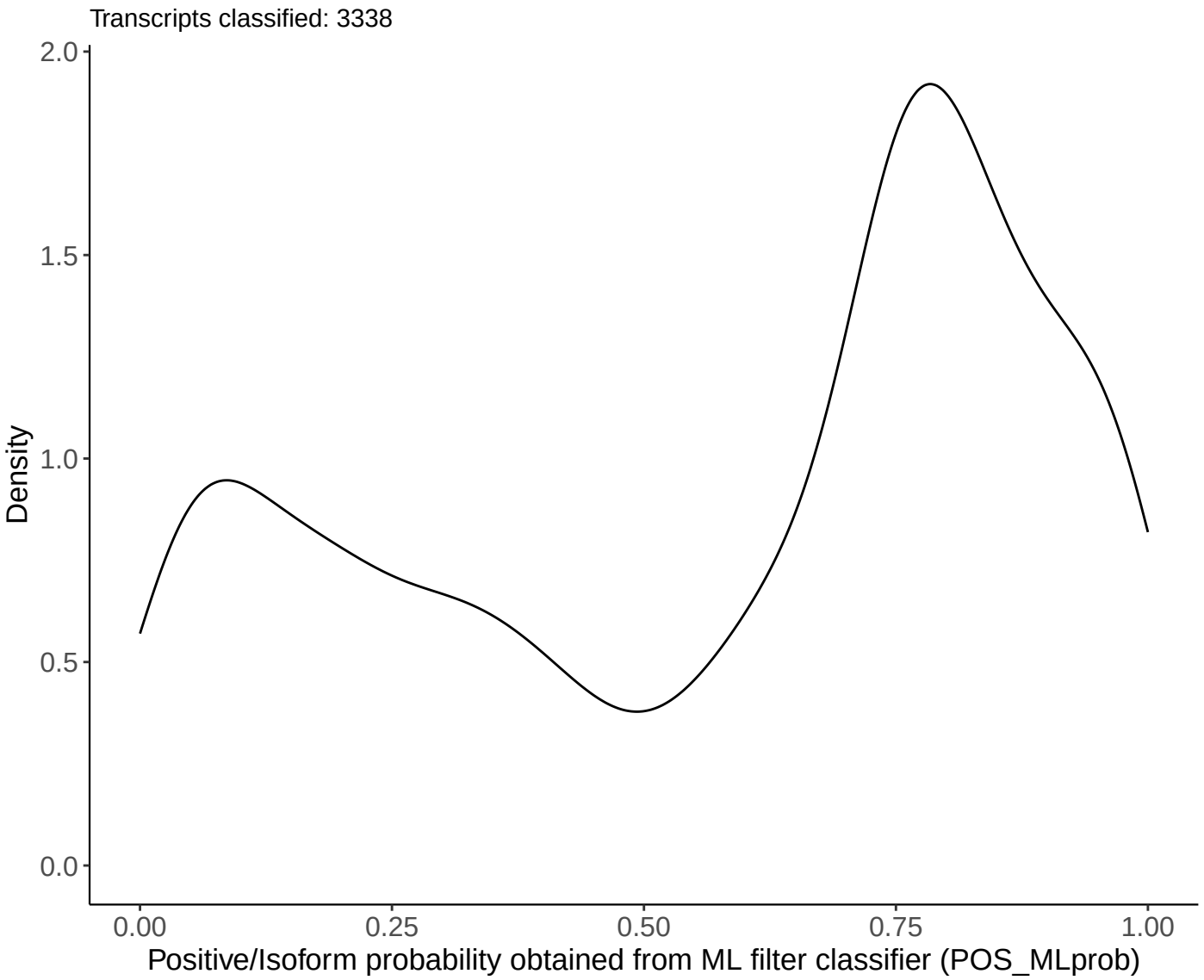
Performance metrics

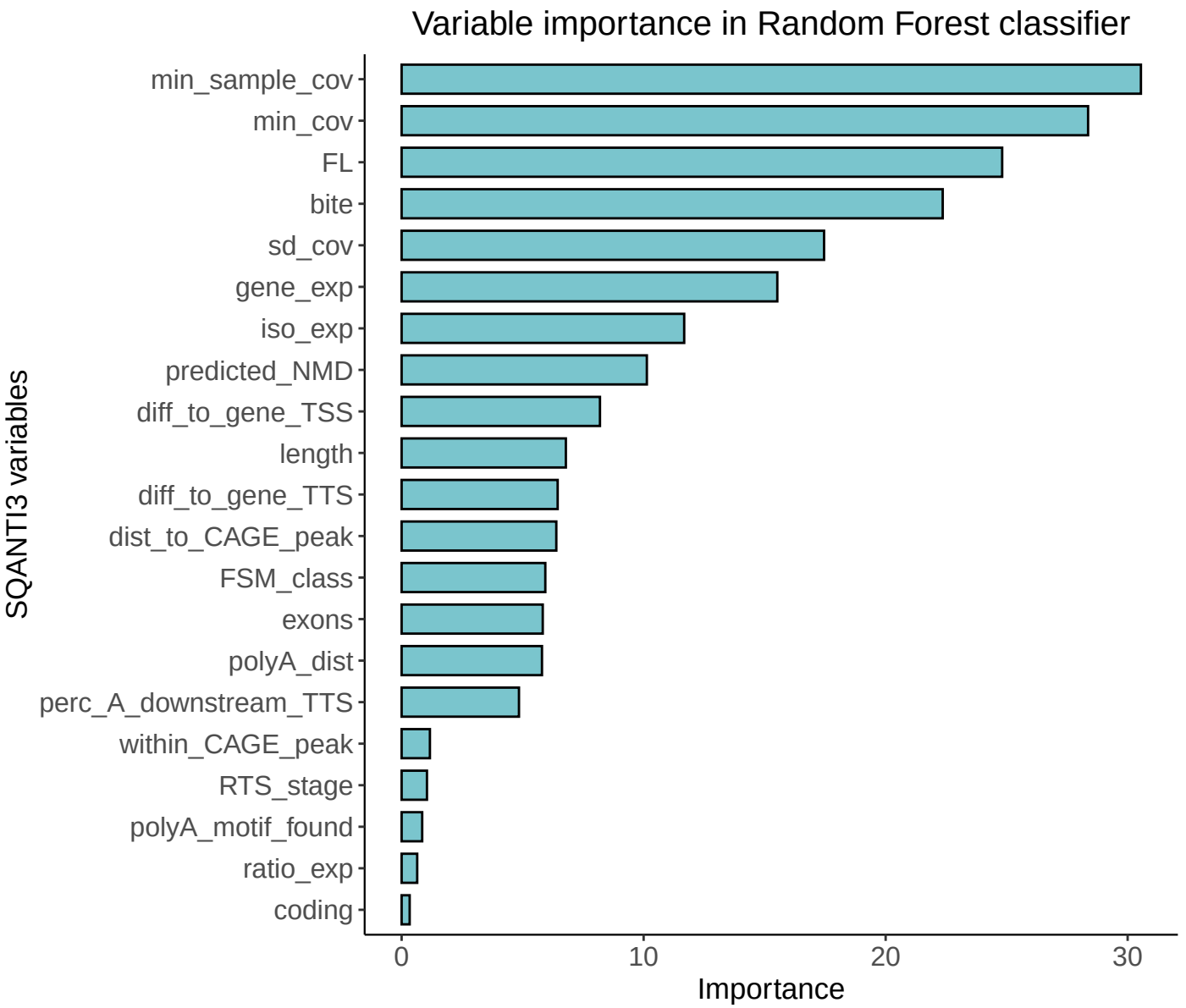
Metric	Value
Accuracy	0.895
Kappa	0.789
AccuracyLower	0.823
AccuracyUpper	0.944
AccuracyNull	0.5
McnemarPValue	0.00937
Sensitivity	0.807
Specificity	0.982
Pos Pred Value	0.979
Neg Pred Value	0.836
Precision	0.979
Recall	0.807
F1	0.885
Prevalence	0.5
Detection Rate	0.404
Detection Prevalence	0.412
Balanced Accuracy	0.895

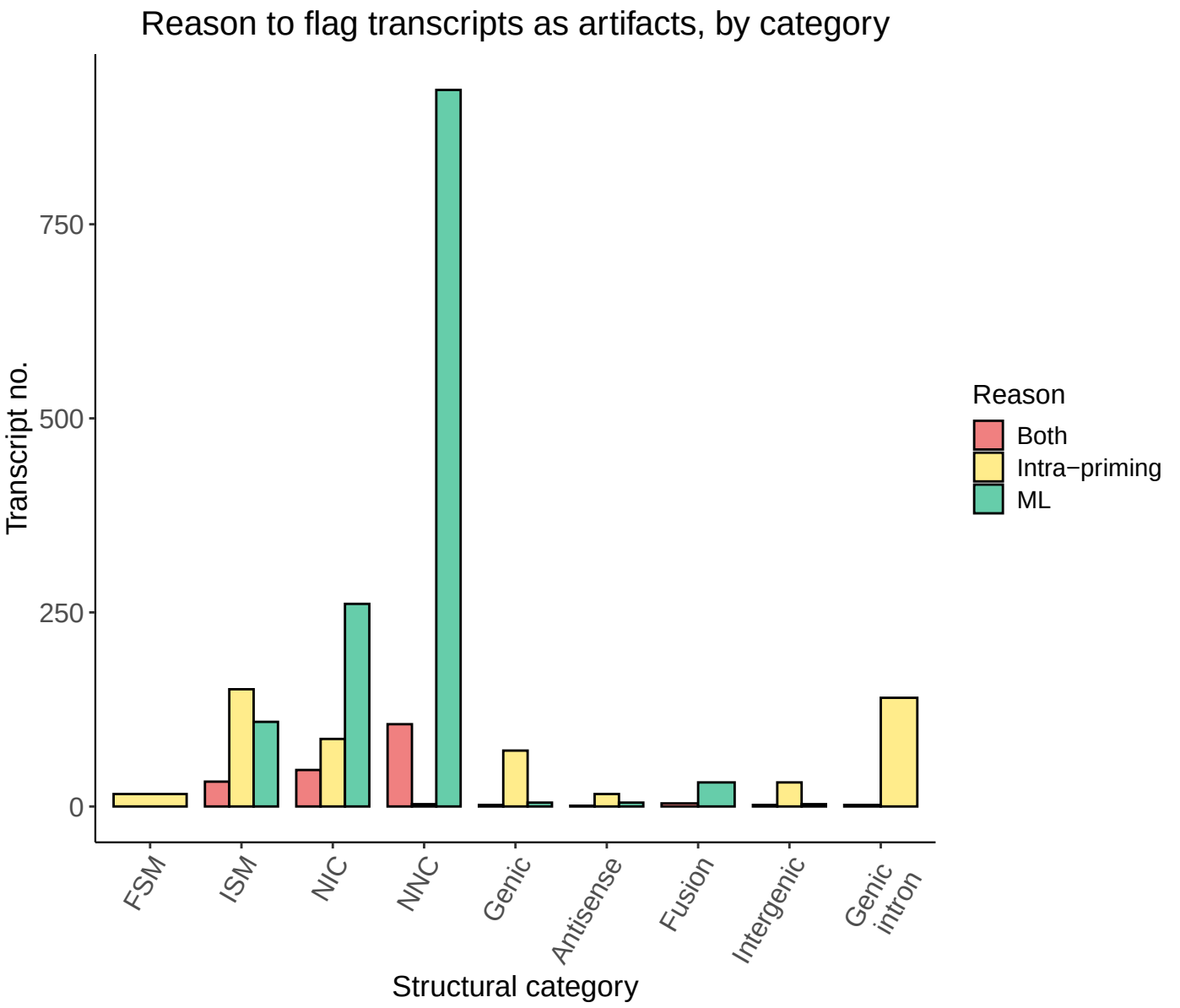
AccuracyPValue	3.005249e−19

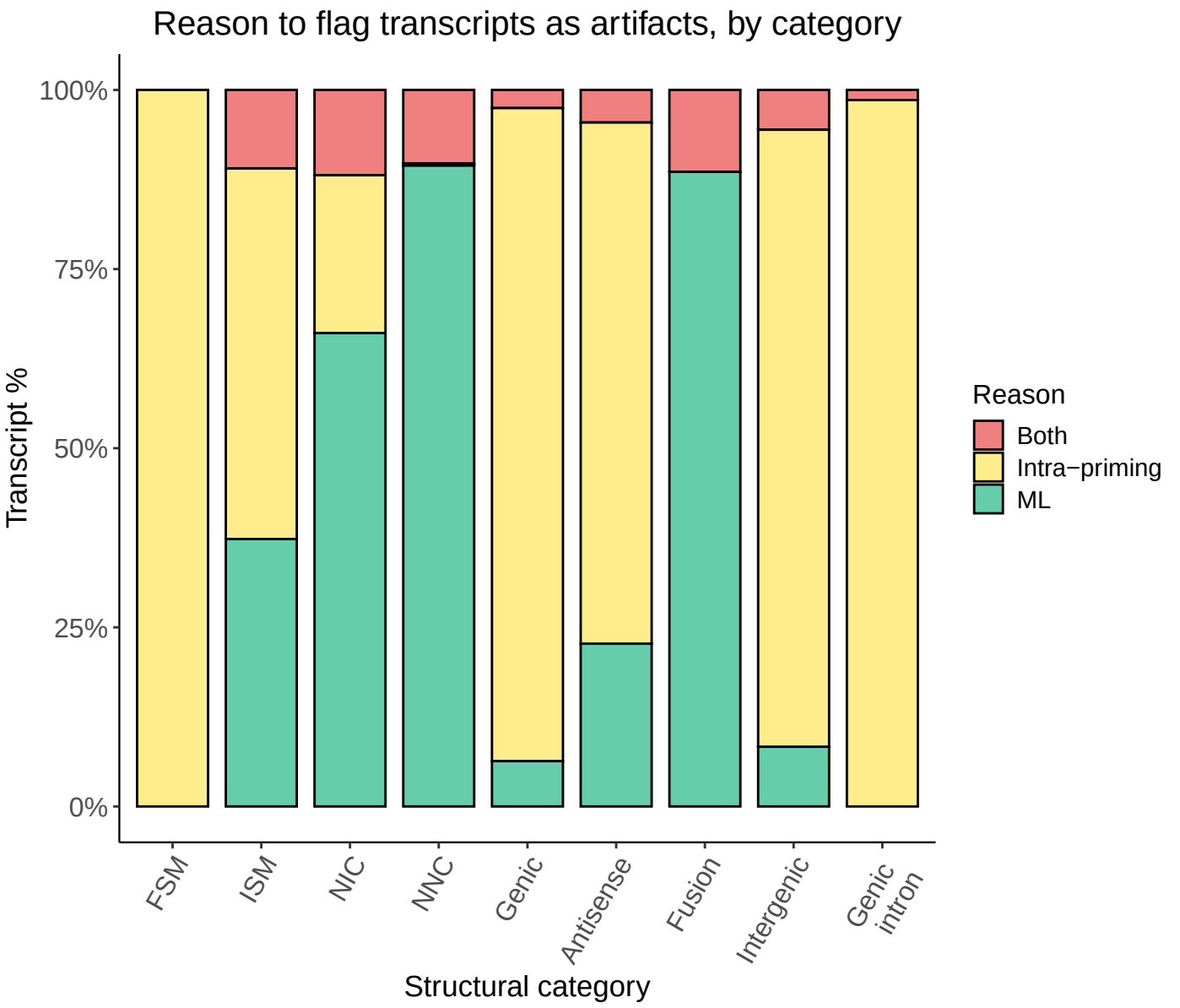
Confusion matrix

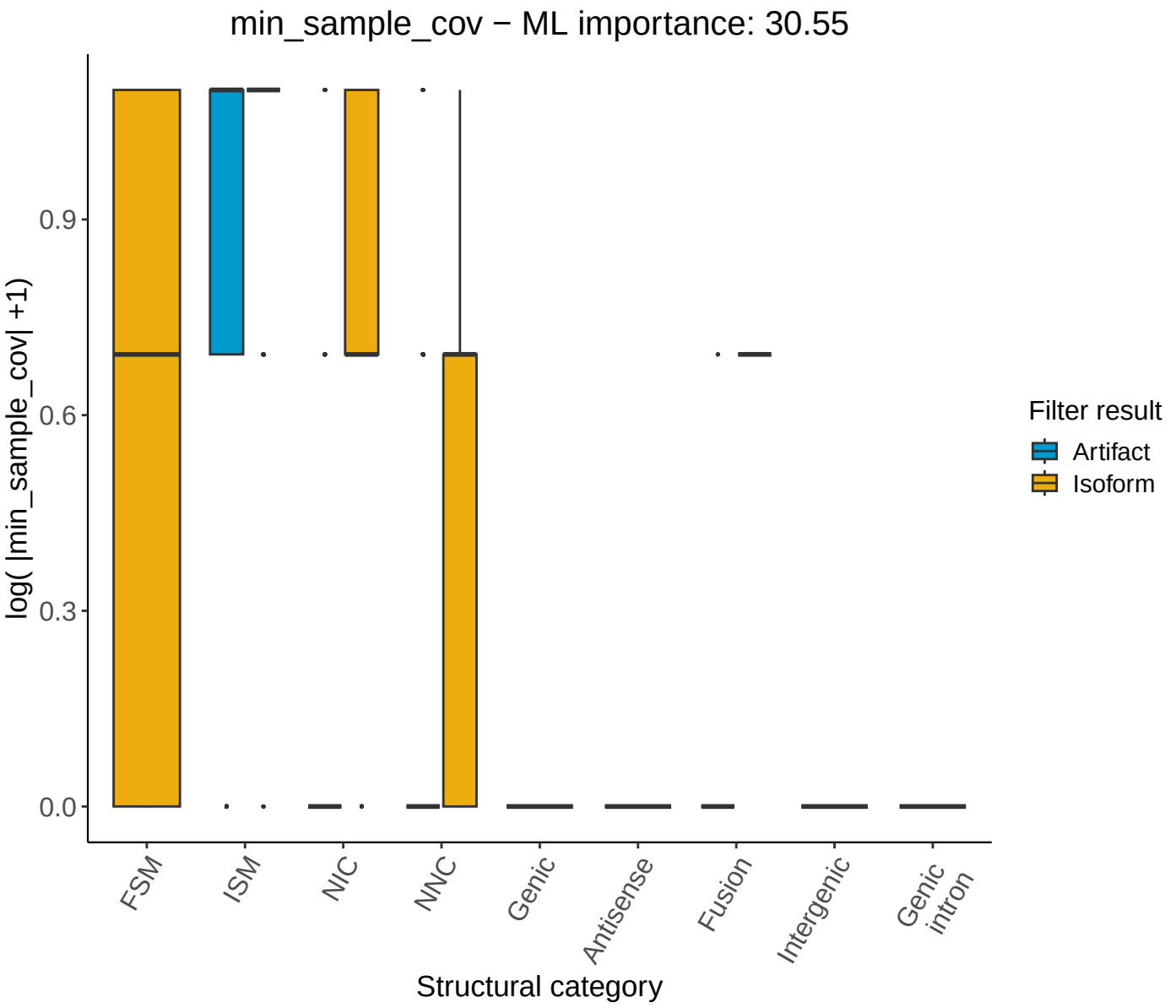
Prediction	Reference	Freq
POS	POS	46
NEG	POS	11
POS	NEG	1
NEG	NEG	56

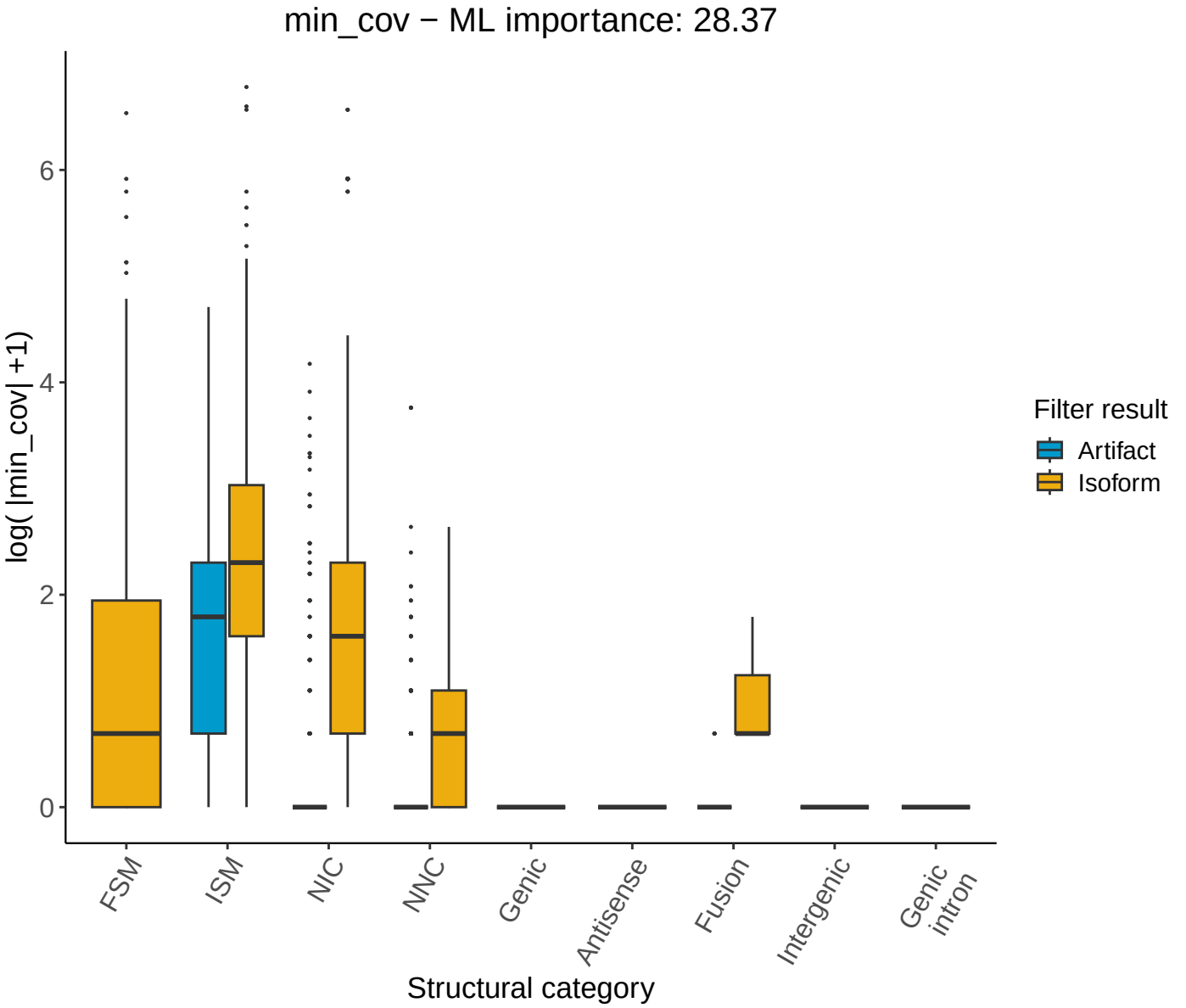

Transcripts classified: 3338
2.0
1.5
1.0
0.5
0.0
Density
0.00
0.25
0.50
0.75
1.00
Positive/Isoform probability obtained from ML filter classifier (POS_MLprob)

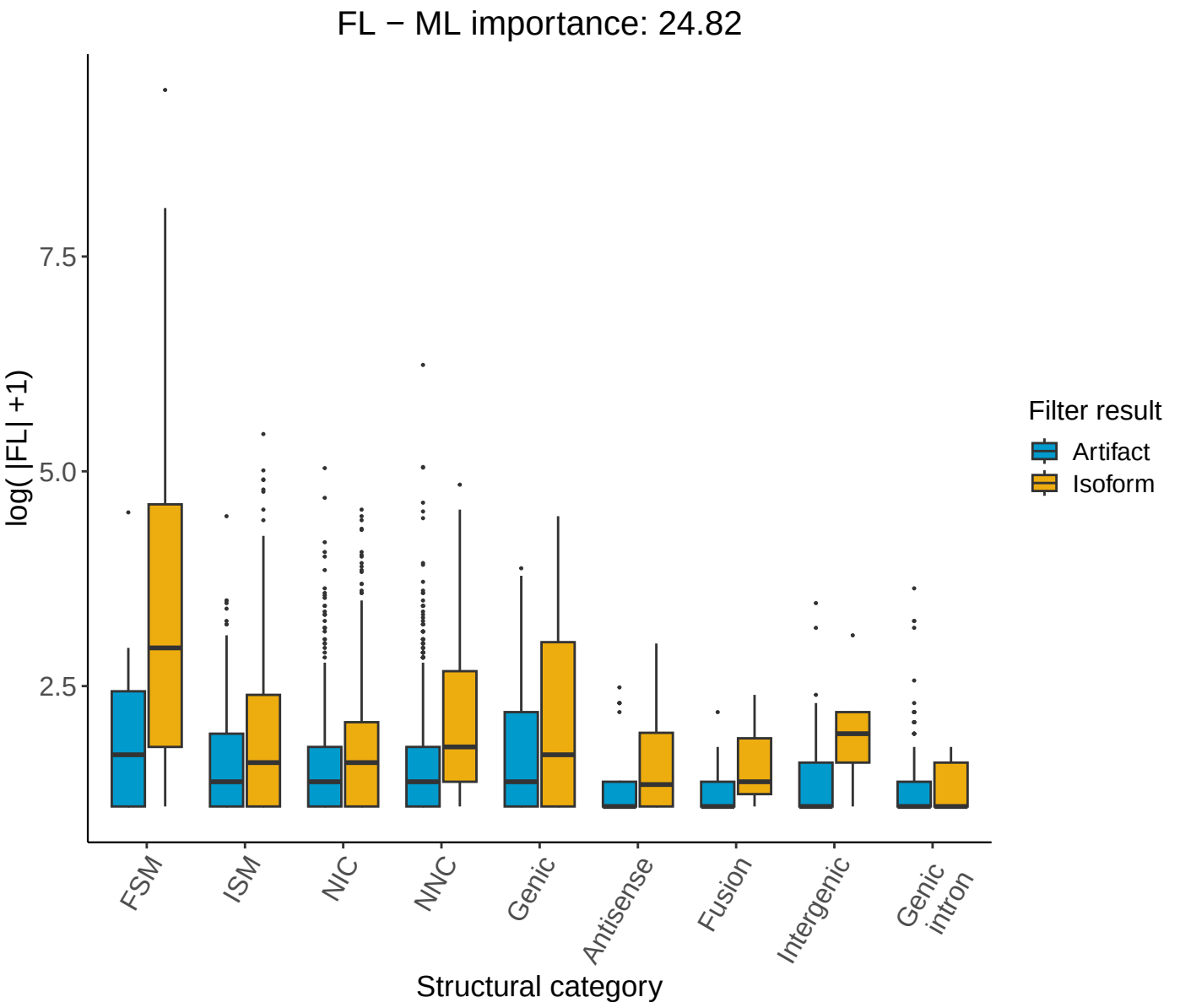

Variable importance in Random Forest classifier
SQANTI3 variables
min_sample_cov
min_cov
FL
bite
sd_cov
gene_exp
iso_exp
predicted_NMD
diff_to_gene_TSS
length
diff_to_gene_TTS
dist_to_CAGE_peak
FSM_class
exons
polyA_dist
perc_A_downstream_TTS
within_CAGE_peak
RTS_stage
polyA_motif_found
ratio_exp
coding
0
10
20
30
Importance

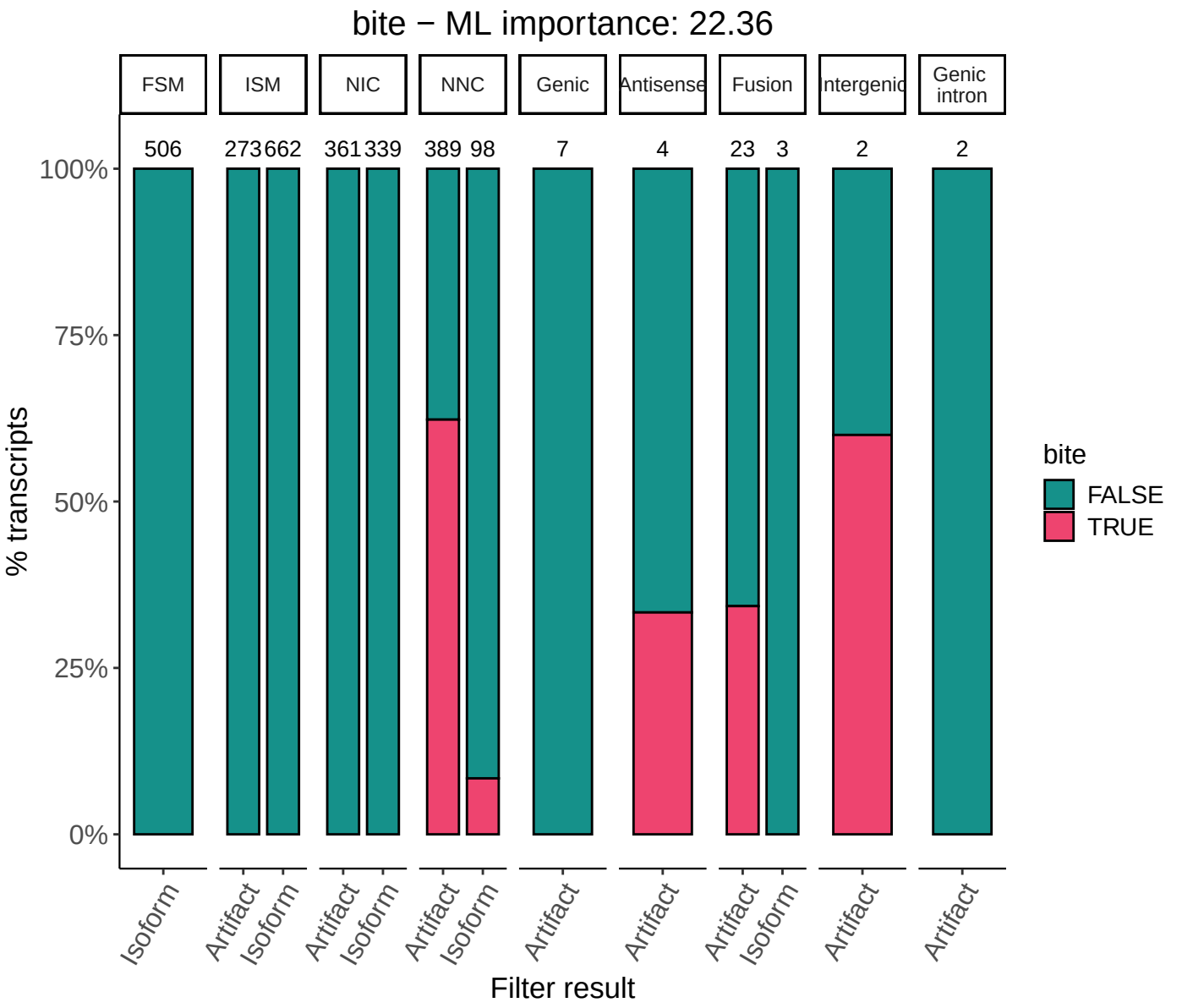

Reason to flag transcripts as artifacts, by category
Transcript no.
0
250
500
750
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Reason
Both
Intra−priming
ML

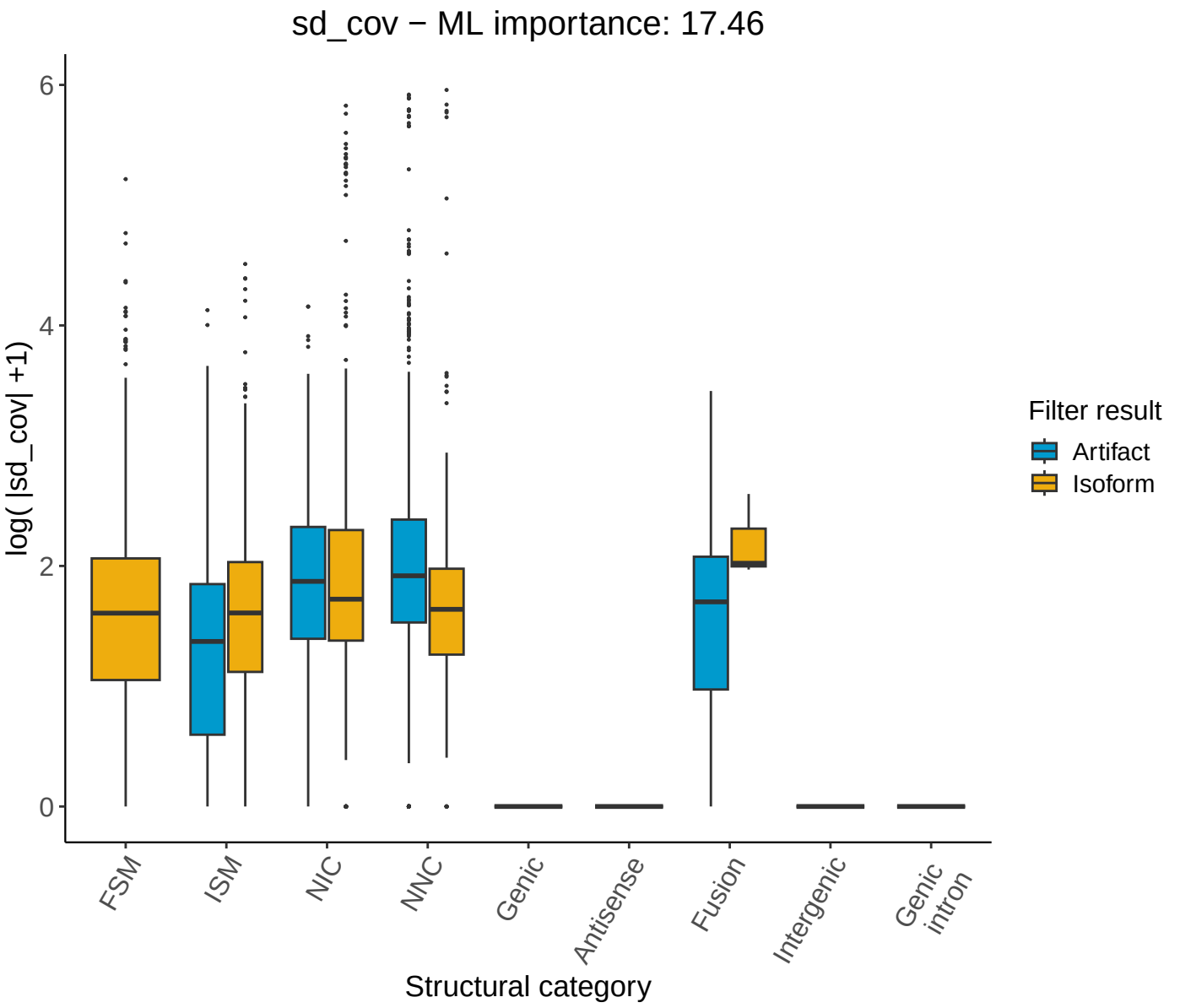

Reason to flag transcripts as artifacts, by category
100%
75%
50%
25%
0%
Transcript %
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Reason
Both
Intra−priming
ML

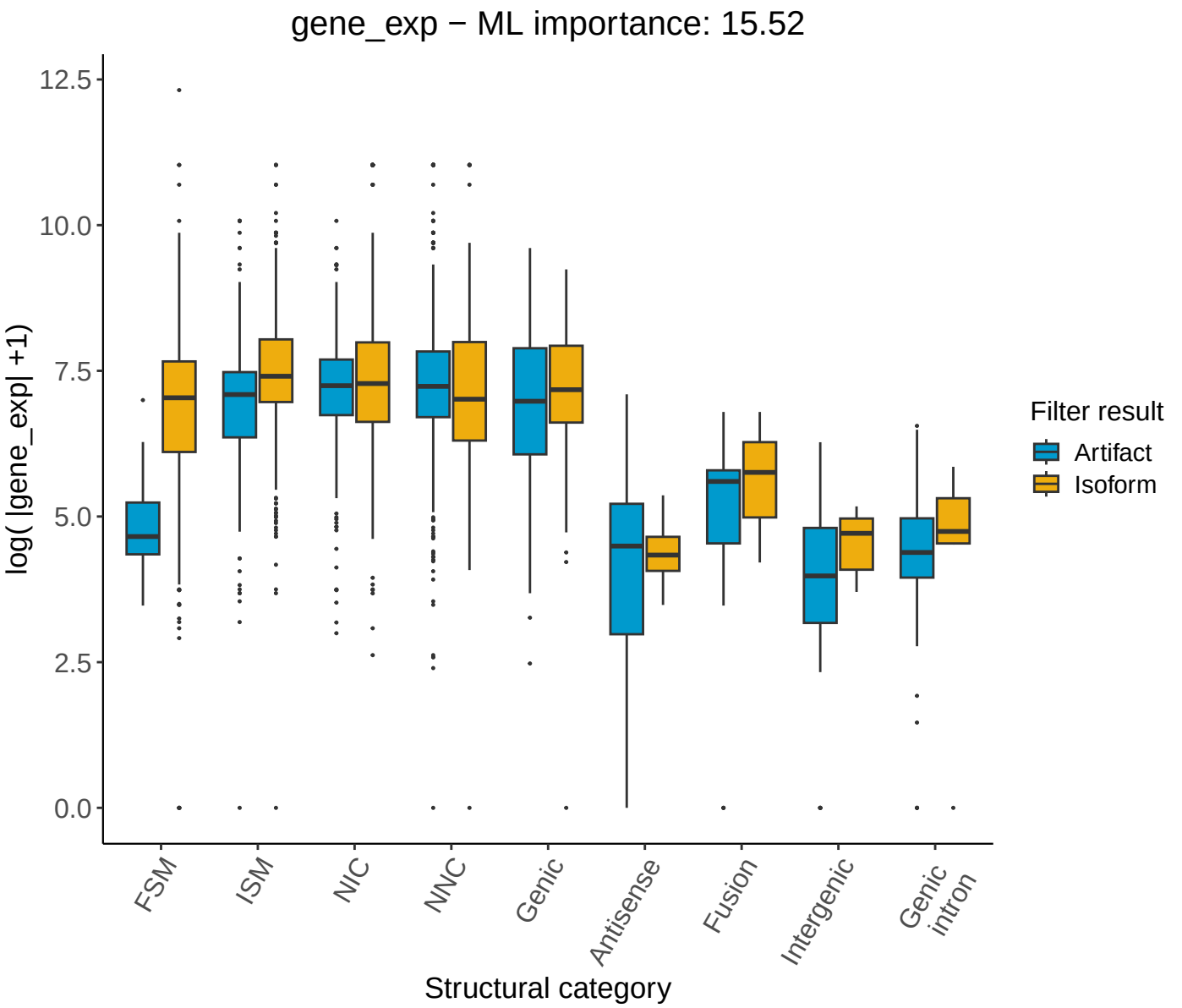

min_sample_cov − ML importance: 30.55
log(|min_sample_cov| +1)
0.0
0.3
0.6
0.9
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

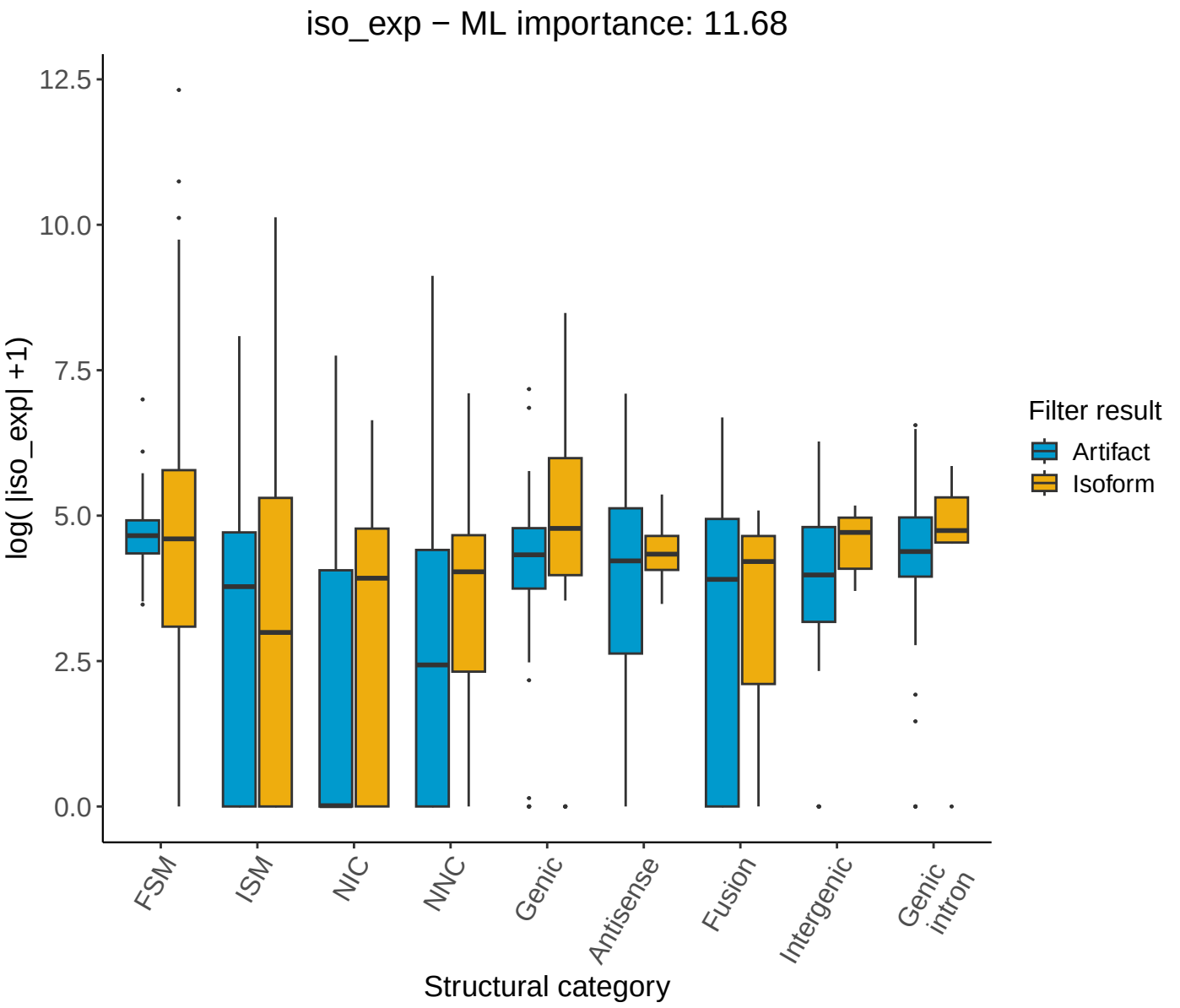

min_cov − ML importance: 28.37
log(|min_cov| +1)
6
4
2
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

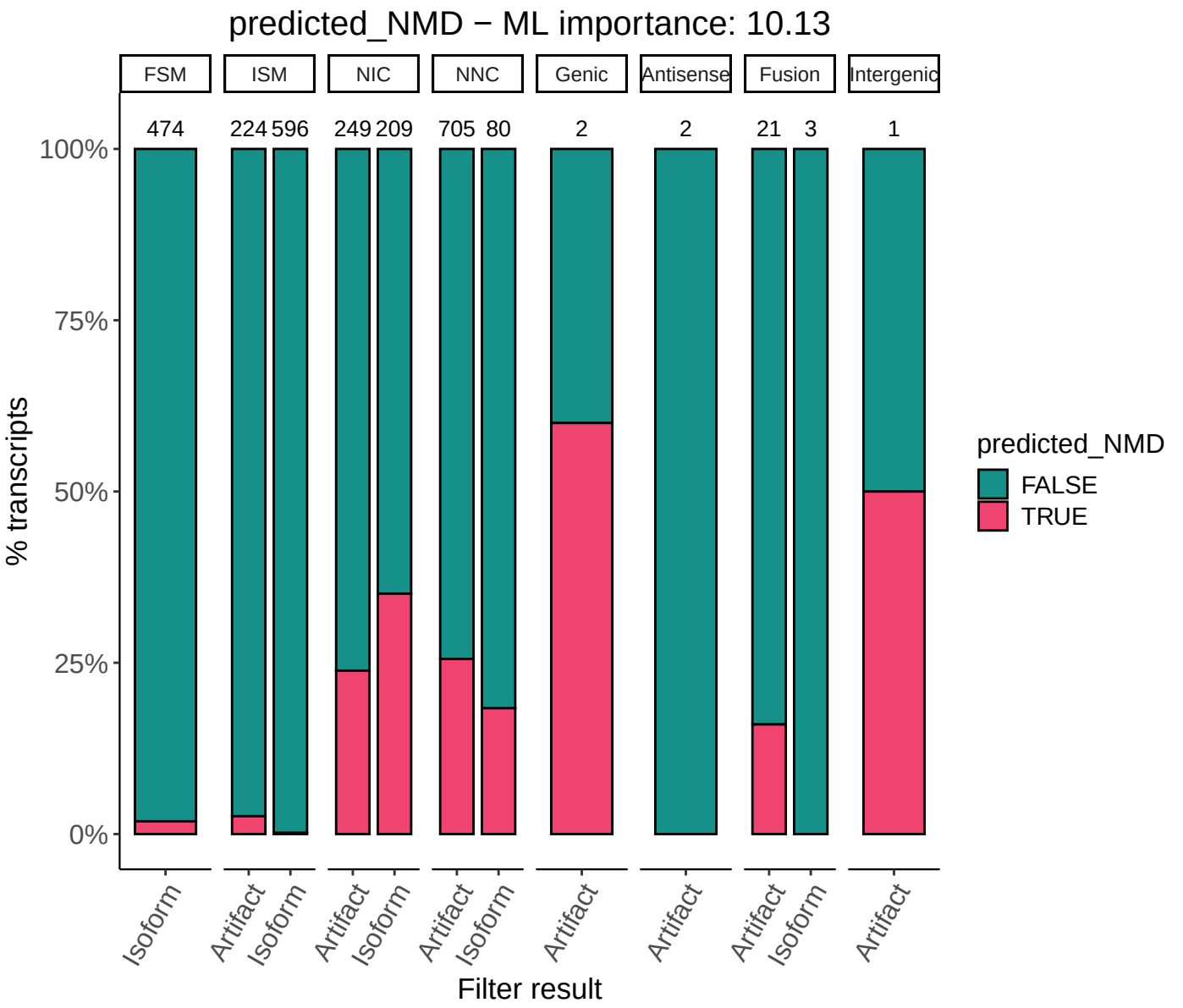

FL − ML importance: 24.82
log(|FL| +1)
7.5
5.0
2.5
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

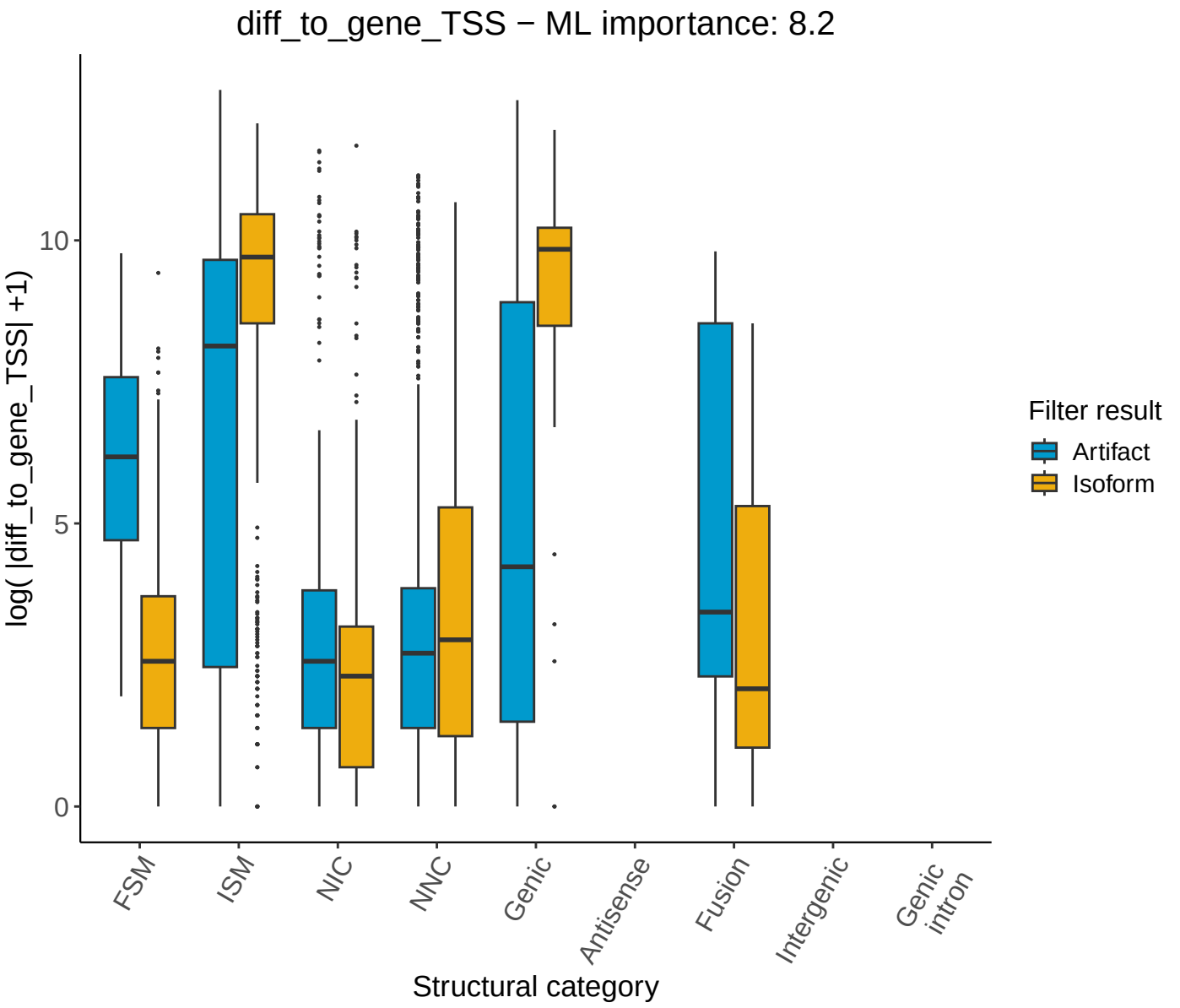

bite − ML importance: 22.36
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
506
273
662
361
339
389
98
7
4
23
3
2
2
100%
75%
50%
25%
0%
% transcripts
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Artifact
Artifact
Isoform
Artifact
Artifact
Filter result
bite
FALSE
TRUE

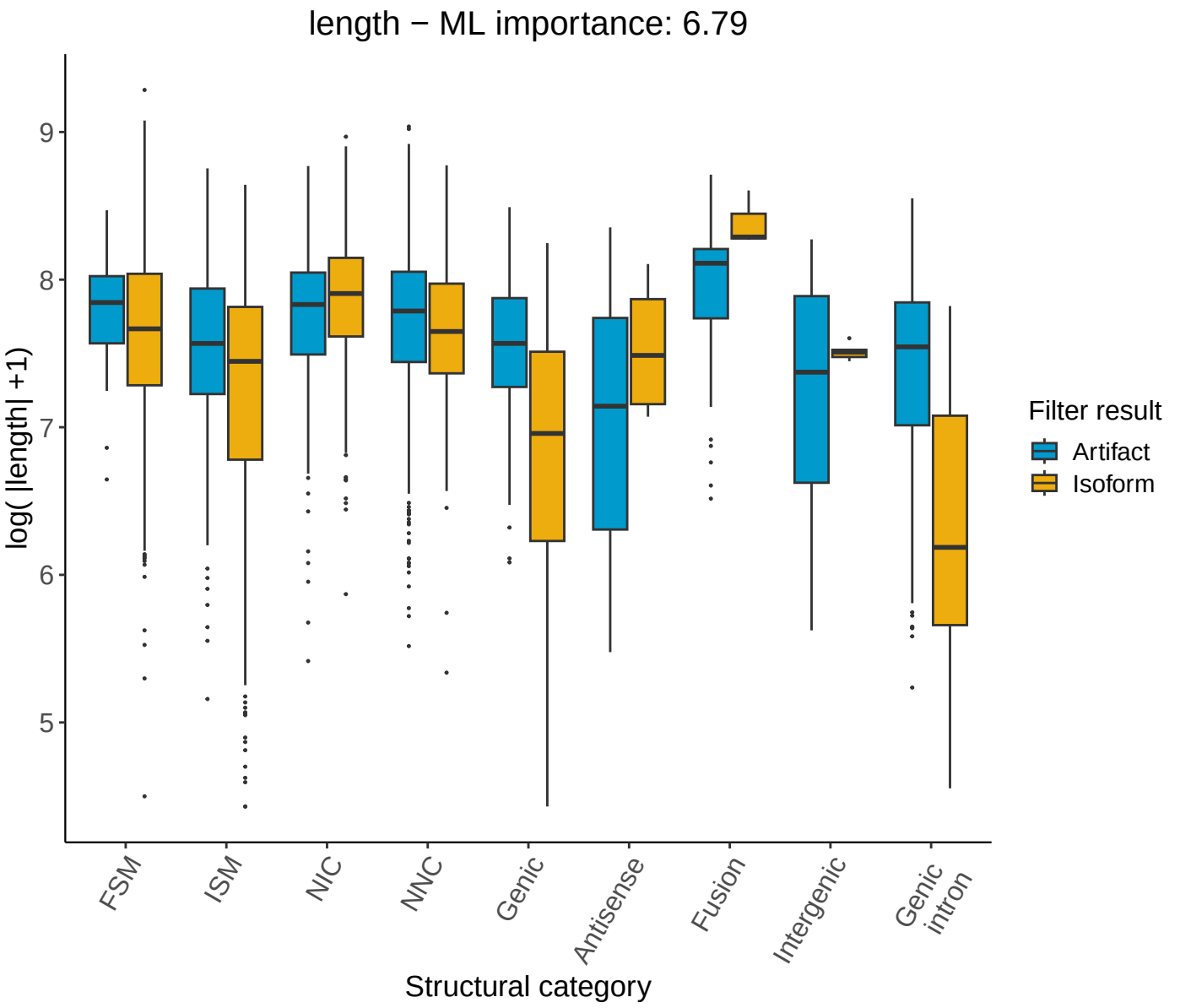

sd_cov − ML importance: 17.46
log(|sd_cov| +1)
6
4
2
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

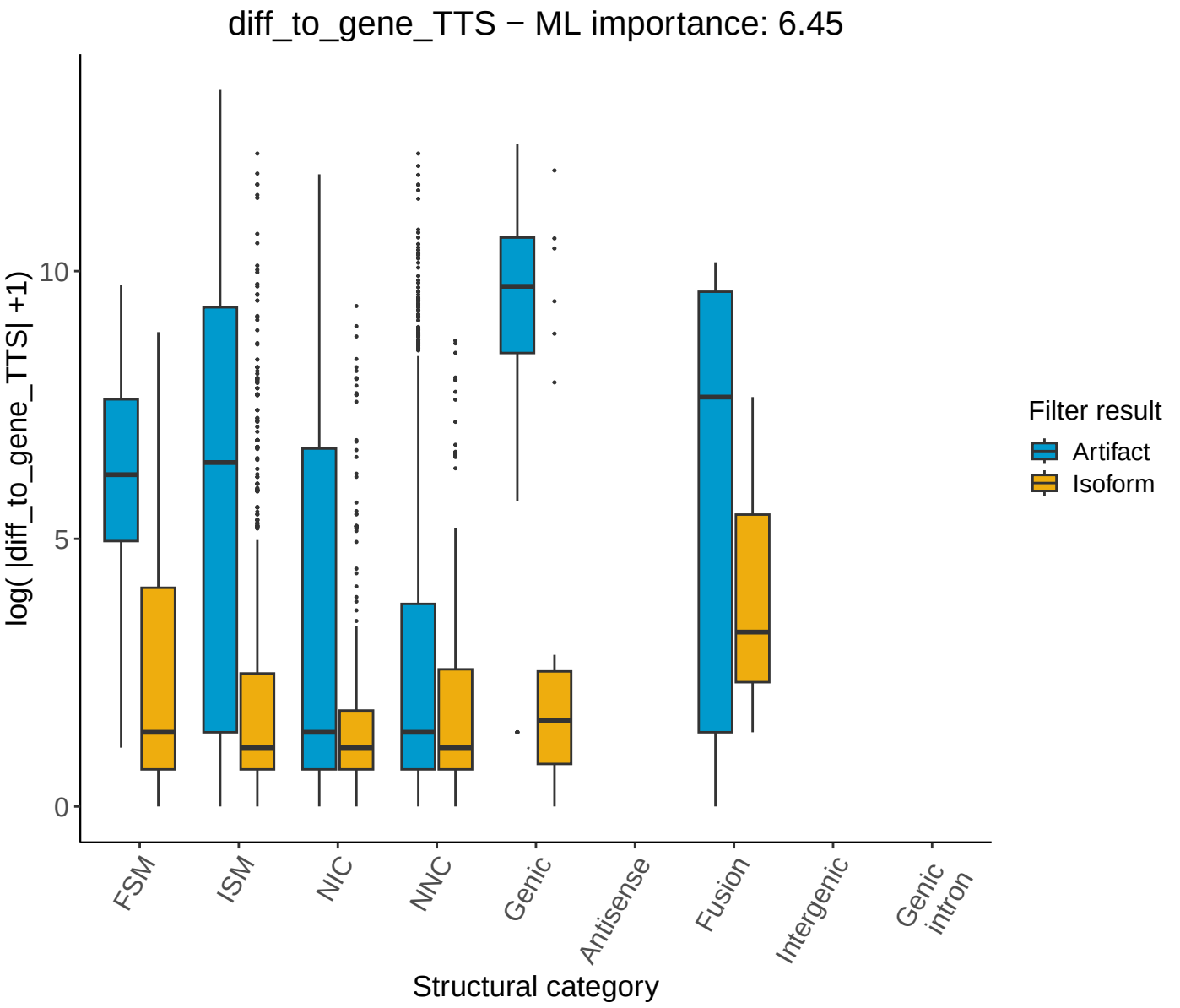

gene_exp − ML importance: 15.52
log(|gene_exp| +1)
12.5
10.0
7.5
5.0
2.5
0.0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

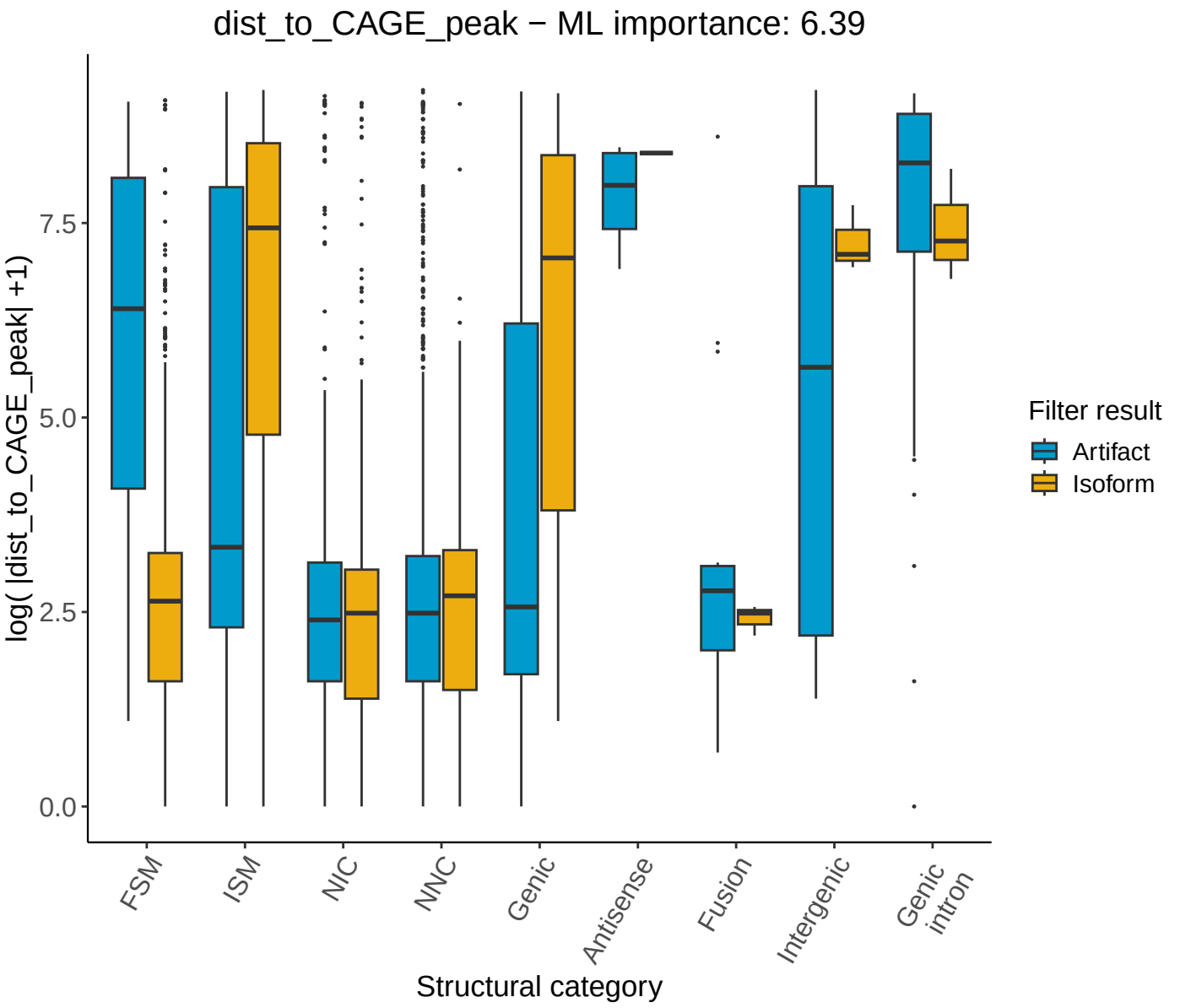

iso_exp − ML importance: 11.68
log(|iso_exp| +1)
12.5
10.0
7.5
5.0
2.5
0.0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

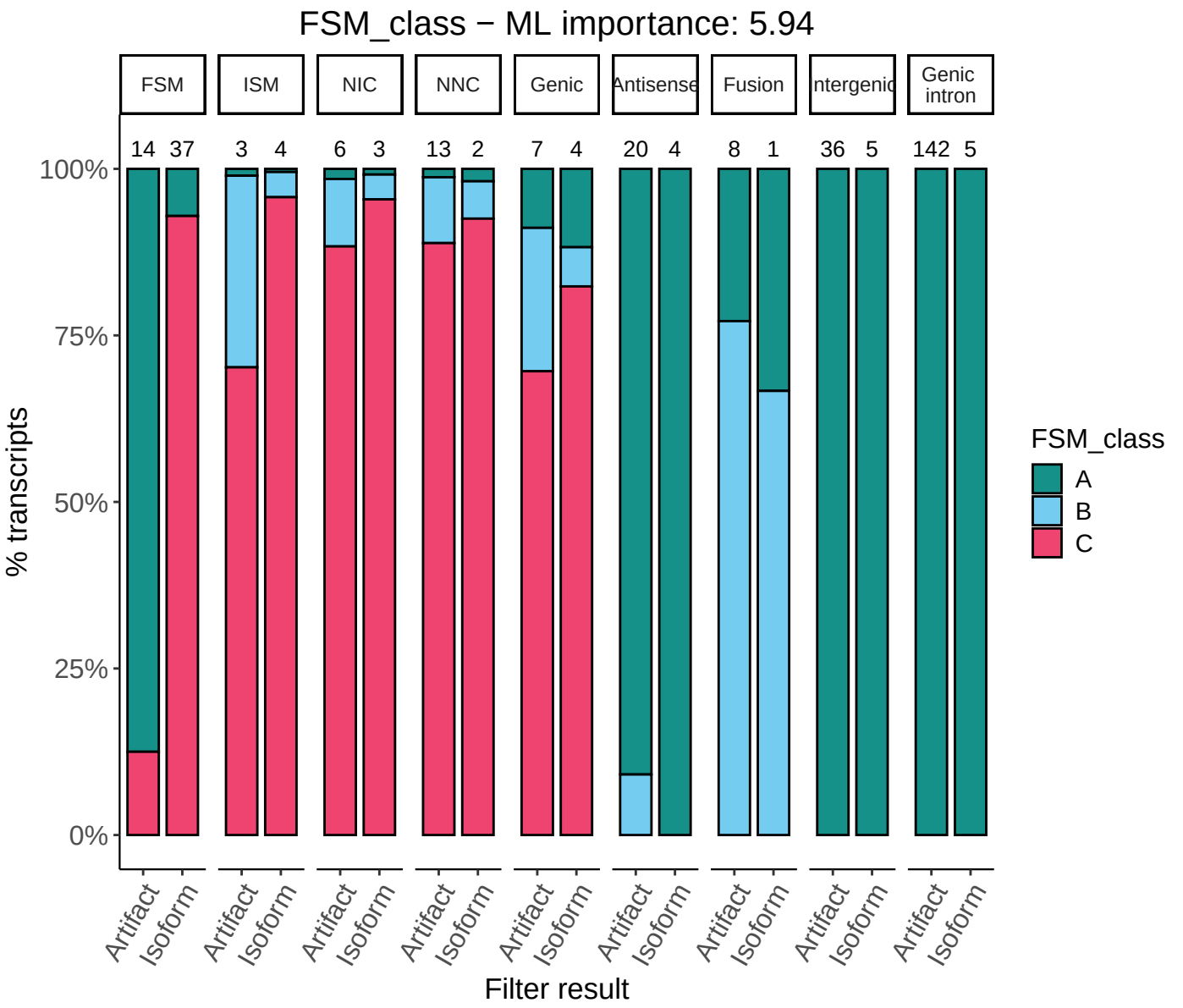

predicted_NMD − ML importance: 10.13
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
474
224
596
249
209
705
80
2
2
21
3
1
100%
75%
50%
0%
% transcripts
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Artifact
Artifact
Isoform
Artifact
Filter result
predicted_NMD
FALSE
TRUE

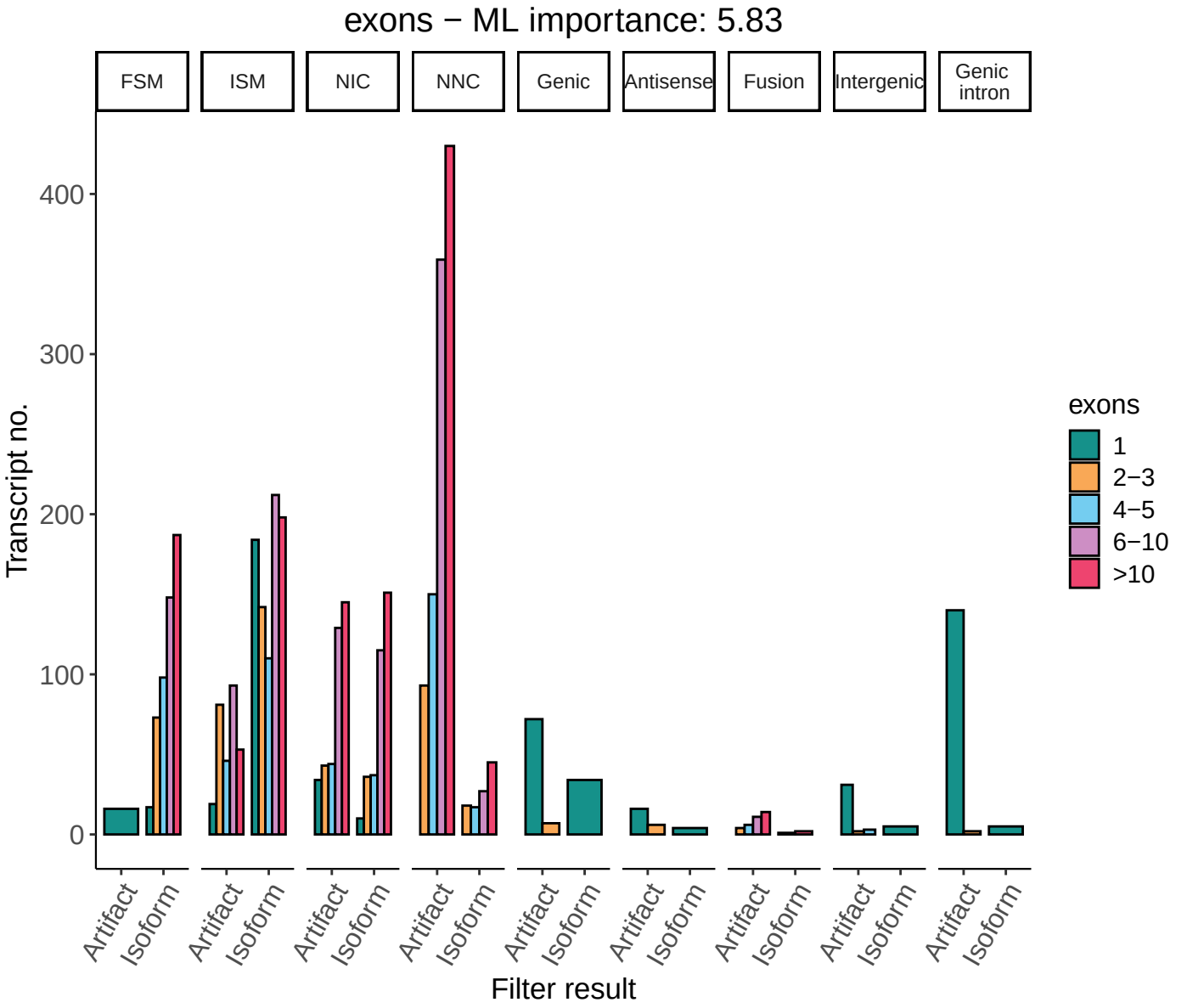

diff_to_gene_TSS − ML importance: 8.2
log(|diff_to_gene_TSS| +1)
10
5
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

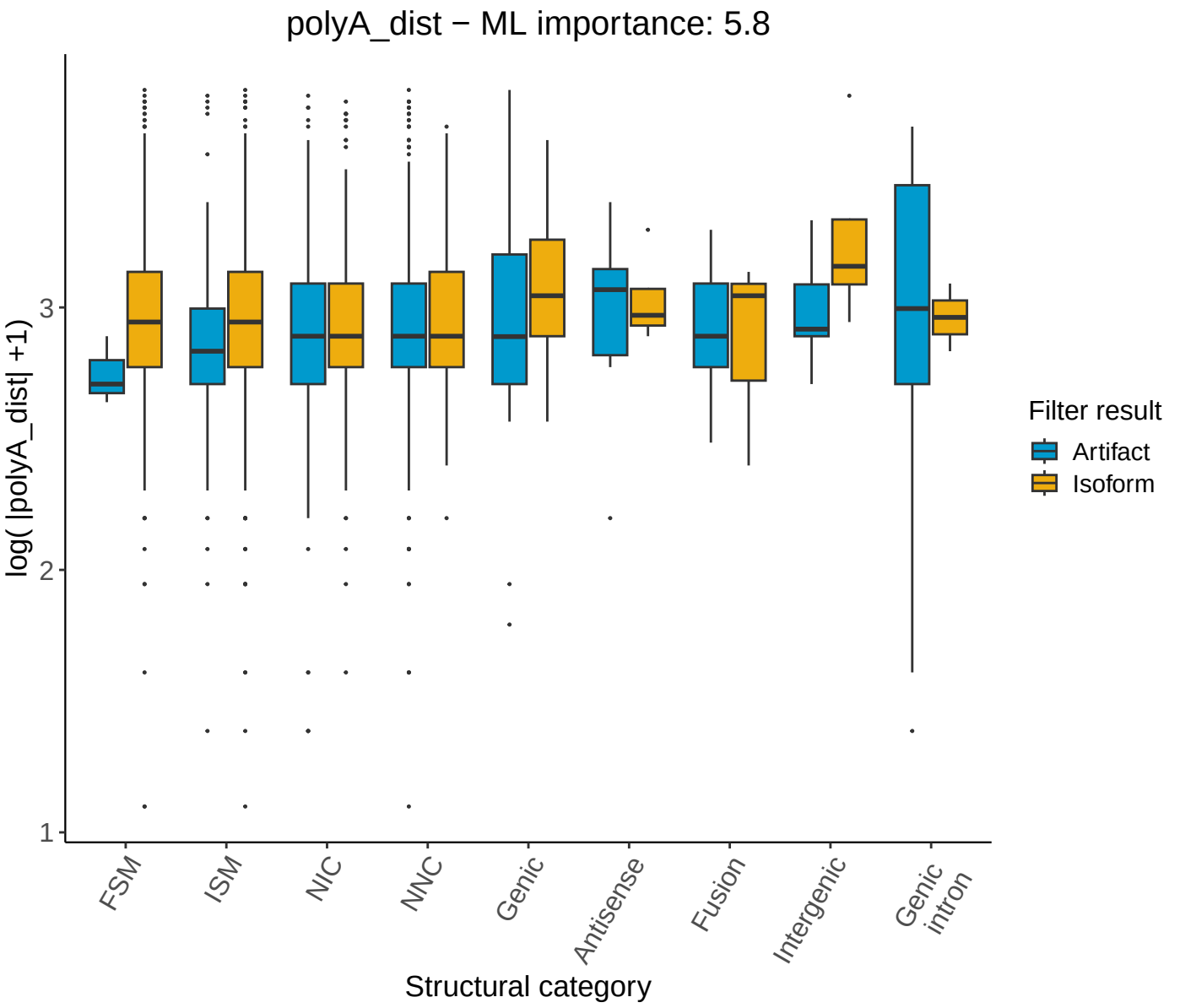

length − ML importance: 6.79
log(|length| +1)
9
8
7
6
5
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

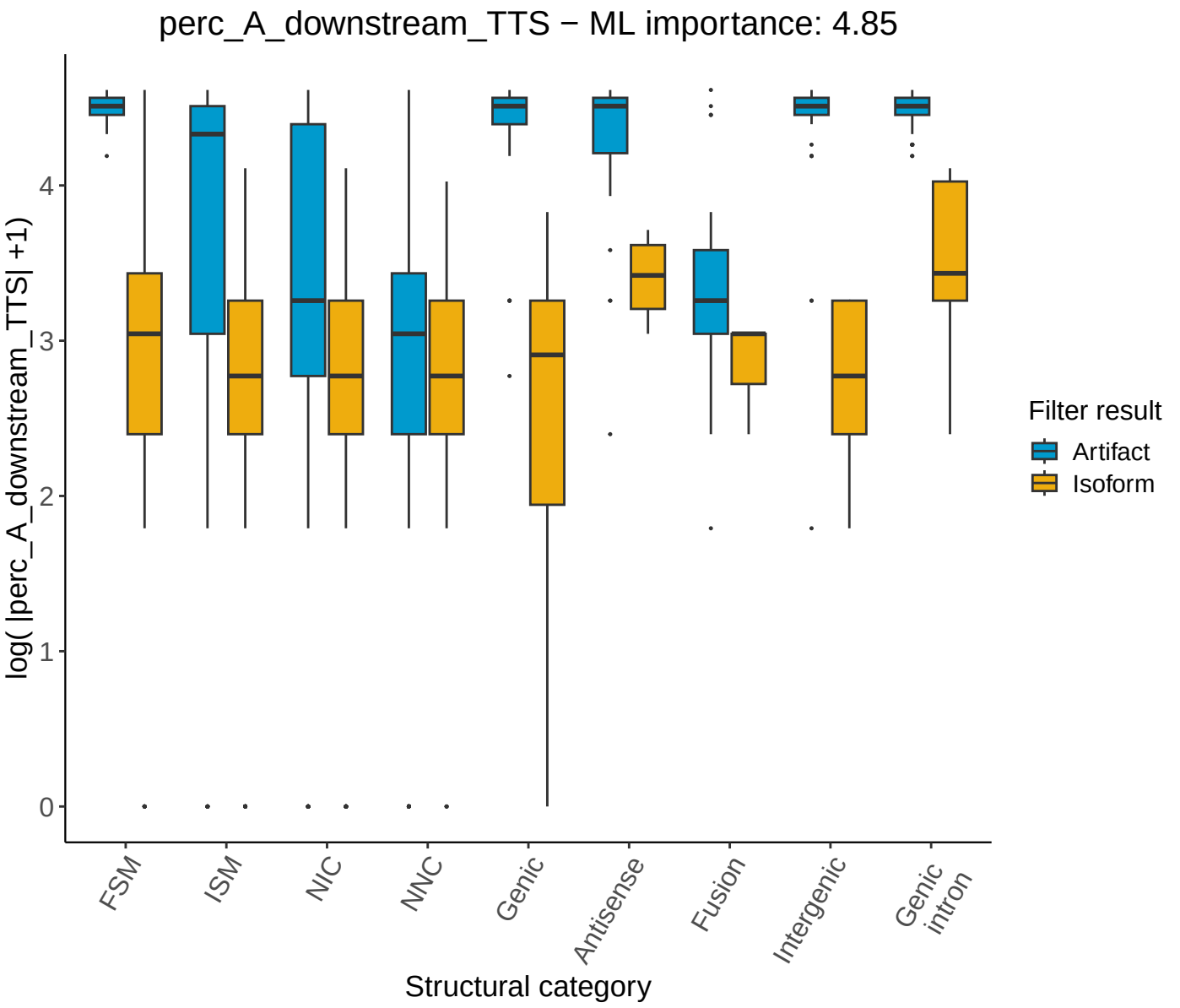

diff_to_gene_TTS − ML importance: 6.45
log(|diff_to_gene_TTS| +1)
10
5
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

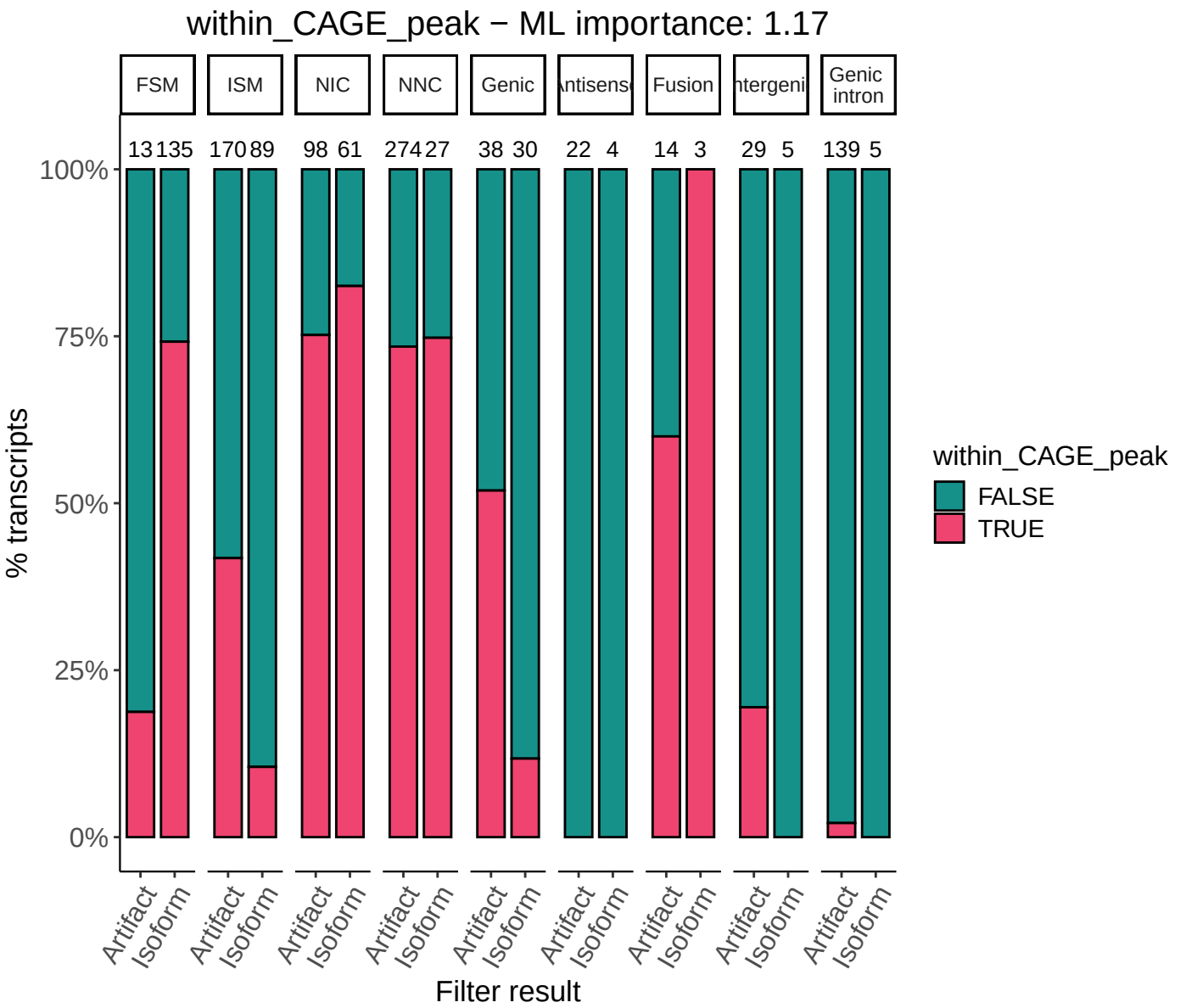

dist_to_CAGE_peak − ML importance: 6.39
log(|dist_to_CAGE_peak| +1)
7.5
5.0
2.5
0.0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

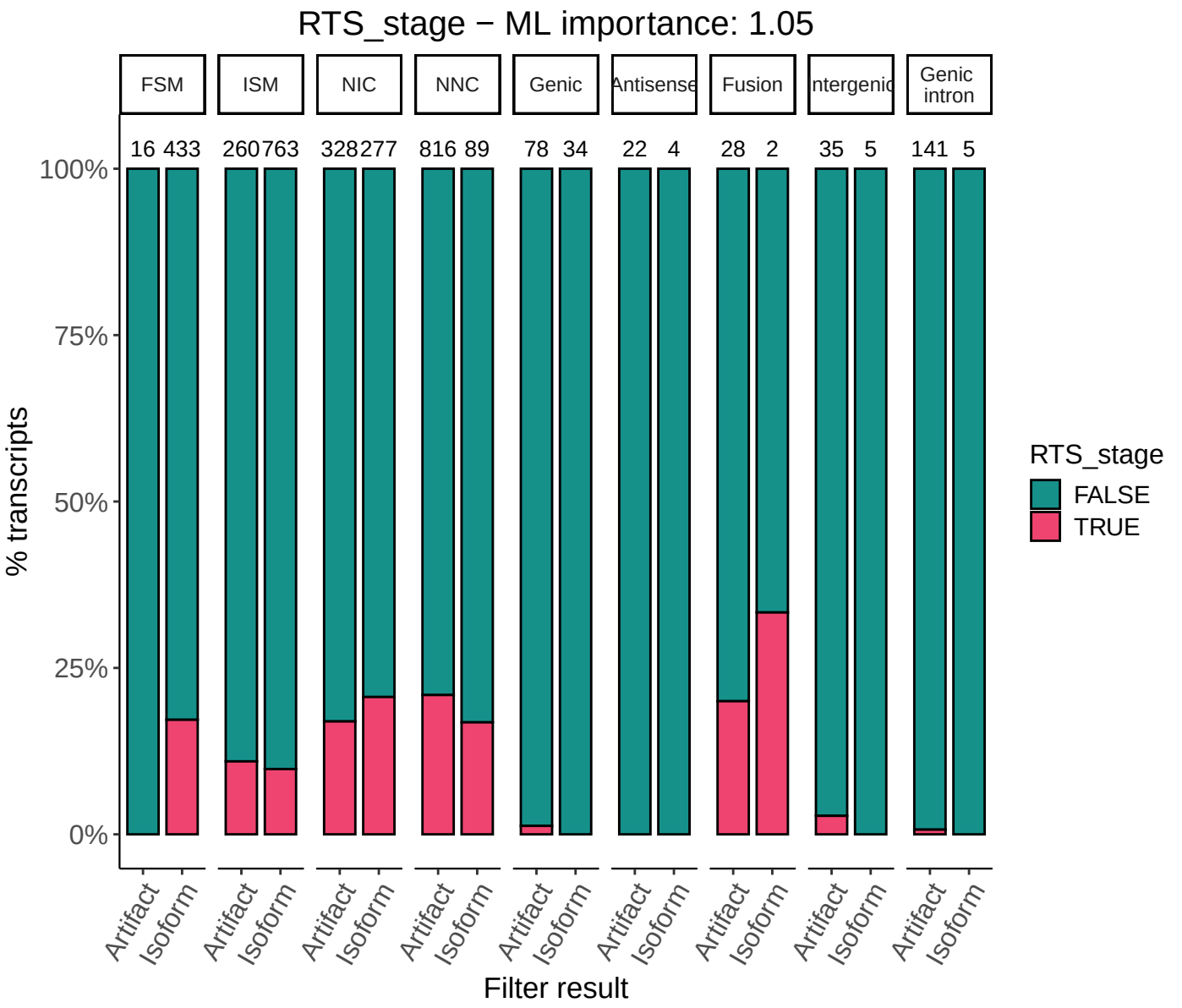

FSM_class − ML importance: 5.94
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
ntergenic
Genic intron
14 37
3 4
6 3
13 2
7 4
20 4
8 1
36 5
142 5
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
FSM_class
A
B
C

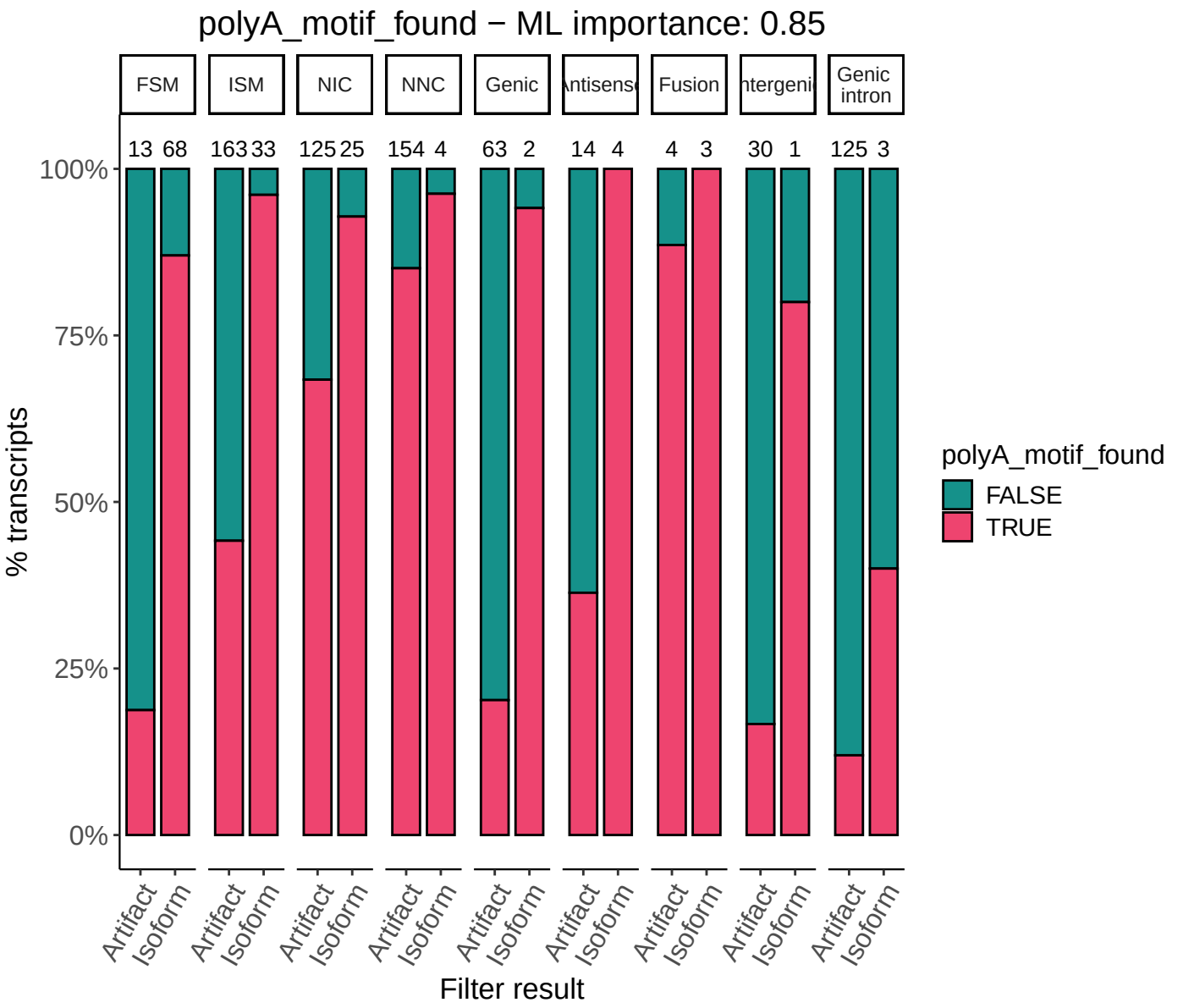

exons − ML importance: 5.83
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Transcript no.
0
100
200
300
400
Artifact
Isoform
Filter result
exons
1
2−3
4−5
6−10
>10

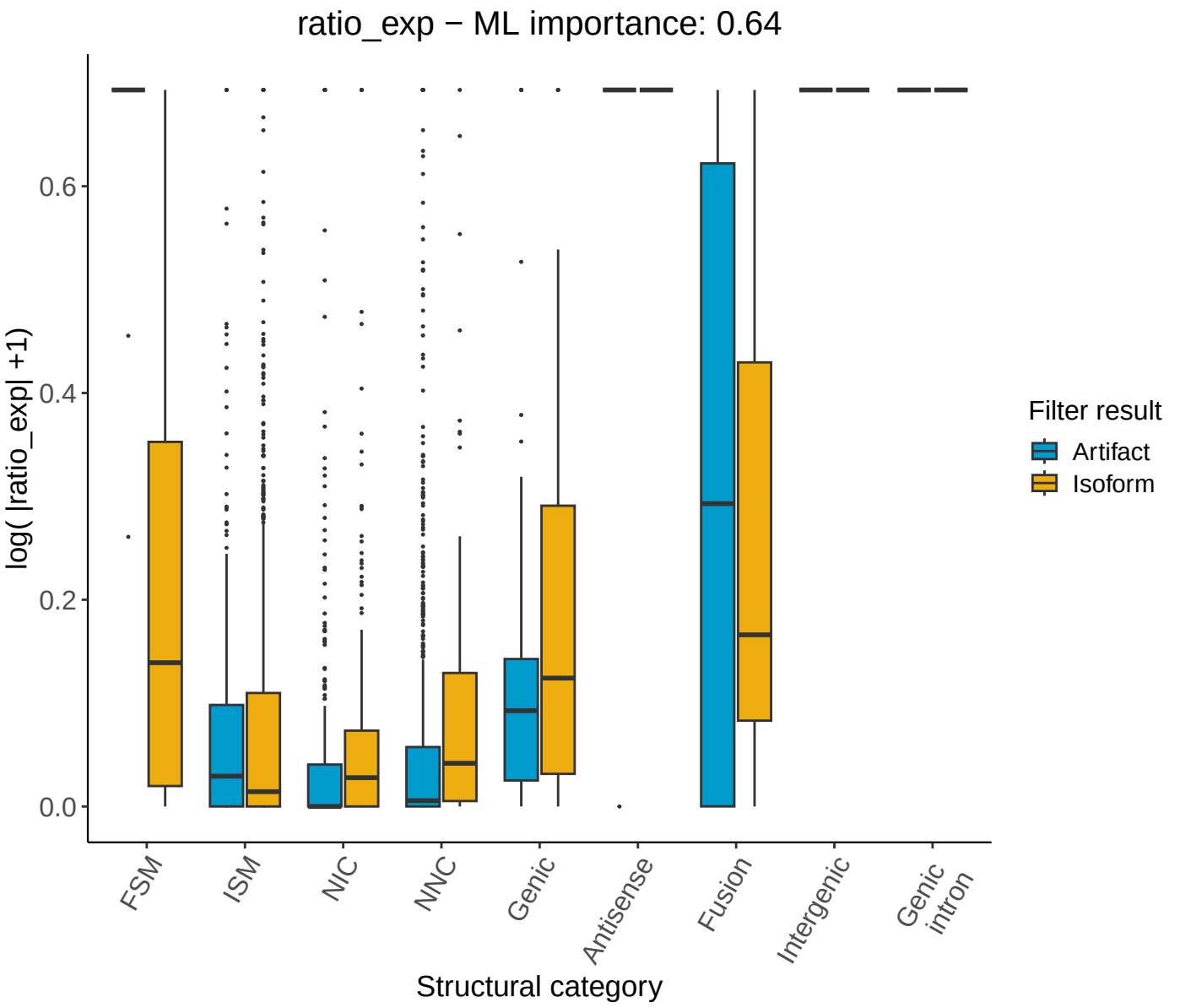

polyA_dist − ML importance: 5.8
log(|polyA_dist| +1)
3
2
1
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

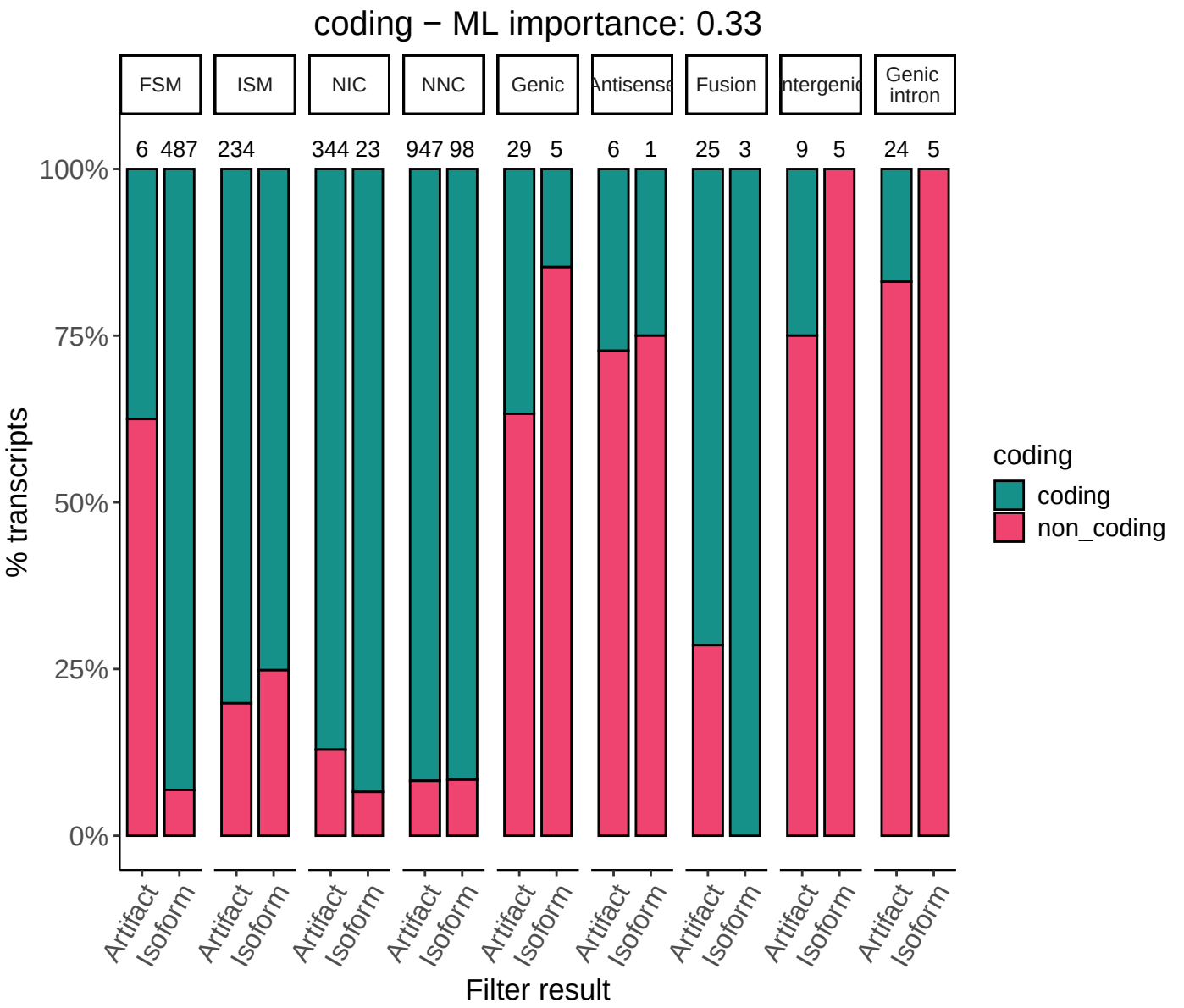

perc_A_downstream_TTS − ML importance: 4.85
log(|perc_A_downstream_TTS| +1)
0
1
2
3
4
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform


within_CAGE_peak − ML importance: 1.17
FSM
ISM
NIC
NNC
Genic
ntisens
Fusion
ntergeni
Genic intron
13 135
170 89
98 61
274 27
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

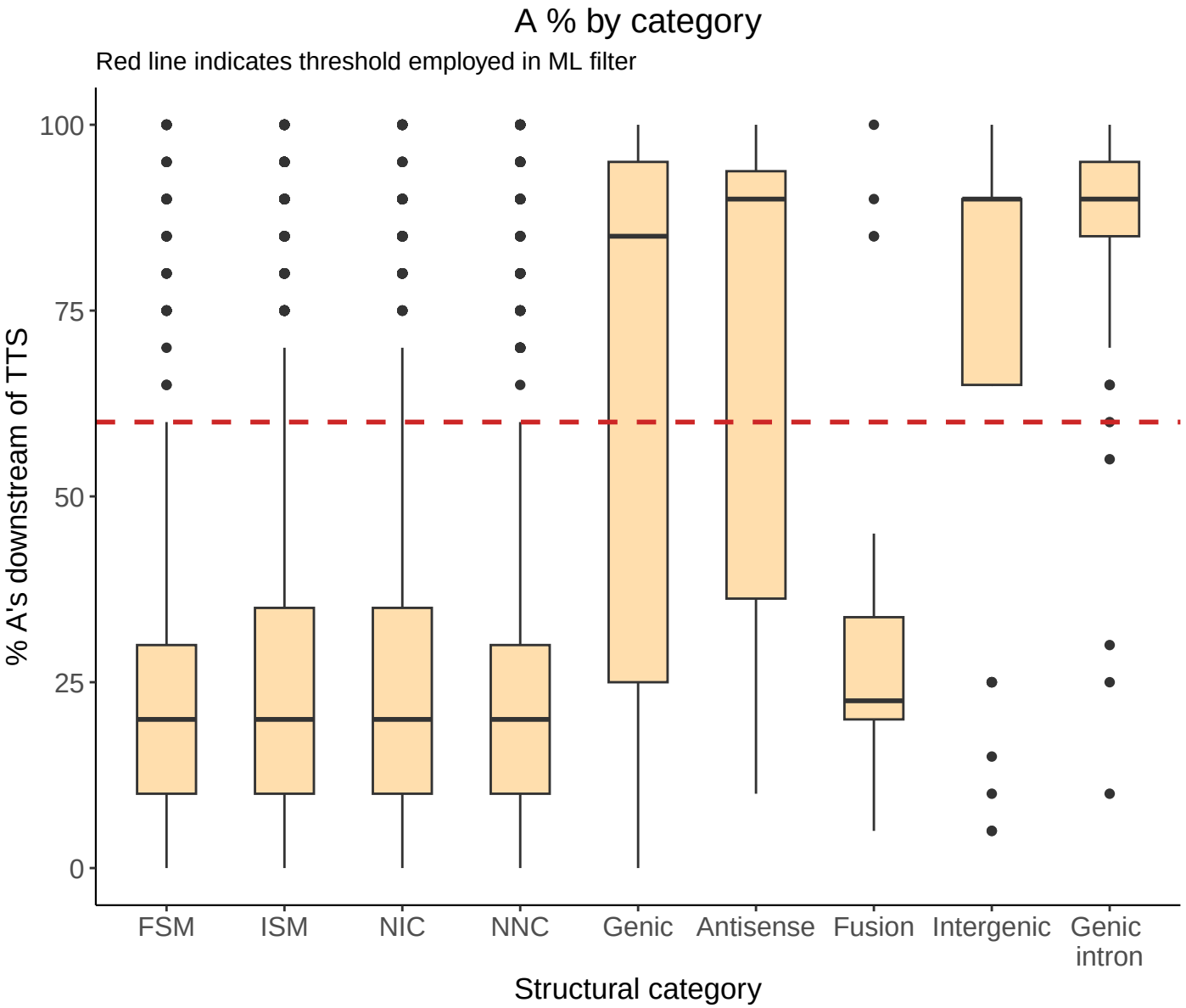

RTS_stage − ML importance: 1.05
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
ntergenic
Genic intron
16 433
260 763
328 277
816 89
78 34
22 4
28 2
35 5
141 5
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
RTS_stage
FALSE
TRUE

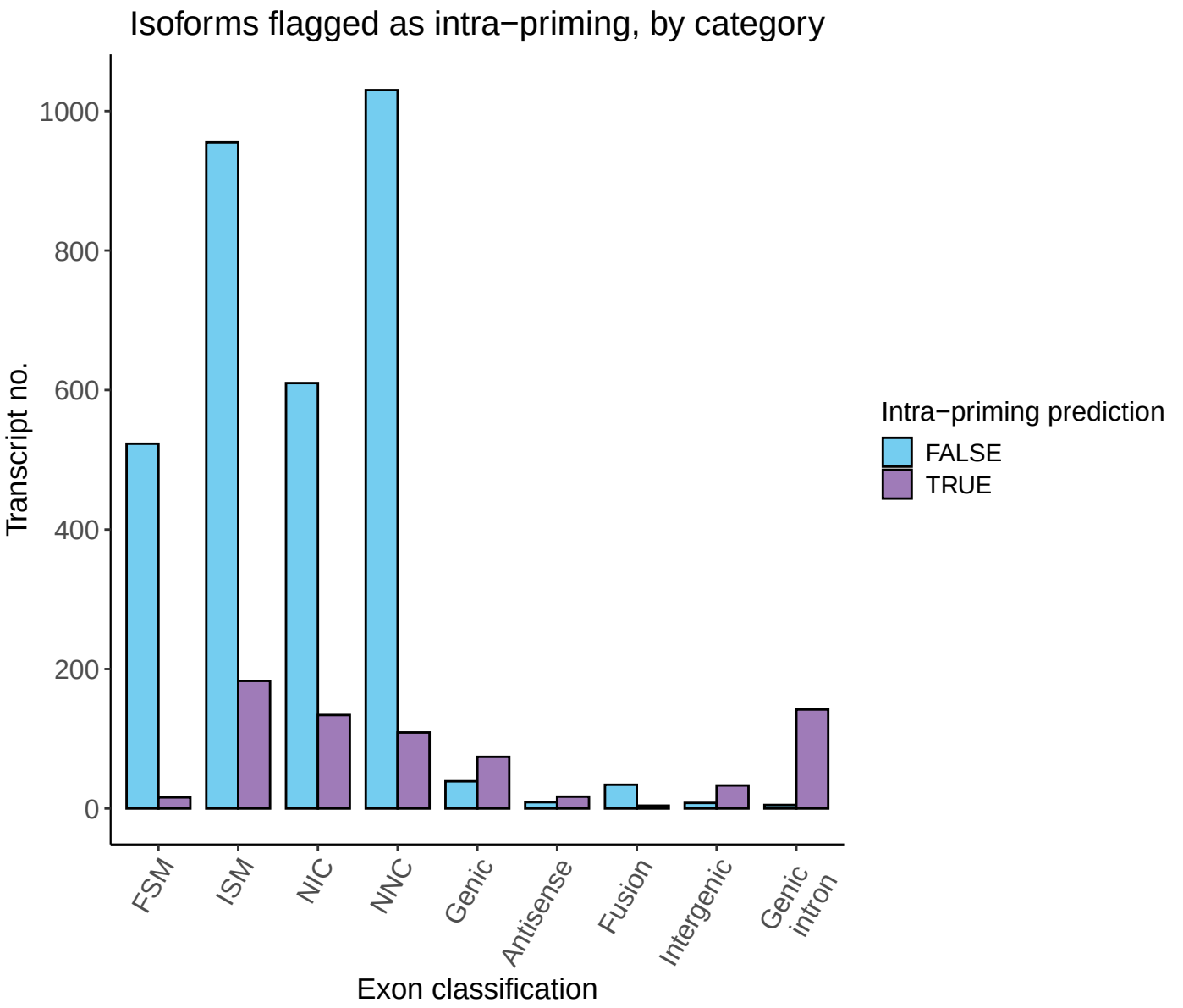

polyA_motif_found − ML importance: 0.85
FSM
ISM
NIC
NNC
Genic
ntisens
Fusion
ntergeni
Genic intron
13 68
163 33
125 25
154 4
63 2
14 4
4 3
30 1
125 3
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
polyA_motif_found
FALSE
TRUE

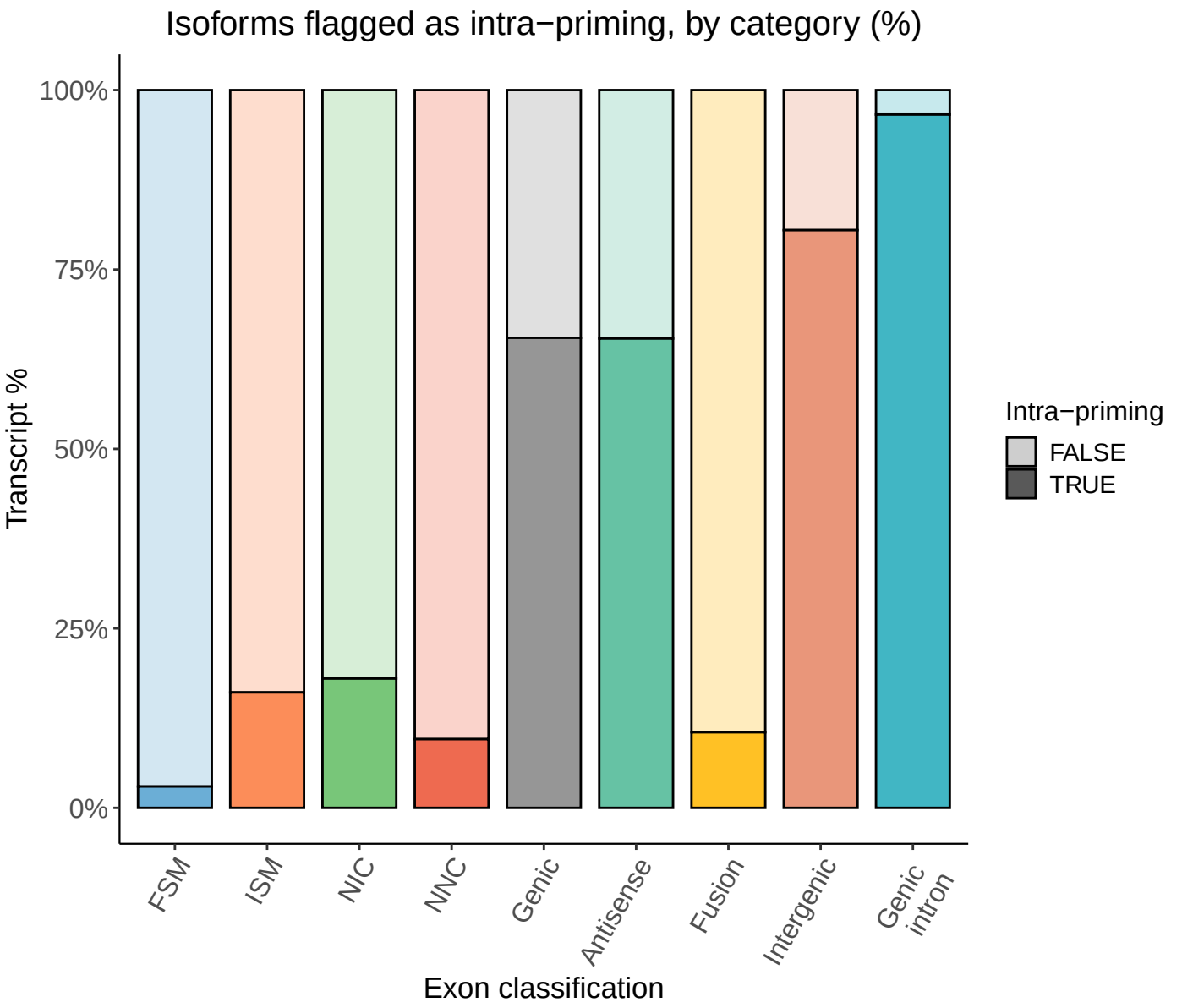

ratio_exp − ML importance: 0.64
log(|ratio_exp| +1)
0.0
0.2
0.4
0.6
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
Structural category
Filter result
Artifact
Isoform

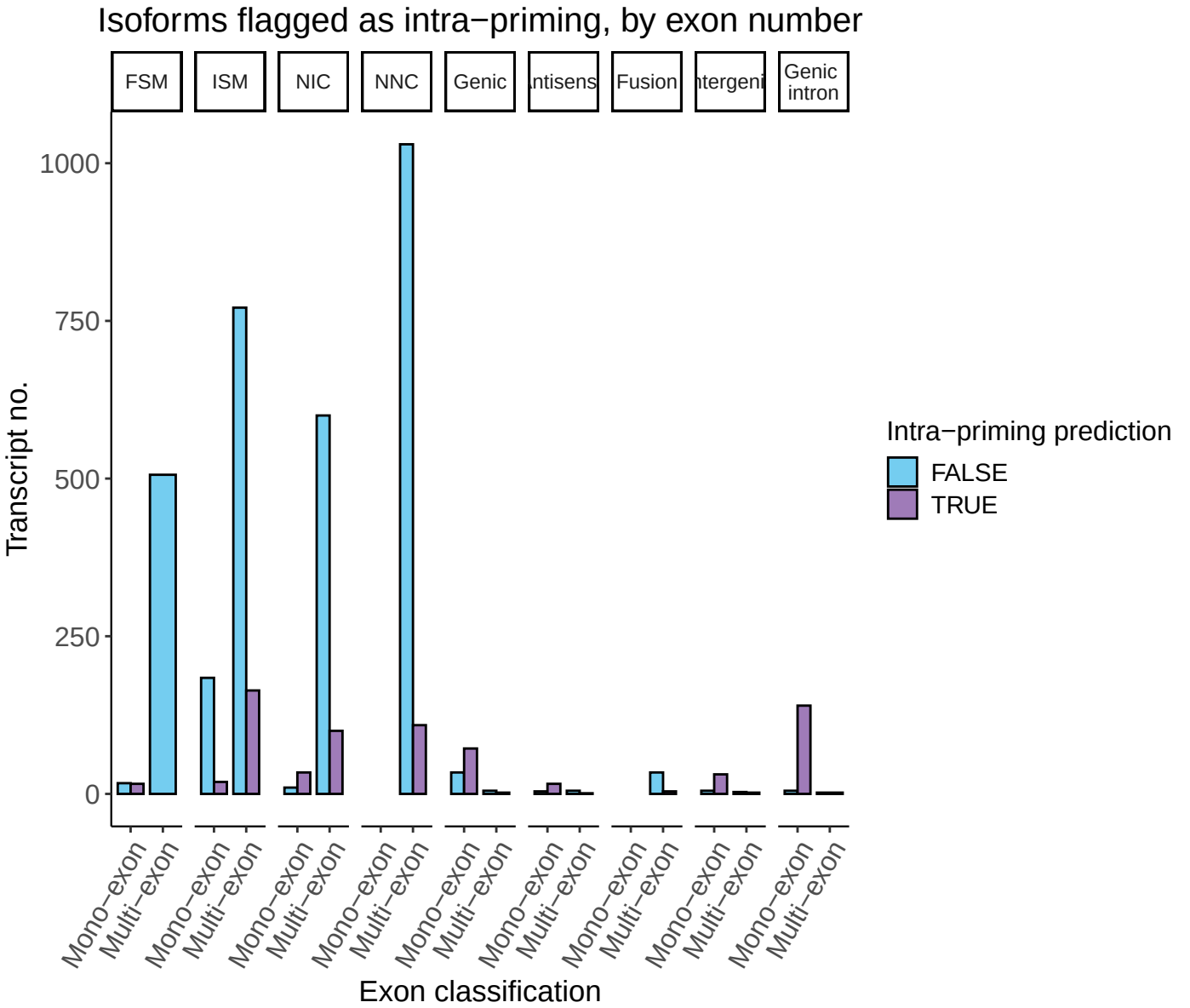

coding − ML importance: 0.33
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Genic intron
6 487
234
344 23
947 98
29 5
6 1
25 3
9 5
24 5
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
coding
coding
non_coding

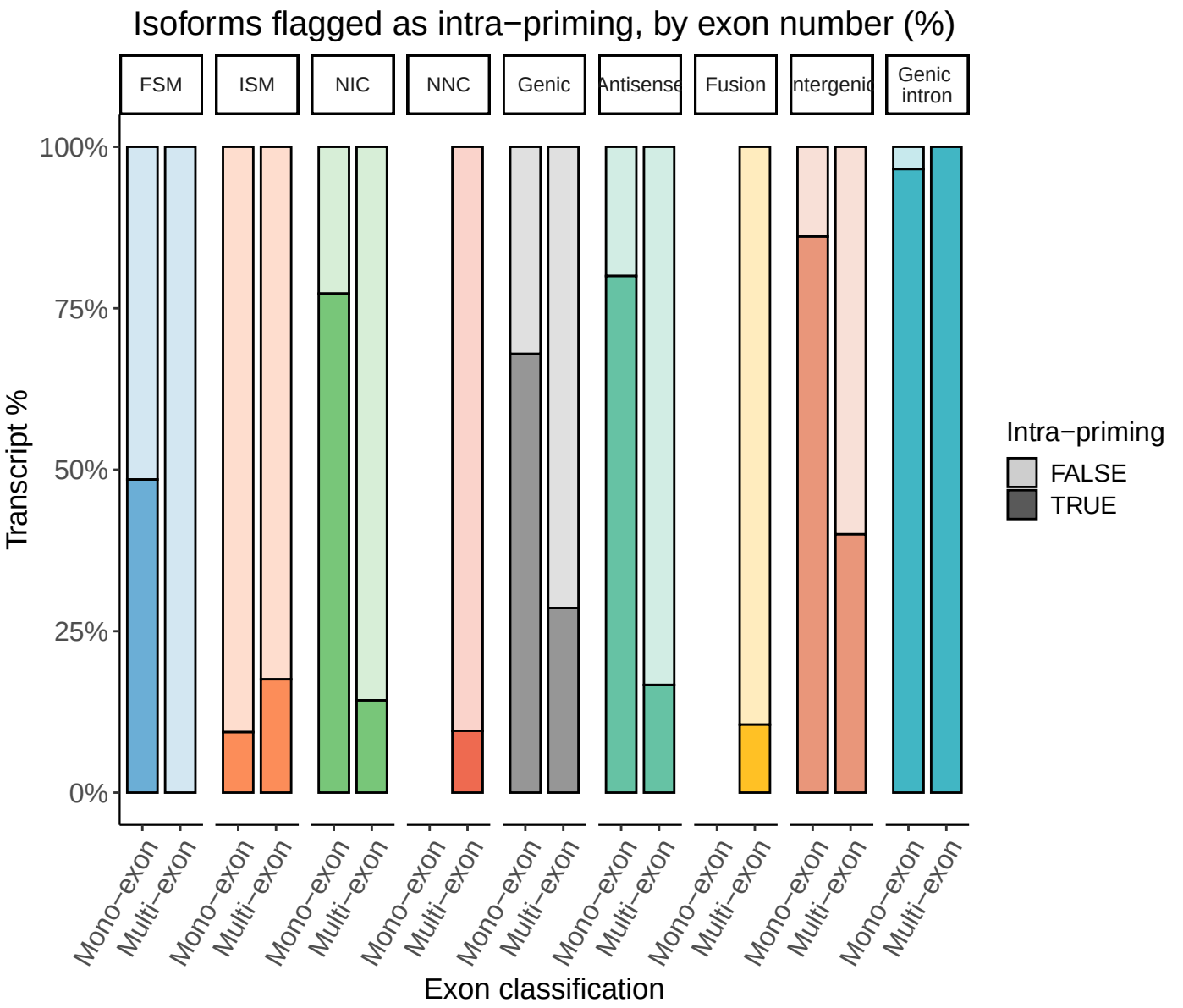
Intra−primming filter report


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 (%)
Transcript %
100%
75%
50%
25%
0%
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