Supplementary Figure Legend

Figure S1: Read count distribution for the (A) Universal Human Reference real data and the (B) The Type I1 simulation data with gradient depth.

Figure S2: Read length distribution of the Type I1 simulation data with gradient median read length.

Figure S3: Root mean squared error (RMSE) between the true and the estimated isoform expression in the Type I1 simulation data. (A) RMSE of isoform expression between the truth and each tool under different sequencing accuracy with 0.5M reads and 1500 bp median length. (B) RMSE of isoform expression between the truth and each tool under different sequencing depth with accurate sequencing and 1500 bp median length. (c) RMSE of isoform expression between the truth and each tool under gradient median read lengths with 0.5M reads and accurate sequencing.

Figure S4: Computational cost for isoform quantification tools in the Type I1 simulation. (A&B) Time cost along different number of reads and read lengths. (C) Memory cost across different read lengths when working on 4M reads.

Figure S5: The percentage contingency table between the true isoform assignments and the annotation by IFDlong in the Type I2 simulation data with gradient median read lengths.

Figure S6: Sensitivity (left) and Specificity (right) of IFDlong when detecting novel isoforms in the Type I3 simulation with ONT (top) and PacBio (bottom) accuracy. The bars are colored by different buffer length setter (0bp or 9bp) and median length (1000 bp or 2000 bp) settings.

Figure S7: Sensitivity (left) and Specificity (right) of TALON when detecting novel isoforms in the Type I3 simulation with ONT (top) and PacBio (bottom) accuracy. The bars are colored by different median length (1000 bp or 2000 bp) settings.

Figure S8: Spearman's correlation of isoform quantification in human MCF7 breast cancer cell line by different ONT sequencing settings.

Figure S9: Pairwise comparison of isoform quantification in mouse C2C12 and Heart samples. The bottom left cells present the pairwise scatter plot of isoform expression. The upper right cells indicate the Spearman's correlation.

Figure S10: Isoform distribution of Calcium Voltage-Gated Channel Subunit Alpha1 C (Cacna1c), Troponin T2 (Tnnt2), Calcium/Calmodulin Dependent Protein Kinase II Gamma (Camk2g) and Myocyte Enhancer Factor 2C (Mef2c) in heart tissue.

Figure S11: Pathway analysis on isoform differential expression analysis in HCC data. (A) Gene set enrichment analysis (GSEA) on Gene ontology (GO) database. (B) Ingenuity pathway analysis (IPA).

Figure S12: Feature plot of isoform expression for gene CCL5 and IL7R in HCC data.

Figure S13: Pairwise comparison of fusion quantification in the Type F1 simulation. (A) The Type F1 simulation with median length of 1500 bp and PacBio accuracy. (B) The Type F1 simulation with median length of 1500 bp and ONT accuracy. The bottom left cells present the pairwise scatter plot of isoform expression. The upper right cells indicate the Spearman's correlation.

Figure S14: Scatter plot of isoform-level fusion quantification comparing the truth the IFDlong estimation in the Type F1 simulation. (A) The Type F1 simulation with median length of 1500 bp and accurate setting. (B) The Type F1 simulation with median length of 1500 bp and PacBio accuracy. (C) The Type F1 simulation with median length of 1500 bp and ONT accuracy.

Figure S15: Precision and Recall curve for fusion transcript detection by IFDlong pipeline using different anchor lengths. (A) The Type F1 simulation with median length of 1500 bp and PacBio accuracy. (B) The Type F1 simulation with median length of 1500 bp and ONT accuracy.

Figure S16: Relative abundance of multiple paired genes in the Type F2 simulation data with different accuracy setting. (A) The Type F2 simulation with 1500 bp median length and PacBio setting. (B) The Type F2 simulation with 1500 bp median length and ONT setting.

Figure S17: Cosine similarity of the relative abundance of multiple paired genes in the Type F2 simulation data. (A) The Type F2 simulation with median length to be 200bp and accurate setting. (B) The Type F2 simulation with median length to be 1000bp and accurate setting. (C) The Type F2 simulation with median length to be 1500bp and PacBio setting. (D) The Type F2 simulation with median length to be 1500bp and ONT setting.

Figure S18: Relative abundance of multiple paired genes in the Type F2 simulation data with gradient median length. (A) The Type F2 simulation with 200 bp median length and accurate setting. (B) The Type F2 simulation with 1000 bp median length and accurate setting.

Supplementary Table

Table S1: Isoform quantification by IFDlong on UHR dataset.
Table S2: Isoform quantification by IFDlong on MCF7 dataset.
Table S3: Isoform quantification by IFDlong on mouse C2C12 dataset.
Table S4: Isoform quantification by IFDlong on mouse heart dataset.
Table S5: Differentially expressed isoforms in HCC dataset comparing tumor and normal.
Table S6: Pathway analysis on the differentially expressed isoforms in HCC dataset.
Table S7: Popular fusion transcripts in TCGA dataset that were employed in the Type F2 simulation.
Table S8: Two-way fusions detected in MCF7 ONT dataset by IFDlong.

Figure S1

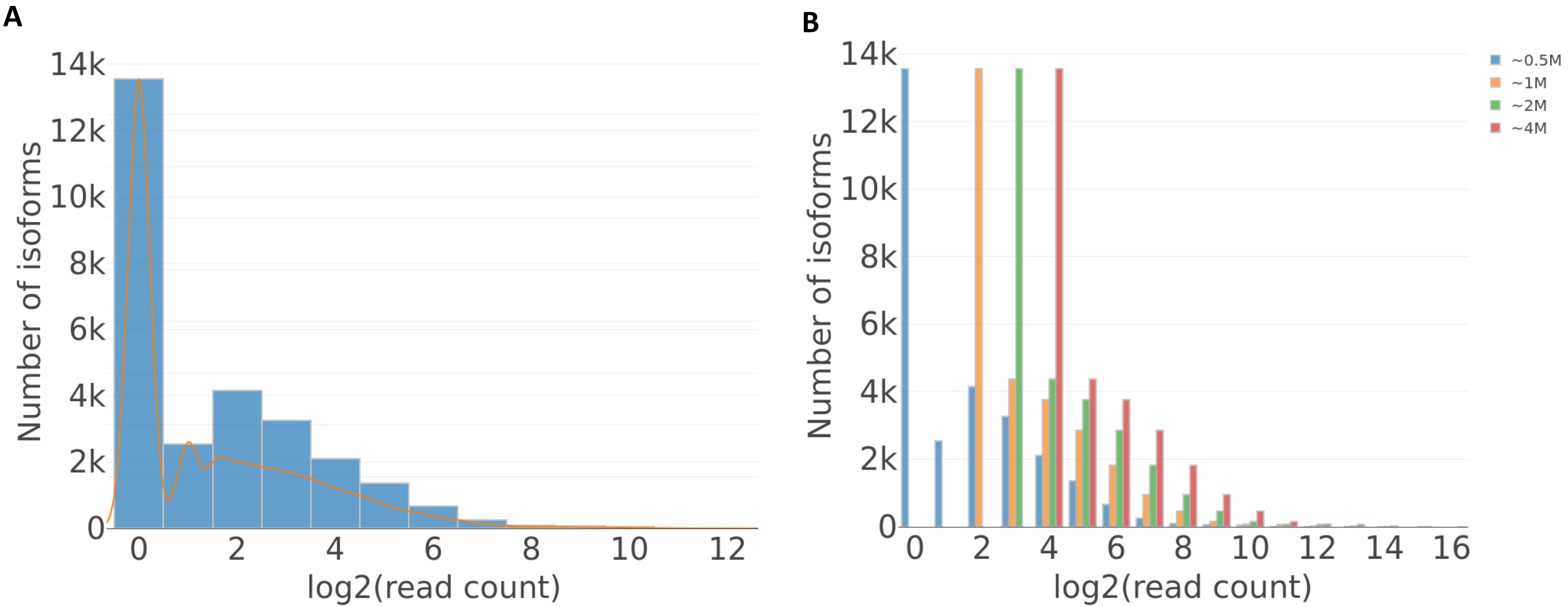
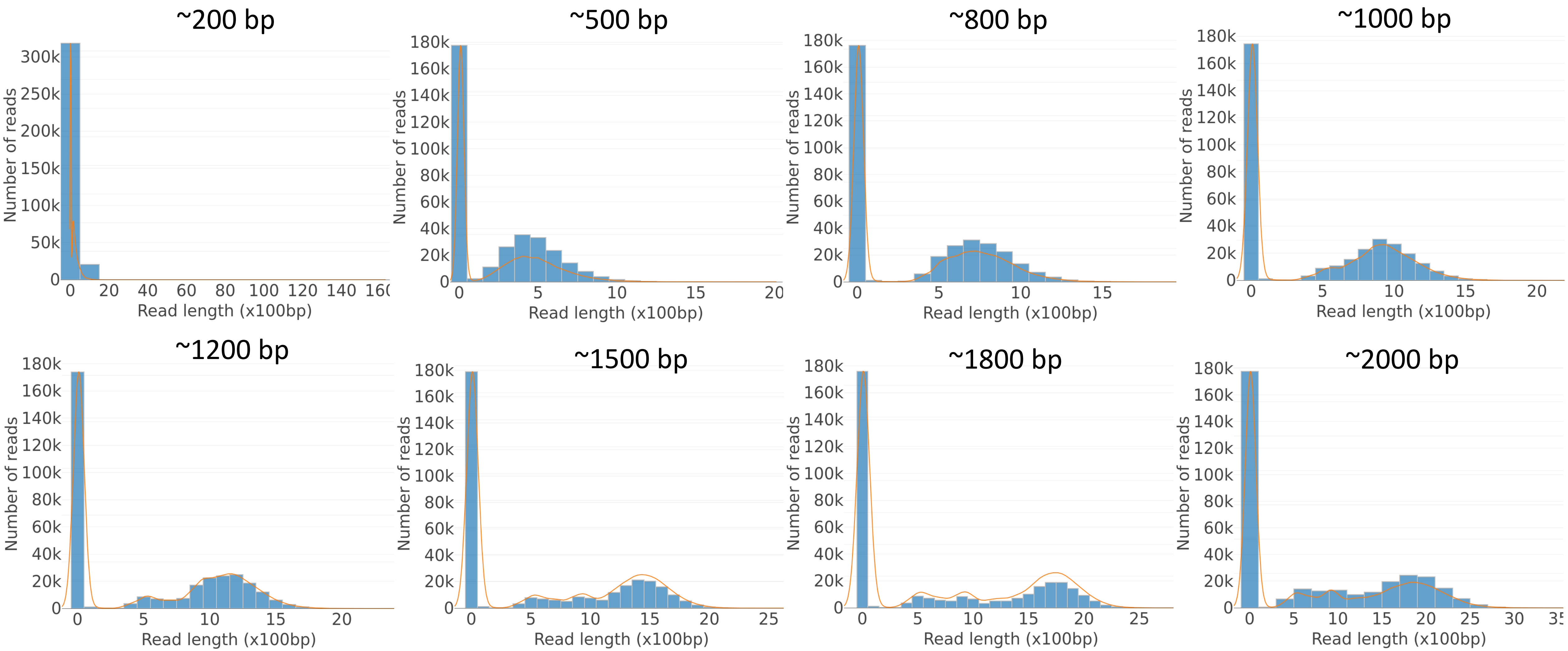

A
14k
12k
10k
8k
6k
4k
2k
0
Number of isoforms
0
2
4
6
8
10
12
log2(read count)
B
14k
12k
10k
8k
6k
4k
2k
0
Number of isoforms
0
2
4
6
8
10
12
14
16
log2(read count)
~0.5M
~1M
~2M
~4M

Figure S2



~200 bp

~500 bp

~800 bp

~1000 bp

~1200 bp

~1500 bp

~1800 bp

~2000 bp

Figure S3

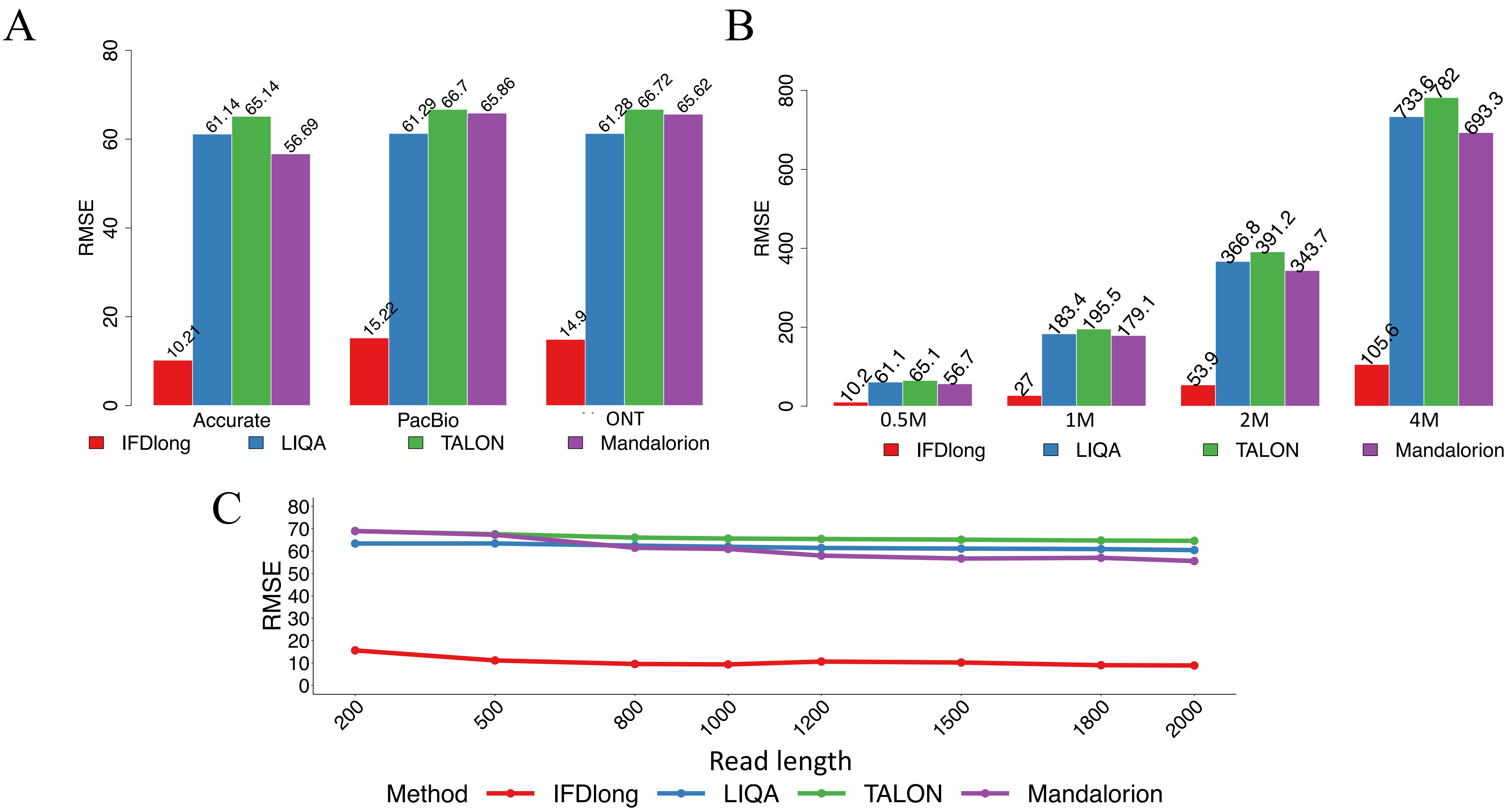

A
RMSE
0
20
40
60
80
10.21
61.14
65.14
56.69
15.22
61.29
66.7
65.86
14.9
61.28
66.72
65.62
Accurate
PacBio
ONT
IFDlong
LIQA
TALON
Mandalorion
B
RMSE
0
200
400
600
800
10.2
61.1
65.1
56.7
27
183.4
195.5
179.1
53.9
366.8
391.2
343.7
105.6
733.6
782
693.3
0.5M
1M
2M
4M
IFDlong
LIQA
TALON
Mandalorion
C
RMSE
0
10
20
30
40
50
60
70
80
200
500
800
1000
1200
1500
1800
2000
Read length
Method
IFDlong
LIQA
TALON
Mandalorion

Figure S4

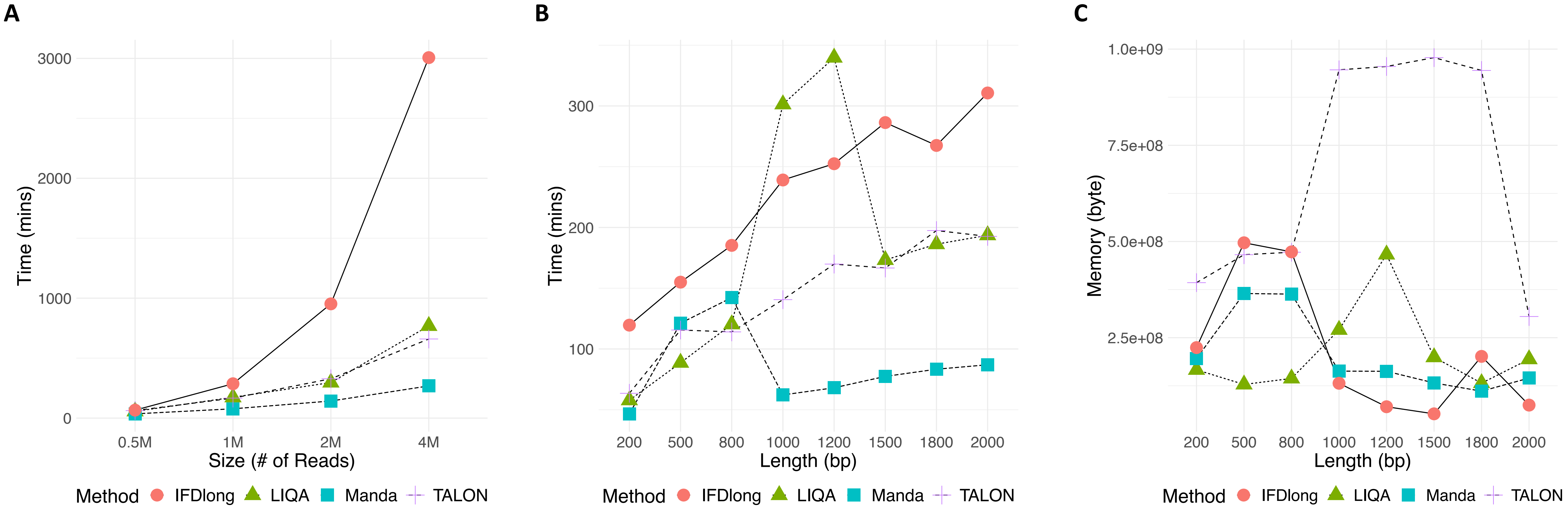

A
Time (mins)
3000
2000
1000
0
0.5M
1M
2M
4M
Size (# of Reads)
Method IFDlong LIQA Manda TALON
B
Time (mins)
300
200
100
200
500
800
1000
1200
1500
1800
2000
Length (bp)
Method IFDlong LIQA Manda TALON
C
Memory (byte)
1.0e+09
7.5e+08
5.0e+08
2.5e+08
200
500
800
1000
1200
1500
1800
2000
Length (bp)
Method IFDlong LIQA Manda TALON

Figure S5

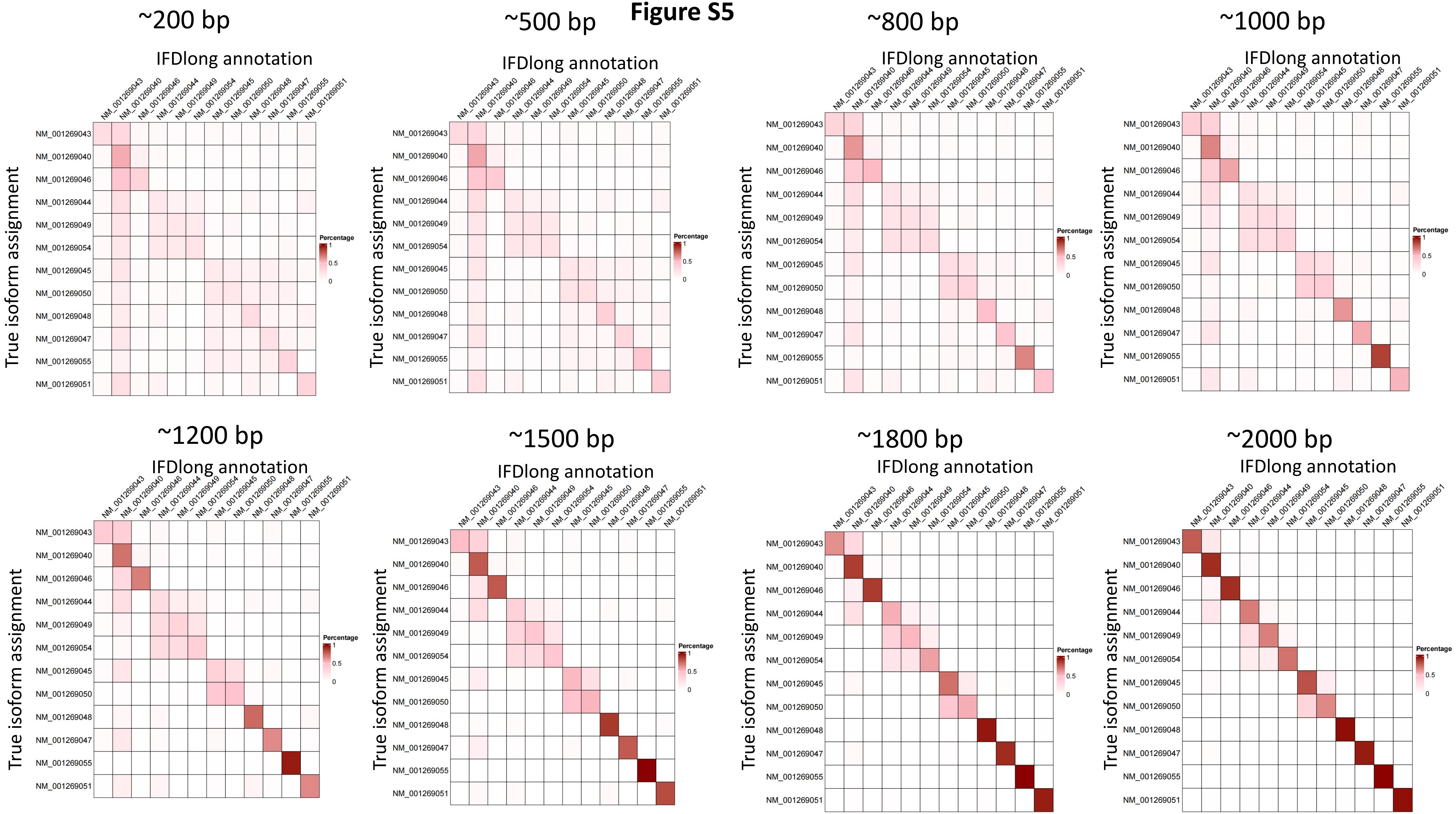

~200 bp
~500 bp
~800 bp
~1000 bp
~1200 bp
~1500 bp
~1800 bp
~2000 bp
IFDlong annotation
True isoform assignment
NM_001269043
NM_001269040
NM_001269046
NM_001269044
NM_001269049
NM_001269054
NM_001269045
NM_001269050
NM_001269048
NM_001269047
NM_001269055
NM_001269051
Percentage
1
0.5
0

Figure S6

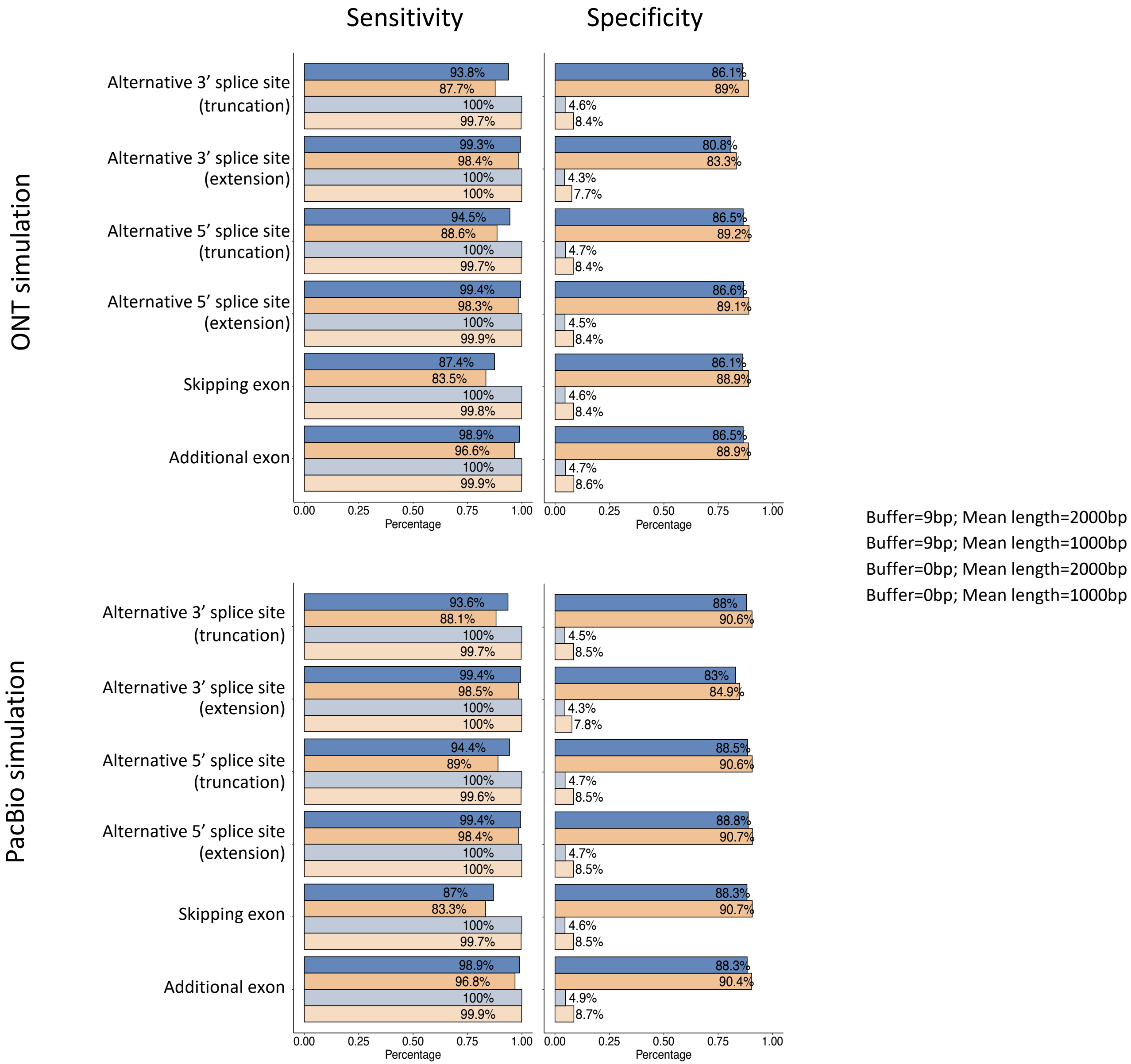

Sensitivity
Specificity
ONT simulation
Alternative 3’ splice site (truncation)
93.8%
87.7%
100%
99.7%
86.1%
89%
4.6%
8.4%
Alternative 3’ splice site (extension)
99.3%
98.4%
100%
100%
80.8%
83.3%
4.3%
7.7%
Alternative 5’ splice site (truncation)
94.5%
88.6%
100%
99.7%
86.5%
89.2%
4.7%
8.4%
Alternative 5’ splice site (extension)
99.4%
98.3%
100%
99.9%
86.6%
89.1%
4.5%
8.4%
Skipping exon
87.4%
83.5%
100%
99.8%
86.1%
88.9%
4.6%
8.4%
Additional exon
98.9%
96.6%
100%
99.9%
86.5%
88.9%
4.7%
8.6%
0.00
0.25
0.50
0.75
1.00
Percentage
PacBio simulation
Alternative 3’ splice site (truncation)
93.6%
88.1%
100%
99.7%
88%
90.6%
4.5%
8.5%
Alternative 3’ splice site (extension)
99.4%
98.5%
100%
100%
83%
84.9%
4.3%
7.8%
Alternative 5’ splice site (truncation)
94.4%
89%
100%
99.6%
88.5%
90.6%
4.7%
8.5%
Alternative 5’ splice site (extension)
99.4%
98.4%
100%
100%
88.8%
90.7%
4.7%
8.5%
Skipping exon
87%
83.3%
100%
99.7%
88.3%
90.7%
4.6%
8.5%
Additional exon
98.9%
96.8%
100%
99.9%
88.3%
90.4%
4.9%
8.7%
0.00
0.25
0.50
0.75
1.00
Percentage
Buffer=9bp; Mean length=2000bp
Buffer=9bp; Mean length=1000bp
Buffer=0bp; Mean length=2000bp
Buffer=0bp; Mean length=1000bp

Figure S7

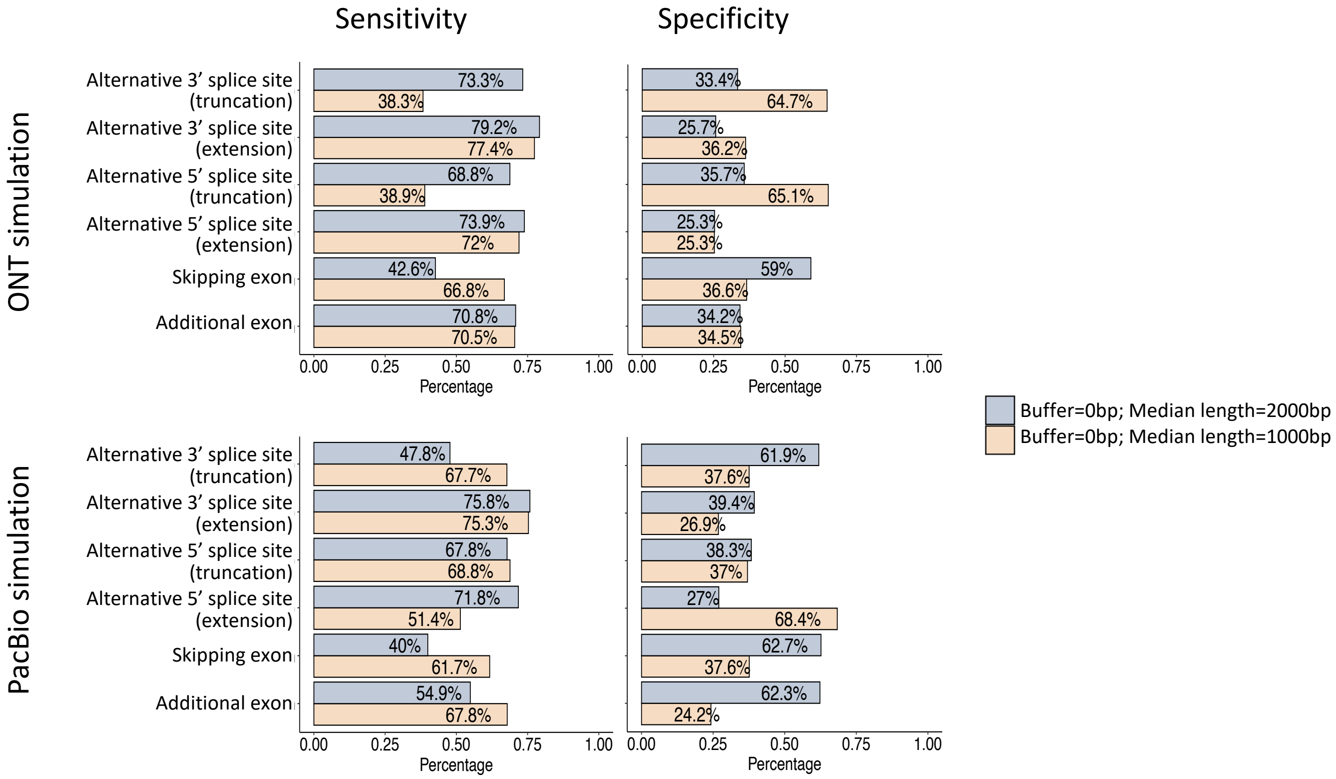

Sensitivity
Specificity
ONT simulation
PacBio simulation
Alternative 3’ splice site (truncation)
Alternative 3’ splice site (extension)
Alternative 5’ splice site (truncation)
Alternative 5’ splice site (extension)
Skipping exon
Additional exon
73.3%
38.3%
79.2%
77.4%
68.8%
38.9%
73.9%
72%
42.6%
66.8%
70.8%
70.5%
33.4%
64.7%
25.7%
36.2%
35.7%
65.1%
25.3%
25.3%
59%
36.6%
34.2%
34.5%
47.8%
67.7%
75.8%
75.3%
67.8%
68.8%
71.8%
51.4%
40%
61.7%
54.9%
67.8%
61.9%
37.6%
39.4%
26.9%
38.3%
37%
27%
68.4%
62.7%
37.6%
62.3%
24.2%
0.00
0.25
0.50
0.75
1.00
Percentage
Buffer=0bp; Median length=2000bp
Buffer=0bp; Median length=1000bp

Figure S8

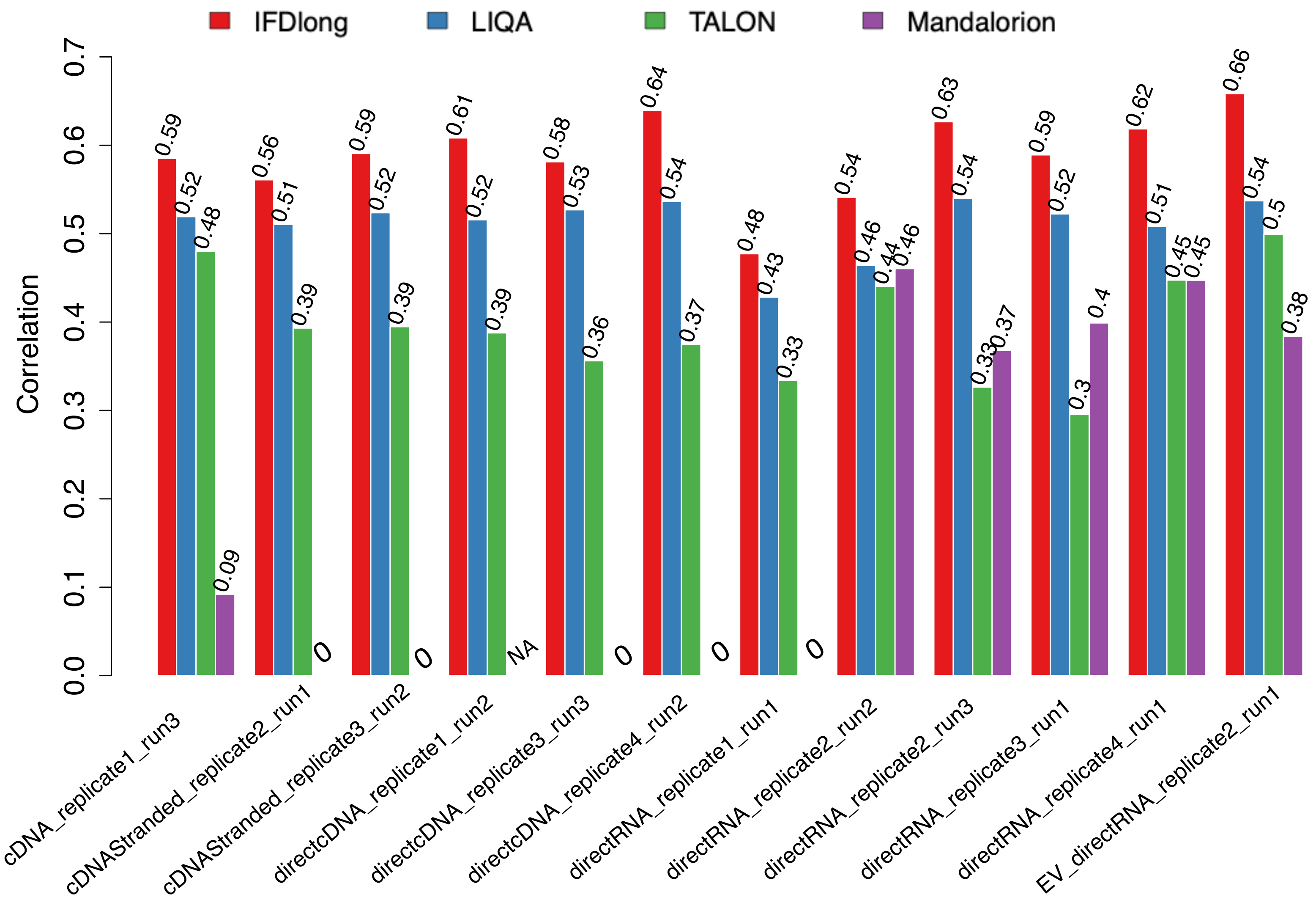

IFDlong
LIQA
TALON
Mandalorion
Correlation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.59
0.52
0.48
0.09
0.56
0.51
0.39
0
0.59
0.52
0.39
0
0.61
0.52
0.39
NA
0.58
0.53
0.36
0
0.64
0.54
0.37
0
0.48
0.43
0.33
0
0.54
0.46
0.44
0.46
0.63
0.54
0.33
0.37
0.59
0.52
0.3
0.4
0.62
0.51
0.45
0.45
0.66
0.54
0.5
0.38
cDNA_replicate1_run3
cDNAStranded_replicate2_run1
cDNAStranded_replicate3_run2
directcDNA_replicate1_run2
directcDNA_replicate3_run3
directcDNA_replicate4_run2
directRNA_replicate1_run1
directRNA_replicate2_run2
directRNA_replicate2_run3
directRNA_replicate3_run1
directRNA_replicate4_run1
EV_directRNA_replicate2_run1

Figure S9

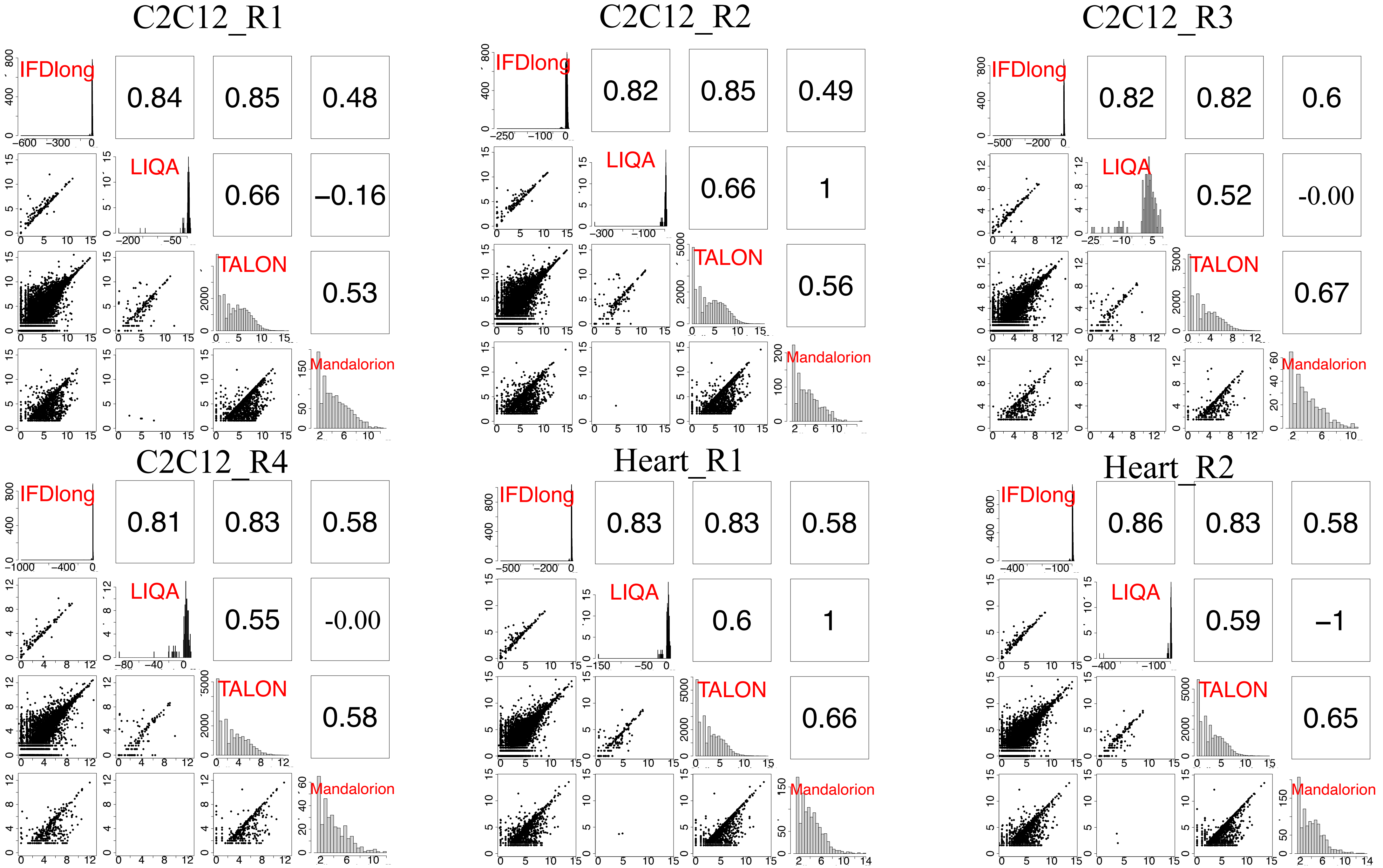

C2C12_R1
IFDlong
0.84
0.85
0.48
LIQA
0.66
−0.16
TALON
0.53
Mandalorion
C2C12_R2
IFDlong
0.82
0.85
0.49
LIQA
0.66
1
TALON
0.56
Mandalorion
C2C12_R3
IFDlong
0.82
0.82
0.6
LIQA
0.52
-0.00
TALON
0.67
Mandalorion
C2C12_R4
IFDlong
0.81
0.83
0.58
LIQA
0.55
-0.00
TALON
0.58
Mandalorion
Heart_R1
IFDlong
0.83
0.83
0.58
LIQA
0.6
1
TALON
0.66
Mandalorion
Heart_R2
IFDlong
0.86
0.83
0.58
LIQA
0.59
−1
TALON
0.65
Mandalorion

Figure S10

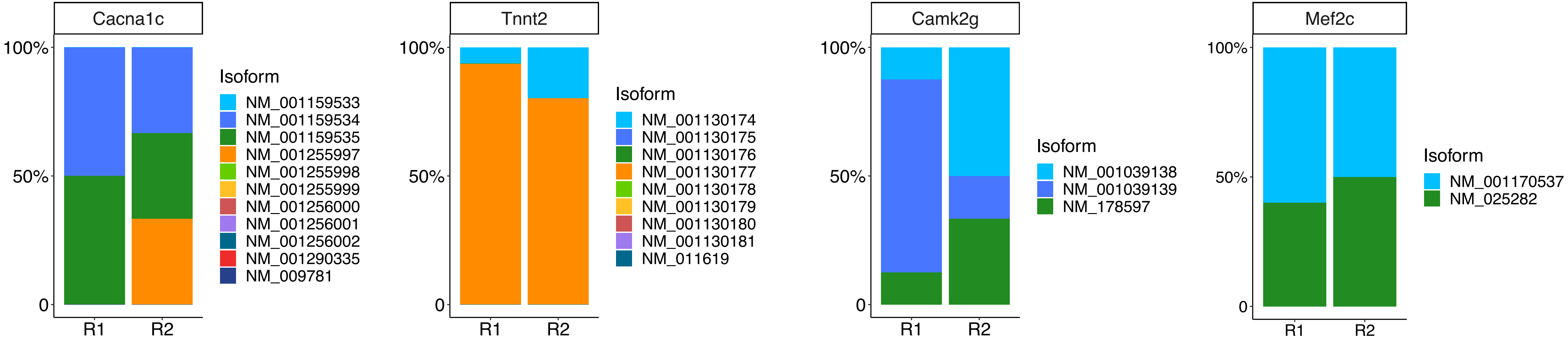

Cacna1c
100%
50%
0
R1
R2
Isoform
NM_001159533
NM_001159534
NM_001159535
NM_001255997
NM_001255998
NM_001255999
NM_001256000
NM_001256001
NM_001256002
NM_001290335
NM_009781
Tnnt2
100%
50%
0
R1
R2
Isoform
NM_001130174
NM_001130175
NM_001130176
NM_001130177
NM_001130178
NM_001130179
NM_001130180
NM_001130181
NM_011619
Camk2g
100%
50%
0
R1
R2
Isoform
NM_001039138
NM_001039139
NM_178597
Mef2c
100%
50%
0
R1
R2
Isoform
NM_001170537
NM_025282

Figure S11

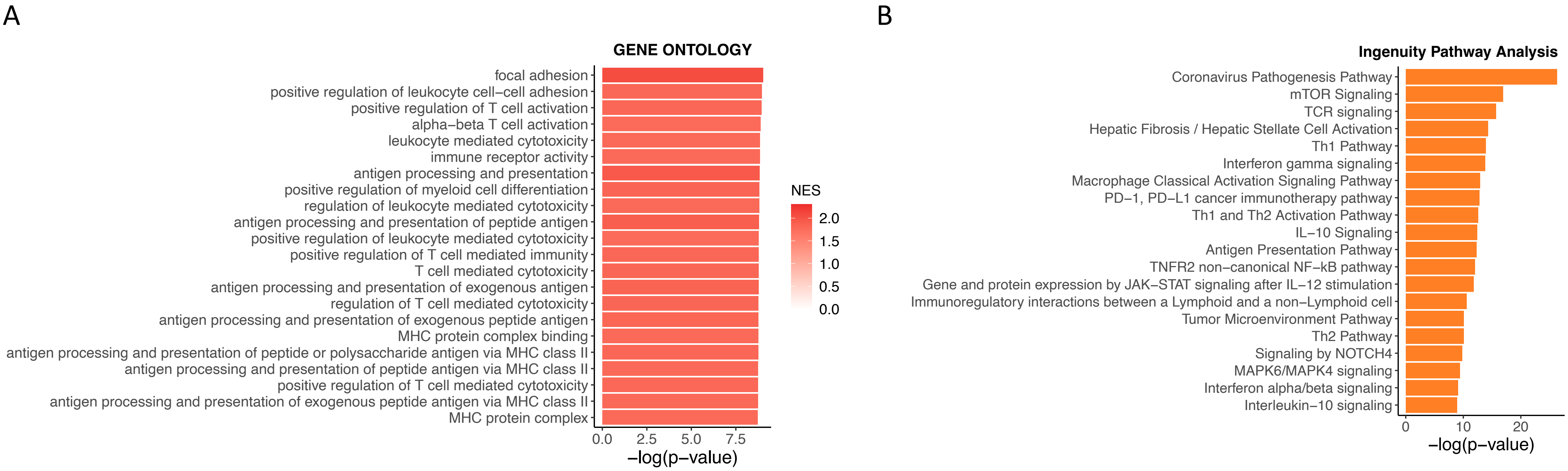

A
GENE ONTOLOGY
focal adhesion
positive regulation of leukocyte cell–cell adhesion
positive regulation of T cell activation
alpha–beta T cell activation
leukocyte mediated cytotoxicity
immune receptor activity
antigen processing and presentation
positive regulation of myeloid cell differentiation
regulation of leukocyte mediated cytotoxicity
antigen processing and presentation of peptide antigen
positive regulation of leukocyte mediated cytotoxicity
positive regulation of T cell mediated immunity
T cell mediated cytotoxicity
antigen processing and presentation of exogenous antigen
regulation of T cell mediated cytotoxicity
antigen processing and presentation of exogenous peptide antigen
MHC protein complex binding
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
antigen processing and presentation of peptide antigen via MHC class II
positive regulation of T cell mediated cytotoxicity
antigen processing and presentation of exogenous peptide antigen via MHC class II
MHC protein complex
0.0
2.5
5.0
7.5
–log(p–value)
NES
2.0
1.5
1.0
0.5
0.0
B
Ingenuity Pathway Analysis
Coronavirus Pathogenesis Pathway
mTOR Signaling
TCR signaling
Hepatic Fibrosis / Hepatic Stellate Cell Activation
Th1 Pathway
Interferon gamma signaling
Macrophage Classical Activation Signaling Pathway
PD–1, PD–L1 cancer immunotherapy pathway
Th1 and Th2 Activation Pathway
IL–10 Signaling
Antigen Presentation Pathway
TNFR2 non–canonical NF–kB pathway
Gene and protein expression by JAK–STAT signaling after IL–12 stimulation
Immunoregulatory interactions between a Lymphoid and a non–Lymphoid cell
Tumor Microenvironment Pathway
Th2 Pathway
Signaling by NOTCH4
MAPK6/MAPK4 signaling
Interferon alpha/beta signaling
Interleukin–10 signaling
0
10
20
–log(p–value)

Figure S12

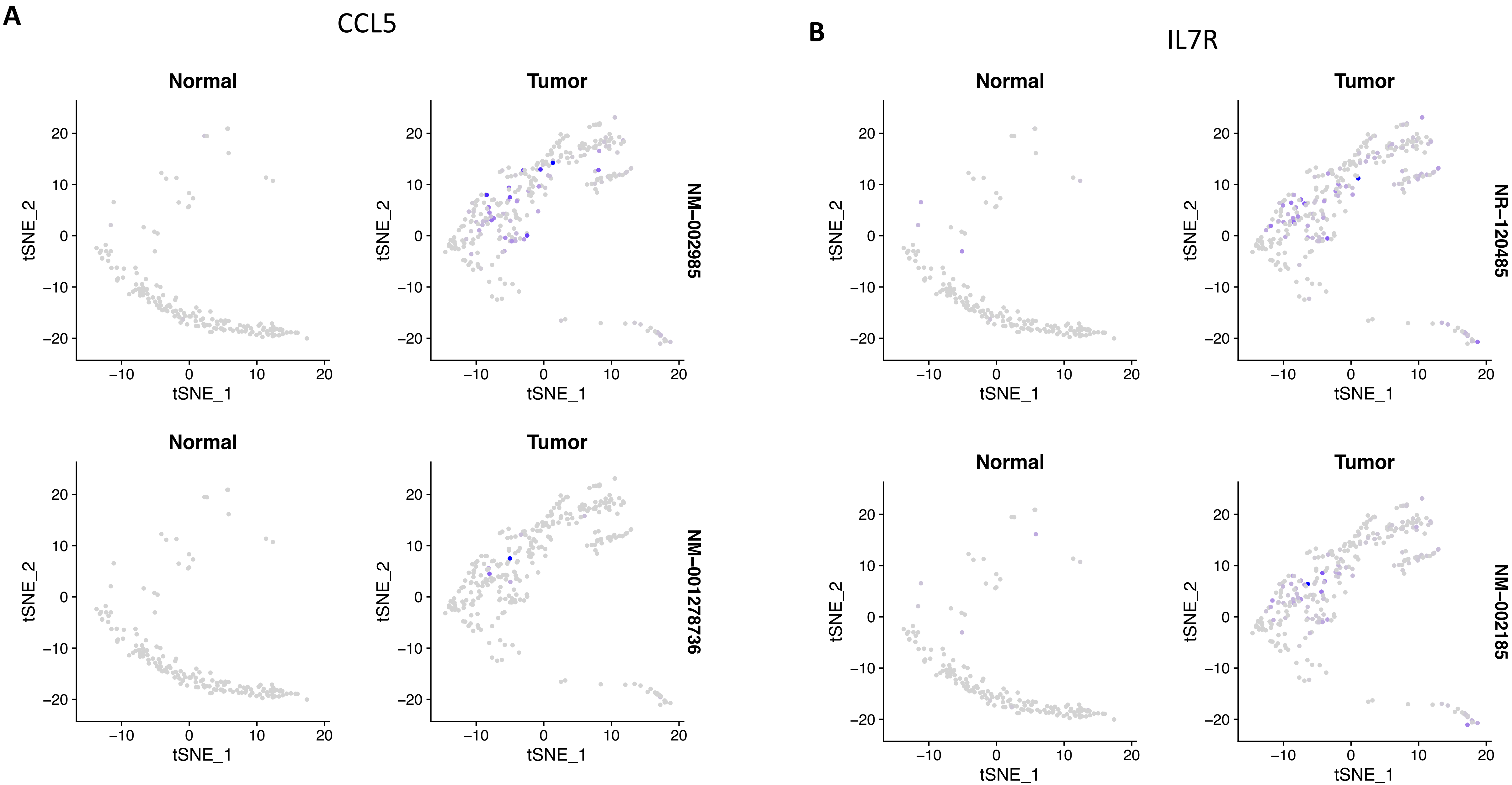

A
CCL5
Normal
Tumor
NM−002985
Normal
Tumor
NM−001278736
B
IL7R
Normal
Tumor
NR−120485
Normal
Tumor
NM−002185
tSNE_1
tSNE_2
−10
0
10
20
−20

Figure S13

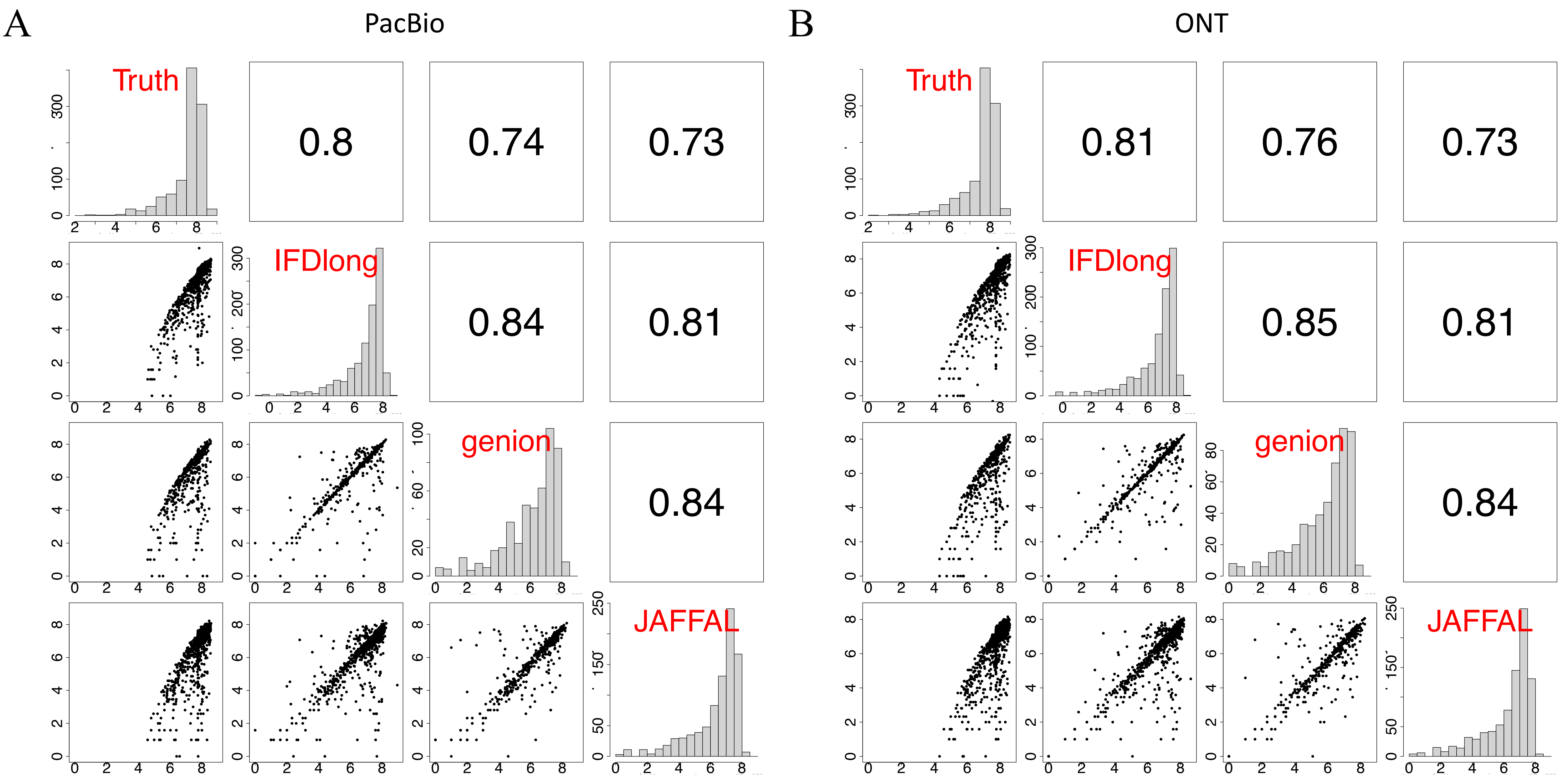

A
PacBio
Truth
0.8
0.74
0.73
IFDlong
0.84
0.81
genion
0.84
JAFFAL
B
ONT
Truth
0.81
0.76
0.73
IFDlong
0.85
0.81
genion
0.84
JAFFAL

Figure S14

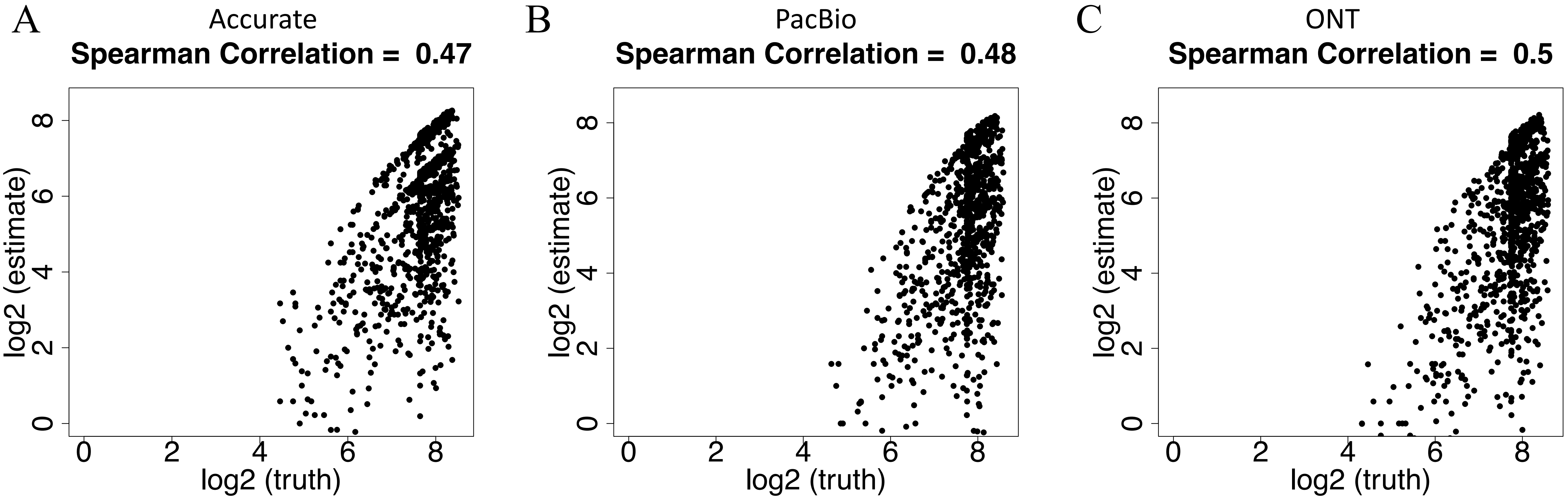

A
Accurate
Spearman Correlation = 0.47
B
PacBio
Spearman Correlation = 0.48
C
ONT
Spearman Correlation = 0.5
log2 (estimate)
0
2
4
6
8
log2 (truth)
0
2
4
6
8

Figure S15

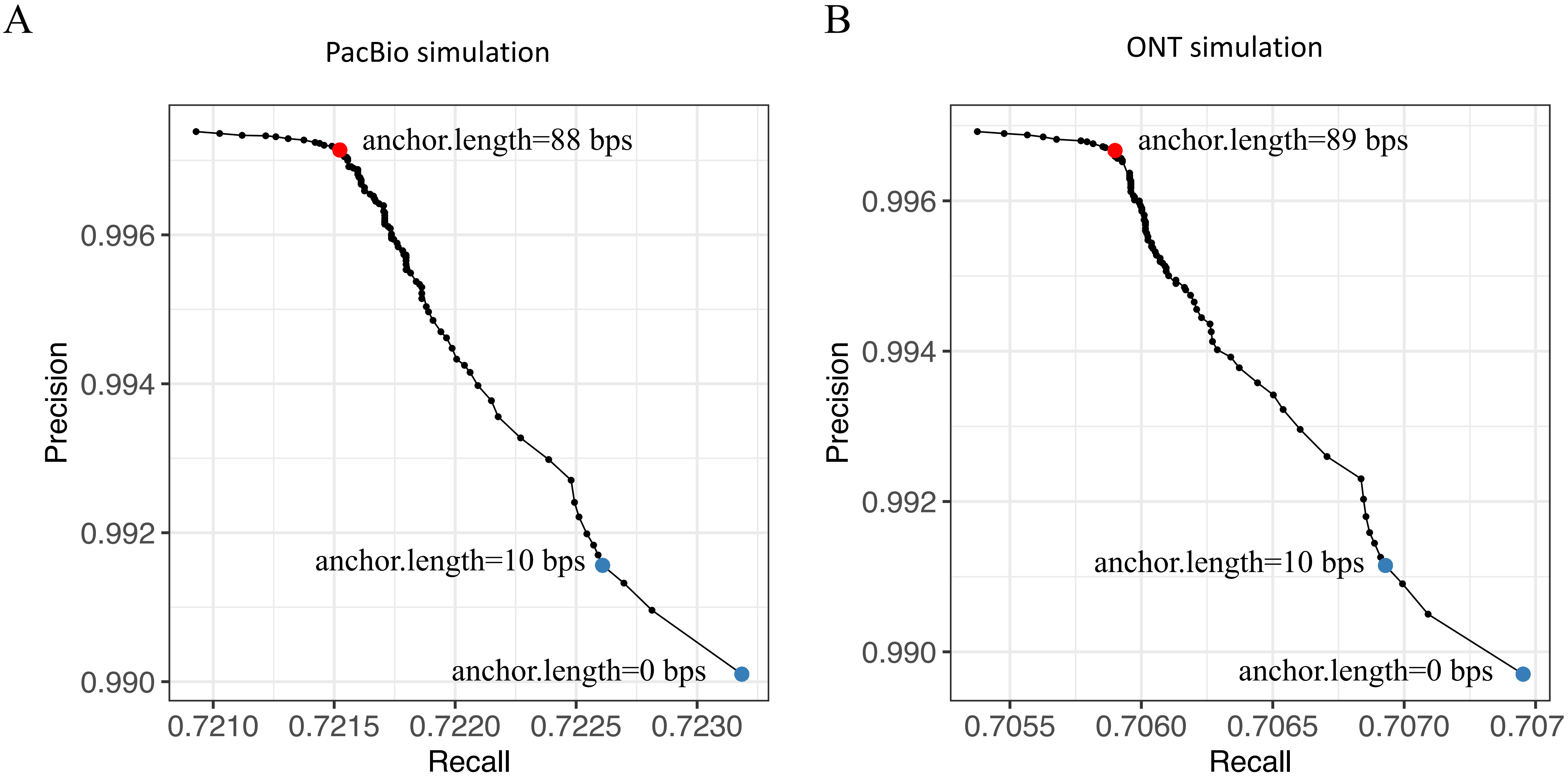

A
PacBio simulation
anchor.length=88 bps
anchor.length=10 bps
anchor.length=0 bps
Precision
0.996
0.994
0.992
0.990
Recall
0.7210
0.7215
0.7220
0.7225
0.7230
B
ONT simulation
anchor.length=89 bps
anchor.length=10 bps
anchor.length=0 bps
Precision
0.996
0.994
0.992
0.990
Recall
0.7055
0.7060
0.7065
0.7070
0.707

Figure S16

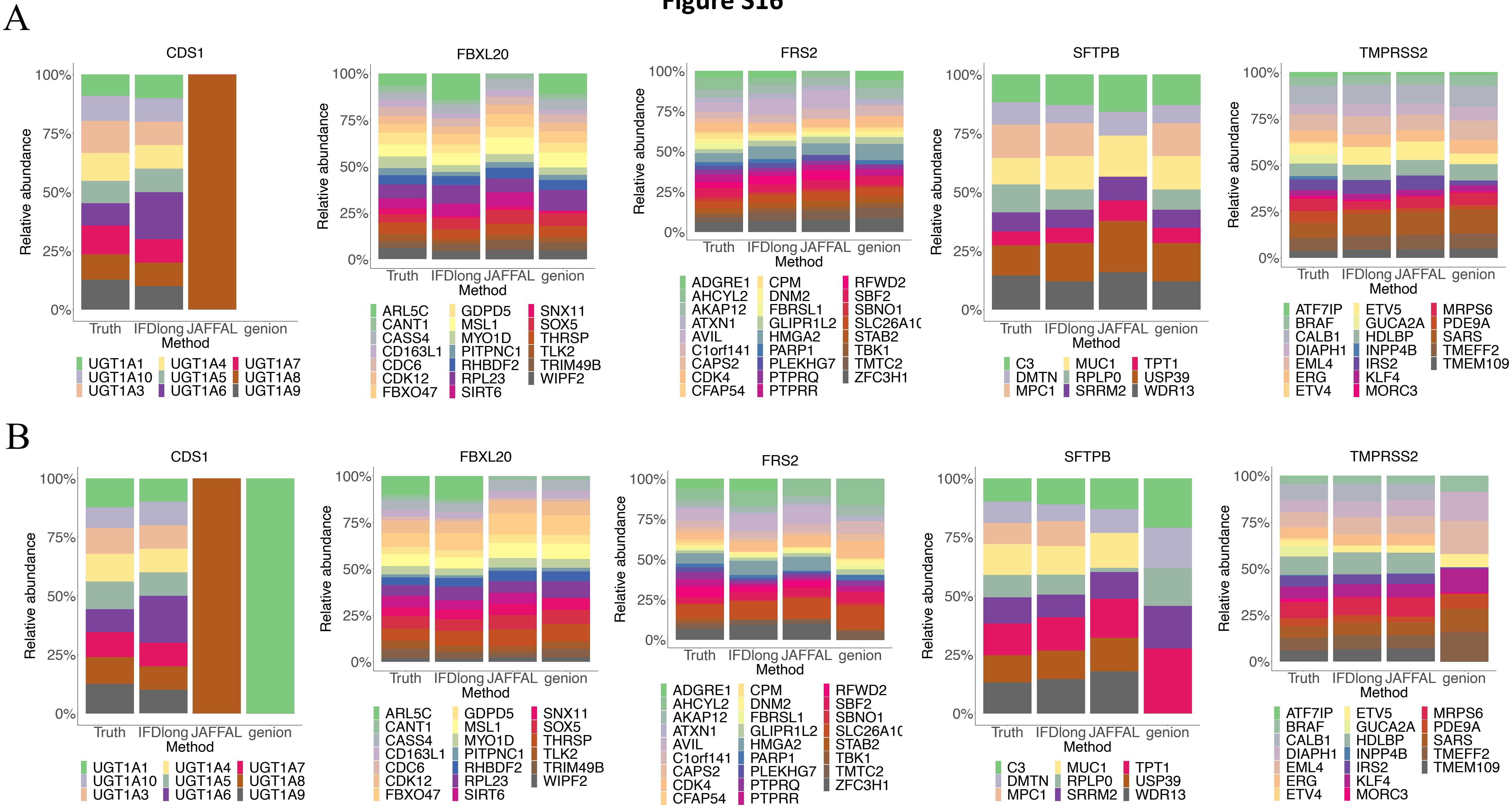

A
CDS1
FBXL20
FRS2
SFTPB
TMPRSS2
Relative abundance
100%
75%
50%
25%
0%
Truth
IFDlong
JAFFAL
genion
Method
UGT1A1
UGT1A10
UGT1A3
UGT1A4
UGT1A5
UGT1A6
UGT1A7
UGT1A8
UGT1A9
ARL5C
CANT1
CASS4
CD163L1
CDC6
CDK12
FBXO47
GDPD5
MSL1
MYO1D
PITPNC1
RHBDF2
RPL23
SIRT6
SNX11
SOX5
THRSP
TLK2
TRIM49B
WIPF2
ADGRE1
AHCYL2
AKAP12
ATXN1
AVIL
C1orf141
CAPS2
CDK4
CFAP54
CPM
DNM2
FBRSL1
GLIPR1L2
HMGA2
PARP1
PLEKHG7
PTPRQ
PTPRR
RFWD2
SBF2
SBNO1
SLC26A1
STAB2
TBK1
TMTC2
ZFC3H1
C3
DMTN
MPC1
MUC1
RPLP0
SRRM2
TPT1
USP39
WDR13
ATF7IP
BRAF
CALB1
DIAPH1
EML4
ERG
ETV4
ETV5
GUCA2A
HDLBP
INPP4B
IRS2
KLF4
MORC3
MRPS6
PDE9A
SARS
TMEFF2
TMEM109
B
CDS1
FBXL20
FRS2
SFTPB
TMPRSS2
Relative abundance
100%
75%
50%
25%
0%
Truth
IFDlong
JAFFAL
genion
Method
UGT1A1
UGT1A10
UGT1A3
UGT1A4
UGT1A5
UGT1A6
UGT1A7
UGT1A8
UGT1A9
ARL5C
CANT1
CASS4
CD163L1
CDC6
CDK12
FBXO47
GDPD5
MSL1
MYO1D
PITPNC1
RHBDF2
RPL23
SIRT6
SNX11
SOX5
THRSP
TLK2
TRIM49B
WIPF2
ADGRE1
AHCYL2
AKAP12
ATXN1
AVIL
C1orf141
CAPS2
CDK4
CFAP54
CPM
DNM2
FBRSL1
GLIPR1L2
HMGA2
PARP1
PLEKHG7
PTPRQ
PTPRR
RFWD2
SBF2
SBNO1
SLC26A1
STAB2
TBK1
TMTC2
ZFC3H1
C3
DMTN
MPC1
MUC1
RPLP0
SRRM2
TPT1
USP39
WDR13
ATF7IP
BRAF
CALB1
DIAPH1
EML4
ERG
ETV4
ETV5
GUCA2A
HDLBP
INPP4B
IRS2
KLF4
MORC3
MRPS6
PDE9A
SARS
TMEFF2
TMEM109

Figure S17

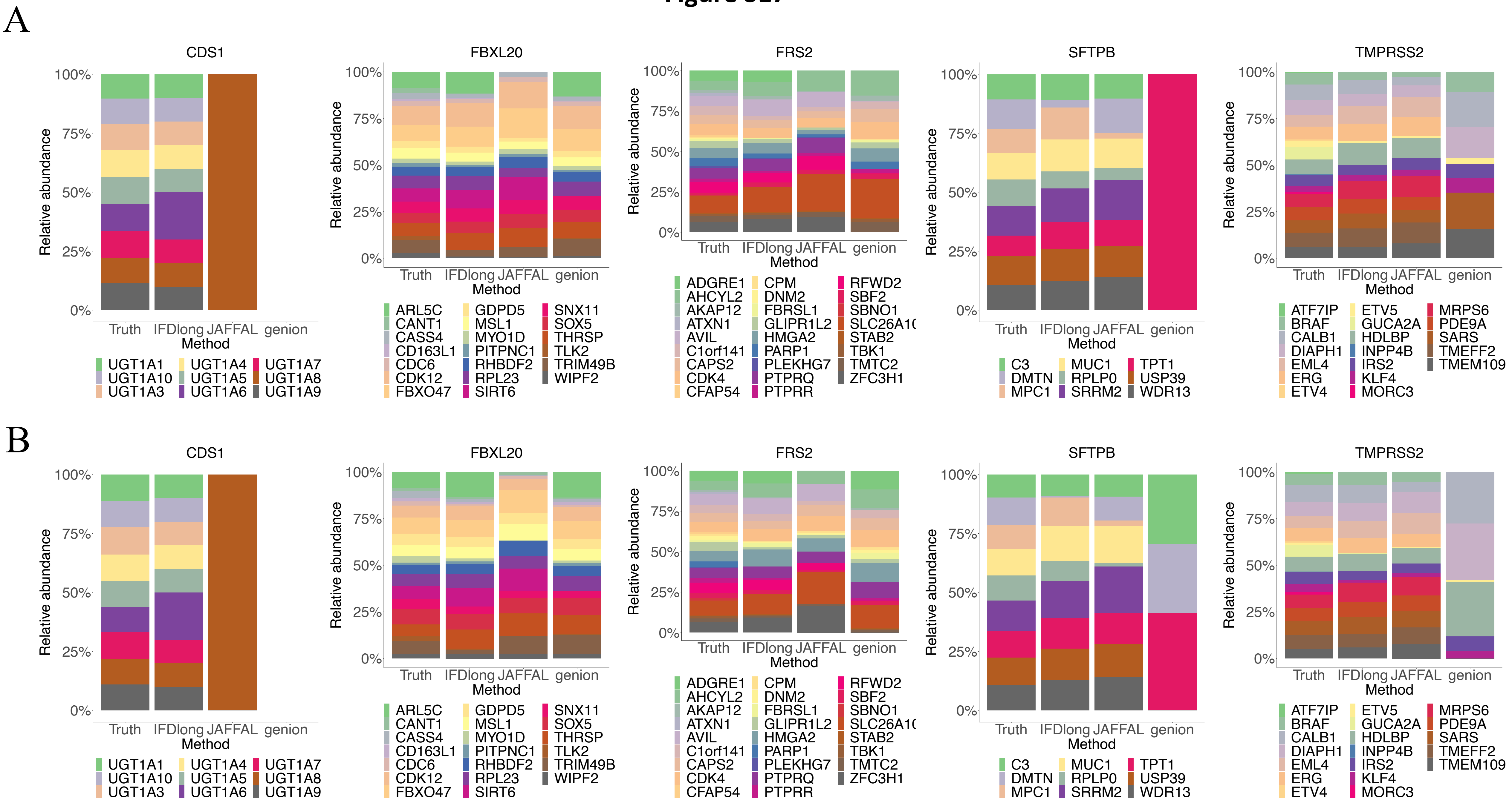
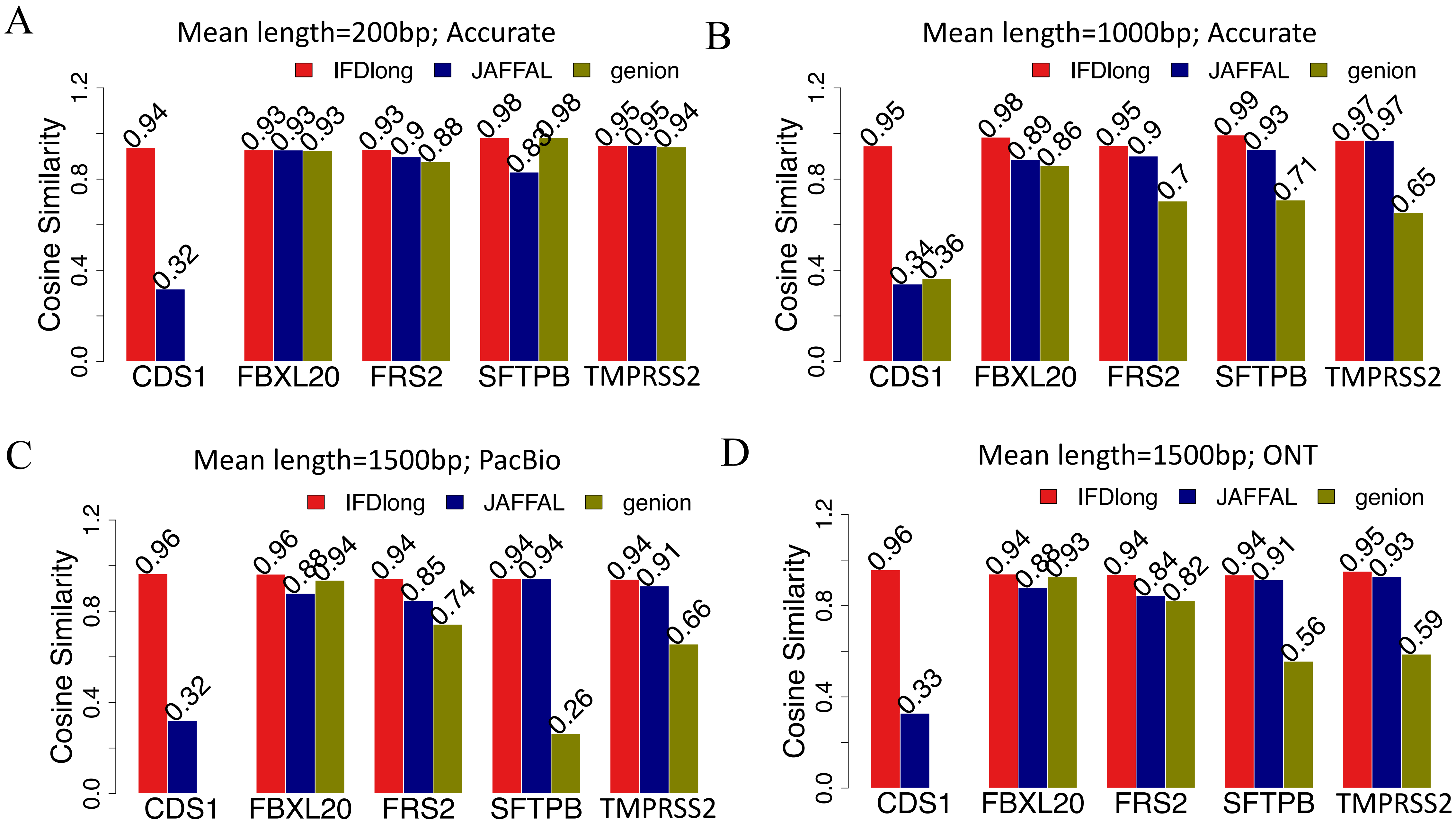

A
CDS1
FBXL20
FRS2
SFTPB
TMPRSS2
Relative abundance
100%
75%
50%
25%
0%
Truth
IFDlong
JAFFAL
genion
Method
UGT1A1
UGT1A10
UGT1A3
UGT1A4
UGT1A5
UGT1A6
UGT1A7
UGT1A8
UGT1A9
ARL5C
CANT1
CASS4
CD163L1
CDC6
CDK12
FBXO47
GDPD5
MSL1
MYO1D
PITPNC1
RHBDF2
RPL23
SIRT6
SNX11
SOX5
THRSP
TLK2
TRIM49B
WIPF2
ADGRE1
AHCYL2
AKAP12
ATXN1
AVIL
C1orf141
CAPS2
CDK4
CFAP54
CPM
DNM2
FBRSL1
GLIPR1L2
HMGA2
PARP1
PLEKHG7
PTPRQ
PTPRR
RFWD2
SBF2
SBNO1
SLC26A1
STAB2
TBK1
TMTC2
ZFC3H1
C3
DMTN
MPC1
MUC1
RPLP0
SRRM2
TPT1
USP39
WDR13
ATF7IP
BRAF
CALB1
DIAPH1
EML4
ERG
ETV4
ETV5
GUCA2A
HDLBP
INPP4B
IRS2
KLF4
MORC3
MRPS6
PDE9A
SARS
TMEFF2
TMEM109
B
CDS1
FBXL20
FRS2
SFTPB
TMPRSS2
Relative abundance
100%
75%
50%
25%
0%
Truth
IFDlong
JAFFAL
genion
Method
UGT1A1
UGT1A10
UGT1A3
UGT1A4
UGT1A5
UGT1A6
UGT1A7
UGT1A8
UGT1A9
ARL5C
CANT1
CASS4
CD163L1
CDC6
CDK12
FBXO47
GDPD5
MSL1
MYO1D
PITPNC1
RHBDF2
RPL23
SIRT6
SNX11
SOX5
THRSP
TLK2
TRIM49B
WIPF2
ADGRE1
AHCYL2
AKAP12
ATXN1
AVIL
C1orf141
CAPS2
CDK4
CFAP54
CPM
DNM2
FBRSL1
GLIPR1L2
HMGA2
PARP1
PLEKHG7
PTPRQ
PTPRR
RFWD2
SBF2
SBNO1
SLC26A1
STAB2
TBK1
TMTC2
ZFC3H1
C3
DMTN
MPC1
MUC1
RPLP0
SRRM2
TPT1
USP39
WDR13
ATF7IP
BRAF
CALB1
DIAPH1
EML4
ERG
ETV4
ETV5
GUCA2A
HDLBP
INPP4B
IRS2
KLF4
MORC3
MRPS6
PDE9A
SARS
TMEFF2
TMEM109

Figure S18


A
Mean length=200bp; Accurate
IFDlong
JAFFAL
genion
Cosine Similarity
0.0
0.4
0.8
1.2
0.94
0.32
0.93
0.93
0.93
0.93
0.9
0.88
0.98
0.83
0.98
0.95
0.95
0.94
CDS1
FBXL20
FRS2
SFTPB
TMPRSS2
B
Mean length=1000bp; Accurate
IFDlong
JAFFAL
genion
Cosine Similarity
0.0
0.4
0.8
1.2
0.95
0.34
0.36
0.98
0.89
0.86
0.95
0.9
0.7
0.99
0.93
0.71
0.97
0.97
0.65
CDS1
FBXL20
FRS2
SFTPB
TMPRSS2
C
Mean length=1500bp; PacBio
IFDlong
JAFFAL
genion
Cosine Similarity
0.0
0.4
0.8
1.2
0.96
0.32
0.96
0.88
0.94
0.94
0.85
0.74
0.94
0.94
0.26
0.94
0.91
0.66
CDS1
FBXL20
FRS2
SFTPB
TMPRSS2
D
Mean length=1500bp; ONT
IFDlong
JAFFAL
genion
Cosine Similarity
0.0
0.4
0.8
1.2
0.96
0.33
0.94
0.88
0.93
0.94
0.84
0.82
0.94
0.91
0.56
0.95
0.93
0.59
CDS1
FBXL20
FRS2
SFTPB
TMPRSS2