SQANTI3 filter report

Total Genes: 650

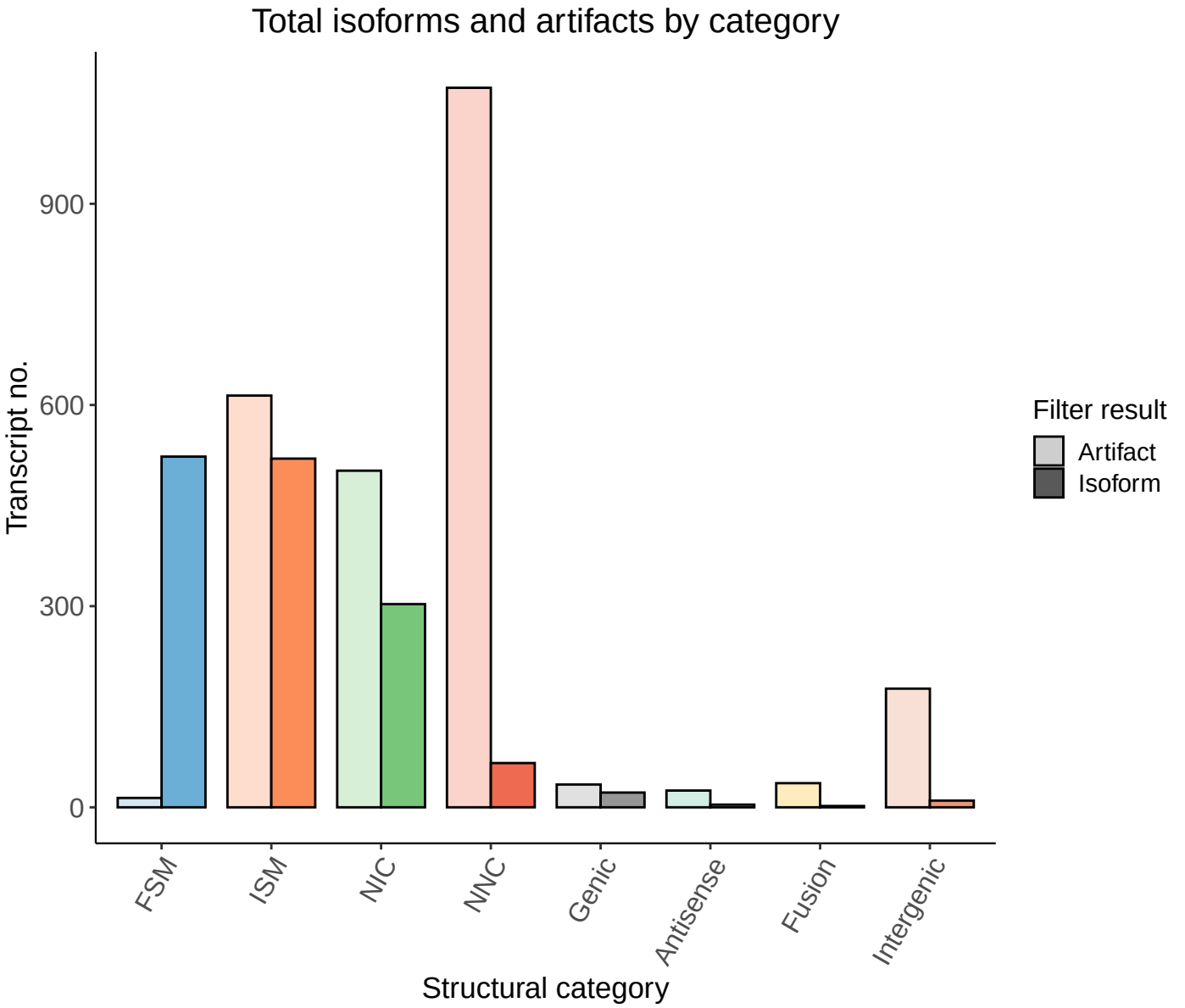
Total Transcripts: 3925

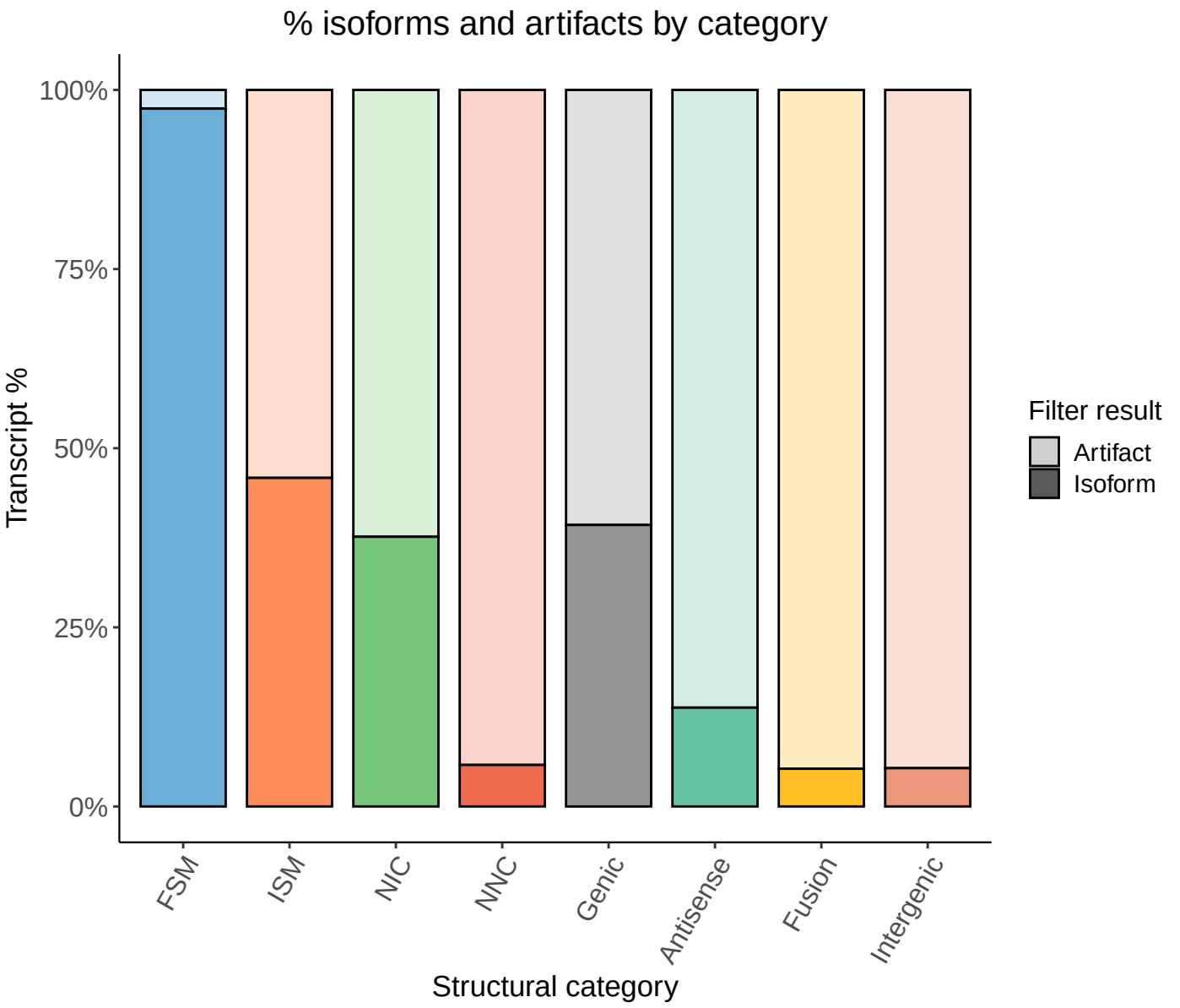
− Isoforms: 1450 (37%)

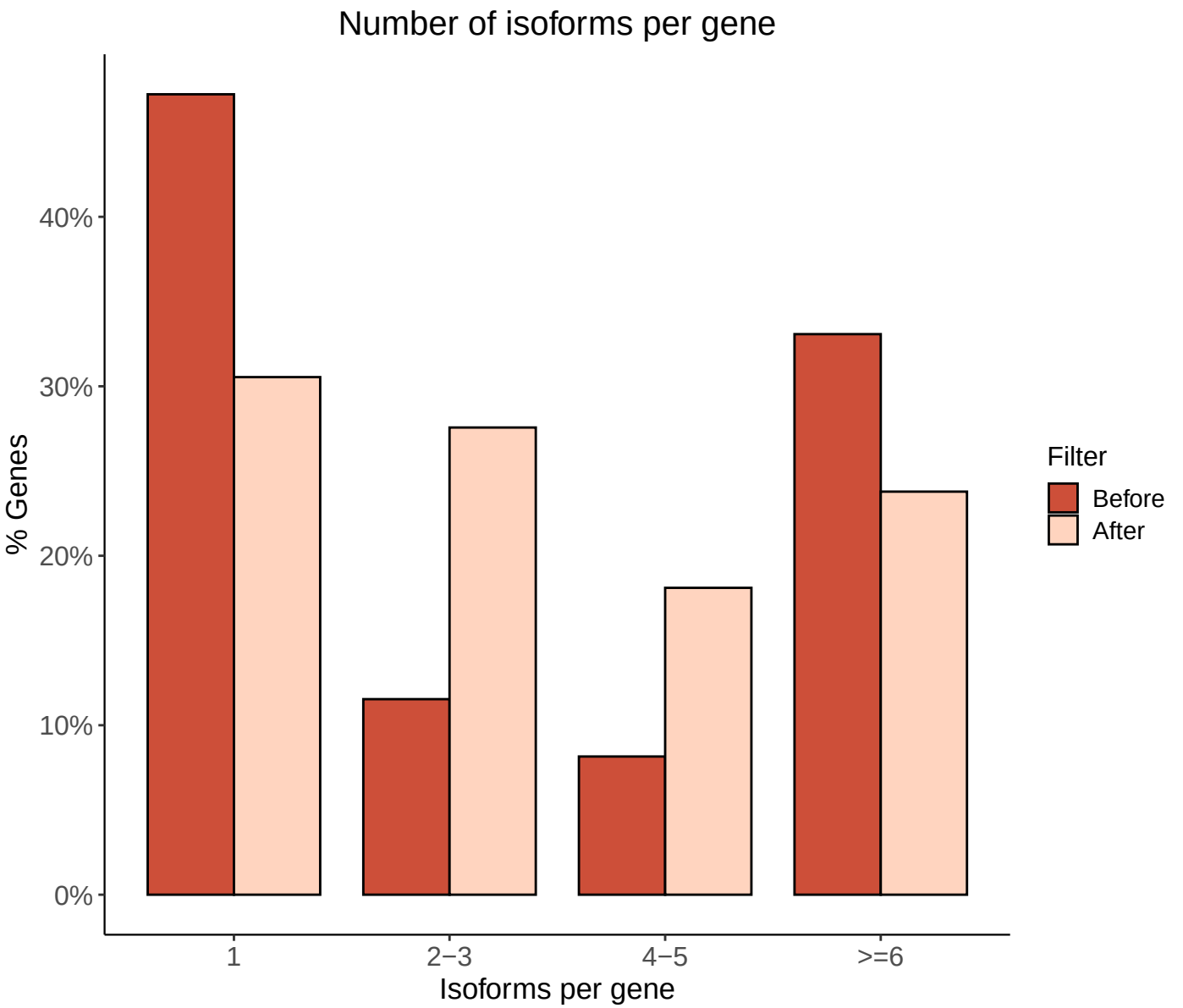
− Artifacts: 2475 (63%)

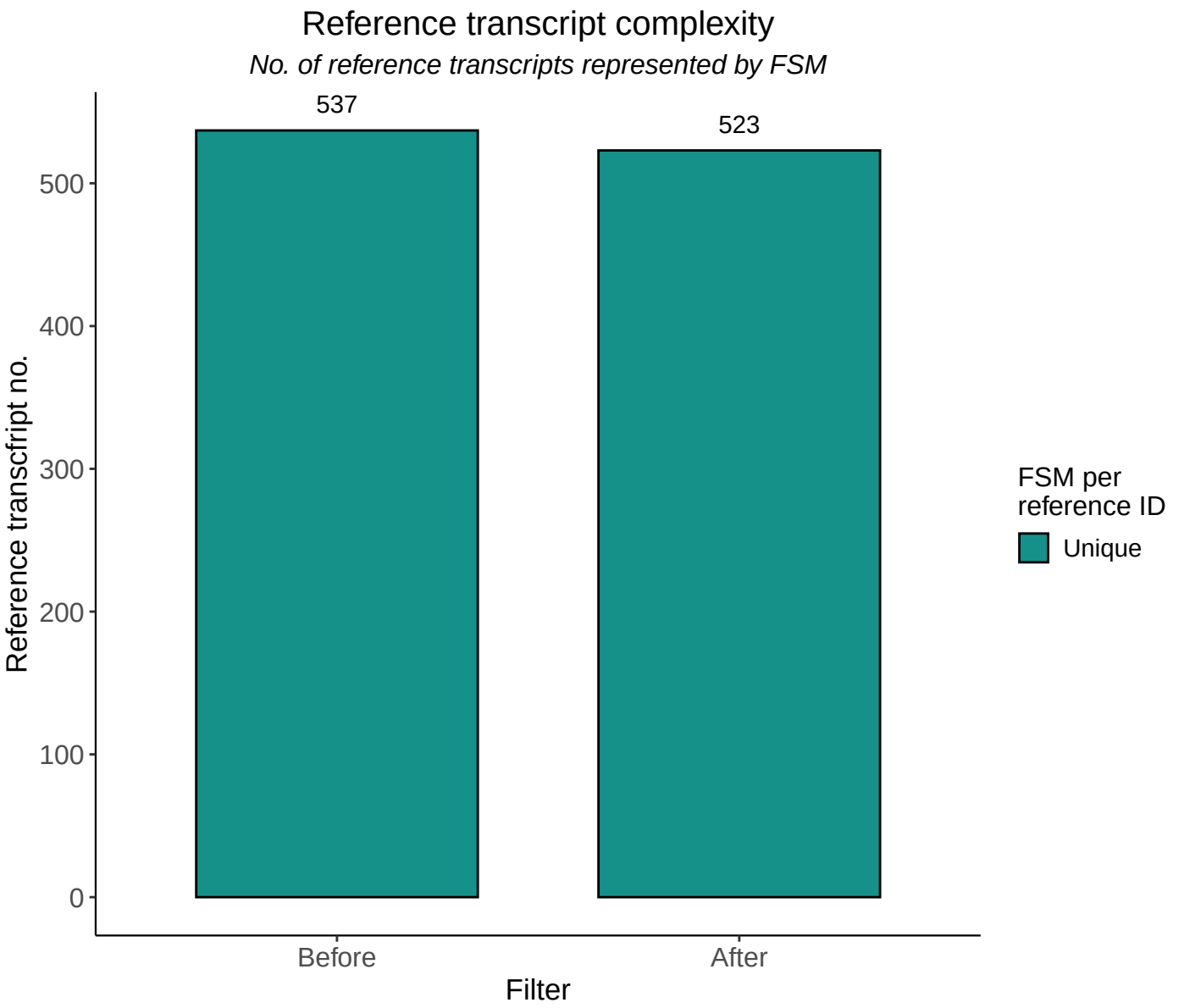
Gene category	Gene no.	No. of genes with artifacts only
Annotated	437	81
Novel	213	199

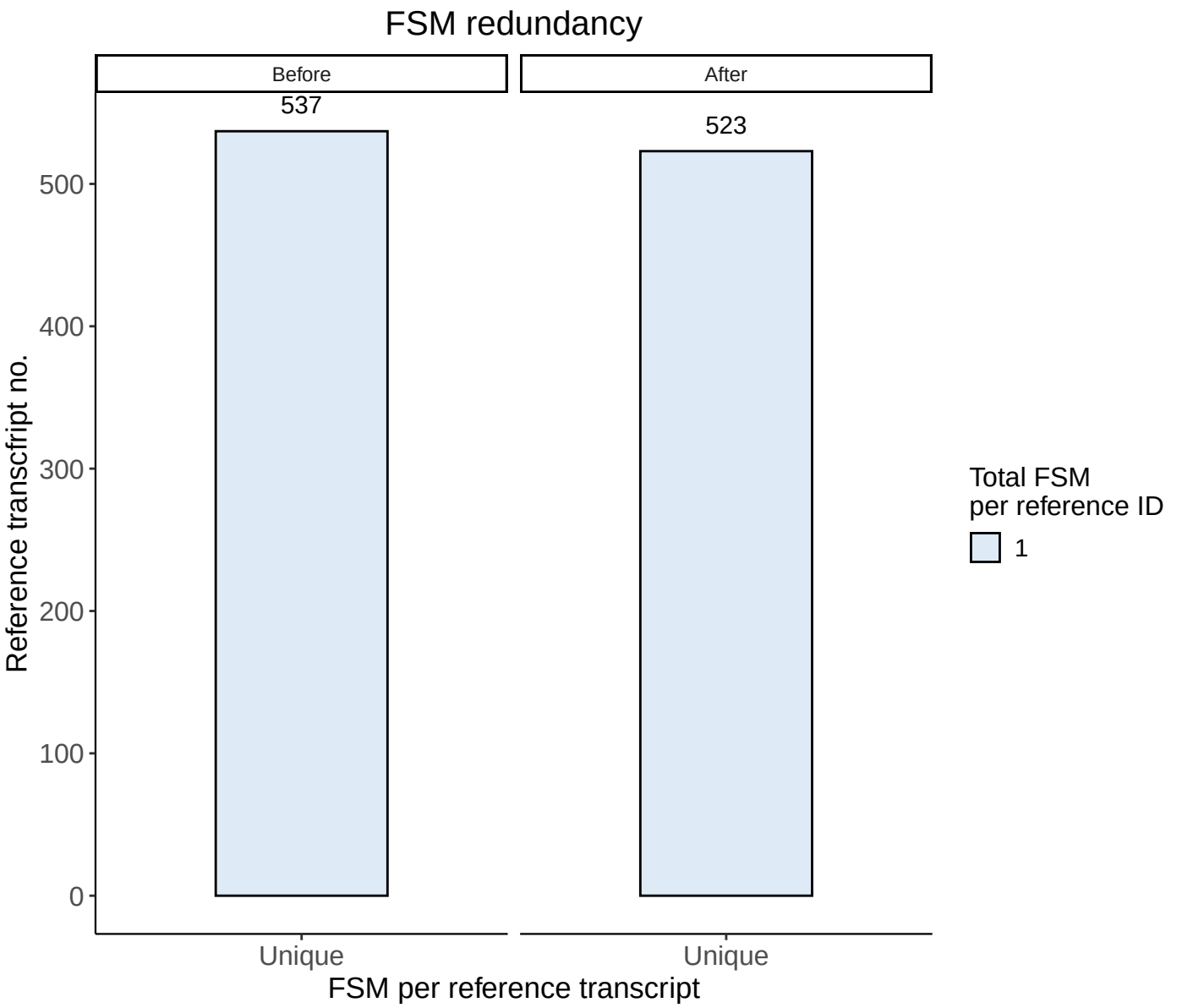
Structural category	Artifact no.	Isoform no.
FSM	14	523
ISM	614	520
NIC	502	303
NNC	1073	66
Genic	34	22
Antisense	25	4
Fusion	36	2
Intergenic	177	10

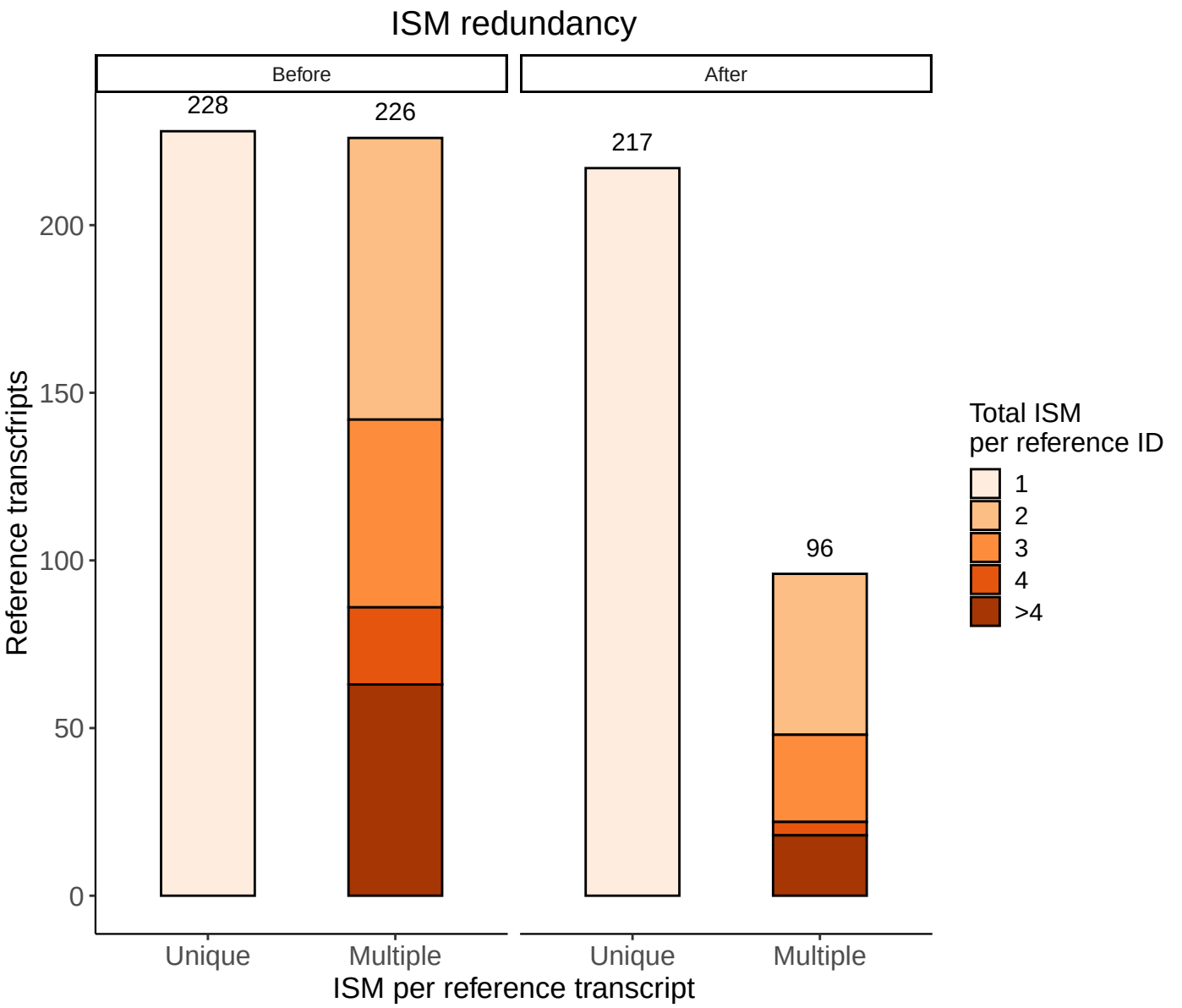

Total isoforms and artifacts by category
Transcript no.
900
600
300
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

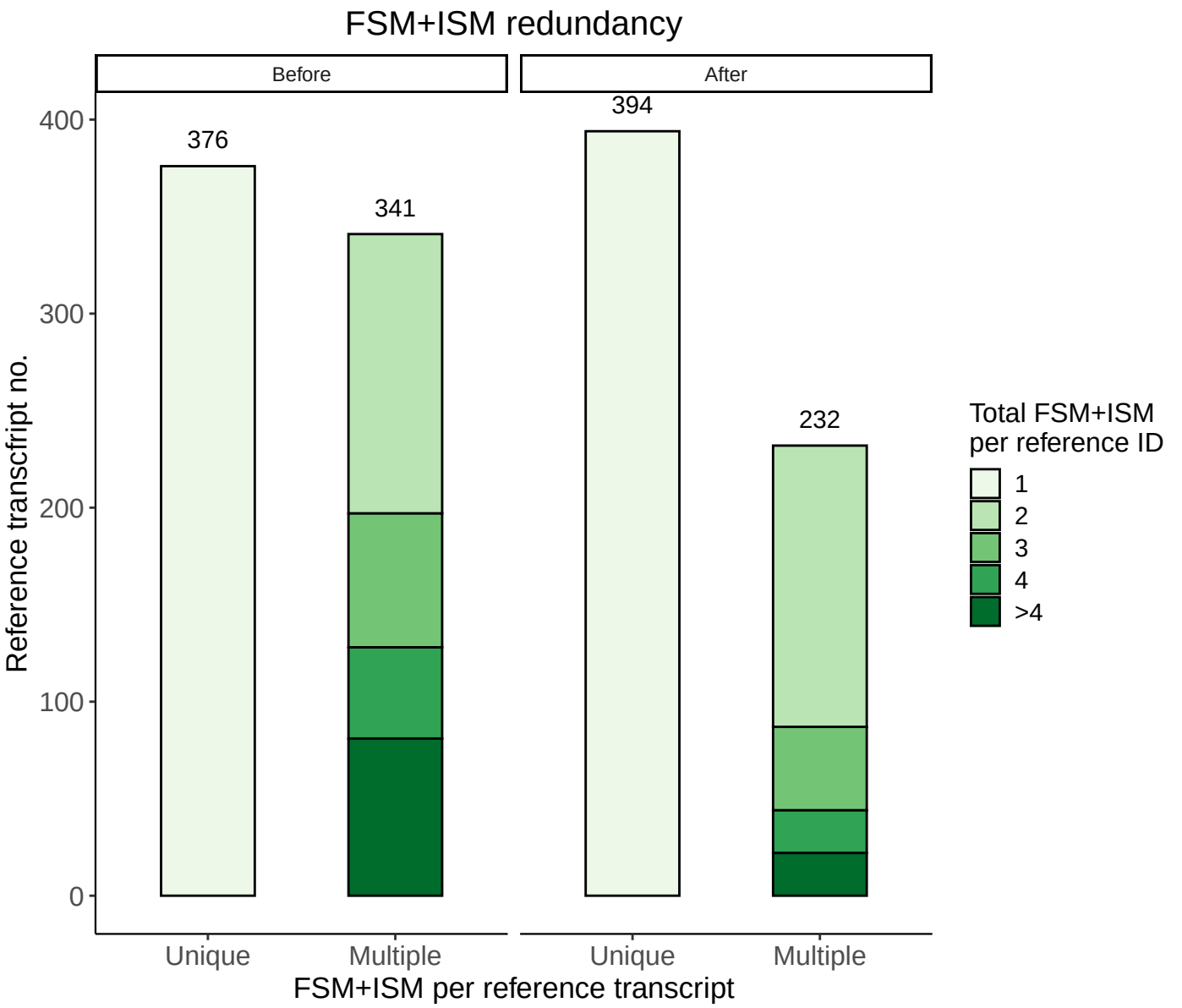

% isoforms and artifacts by category
100%
75%
50%
25%
0%
Transcript %
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform


Number of isoforms per gene
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
537
523
500
400
300
200
100
0
Reference transcfript no.
Before
After
Filter
FSM per reference ID
Unique


FSM redundancy
Before
After
537
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
228
226
217
96
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
341
394
232
400
300
200
100
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

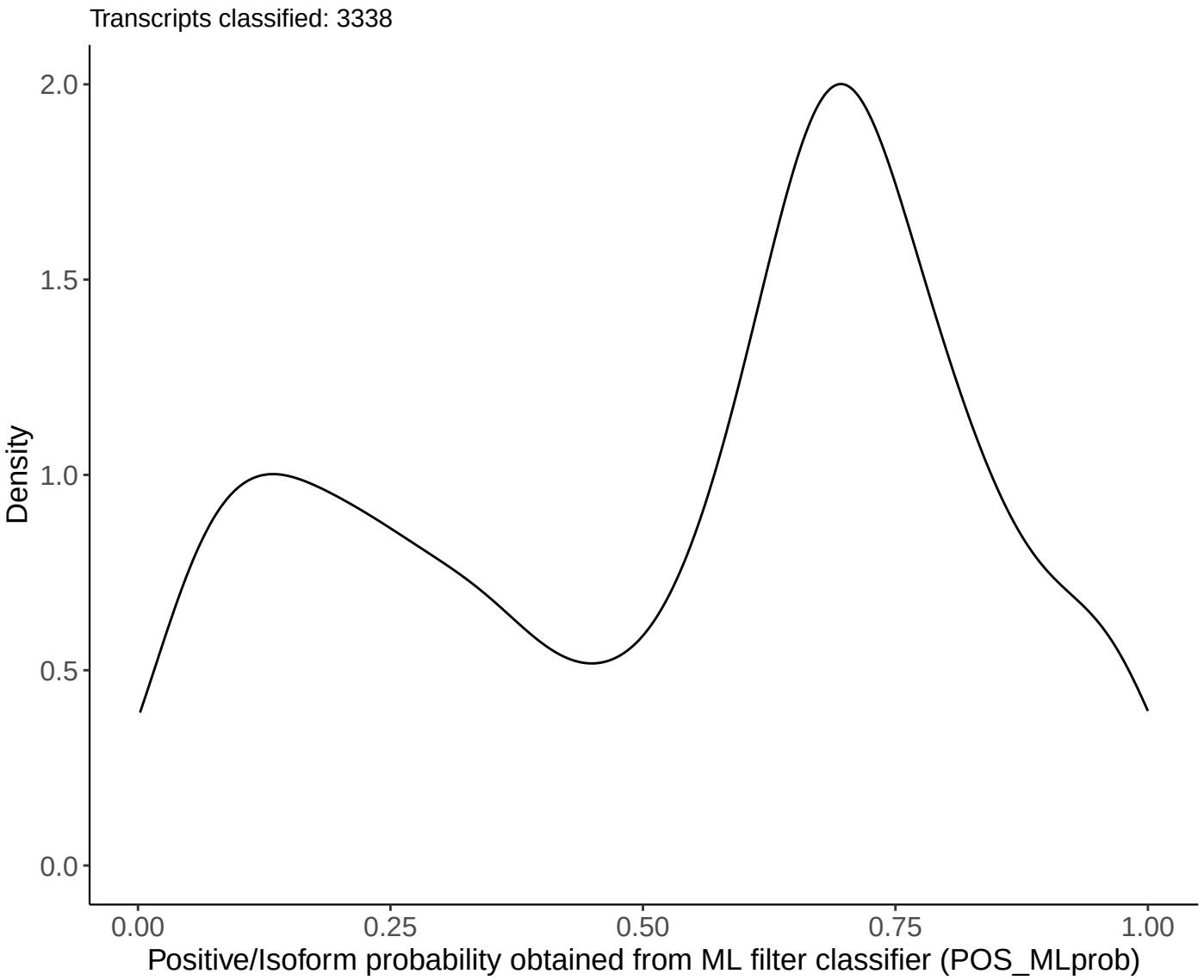
Performance metrics

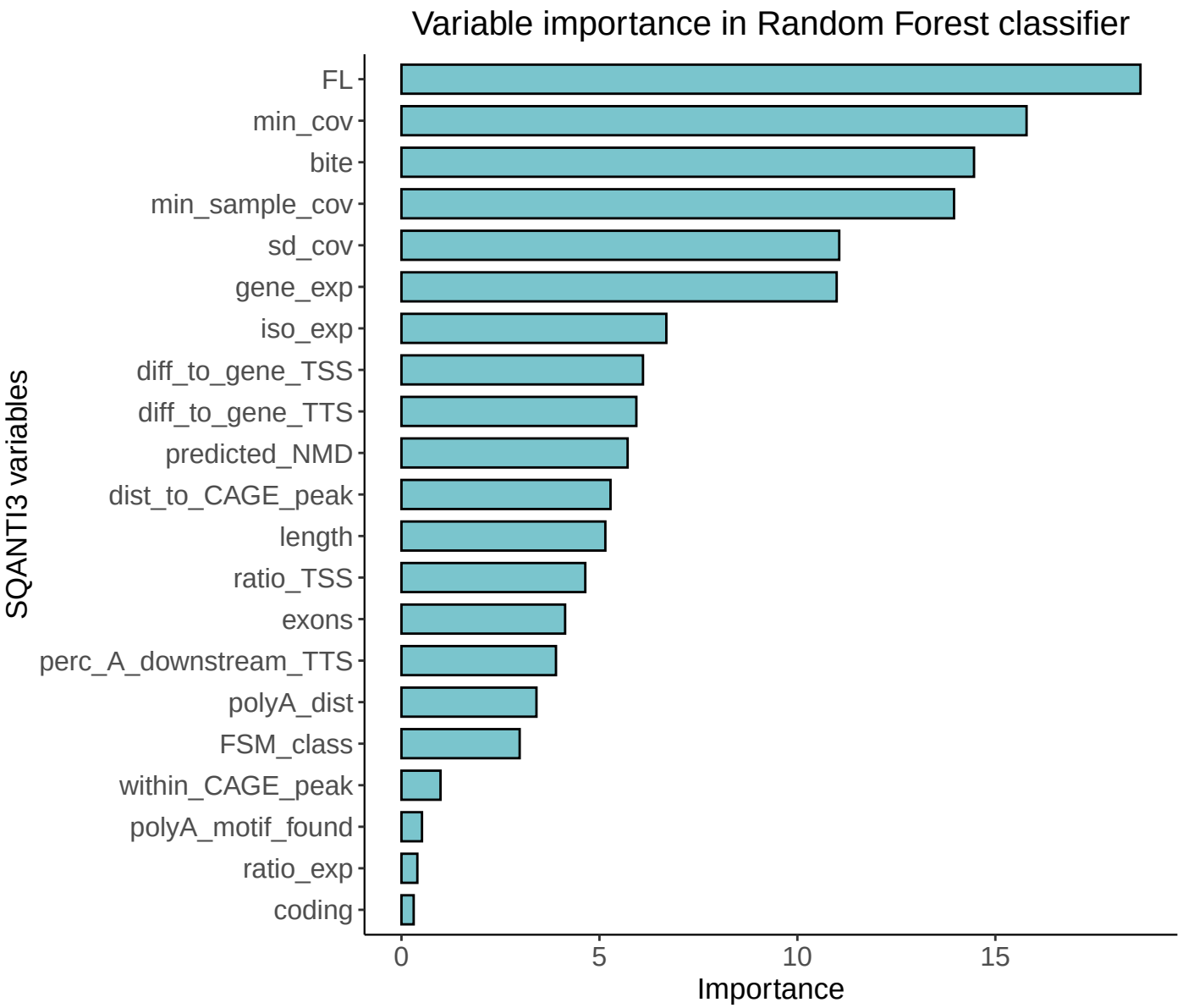
Metric	Value
Accuracy	0.959
Kappa	0.919
AccuracyLower	0.886
AccuracyUpper	0.992
AccuracyNull	0.5
McnemarPValue	0.248
Sensitivity	0.919
Specificity	1
Pos Pred Value	1
Neg Pred Value	0.925
Precision	1
Recall	0.919
F1	0.958
Prevalence	0.5
Detection Rate	0.459
Detection Prevalence	0.459
Balanced Accuracy	0.959

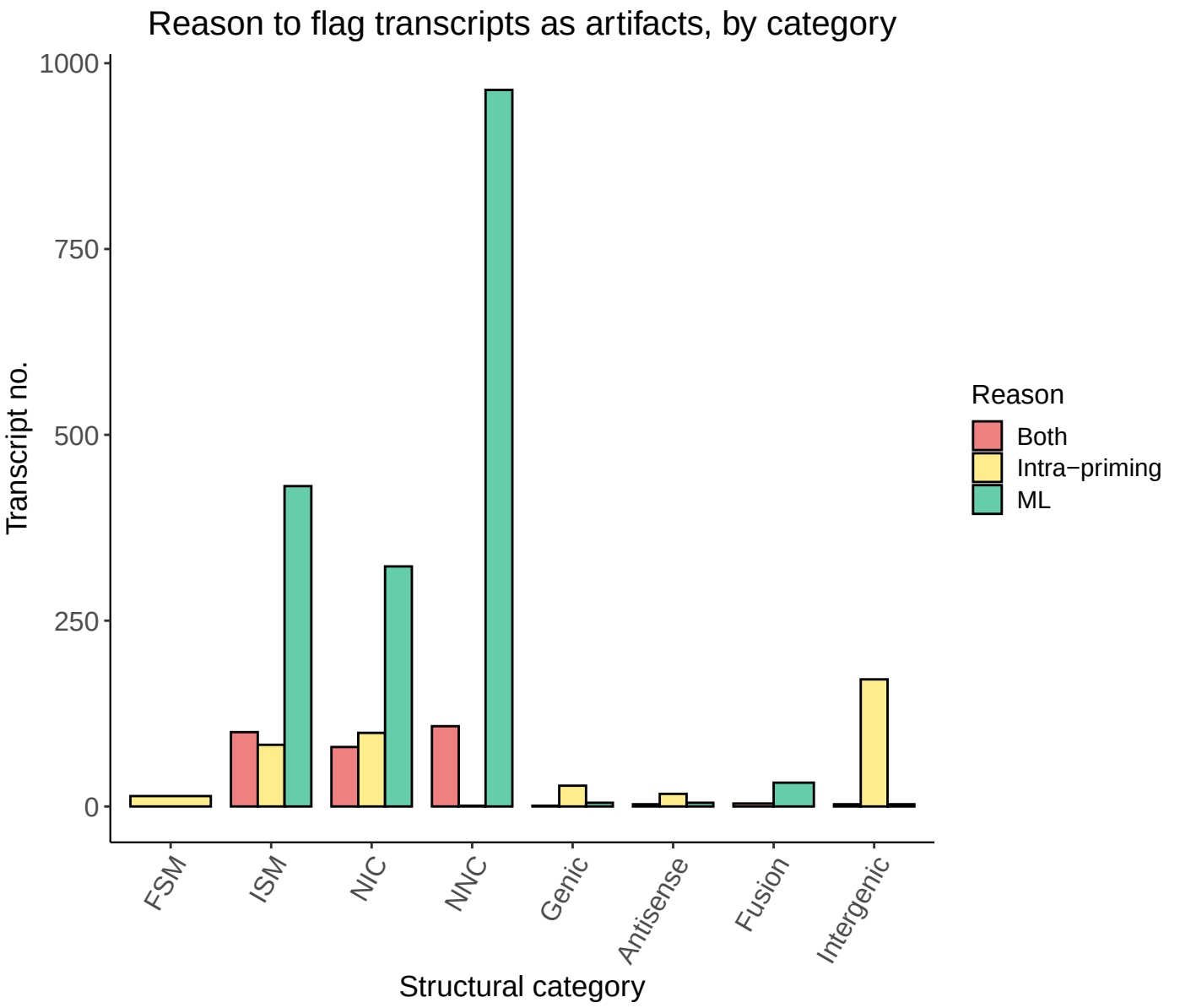
AccuracyPValue	3.578714e−18

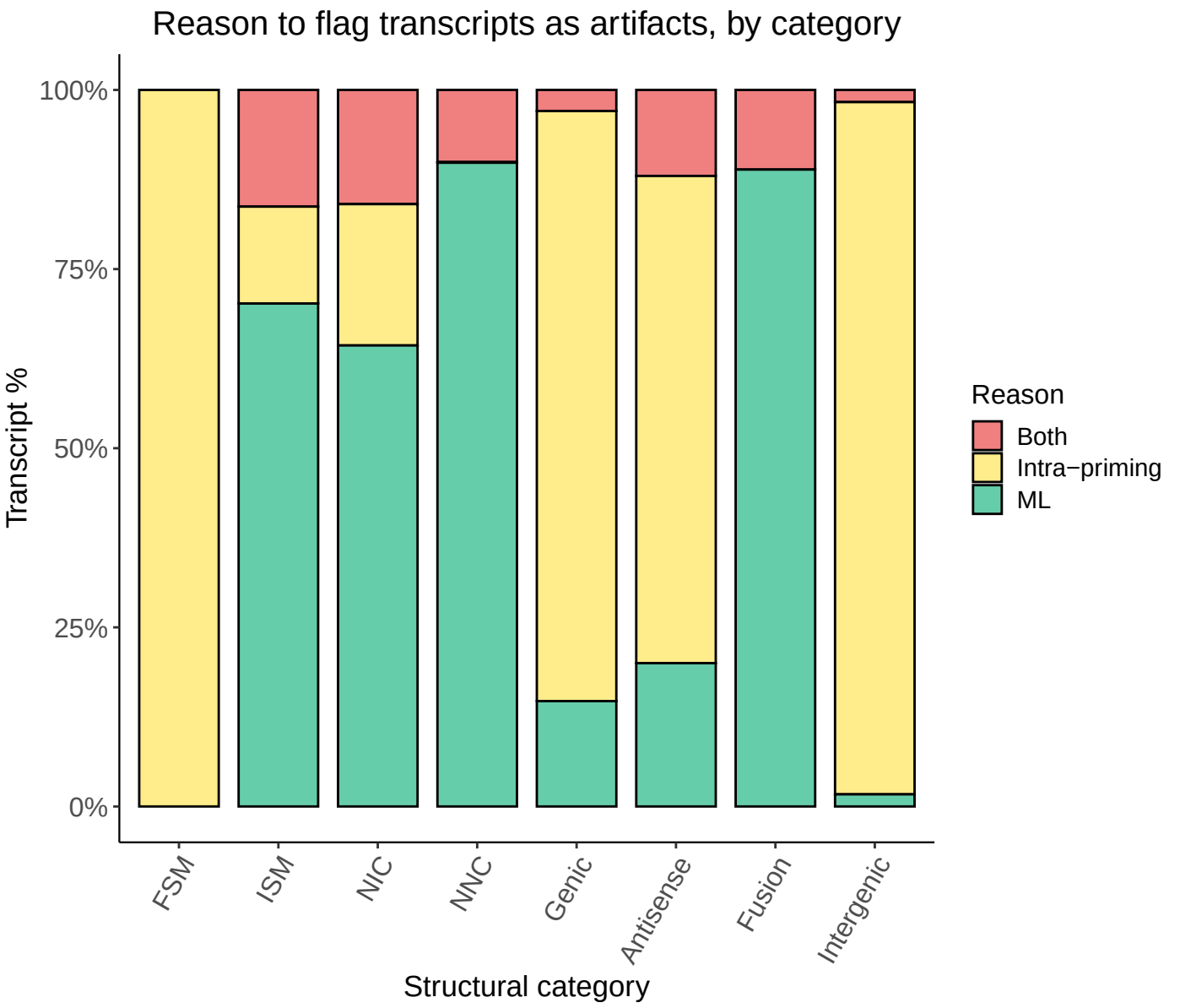
Confusion matrix

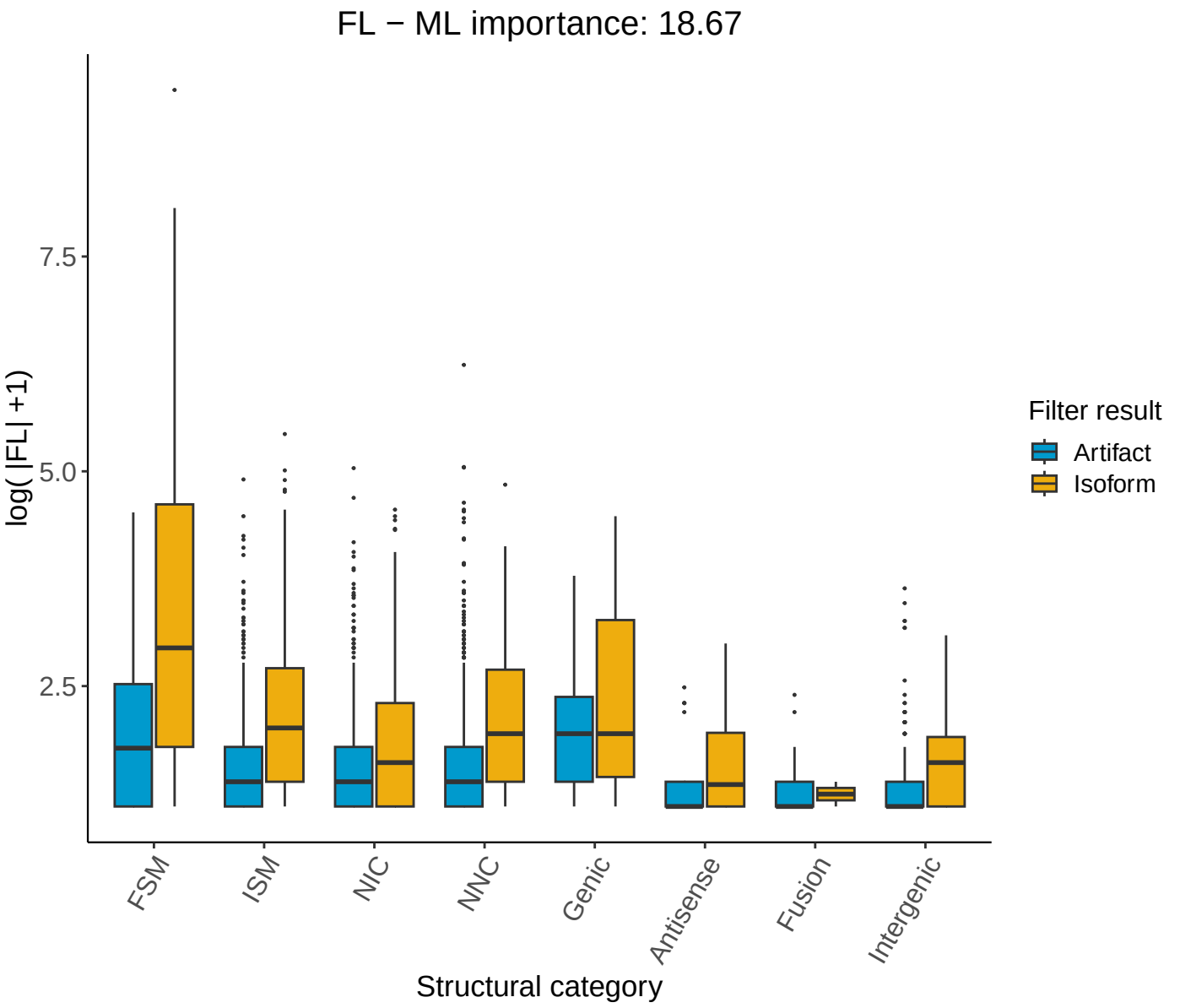
Prediction	Reference	Freq
POS	POS	34
NEG	POS	3
POS	NEG	0
NEG	NEG	37

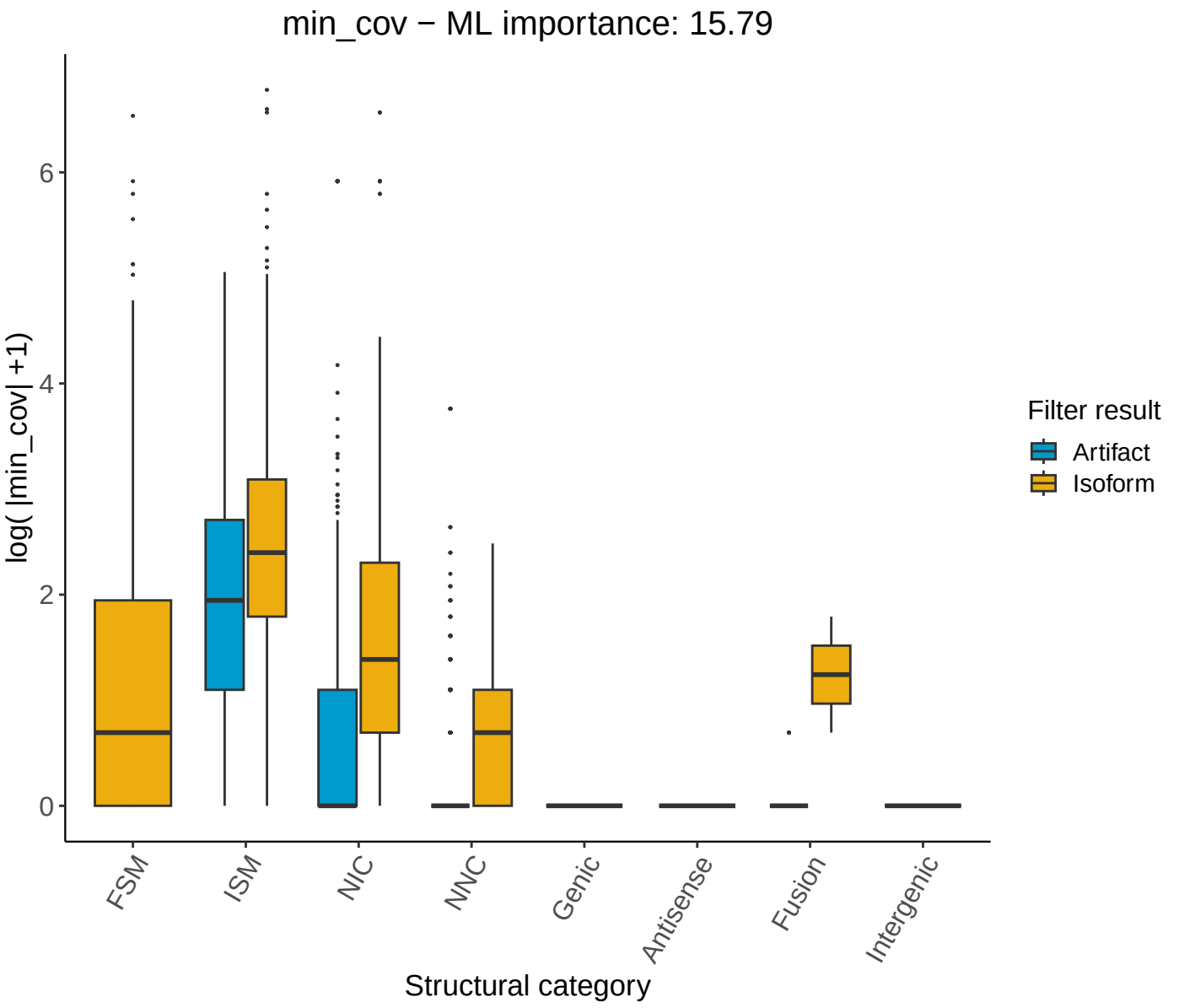

Transcripts classified: 3338
2.0
1.5
1.0
0.5
0.0
Density
0.00
0.25
0.50
0.75
1.00
Positive/Isoform probability obtained from ML filter classifier (POS_MLprob)

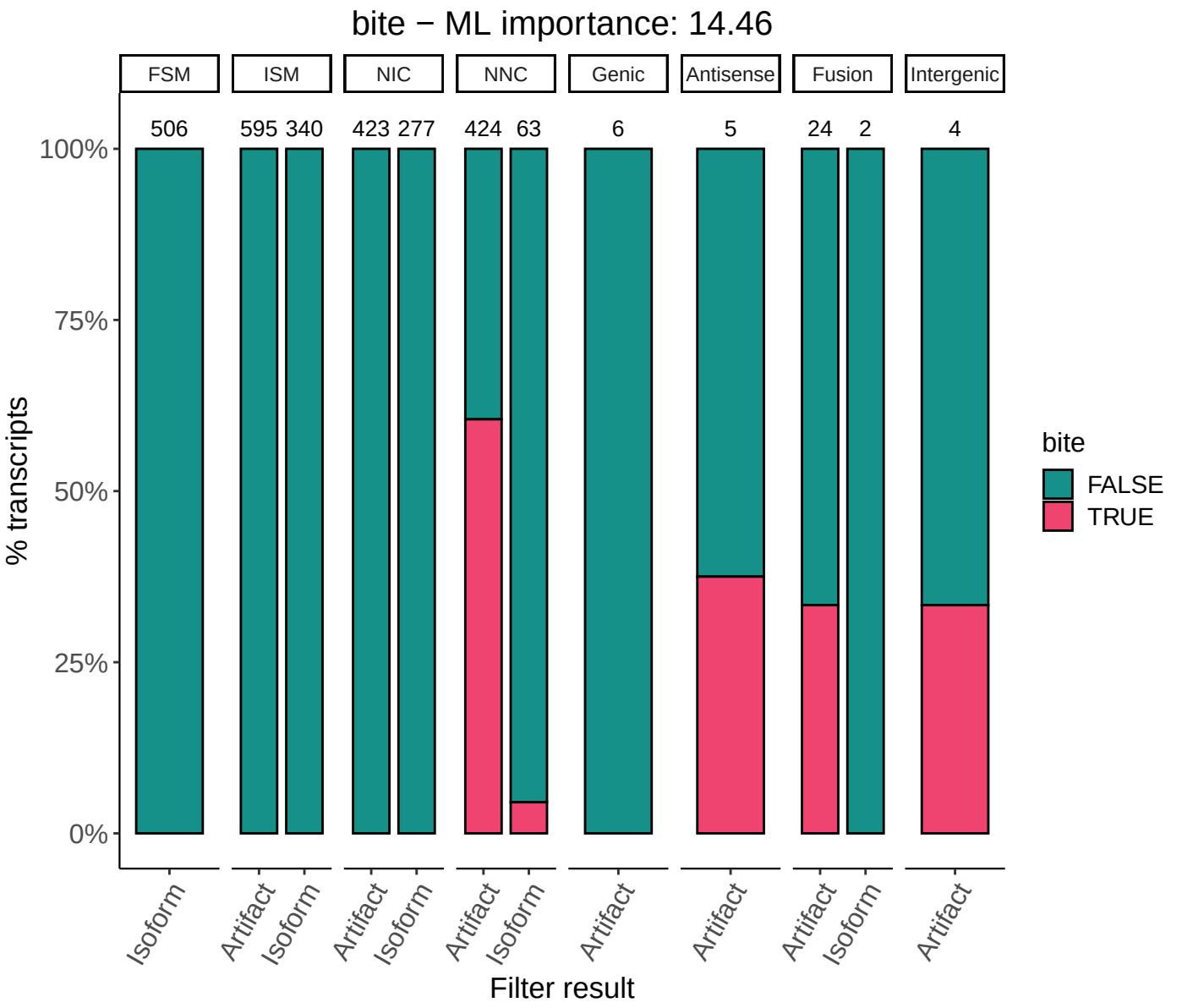

Variable importance in Random Forest classifier
SQANTI3 variables
FL
min_cov
bite
min_sample_cov
sd_cov
gene_exp
iso_exp
diff_to_gene_TSS
diff_to_gene_TTS
predicted_NMD
dist_to_CAGE_peak
length
ratio_TSS
exons
perc_A_downstream_TTS
polyA_dist
FSM_class
within_CAGE_peak
polyA_motif_found
ratio_exp
coding
0
5
10
15
Importance

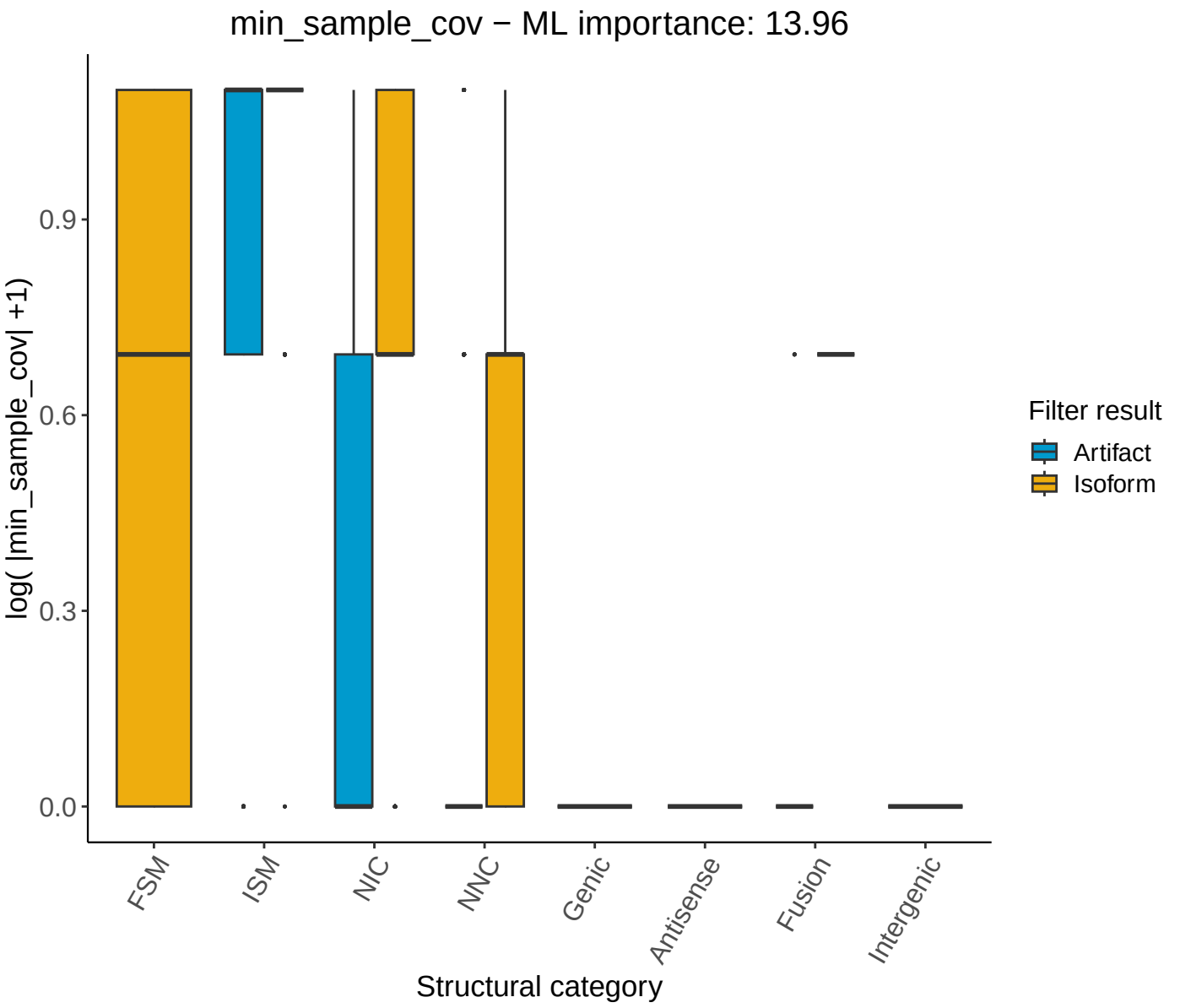

Reason to flag transcripts as artifacts, by category
Transcript no.
1000
750
500
250
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Reason
Both
Intra−priming
ML

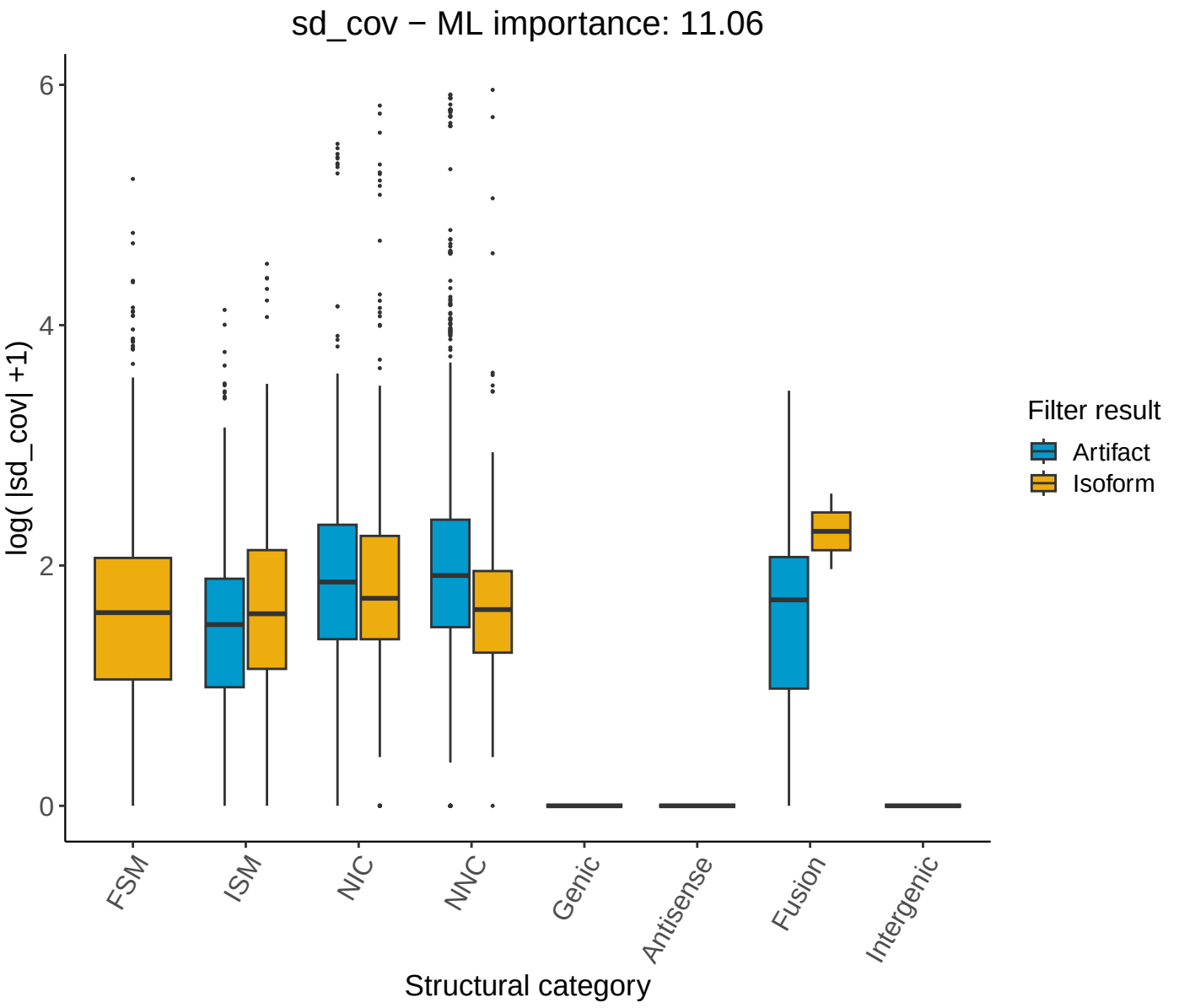

Reason to flag transcripts as artifacts, by category
100%
75%
50%
25%
0%
Transcript %
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Reason
Both
Intra−priming
ML

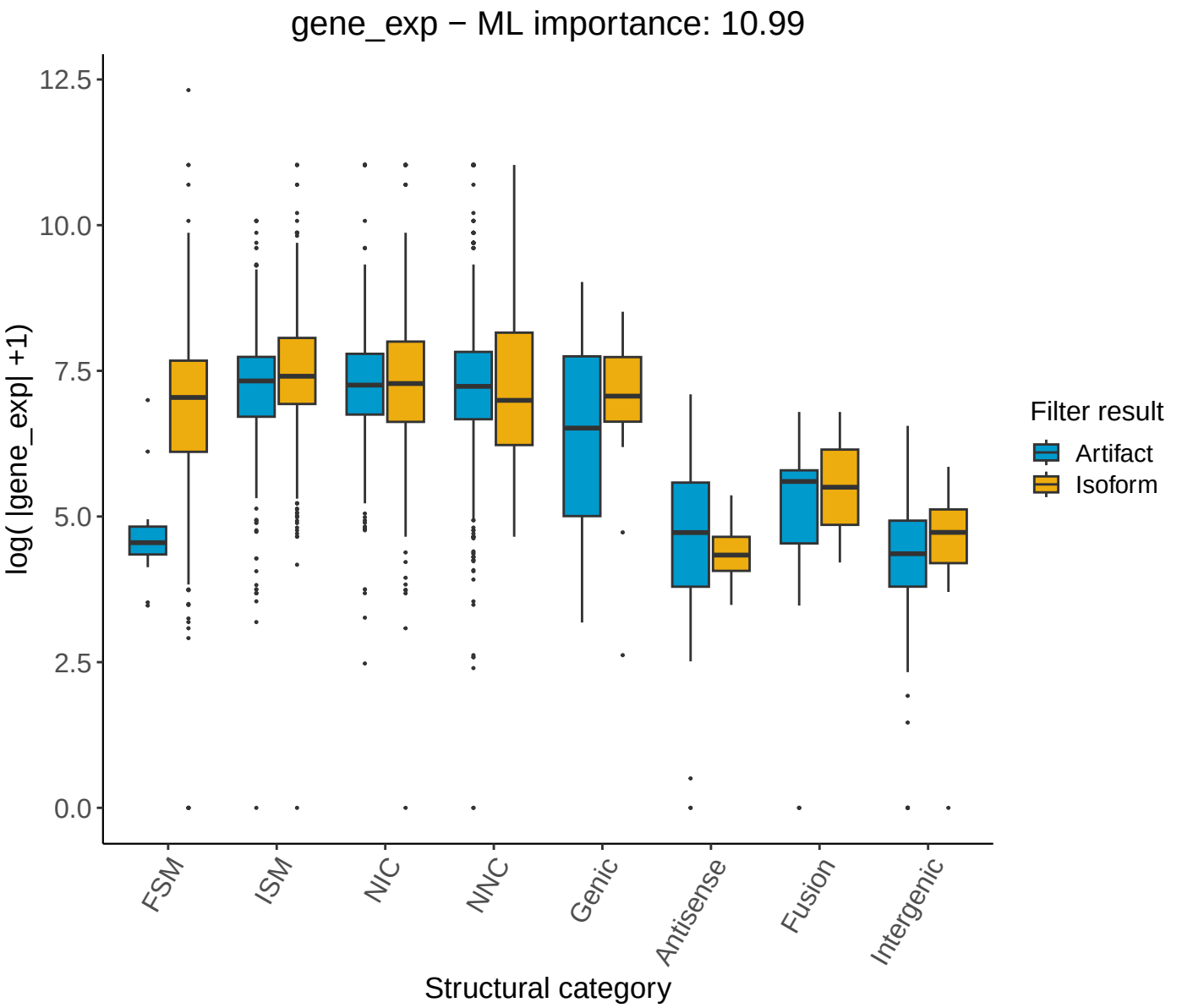

FL − ML importance: 18.67
log(|FL| +1)
7.5
5.0
2.5
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

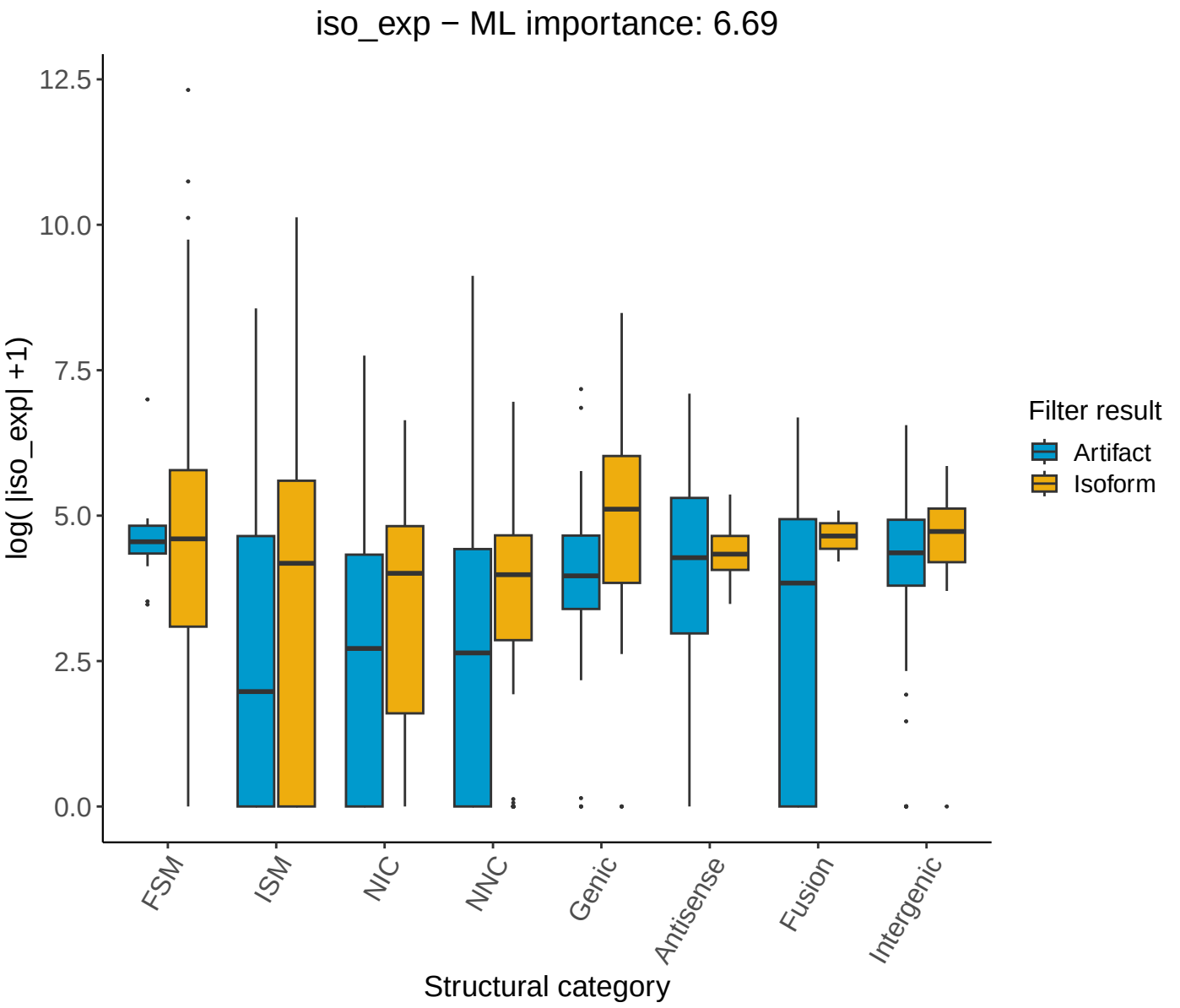

min_cov − ML importance: 15.79
log(|min_cov| +1)
0
2
4
6
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

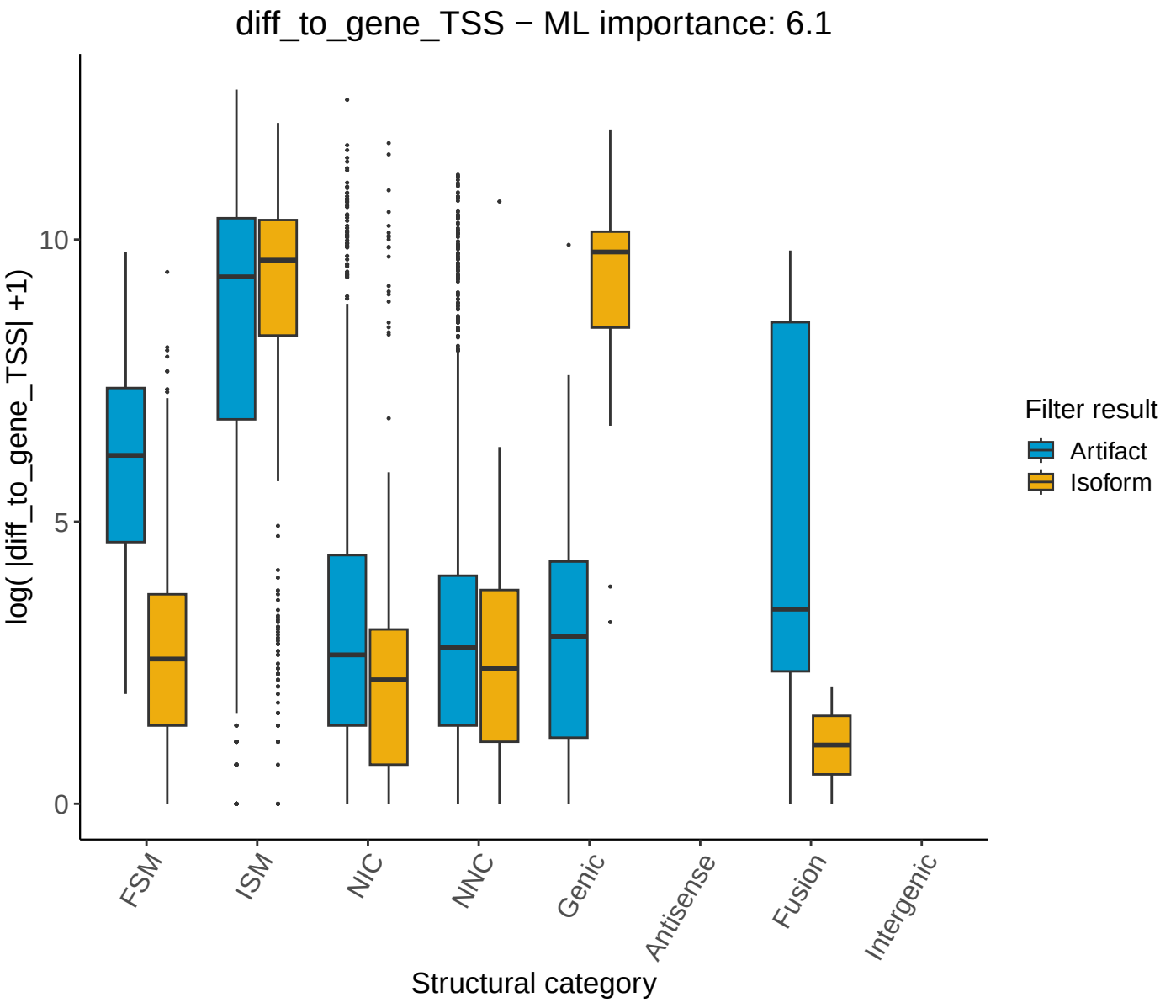

bite − ML importance: 14.46
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
506
595
340
423
277
424
63
6
5
24
2
4
100%
75%
50%
25%
0%
% transcripts
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Artifact
Artifact
Isoform
Artifact
Filter result
bite
FALSE
TRUE

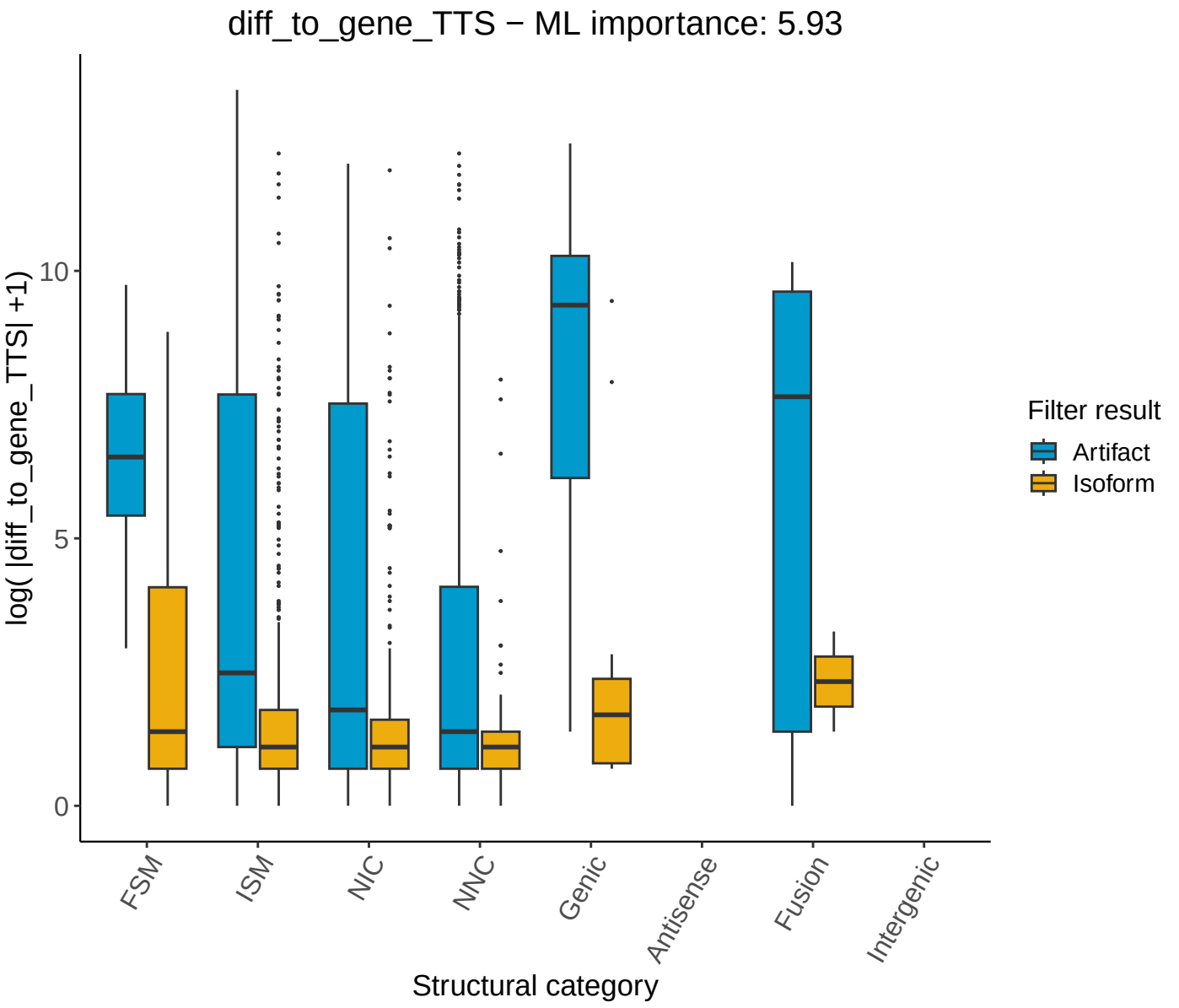

min_sample_cov − ML importance: 13.96
log(|min_sample_cov| +1)
0.0
0.3
0.6
0.9
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

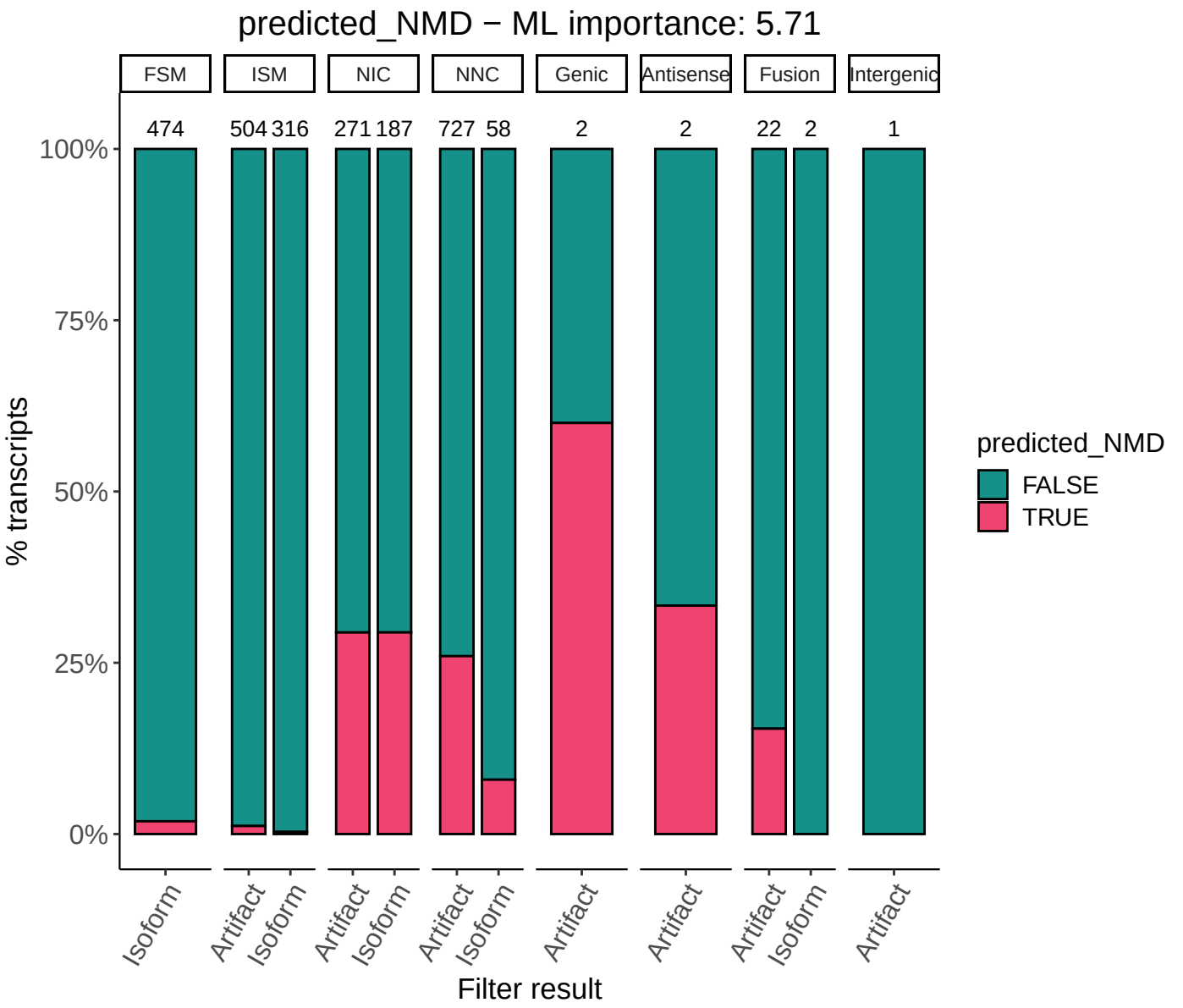

sd_cov − ML importance: 11.06
log(|sd_cov| +1)
6
4
2
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

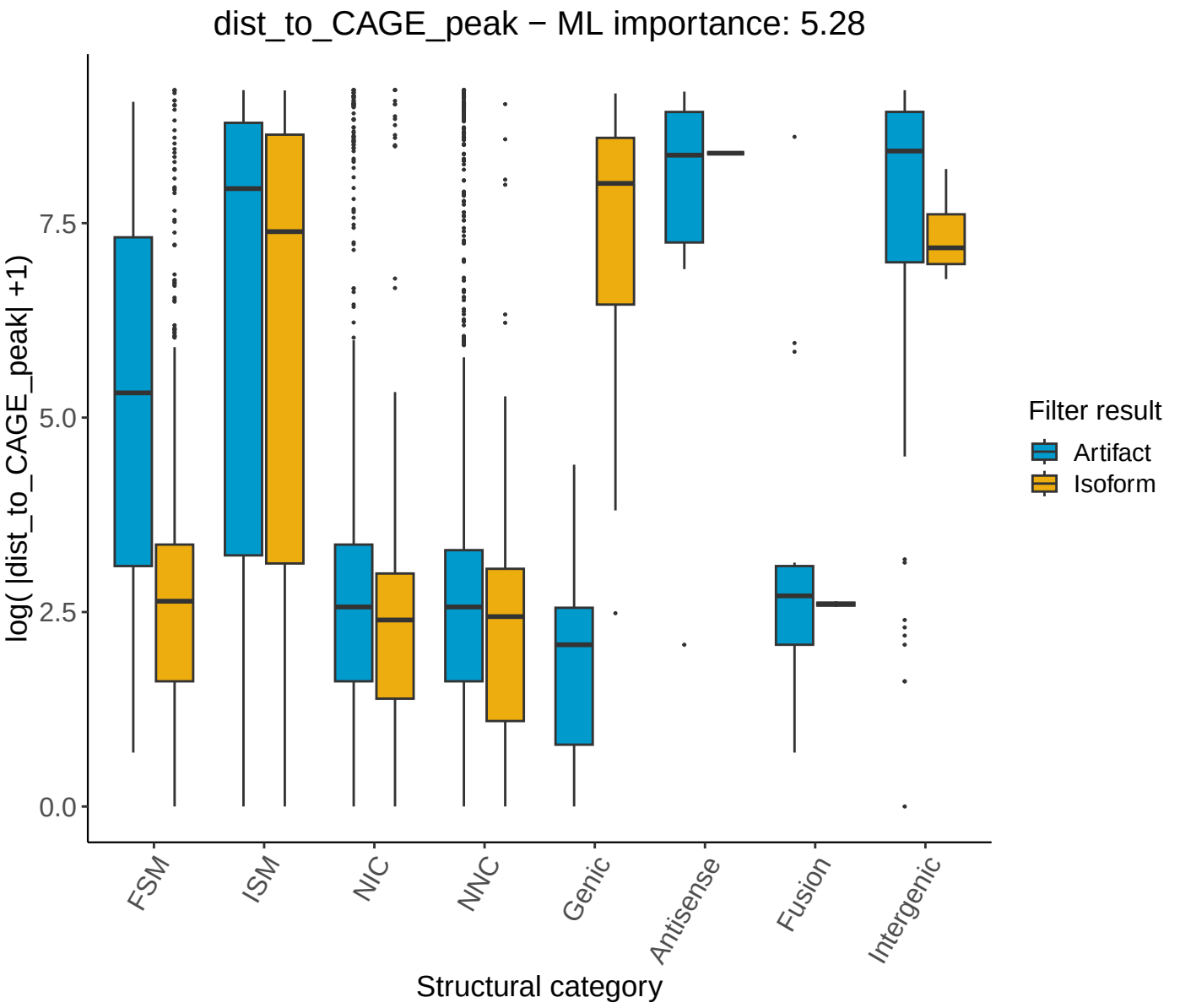

gene_exp − ML importance: 10.99
log(|gene_exp| +1)
12.5
10.0
7.5
5.0
2.5
0.0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

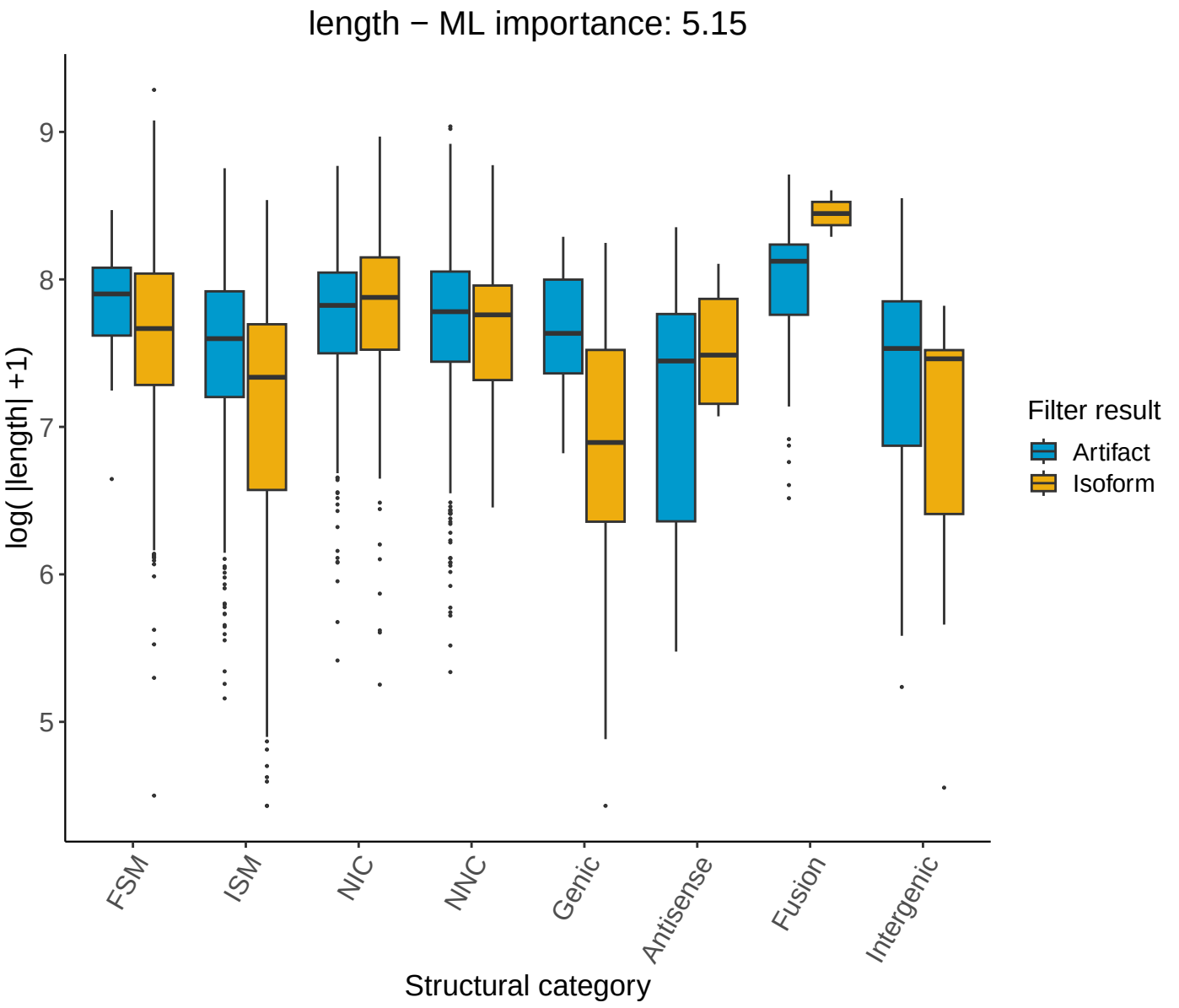

iso_exp − ML importance: 6.69
log(|iso_exp| +1)
12.5
10.0
7.5
5.0
2.5
0.0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

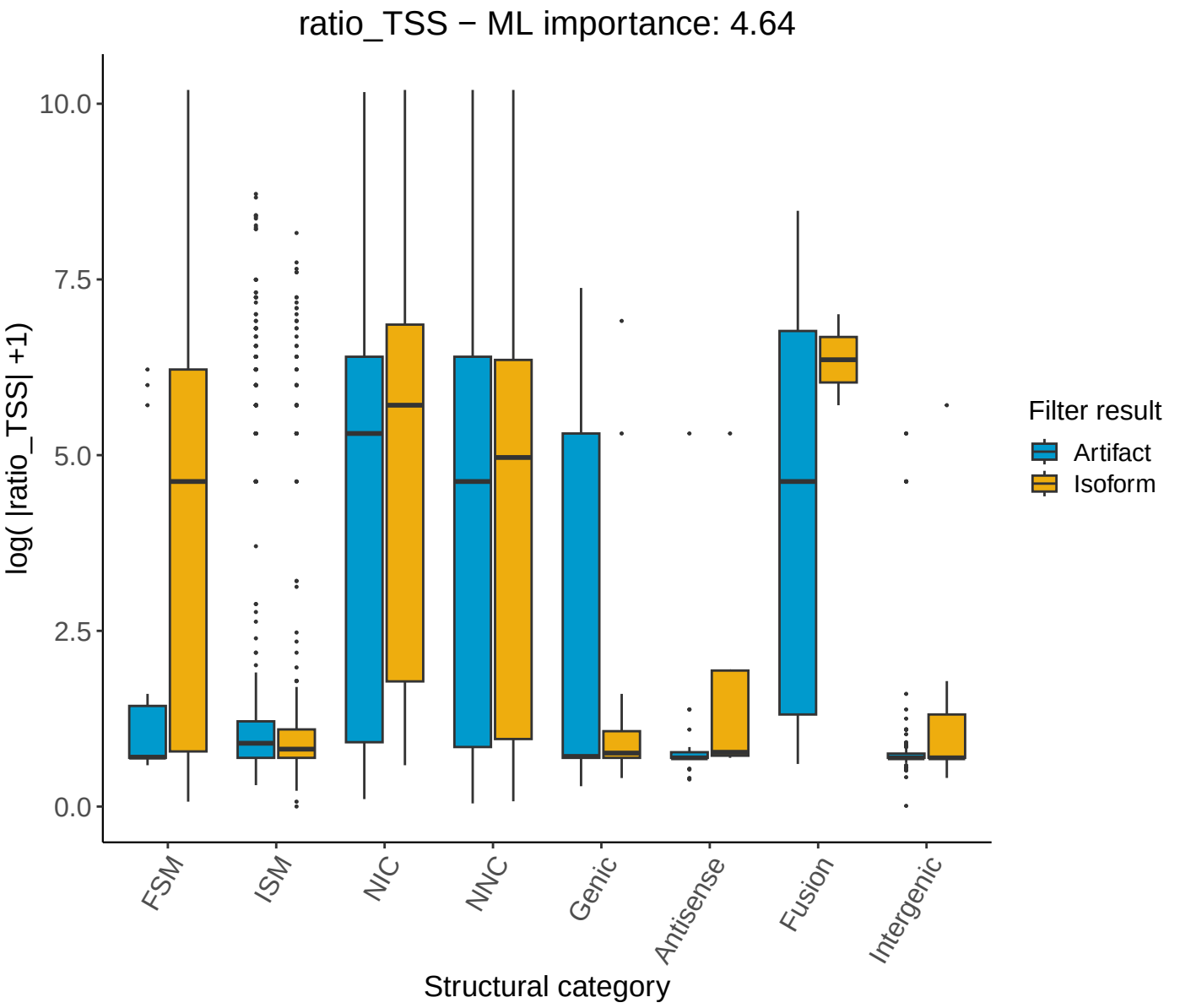

diff_to_gene_TSS − ML importance: 6.1
log(|diff_to_gene_TSS| +1)
10
5
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

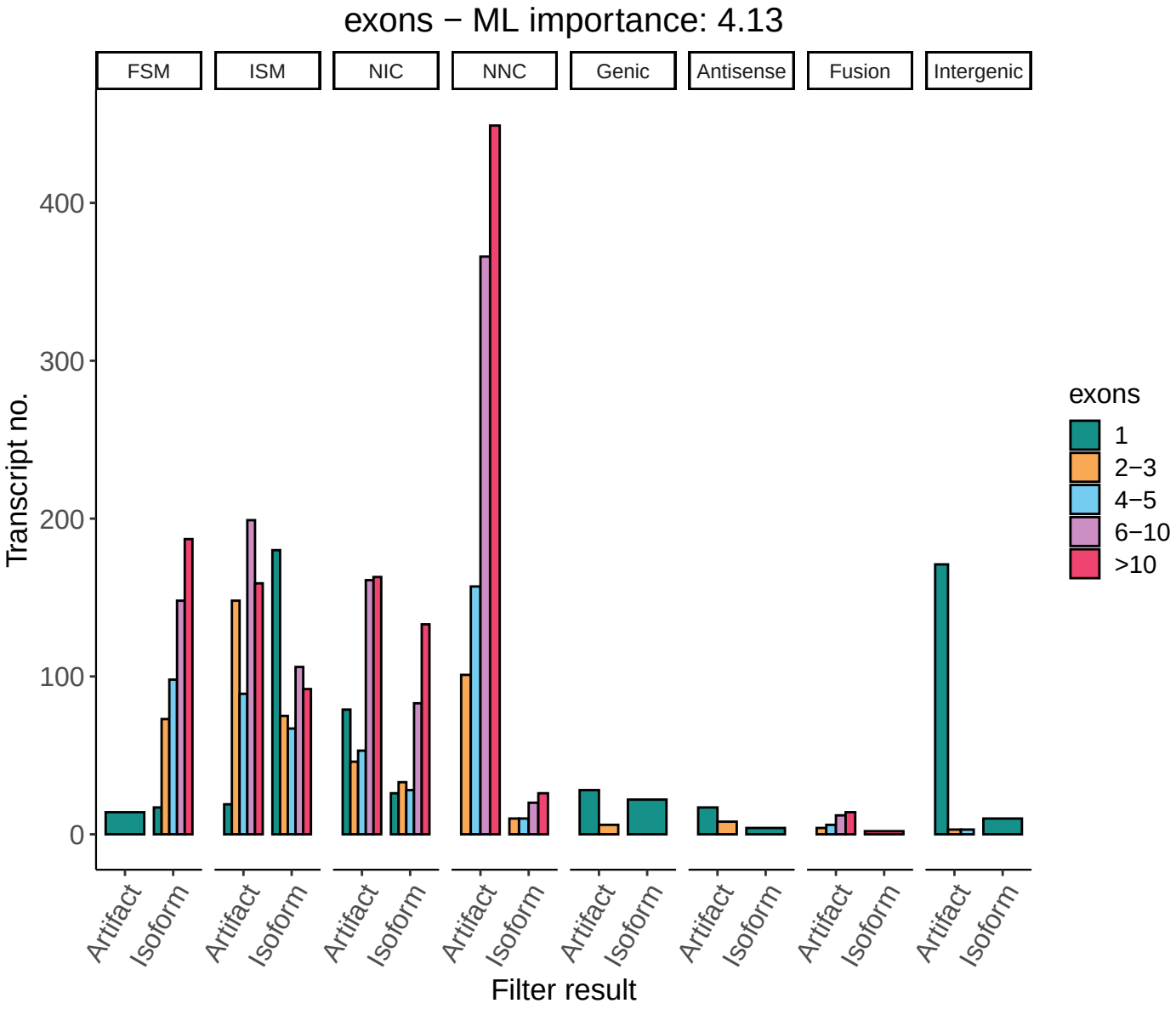

diff_to_gene_TTS − ML importance: 5.93
log(|diff_to_gene_TTS| +1)
10
5
0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

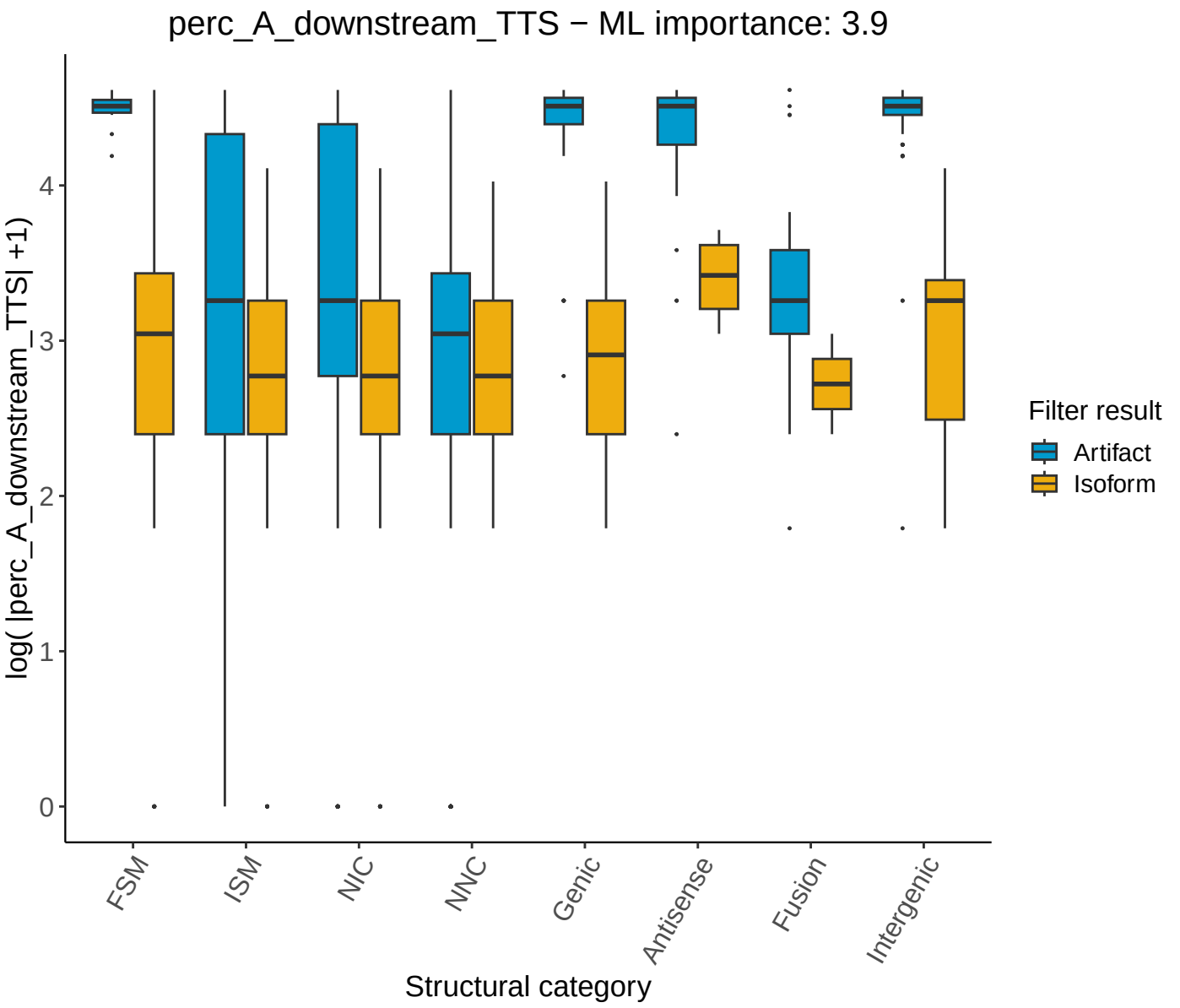

predicted_NMD − ML importance: 5.71
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
474
504
316
271
187
727
58
2
2
22
2
1
100%
75%
50%
25%
0%
% transcripts
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Isoform
Artifact
Artifact
Artifact
Isoform
Artifact
Filter result
predicted_NMD
FALSE
TRUE

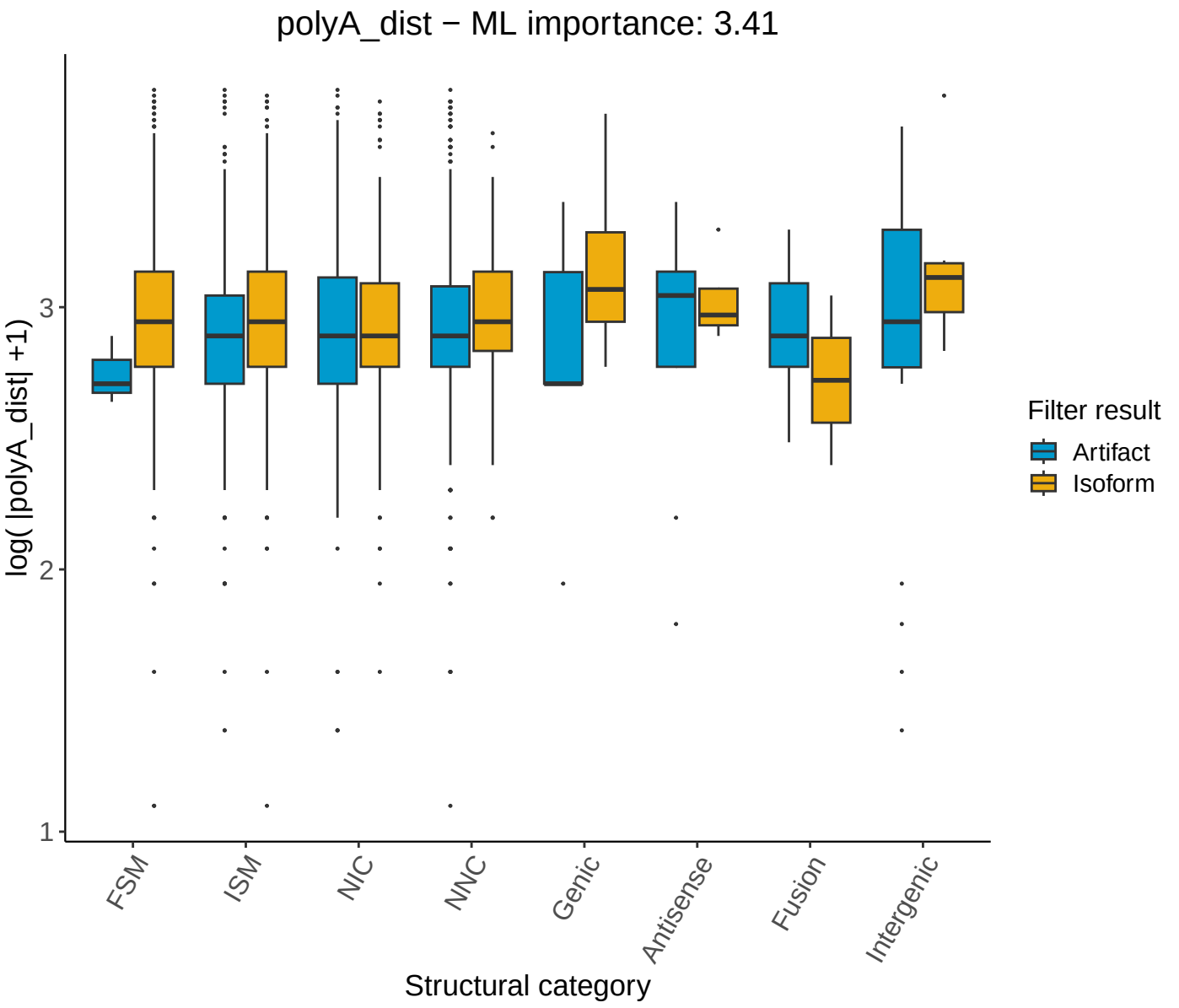

dist_to_CAGE_peak − ML importance: 5.28
log(|dist_to_CAGE_peak| +1)
7.5
5.0
2.5
0.0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

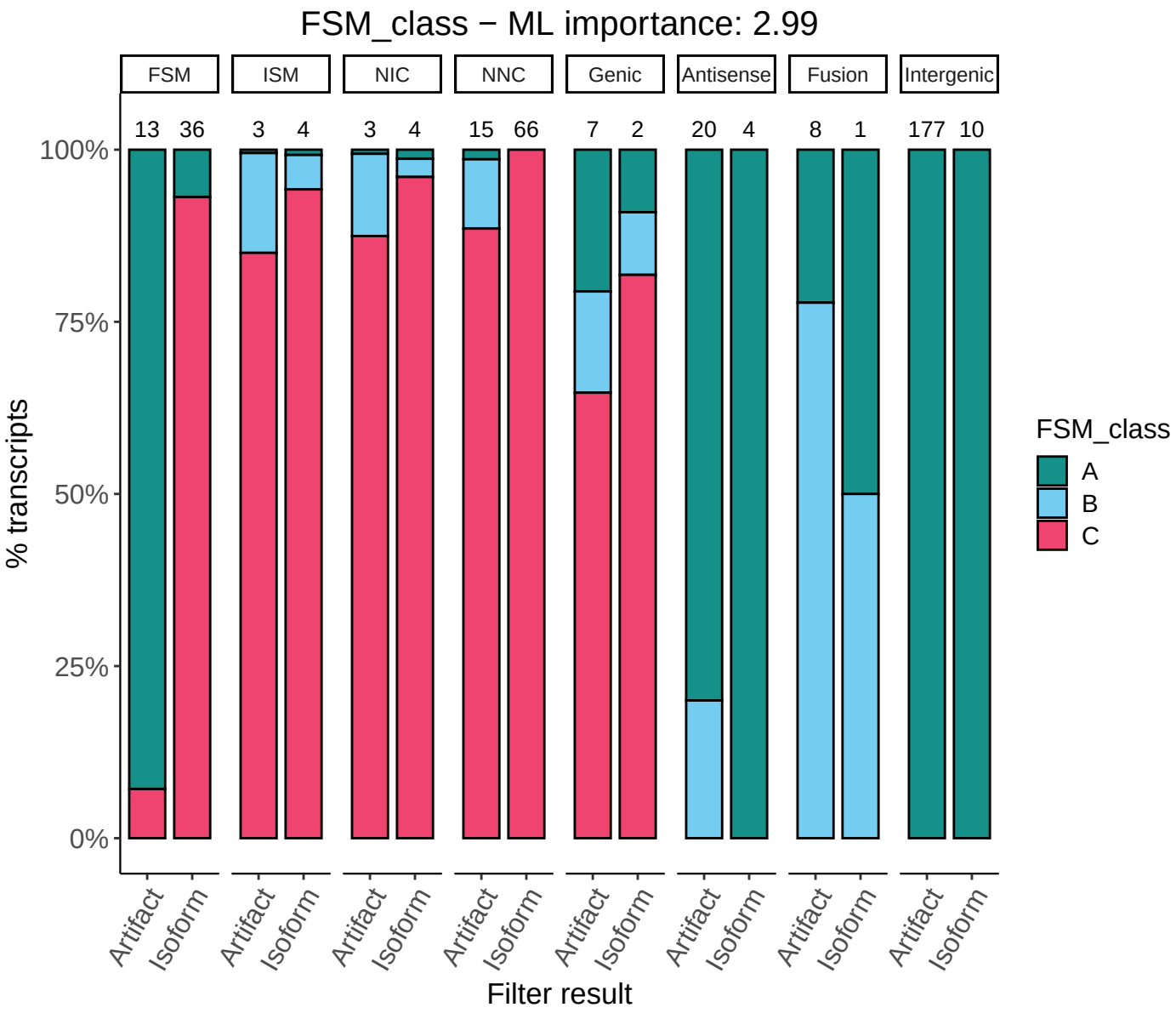

length − ML importance: 5.15
log(|length| +1)
9
8
7
6
5
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

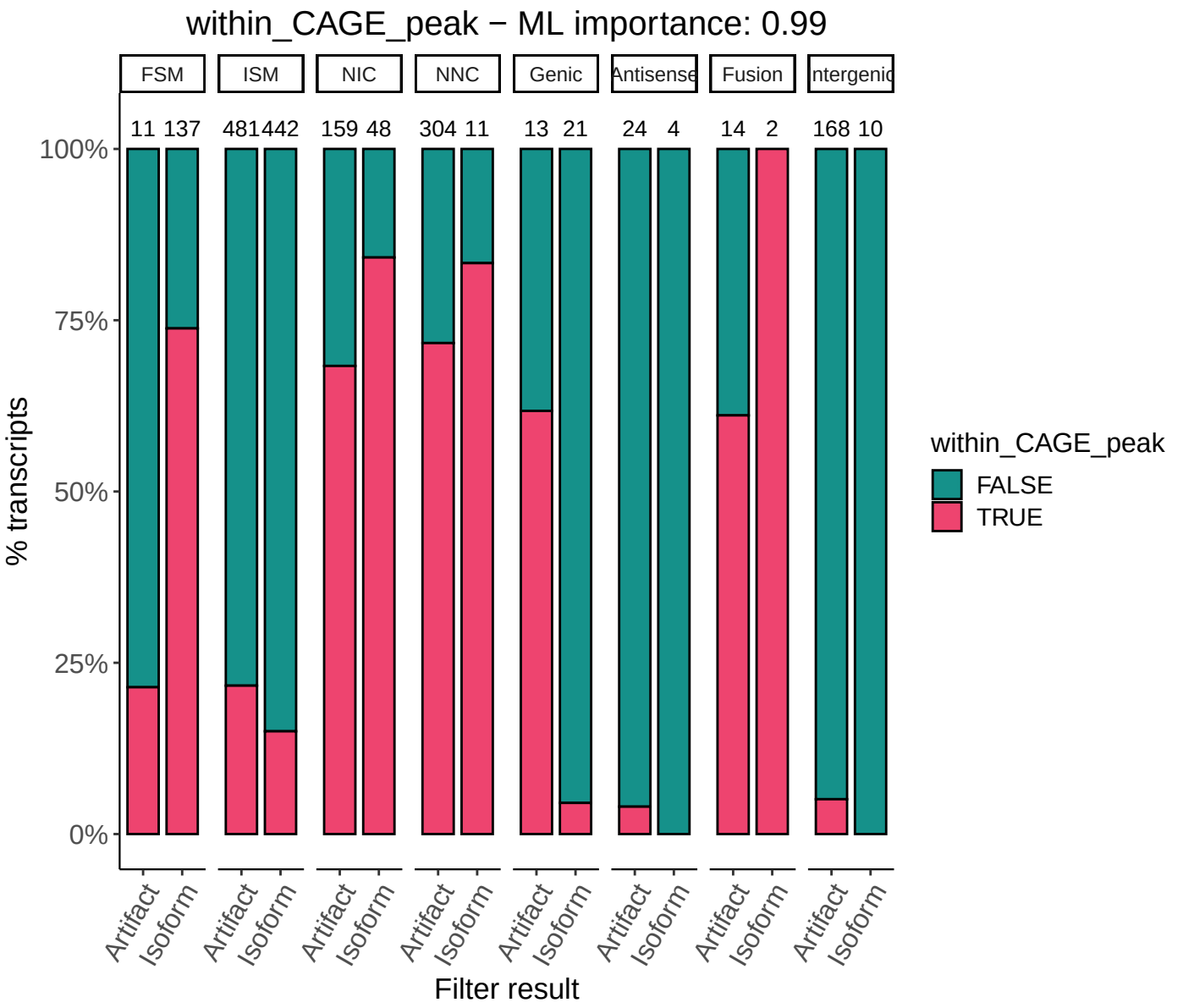

ratio_TSS − ML importance: 4.64
log(|ratio_TSS| +1)
10.0
7.5
5.0
2.5
0.0
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

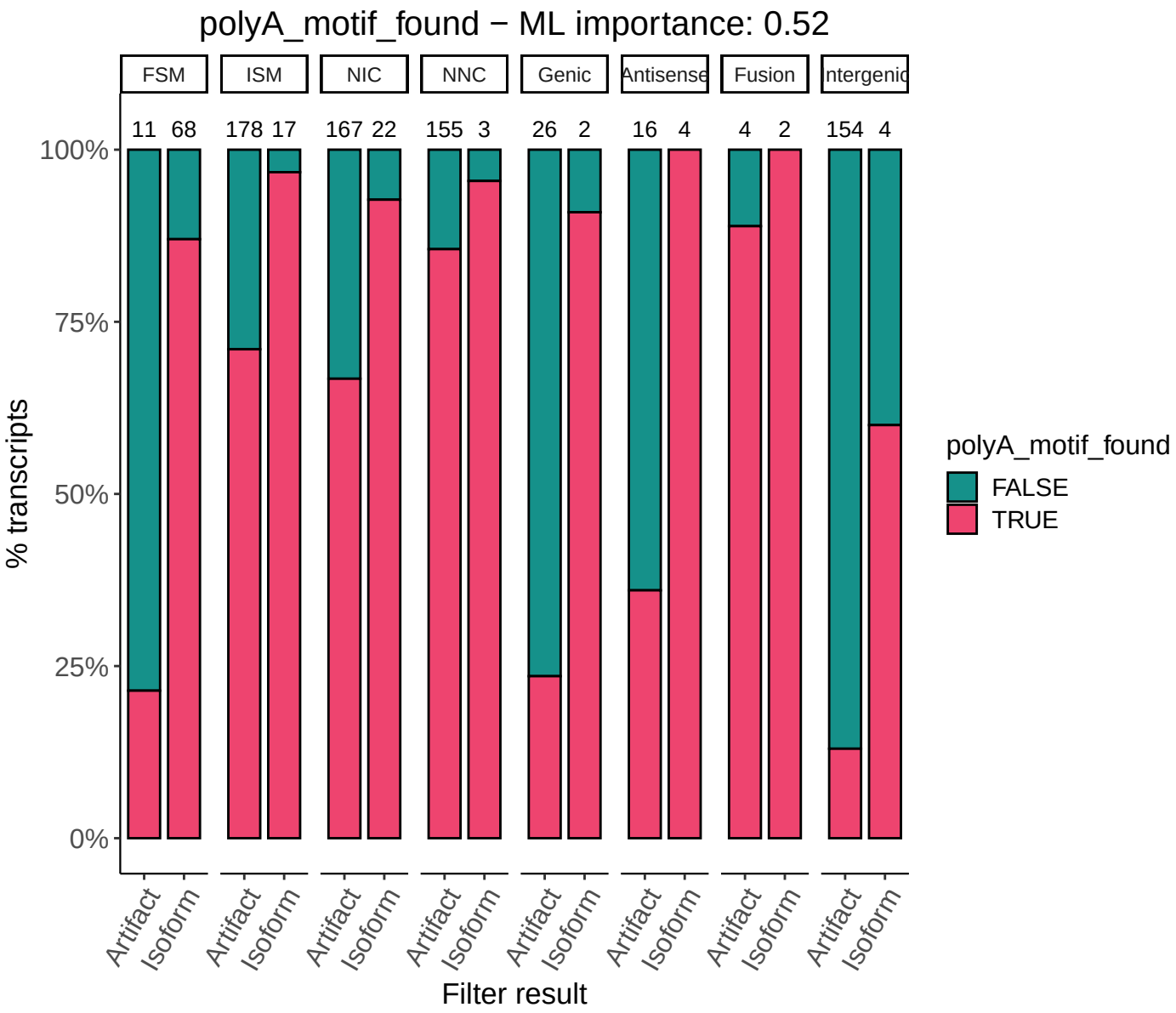

exons − ML importance: 4.13
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Transcript no.
0
100
200
300
400
Artifact
Isoform
Filter result
exons
1
2−3
4−5
6−10
>10

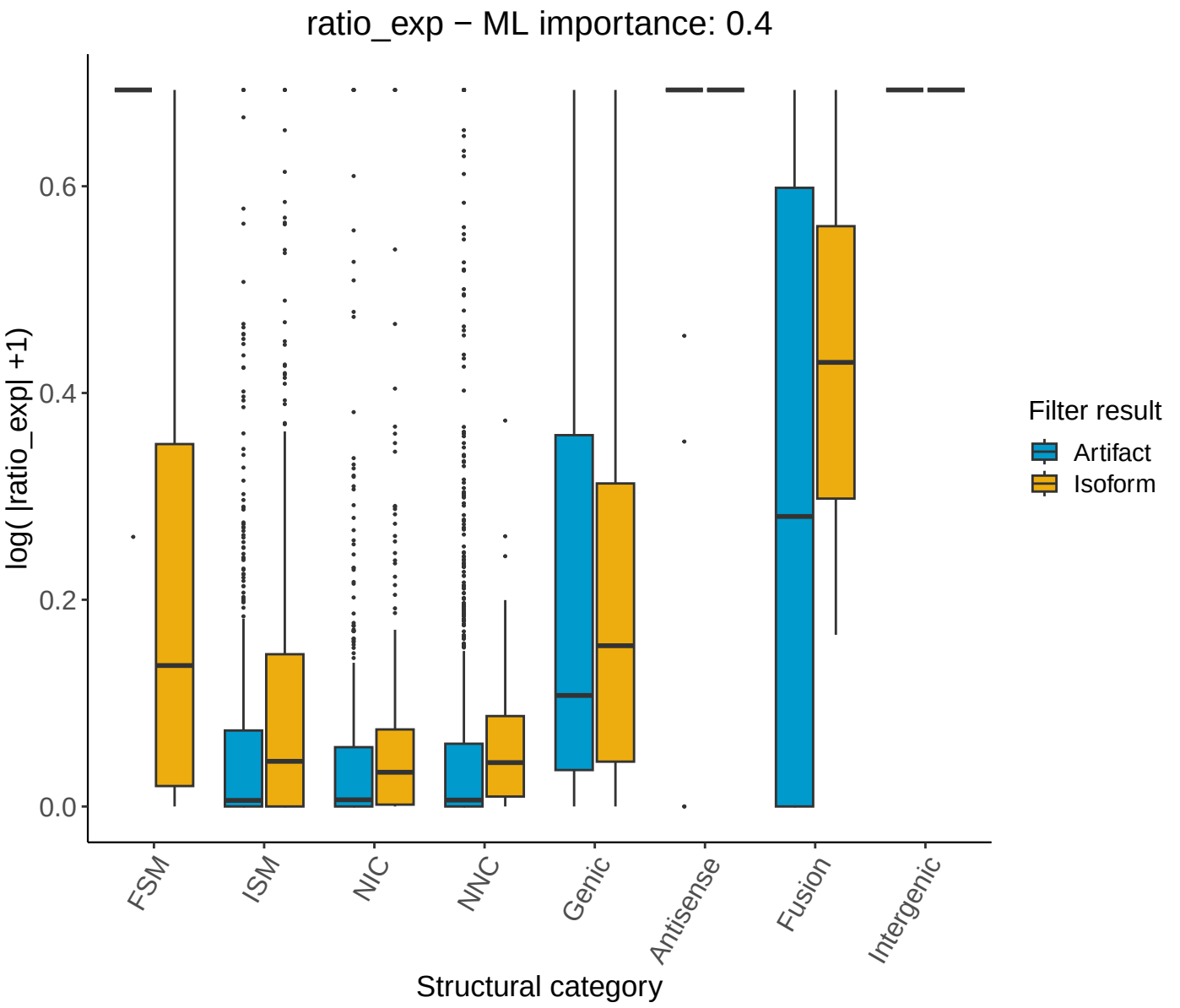

perc_A_downstream_TTS − ML importance: 3.9
log(|perc_A_downstream_TTS| +1)
0
1
2
3
4
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

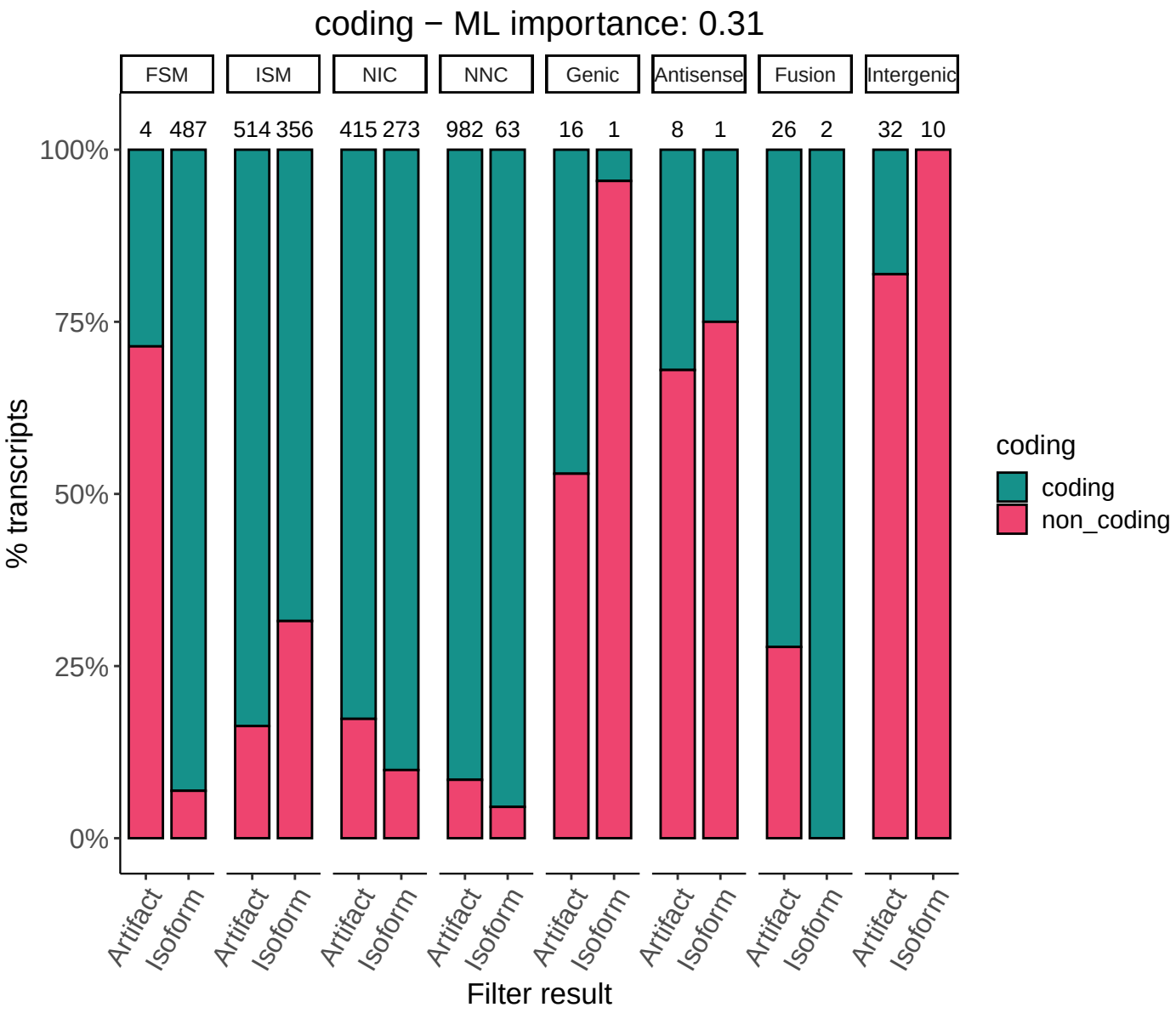

polyA_dist − ML importance: 3.41
log(|polyA_dist| +1)
3
2
1
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform


FSM_class − ML importance: 2.99
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
13 36
3 4
3 4
15 66
7 2
20 4
8 1
177 10
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
FSM_class
A
B
C

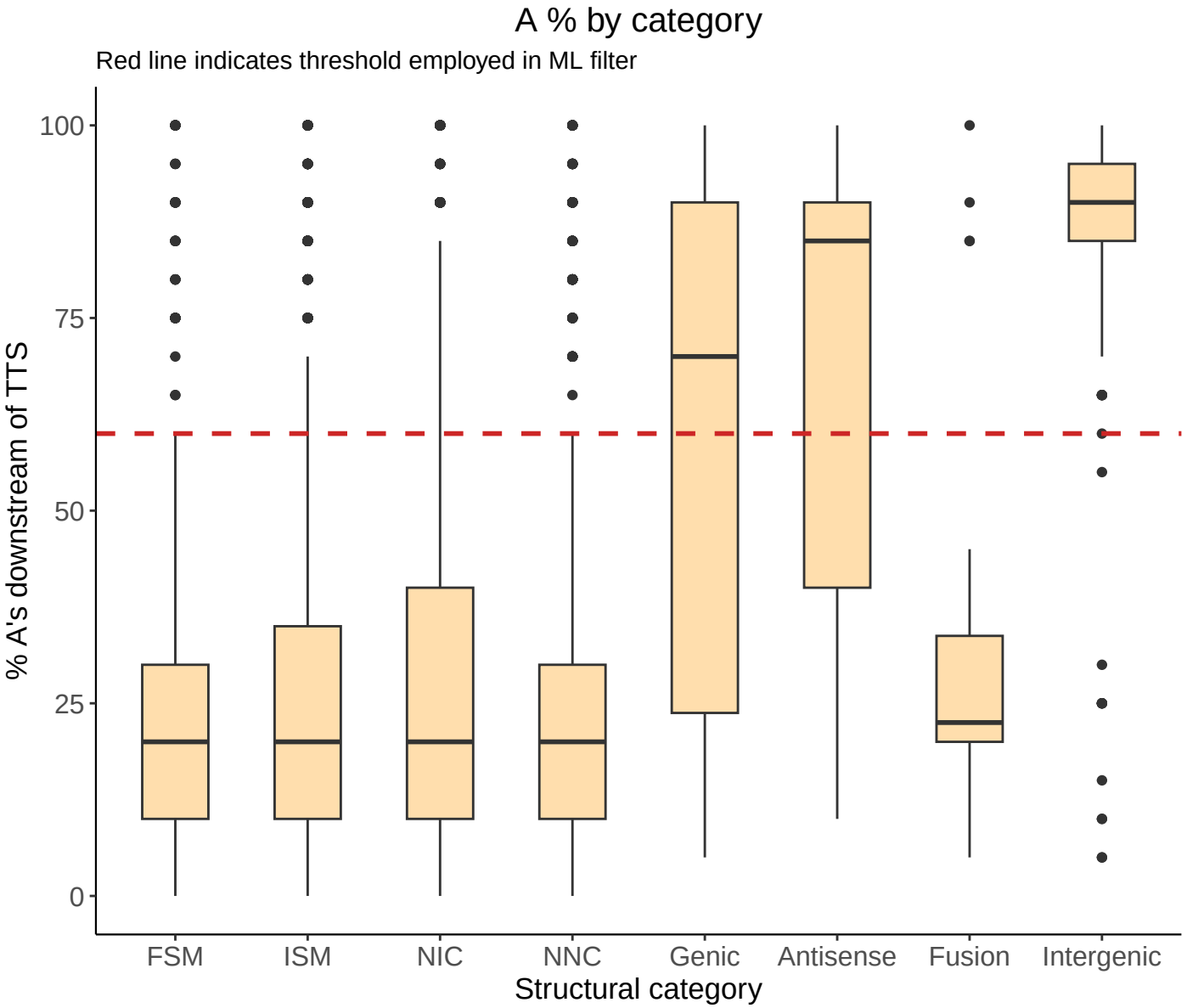

within_CAGE_peak − ML importance: 0.99
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
ntergenic
11 137
481 442
159 48
304 11
13 21
24 4
14 2
168 10
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
within_CAGE_peak
FALSE
TRUE

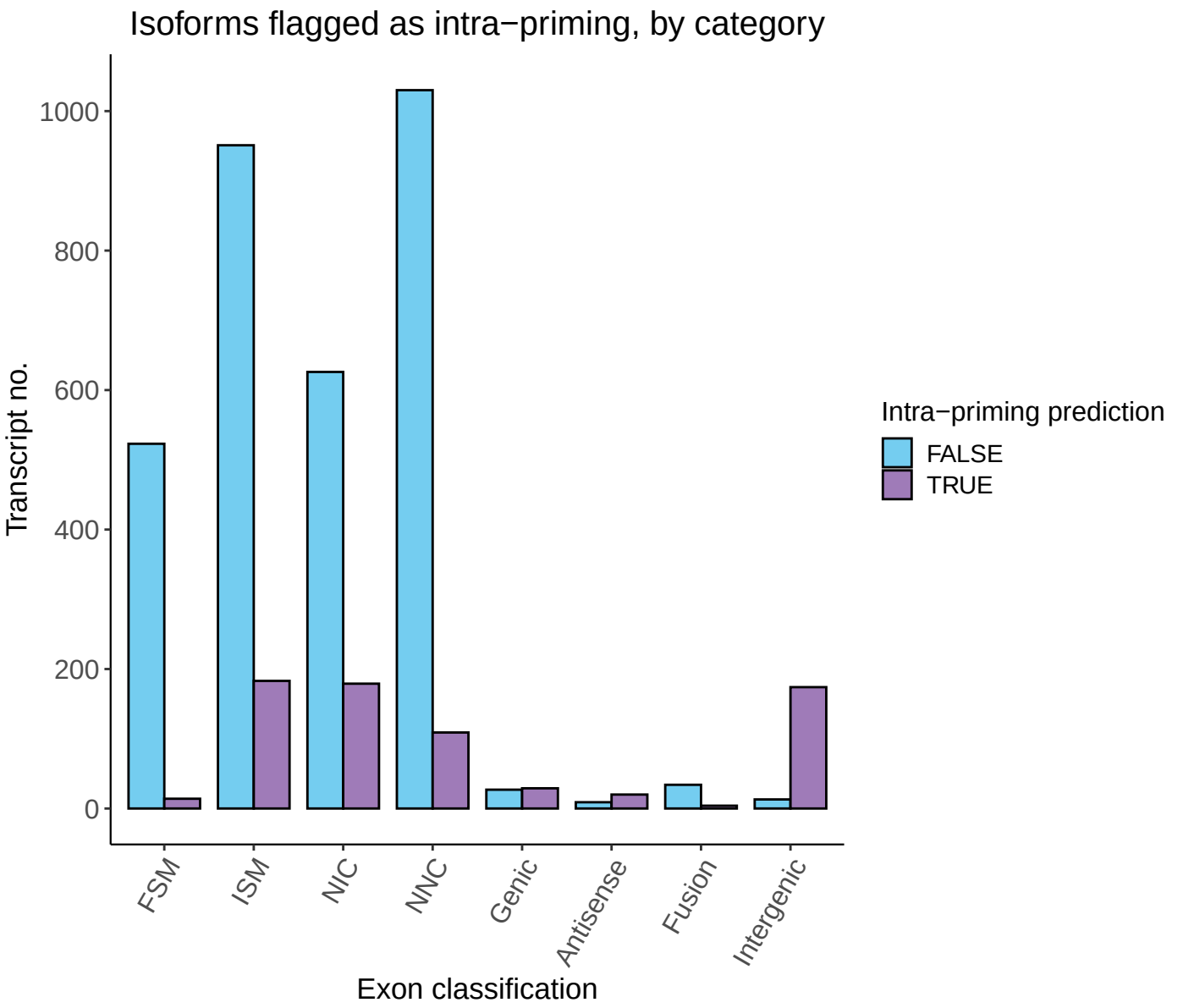

polyA_motif_found − ML importance: 0.52
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
ntergenic
11 68
178 17
167 22
155 3
26 2
16 4
4 2
154 4
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
polyA_motif_found
FALSE
TRUE

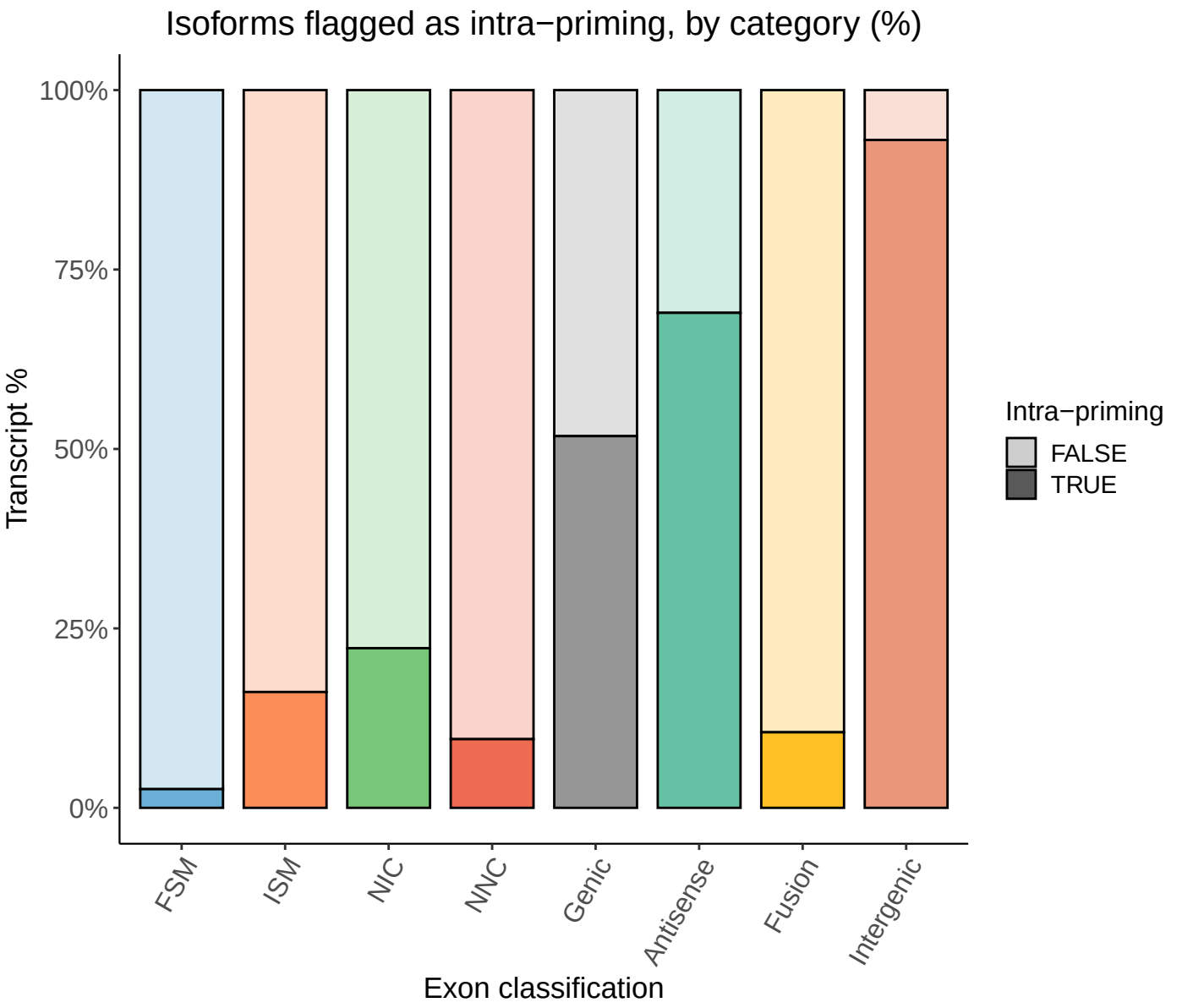

ratio_exp − ML importance: 0.4
log(|ratio_exp| +1)
0.0
0.2
0.4
0.6
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
Structural category
Filter result
Artifact
Isoform

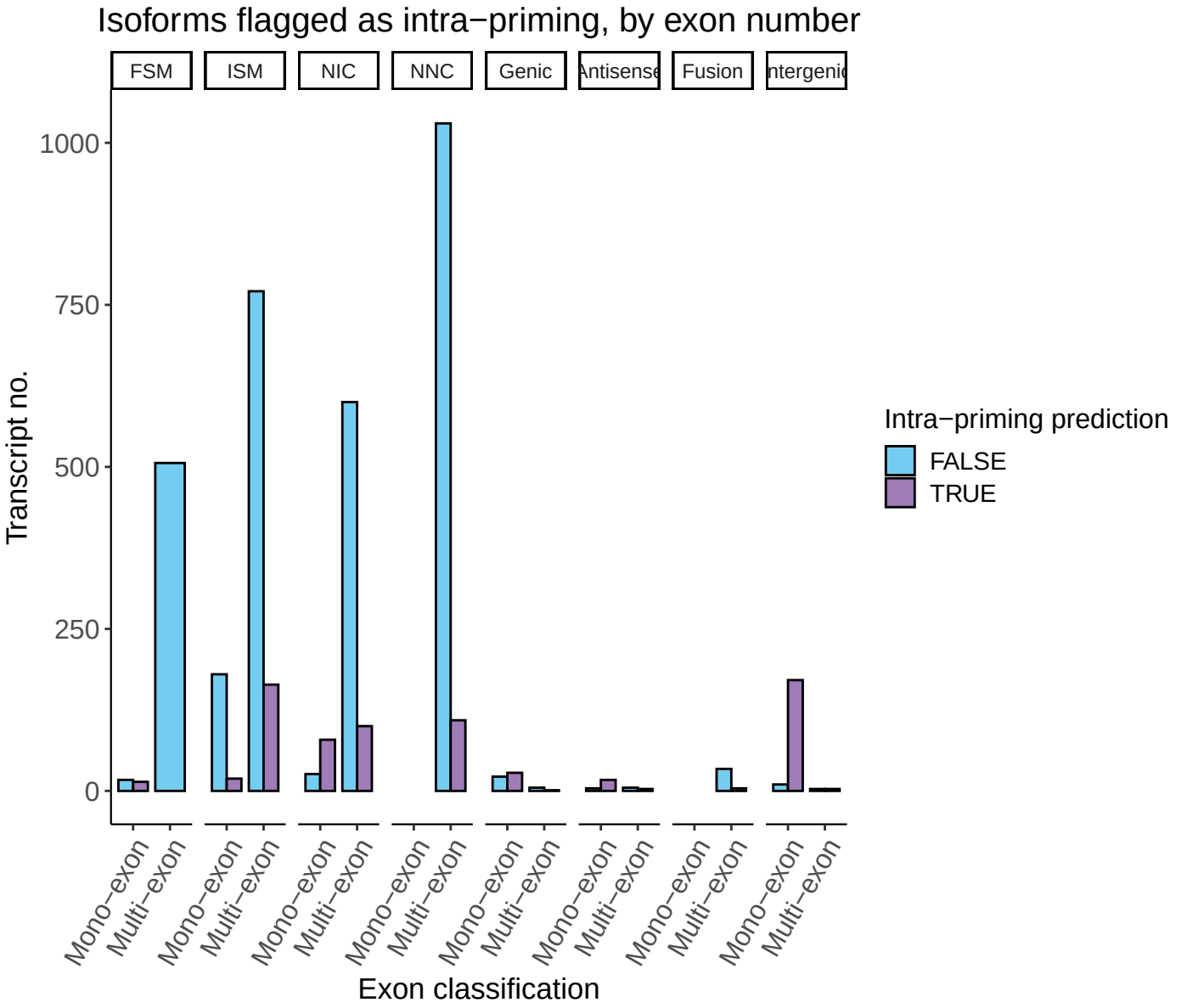

coding − ML importance: 0.31
FSM
ISM
NIC
NNC
Genic
Antisense
Fusion
Intergenic
4 487
514 356
415 273
982 63
16 1
8 1
26 2
32 10
100%
75%
50%
25%
0%
% transcripts
Artifact
Isoform
Filter result
coding
coding
non_coding

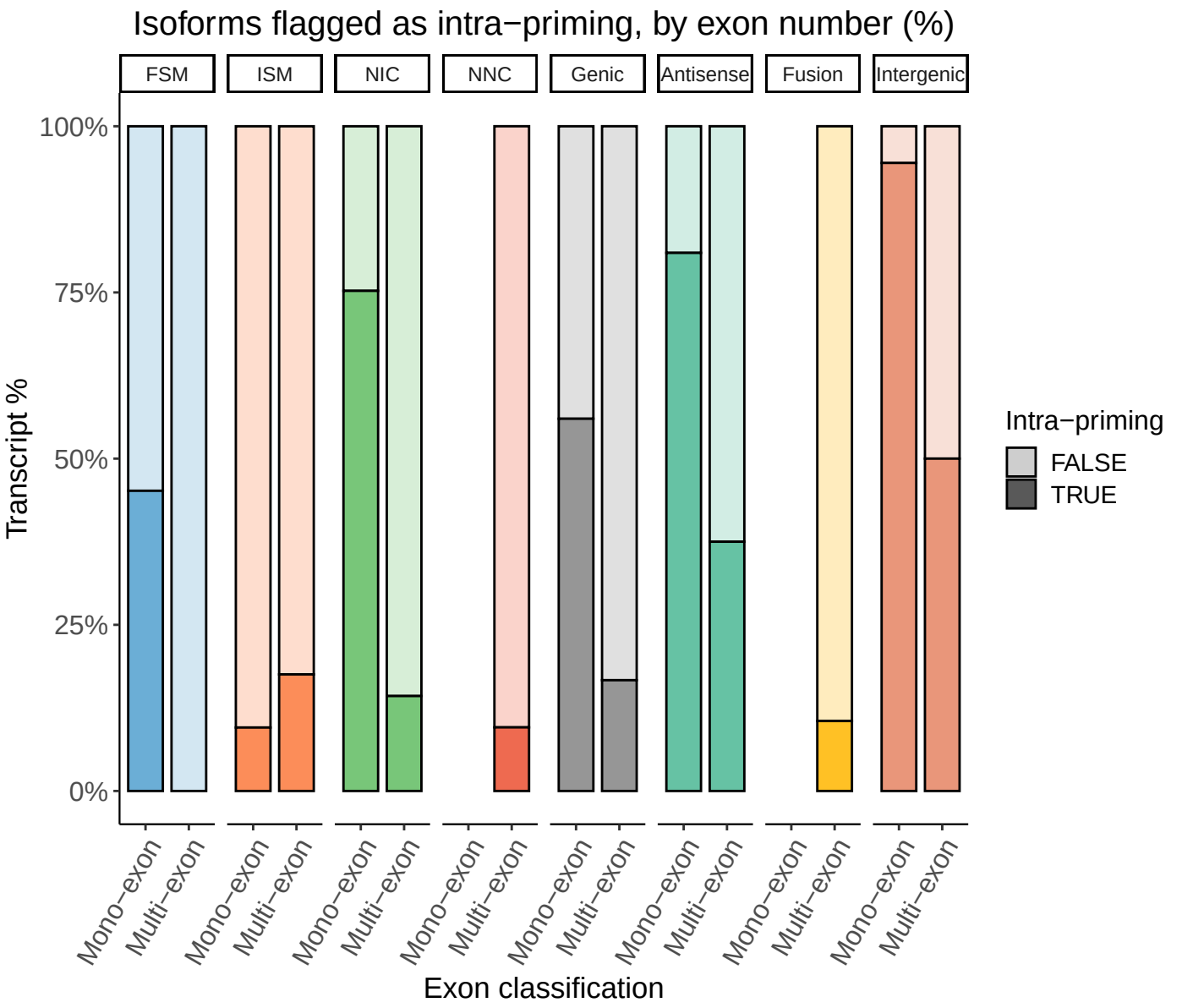
Intra−primming filter report


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
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
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
Exon classification
Intra−priming
FALSE
TRUE


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