
Annotating the genome at single-nucleotide resolution with DNA foundation models

In the format provided by the
authors and unedited

Annotating the genome at single-nucleotide resolution with DNA foundation models

Bernardo P. de Almeida^{*,1}, Hugo Dalla-Torre^{*,1},

Guillaume Richard¹, Christopher Blum², Lorenz Hexemer², Maxence Gélard¹, Javier Mendoza-Revilla¹, Ziqi Tang^{1,3}, Frederikke I. Marin^{1,4}, David M. Emms¹, Priyanka Pandey², Stefan Laurent², Marie Lopez¹,

Alexandre Laterre¹, Maren Lang², Uğur Şahin², Karim Beguir¹ and Thomas Pierrot¹

^{*}Equal contributions, ¹InstaDeep, ²BioNTech, ³Cold Spring Harbor Laboratory, ⁴University of Copenhagen

Supplementary Tables

Model	Pre-trained DNA encoder	Number of parameters	Number of training sequences seen (M)
BPNet arch.	No	120k	53
BPNet arch. Large	No	29M	49
SpliceAI arch.	No	700k	125
SpliceAI arch. large	No	44M	61
SpliceAI arch. extra-large	No	573M	45
UNet one-hot	No	63M	14
UNet 1024-embedding	No	66M	20
UNet large one-hot	No	252M	18
SegmentNT-3kb (NTv1 human; 2.5B)	NTv1-1000G (2.5B)	2.6B	22
SegmentNT-3kb (random-init)	No	563M	79
SegmentNT-3kb (random-init; only head)	No	563M	18
SegmentNT-3kb (only head)	NTv2 500M	563M	12
SegmentNT-3kb	NTv2 500M	563M	20
SegmentEnformer	Enformer	405M	14
SegmentBorzoi	Borzoi	340M	14

Supplementary Table 1 | Baselines models training details

Model	3UTR	5UTR	CTCF-bound	enhancer tissue-invariant	enhancer tissue-specific	exon	intron
BPNet arch.	0.00 ($\pm$ 0.002)	0.17 ($\pm$ 0.003)	0.00 ($\pm$ 0.000)	0.00 ($\pm$ 0.000)	0.04 ($\pm$ 0.001)	0.24 ($\pm$ 0.002)	0.22 ($\pm$ 0.001)
BPNet arch. large	0.10 ($\pm$ 0.004)	0.26 ($\pm$ 0.004)	0.06 ($\pm$ 0.003)	0.12 ($\pm$ 0.003)	0.20 ($\pm$ 0.001)	0.27 ($\pm$ 0.001)	0.21 ($\pm$ 0.001)
SpliceAI arch.	0.27 ($\pm$ 0.009)	0.23 ($\pm$ 0.006)	-0.0 ($\pm$ 0.000)	0.00 ($\pm$ 0.002)	0.13 ($\pm$ 0.002)	0.38 ($\pm$ 0.003)	0.31 ($\pm$ 0.002)
SpliceAI arch. large	0.27 ($\pm$ 0.006)	0.30 ($\pm$ 0.006)	0.06 ($\pm$ 0.004)	0.17 ($\pm$ 0.008)	0.26 ($\pm$ 0.001)	0.37 ($\pm$ 0.005)	0.28 ($\pm$ 0.003)
SpliceAI arch. extra-large	0.25 ($\pm$ 0.006)	0.29 ($\pm$ 0.003)	0.01 ($\pm$ 0.005)	0.21 ($\pm$ 0.003)	0.20 ($\pm$ 0.001)	0.37 ($\pm$ 0.002)	0.20 ($\pm$ 0.003)
UNet	-0.0 ($\pm$ 0.001)	0.08 ($\pm$ 0.001)	0.00 ($\pm$ 0.000)	0.03 ($\pm$ 0.001)	0.02 ($\pm$ 0.000)	0.17 ($\pm$ 0.001)	0.18 ($\pm$ 0.001)
UNet 1024-embedding	0.00 ($\pm$ 0.000)	0.09 ($\pm$ 0.001)	0.00 ($\pm$ 0.001)	0.04 ($\pm$ 0.001)	0.03 ($\pm$ 0.000)	0.14 ($\pm$ 0.001)	0.17 ($\pm$ 0.001)
UNet large	0.01 ($\pm$ 0.000)	0.13 ($\pm$ 0.002)	0.02 ($\pm$ 0.000)	0.05 ($\pm$ 0.001)	0.09 ($\pm$ 0.000)	0.16 ($\pm$ 0.000)	0.12 ($\pm$ 0.001)
Random-Init (only head)	-0.0 ($\pm$ 0.000)	0.00 ($\pm$ 0.000)	0.00 ($\pm$ 0.000)	-0.0 ($\pm$ 0.000)	0.01 ($\pm$ 0.000)	0.03 ($\pm$ 0.001)	0.06 ($\pm$ 0.001)
Random-Init	0.14 ($\pm$ 0.007)	0.23 ($\pm$ 0.007)	0.00 ($\pm$ 0.002)	0.05 ($\pm$ 0.005)	0.14 ($\pm$ 0.002)	0.27 ($\pm$ 0.004)	0.20 ($\pm$ 0.004)
SegmentNT-3kb (only head)	0.34 ($\pm$ 0.004)	0.41 ($\pm$ 0.008)	0.02 ($\pm$ 0.001)	0.16 ($\pm$ 0.005)	0.18 ($\pm$ 0.002)	0.48 ($\pm$ 0.002)	0.28 ($\pm$ 0.001)
SegmentNT-3kb (NTv1 human; 2.5B)	0.42 ($\pm$ 0.015)	0.38 ($\pm$ 0.006)	0.08 ($\pm$ 0.006)	0.13 ($\pm$ 0.006)	0.20 ($\pm$ 0.002)	0.48 ($\pm$ 0.006)	0.29 ($\pm$ 0.003)
SegmentNT-3kb	0.51 ($\pm$ 0.015)	0.44 ($\pm$ 0.008)	0.08 ($\pm$ 0.004)	0.19 ($\pm$ 0.004)	0.25 ($\pm$ 0.001)	0.52 ($\pm$ 0.005)	0.30 ($\pm$ 0.003)
SegmentNT-3kb-single-task (x14)	0.49 ($\pm$ 0.013)	0.42 ($\pm$ 0.006)	0.09 ($\pm$ 0.002)	0.17 ($\pm$ 0.006)	0.29 ($\pm$ 0.001)	0.55 ($\pm$ 0.003)	0.33 ($\pm$ 0.003)
SegmentNT-10kb	0.66 ($\pm$ 0.007)	0.49 ($\pm$ 0.007)	0.09 ($\pm$ 0.003)	0.19 ($\pm$ 0.007)	0.27 ($\pm$ 0.001)	0.59 ($\pm$ 0.003)	0.40 ($\pm$ 0.004)
SegmentNT-20kb	0.69 ($\pm$ 0.009)	0.49 ($\pm$ 0.007)	0.09 ($\pm$ 0.005)	0.18 ($\pm$ 0.009)	0.27 ($\pm$ 0.002)	0.62 ($\pm$ 0.003)	0.47 ($\pm$ 0.008)
SegmentNT-30kb	0.70 ($\pm$ 0.007)	0.50 ($\pm$ 0.003)	0.08 ($\pm$ 0.002)	0.19 ($\pm$ 0.005)	0.27 ($\pm$ 0.002)	0.62 ($\pm$ 0.003)	0.52 ($\pm$ 0.005)
SegmentEnformer-30kb	0.65 ($\pm$ 0.008)	0.40 ($\pm$ 0.006)	0.17 ($\pm$ 0.006)	0.29 ($\pm$ 0.007)	0.44 ($\pm$ 0.002)	0.59 ($\pm$ 0.004)	0.49 ($\pm$ 0.011)
SegmentEnformer-196kb	0.68 ($\pm$ 0.039)	0.44 ($\pm$ 0.015)	0.19 ($\pm$ 0.014)	0.28 ($\pm$ 0.023)	0.42 ($\pm$ 0.005)	0.56 ($\pm$ 0.015)	0.60 ($\pm$ 0.017)
SegmentBorzo-30kb	0.68 ($\pm$ 0.011)	0.43 ($\pm$ 0.008)	0.12 ($\pm$ 0.007)	0.25 ($\pm$ 0.008)	0.42 ($\pm$ 0.001)	0.59 ($\pm$ 0.004)	0.46 ($\pm$ 0.007)
SegmentBorzo-524kb	0.85 ($\pm$ 0.127)	0.60 ($\pm$ 0.068)	0.16 ($\pm$ 0.023)	0.24 ($\pm$ 0.050)	0.37 ($\pm$ 0.022)	0.49 ($\pm$ 0.089)	0.52 ($\pm$ 0.059)

Model	lncRNA	polyA signal	promoter tissue-invariant	promoter tissue-specific	protein coding gene	splice acceptor	splice donor
BPNet arch.	0.00 ($\pm$ 0.000)	0.00 ($\pm$ 0.000)	0.26 ($\pm$ 0.005)	0.05 ($\pm$ 0.003)	0.30 ($\pm$ 0.001)	0.03 ($\pm$ 0.004)	0.05 ($\pm$ 0.002)
BPNet arch. large	0.00 ($\pm$ 0.000)	0.04 ($\pm$ 0.005)	0.40 ($\pm$ 0.005)	0.06 ($\pm$ 0.004)	0.29 ($\pm$ 0.001)	0.22 ($\pm$ 0.003)	0.29 ($\pm$ 0.002)
SpliceAI arch.	0.00 ($\pm$ 0.000)	0.02 ($\pm$ 0.006)	0.41 ($\pm$ 0.007)	0.05 ($\pm$ 0.008)	0.42 ($\pm$ 0.003)	0.09 ($\pm$ 0.005)	0.22 ($\pm$ 0.003)
SpliceAI arch. large	0.00 ($\pm$ 0.002)	0.16 ($\pm$ 0.006)	0.50 ($\pm$ 0.008)	0.13 ($\pm$ 0.013)	0.40 ($\pm$ 0.002)	0.38 ($\pm$ 0.003)	0.50 ($\pm$ 0.003)
SpliceAI arch. extra-large	0.01 ($\pm$ 0.002)	0.12 ($\pm$ 0.006)	0.53 ($\pm$ 0.007)	0.21 ($\pm$ 0.007)	0.20 ($\pm$ 0.001)	0.45 ($\pm$ 0.002)	0.52 ($\pm$ 0.003)
UNet	0.00 ($\pm$ 0.000)	0.00 ($\pm$ 0.004)	0.13 ($\pm$ 0.002)	0.01 ($\pm$ 0.001)	0.14 ($\pm$ 0.001)	0.08 ($\pm$ 0.002)	0.19 ($\pm$ 0.003)
UNet 1024-embedding	-0.0 ($\pm$ 0.000)	0.00 ($\pm$ 0.002)	0.14 ($\pm$ 0.003)	0.00 ($\pm$ 0.001)	0.13 ($\pm$ 0.001)	0.08 ($\pm$ 0.002)	0.20 ($\pm$ 0.002)
UNet large	-0.0 ($\pm$ 0.001)	0.02 ($\pm$ 0.002)	0.23 ($\pm$ 0.004)	0.01 ($\pm$ 0.000)	0.12 ($\pm$ 0.001)	0.22 ($\pm$ 0.002)	0.28 ($\pm$ 0.002)
Random-Init (only head)	-0.0 ($\pm$ 0.000)	-0.0 ($\pm$ 0.001)	0.00 ($\pm$ 0.001)	0.00 ($\pm$ 0.000)	0.06 ($\pm$ 0.000)	0.00 ($\pm$ 0.001)	-0.0 ($\pm$ 0.002)
Random-Init	0.04 ($\pm$ 0.006)	0.09 ($\pm$ 0.006)	0.40 ($\pm$ 0.011)	0.13 ($\pm$ 0.013)	0.26 ($\pm$ 0.004)	0.09 ($\pm$ 0.003)	0.19 ($\pm$ 0.004)
SegmentNT-3kb (only head)	0.01 ($\pm$ 0.001)	0.25 ($\pm$ 0.011)	0.54 ($\pm$ 0.006)	0.20 ($\pm$ 0.006)	0.40 ($\pm$ 0.001)	0.67 ($\pm$ 0.003)	0.70 ($\pm$ 0.002)
SegmentNT-3kb (NTv1 human; 2.5B)	0.04 ($\pm$ 0.004)	0.28 ($\pm$ 0.006)	0.56 ($\pm$ 0.007)	0.22 ($\pm$ 0.010)	0.41 ($\pm$ 0.002)	0.62 ($\pm$ 0.002)	0.64 ($\pm$ 0.003)
SegmentNT-3kb	0.05 ($\pm$ 0.004)	0.30 ($\pm$ 0.006)	0.58 ($\pm$ 0.006)	0.22 ($\pm$ 0.011)	0.42 ($\pm$ 0.003)	0.69 ($\pm$ 0.004)	0.70 ($\pm$ 0.002)
SegmentNT-3kb-single-task (x14)	0.02 ($\pm$ 0.004)	0.34 ($\pm$ 0.005)	0.51 ($\pm$ 0.005)	0.16 ($\pm$ 0.006)	0.43 ($\pm$ 0.004)	0.71 ($\pm$ 0.003)	0.71 ($\pm$ 0.003)
SegmentNT-10kb	0.06 ($\pm$ 0.013)	0.36 ($\pm$ 0.007)	0.58 ($\pm$ 0.009)	0.23 ($\pm$ 0.009)	0.58 ($\pm$ 0.008)	0.71 ($\pm$ 0.006)	0.73 ($\pm$ 0.005)
SegmentNT-20kb	0.07 ($\pm$ 0.012)	0.38 ($\pm$ 0.008)	0.59 ($\pm$ 0.009)	0.25 ($\pm$ 0.007)	0.66 ($\pm$ 0.006)	0.71 ($\pm$ 0.003)	0.73 ($\pm$ 0.003)
SegmentNT-30kb	0.06 ($\pm$ 0.014)	0.40 ($\pm$ 0.004)	0.58 ($\pm$ 0.007)	0.23 ($\pm$ 0.006)	0.70 ($\pm$ 0.003)	0.73 ($\pm$ 0.002)	0.74 ($\pm$ 0.003)
SegmentEnformer-30kb	0.13 ($\pm$ 0.011)	0.00 ($\pm$ 0.000)	0.65 ($\pm$ 0.011)	0.33 ($\pm$ 0.008)	0.68 ($\pm$ 0.009)	0.00 ($\pm$ 0.000)	0.00 ($\pm$ 0.000)
SegmentEnformer-196kb	0.20 ($\pm$ 0.025)	0.01 ($\pm$ 0.015)	0.59 ($\pm$ 0.042)	0.26 ($\pm$ 0.025)	0.83 ($\pm$ 0.018)	0.00 ($\pm$ 0.000)	0.00 ($\pm$ 0.000)
SegmentBorzo-30kb	0.17 ($\pm$ 0.011)	0.18 ($\pm$ 0.007)	0.64 ($\pm$ 0.008)	0.29 ($\pm$ 0.006)	0.65 ($\pm$ 0.007)	0.00 ($\pm$ 0.012)	0.00 ($\pm$ 0.006)
SegmentBorzo-524kb	0.19 ($\pm$ 0.121)	0.16 ($\pm$ 0.031)	0.28 ($\pm$ 0.306)	0.33 ($\pm$ 0.064)	0.88 ($\pm$ 0.064)	0.04 ($\pm$ 0.046)	0.08 ($\pm$ 0.035)

Supplementary Table 2 | Performance of all models across the 14 genomic elements (MCC). We used as metric the Matthews correlation coefficient (MCC). Data are presented as mean MCC values $\pm$ standard deviation from 10 different samplings of the test set.

Model	3UTR	5UTR	CTCF-bound	enhancer tissue-invariant	enhancer tissue-specific	exon	intron
BPNet arch.	0.06 (± 0.002)	0.15 (± 0.002)	0.01 (± 0.000)	0.04 (± 0.001)	0.23 (± 0.001)	0.24 (± 0.002)	0.62 (± 0.001)
BPNet arch. large	0.07 (± 0.002)	0.21 (± 0.003)	0.06 (± 0.001)	0.11 (± 0.002)	0.30 (± 0.000)	0.26 (± 0.001)	0.62 (± 0.001)
SpliceAI arch.	0.23 (± 0.006)	0.21 (± 0.004)	0.02 (± 0.000)	0.05 (± 0.001)	0.29 (± 0.001)	0.37 (± 0.003)	0.69 (± 0.001)
SpliceAI arch. large	0.22 (± 0.007)	0.21 (± 0.006)	0.07 (± 0.002)	0.12 (± 0.003)	0.33 (± 0.001)	0.35 (± 0.005)	0.67 (± 0.001)
SpliceAI arch. extra-large	0.18 (± 0.006)	0.21 (± 0.003)	0.07 (± 0.001)	0.14 (± 0.002)	0.31 (± 0.001)	0.34 (± 0.002)	0.59 (± 0.001)
UNet	0.02 (± 0.000)	0.08 (± 0.001)	0.02 (± 0.000)	0.03 (± 0.000)	0.17 (± 0.001)	0.15 (± 0.001)	0.52 (± 0.001)
UNet 1024-embedding	0.01 (± 0.000)	0.05 (± 0.001)	0.01 (± 0.000)	0.03 (± 0.000)	0.17 (± 0.000)	0.12 (± 0.001)	0.51 (± 0.001)
UNet large	0.01 (± 0.000)	0.06 (± 0.001)	0.01 (± 0.000)	0.02 (± 0.000)	0.17 (± 0.000)	0.13 (± 0.001)	0.51 (± 0.001)
Random-Init (only head)	0.01 (± 0.000)	0.00 (± 0.000)	0.01 (± 0.000)	0.00 (± 0.000)	0.08 (± 0.000)	0.05 (± 0.000)	0.48 (± 0.000)
Random-Init	0.08 (± 0.003)	0.13 (± 0.004)	0.01 (± 0.000)	0.02 (± 0.001)	0.18 (± 0.001)	0.23 (± 0.004)	0.59 (± 0.002)
SegmentNT-3kb (only head)	0.29 (± 0.003)	0.33 (± 0.006)	0.03 (± 0.000)	0.10 (± 0.002)	0.27 (± 0.001)	0.44 (± 0.001)	0.68 (± 0.001)
SegmentNT-3kb (NTv1 human; 2.5B)	0.36 (± 0.014)	0.31 (± 0.005)	0.04 (± 0.002)	0.10 (± 0.003)	0.30 (± 0.002)	0.45 (± 0.005)	0.69 (± 0.001)
SegmentNT-3kb	0.43 (± 0.010)	0.38 (± 0.007)	0.03 (± 0.001)	0.10 (± 0.003)	0.30 (± 0.001)	0.48 (± 0.005)	0.71 (± 0.001)
SegmentNT-3kb-single-task (x14)	0.46 (± 0.005)	0.36 (± 0.005)	0.03 (± 0.001)	0.10 (± 0.003)	0.32 (± 0.001)	0.53 (± 0.002)	0.73 (± 0.001)
SegmentNT-10kb	0.64 (± 0.008)	0.43 (± 0.009)	0.04 (± 0.001)	0.10 (± 0.004)	0.31 (± 0.001)	0.56 (± 0.003)	0.79 (± 0.003)
SegmentNT-20kb	0.70 (± 0.008)	0.44 (± 0.007)	0.04 (± 0.001)	0.09 (± 0.003)	0.30 (± 0.001)	0.58 (± 0.003)	0.81 (± 0.003)
SegmentNT-30kb	0.70 (± 0.006)	0.44 (± 0.004)	0.04 (± 0.001)	0.10 (± 0.004)	0.29 (± 0.002)	0.58 (± 0.004)	0.84 (± 0.002)
SegmentEnformer-30kb	0.67 (± 0.011)	0.37 (± 0.006)	0.16 (± 0.003)	0.27 (± 0.005)	0.53 (± 0.003)	0.57 (± 0.006)	0.83 (± 0.004)
SegmentEnformer-196kb	0.65 (± 0.051)	0.37 (± 0.020)	0.09 (± 0.008)	0.16 (± 0.020)	0.45 (± 0.006)	0.50 (± 0.020)	0.88 (± 0.010)
SegmentBorzo-30kb	0.67 (± 0.010)	0.39 (± 0.010)	0.15 (± 0.003)	0.22 (± 0.005)	0.50 (± 0.003)	0.60 (± 0.005)	0.83 (± 0.002)
SegmentBorzo-524kb	0.90 (± 0.103)	0.55 (± 0.093)	0.08 (± 0.010)	0.13 (± 0.038)	0.38 (± 0.025)	0.41 (± 0.096)	0.81 (± 0.055)

Model	lncRNA	polyA signal	promoter tissue-invariant	promoter tissue-specific	protein coding gene	splice acceptor	splice donor
BPNet arch.	0.22 (± 0.001)	0.00 (± 0.000)	0.29 (± 0.006)	0.10 (± 0.002)	0.59 (± 0.001)	0.02 (± 0.000)	0.02 (± 0.000)
BPNet arch. large	0.21 (± 0.001)	0.03 (± 0.001)	0.45 (± 0.007)	0.12 (± 0.004)	0.57 (± 0.001)	0.21 (± 0.001)	0.28 (± 0.001)
SpliceAI arch.	0.23 (± 0.001)	0.01 (± 0.000)	0.43 (± 0.009)	0.11 (± 0.002)	0.69 (± 0.001)	0.06 (± 0.002)	0.16 (± 0.004)
SpliceAI arch. large	0.18 (± 0.001)	0.09 (± 0.003)	0.49 (± 0.010)	0.11 (± 0.007)	0.67 (± 0.001)	0.35 (± 0.004)	0.47 (± 0.003)
SpliceAI arch. extra-large	0.20 (± 0.001)	0.06 (± 0.003)	0.49 (± 0.012)	0.15 (± 0.004)	0.54 (± 0.002)	0.41 (± 0.001)	0.49 (± 0.002)
UNet	0.20 (± 0.001)	0.01 (± 0.000)	0.12 (± 0.003)	0.06 (± 0.001)	0.44 (± 0.002)	0.07 (± 0.001)	0.15 (± 0.003)
UNet 1024-embedding	0.20 (± 0.001)	0.00 (± 0.000)	0.12 (± 0.004)	0.04 (± 0.001)	0.43 (± 0.001)	0.06 (± 0.001)	0.15 (± 0.002)
UNet large	0.20 (± 0.001)	0.01 (± 0.000)	0.16 (± 0.004)	0.04 (± 0.001)	0.42 (± 0.001)	0.14 (± 0.002)	0.21 (± 0.002)
Random-Init (only head)	0.17 (± 0.000)	0.00 (± 0.000)	0.00 (± 0.000)	0.00 (± 0.000)	0.37 (± 0.000)	0.00 (± 0.000)	0.00 (± 0.000)
Random-Init	0.21 (± 0.002)	0.03 (± 0.002)	0.35 (± 0.015)	0.07 (± 0.008)	0.55 (± 0.003)	0.06 (± 0.002)	0.10 (± 0.003)
SegmentNT-3kb (only head)	0.22 (± 0.001)	0.16 (± 0.007)	0.55 (± 0.006)	0.13 (± 0.005)	0.67 (± 0.001)	0.58 (± 0.004)	0.63 (± 0.003)
SegmentNT-3kb (NTv1 human; 2.5B)	0.21 (± 0.002)	0.19 (± 0.009)	0.55 (± 0.007)	0.17 (± 0.006)	0.69 (± 0.001)	0.57 (± 0.002)	0.60 (± 0.003)
SegmentNT-3kb	0.21 (± 0.001)	0.19 (± 0.005)	0.55 (± 0.014)	0.14 (± 0.007)	0.72 (± 0.002)	0.62 (± 0.003)	0.65 (± 0.003)
SegmentNT-3kb-single-task (x14)	0.22 (± 0.001)	0.26 (± 0.004)	0.50 (± 0.008)	0.11 (± 0.004)	0.72 (± 0.001)	0.68 (± 0.002)	0.69 (± 0.002)
SegmentNT-10kb	0.23 (± 0.007)	0.28 (± 0.006)	0.57 (± 0.012)	0.15 (± 0.007)	0.83 (± 0.004)	0.65 (± 0.006)	0.67 (± 0.005)
SegmentNT-20kb	0.23 (± 0.005)	0.30 (± 0.006)	0.58 (± 0.013)	0.16 (± 0.006)	0.86 (± 0.002)	0.64 (± 0.003)	0.66 (± 0.003)
SegmentNT-30kb	0.25 (± 0.003)	0.31 (± 0.005)	0.55 (± 0.012)	0.15 (± 0.004)	0.89 (± 0.001)	0.65 (± 0.002)	0.66 (± 0.002)
SegmentEnformer-30kb	0.27 (± 0.008)	0.01 (± 0.001)	0.66 (± 0.012)	0.31 (± 0.011)	0.86 (± 0.005)	0.01 (± 0.000)	0.01 (± 0.000)
SegmentEnformer-196kb	0.35 (± 0.017)	0.02 (± 0.003)	0.55 (± 0.041)	0.17 (± 0.016)	0.93 (± 0.009)	0.02 (± 0.001)	0.02 (± 0.001)
SegmentBorzo-30kb	0.29 (± 0.008)	0.15 (± 0.006)	0.66 (± 0.007)	0.26 (± 0.006)	0.86 (± 0.003)	0.10 (± 0.002)	0.12 (± 0.003)
SegmentBorzo-524kb	0.40 (± 0.100)	0.10 (± 0.038)	0.46 (± 0.305)	0.22 (± 0.075)	0.98 (± 0.021)	0.08 (± 0.032)	0.10 (± 0.024)

Supplementary Table 3 | Performance of all models across the 14 genomic elements (auPRC). We used as metric the area under the precision recall curve (auPRC). Data are presented as mean auPRC values +/- standard deviation from 10 different samplings of the test set.

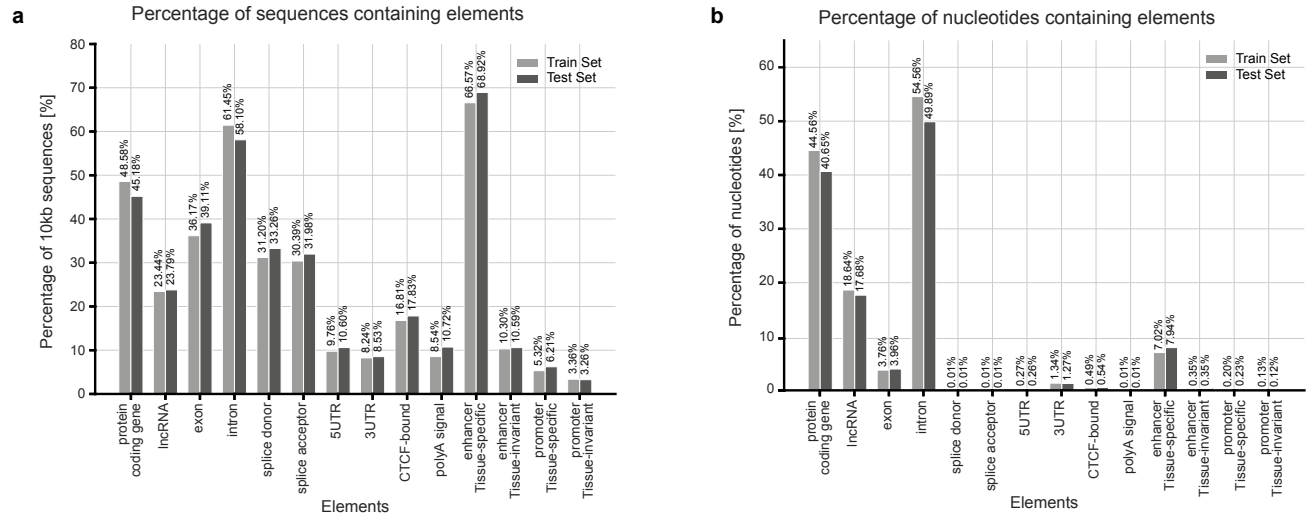
Category	3UTR	5UTR	CTCF-bound	enhancer tissue-invariant
Total edge nucleotides	103780 (0.14%)	152234 (0.21%)	330749 (0.45%)	170691 (0.23%)
Total interior nucleotides	681711 (0.93%)	74139 (0.10%)	251660 (0.34%)	154063 (0.21%)
Total spurious nucleotides	72174509 (98.92%)	72733627 (99.69%)	72377591 (99.20%)	72635246 (99.55%)
Total mispredictions	392336	111869	478149	266853
Edge false positives	5233 (1.33%)	6465 (5.78%)	1313 (0.27%)	2527 (0.95%)
Edge false negatives	23686 (6.04%)	40583 (36.28%)	161933 (33.87%)	82367 (30.87%)
Edge mispredictions	28919 (7.37%)	47048 (42.06%)	163246 (34.14%)	84894 (31.81%)
Enrichment (edge mispred.)	51.82	201.56	75.31	135.98
Interior false positives	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Interior false negatives	287100 (73.18%)	48142 (43.03%)	239029 (49.99%)	135159 (50.65%)
Interior mispredictions	287100 (73.18%)	48142 (43.03%)	239029 (49.99%)	135159 (50.65%)
Enrichment (interior mispred.)	78.32	423.50	144.93	239.86
Spurious false positives	76317 (19.45%)	16679 (14.91%)	75874 (15.87%)	46800 (17.54%)
Spurious false negatives	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Spurious mispredictions	76317 (19.45%)	16679 (14.91%)	75874 (15.87%)	46800 (17.54%)
Enrichment (spurious mispred.)	0.20	0.15	0.16	0.18
chi-squared test	512898.46 (p<2.2e-16)	164319.08 (p<2.2e-16)	679545.79 (p<2.2e-16)	369820.60 (p<2.2e-16)
Precision	0.84	0.69	0.18	0.33
Recall	0.58	0.36	0.04	0.10

Category	enhancer tissue-specific	exon	intron	lncRNA
Total edge nucleotides	3761614 (5.16%)	1184147 (1.62%)	874761 (1.20%)	143190 (0.20%)
Total interior nucleotides	3457869 (4.74%)	1888100 (2.59%)	33276862 (45.61%)	13061758 (17.90%)
Total spurious nucleotides	65740517 (90.10%)	69887753 (95.79%)	38808377 (53.19%)	59755052 (81.90%)
Total mispredictions	6243029	1518770	17587045	14000883
Edge false positives	230720 (3.70%)	15464 (1.02%)	40223 (0.23%)	2527 (0.02%)
Edge false negatives	1587688 (25.43%)	242582 (15.97%)	120283 (0.68%)	85752 (0.61%)
Edge mispredictions	1818408 (29.13%)	258046 (16.99%)	160506 (0.91%)	88279 (0.63%)
Enrichment (edge mispred.)	5.65	10.47	0.76	3.21
Interior false positives	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Interior false negatives	2481513 (39.75%)	1091074 (71.84%)	11982359 (68.13%)	12151517 (86.79%)
Interior mispredictions	2481513 (39.75%)	1091074 (71.84%)	11982359 (68.13%)	12151517 (86.79%)
Enrichment (interior mispred.)	8.39	27.76	1.49	4.85
Spurious false positives	1943108 (31.12%)	169650 (11.17%)	5444180 (30.96%)	1761087 (12.58%)
Spurious false negatives	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Spurious mispredictions	1943108 (31.12%)	169650 (11.17%)	5444180 (30.96%)	1761087 (12.58%)
Enrichment (spurious mispred.)	0.35	0.12	0.58	0.15
chi-squared test	4557807.85 (p<2.2e-16)	2188059.47 (p<2.2e-16)	1824507.34 (p<2.2e-16)	13499796.26 (p<2.2e-16)
Precision	0.39	0.86	0.80	0.34
Recall	0.25	0.46	0.64	0.07

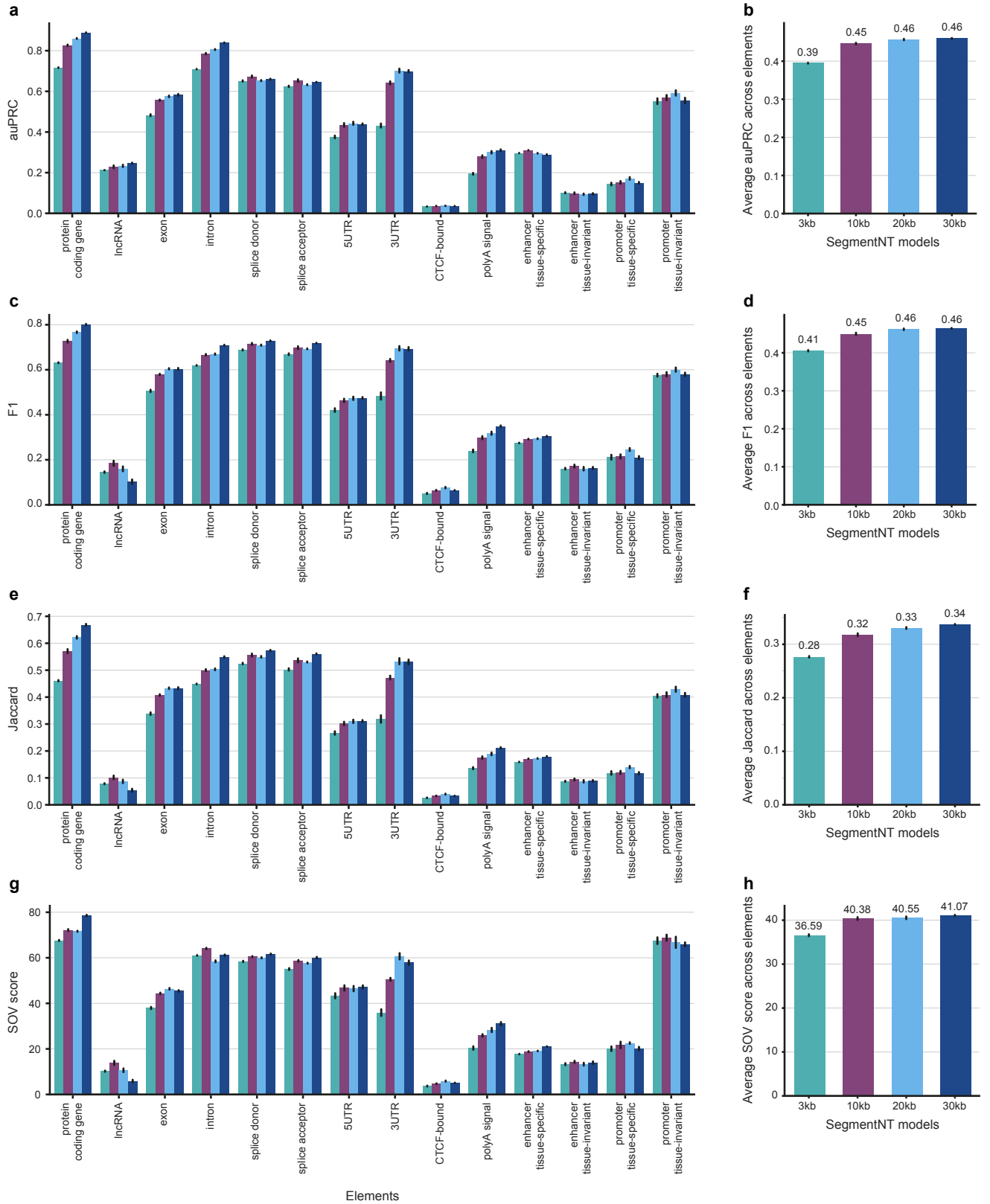
Category	promoter tissue-invariant	promoter tissue-specific	protein coding gene
Total edge nucleotides	40205 (0.06%)	92606 (0.13%)	158375 (0.22%)
Total interior nucleotides	43686 (0.06%)	84289 (0.12%)	25547145 (35.02%)
Total spurious nucleotides	72876109 (99.89%)	72783105 (99.76%)	47254480 (64.77%)
Total mispredictions	51205	148717	10051703
Edge false positives	4486 (8.76%)	3380 (2.27%)	12269 (0.12%)
Edge false negatives	11853 (23.15%)	42731 (28.73%)	28588 (0.28%)
Edge mispredictions	16339 (31.91%)	46111 (31.01%)	40857 (0.41%)
Enrichment (edge mispred.)	579.05	244.28	1.87
Interior false positives	0 (0.00%)	0 (0.00%)	0 (0.00%)
Interior false negatives	17789 (34.74%)	70281 (47.26%)	5016481 (49.91%)
Interior mispredictions	17789 (34.74%)	70281 (47.26%)	5016481 (49.91%)
Enrichment (interior mispred.)	580.21	409.06	1.43
Spurious false positives	17077 (33.35%)	32325 (21.74%)	4994365 (49.69%)
Spurious false negatives	0 (0.00%)	0 (0.00%)	0 (0.00%)
Spurious mispredictions	17077 (33.35%)	32325 (21.74%)	4994365 (49.69%)
Enrichment (spurious mispred.)	0.33	0.22	0.77
chi-squared test	50965.14 (p<2.2e-16)	189829.02 (p<2.2e-16)	468000.97 (p<2.2e-16)
Precision	0.62	0.36	0.80
Recall	0.54	0.15	0.80

Supplementary Table 4 | Statistics of types of mispredicted nucleotides by the model SegmentNT-30kb. For each element type, we splatted all nucleotides from the test set according to their distance to the boundaries of the ground-truth regions: a nucleotide is considered part of the edge if closer than 50nt from a region boundary (edge), part of the interior of a region if further way but inside a positive region (interior), and part of spurious regions otherwise. We include the chi-squared test statistics and a p-value for the three classes based on their expected frequencies. Splice sites and polyA sites were not included as they are too short to define edges.

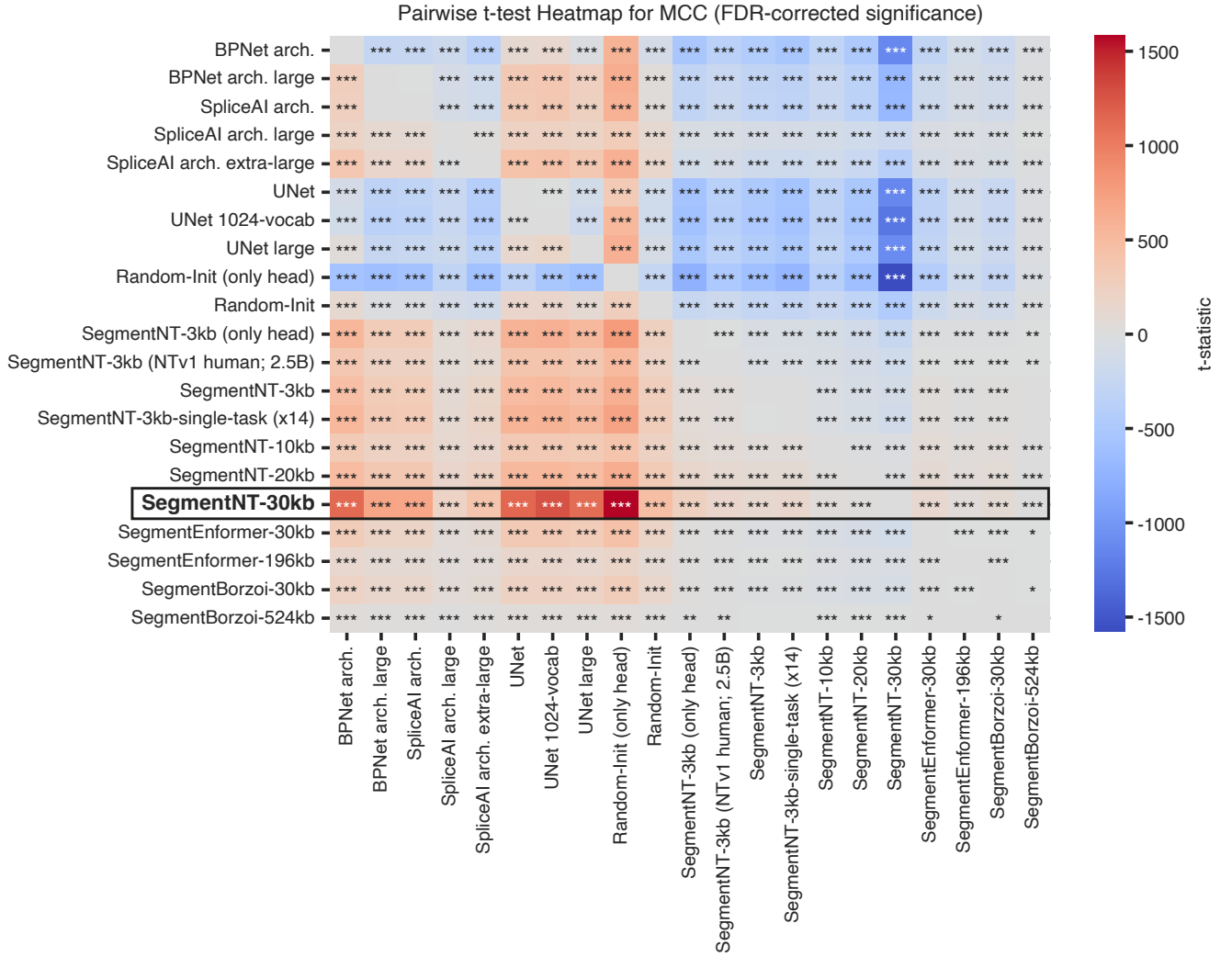
Supplementary Figures



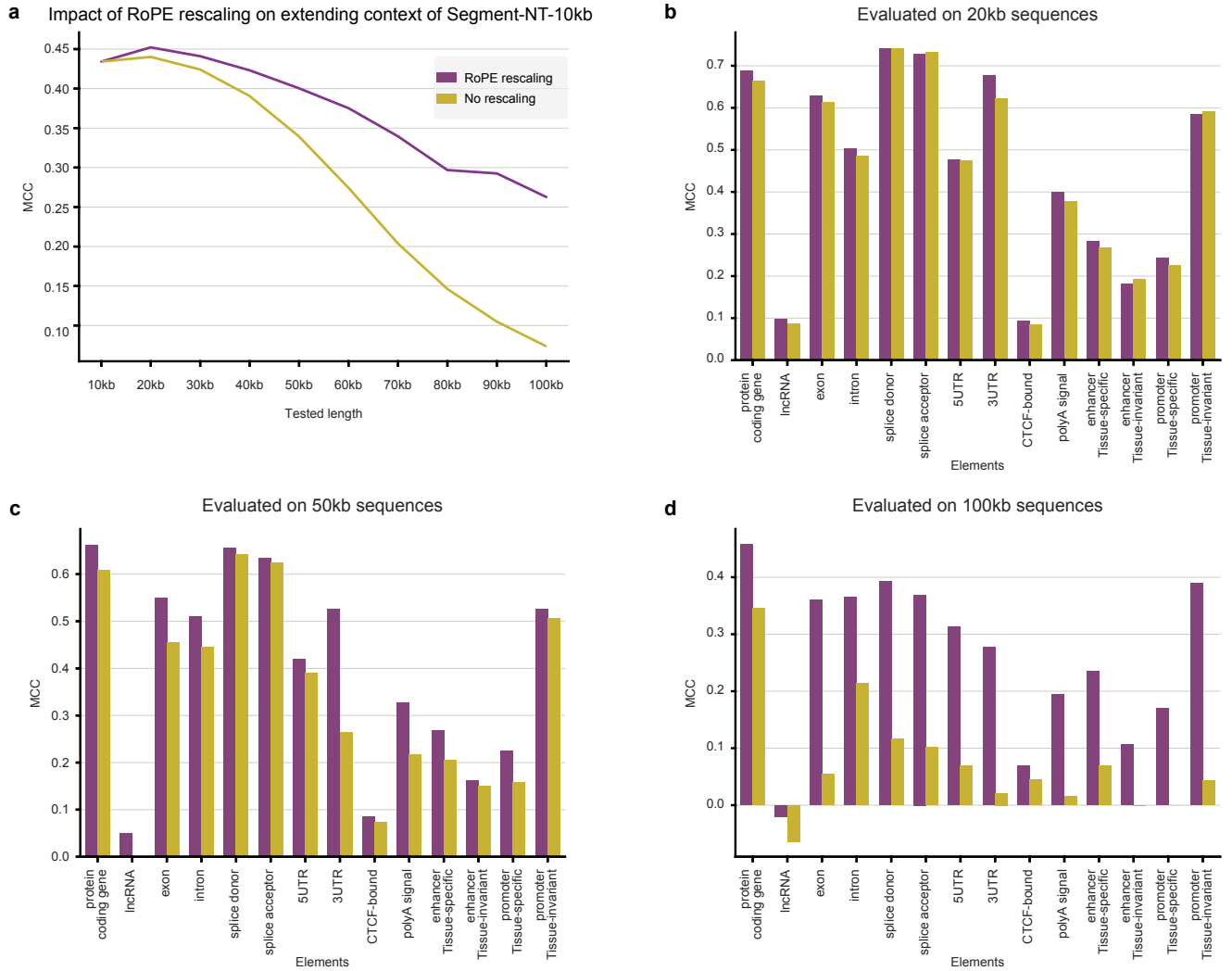
Supplementary Figure 1 | Data distribution per element type. a) Percentage of sequences containing each element type in train and test 10kb dataset. b) Percentage of nucleotides containing each element type in train and test 10kb dataset.



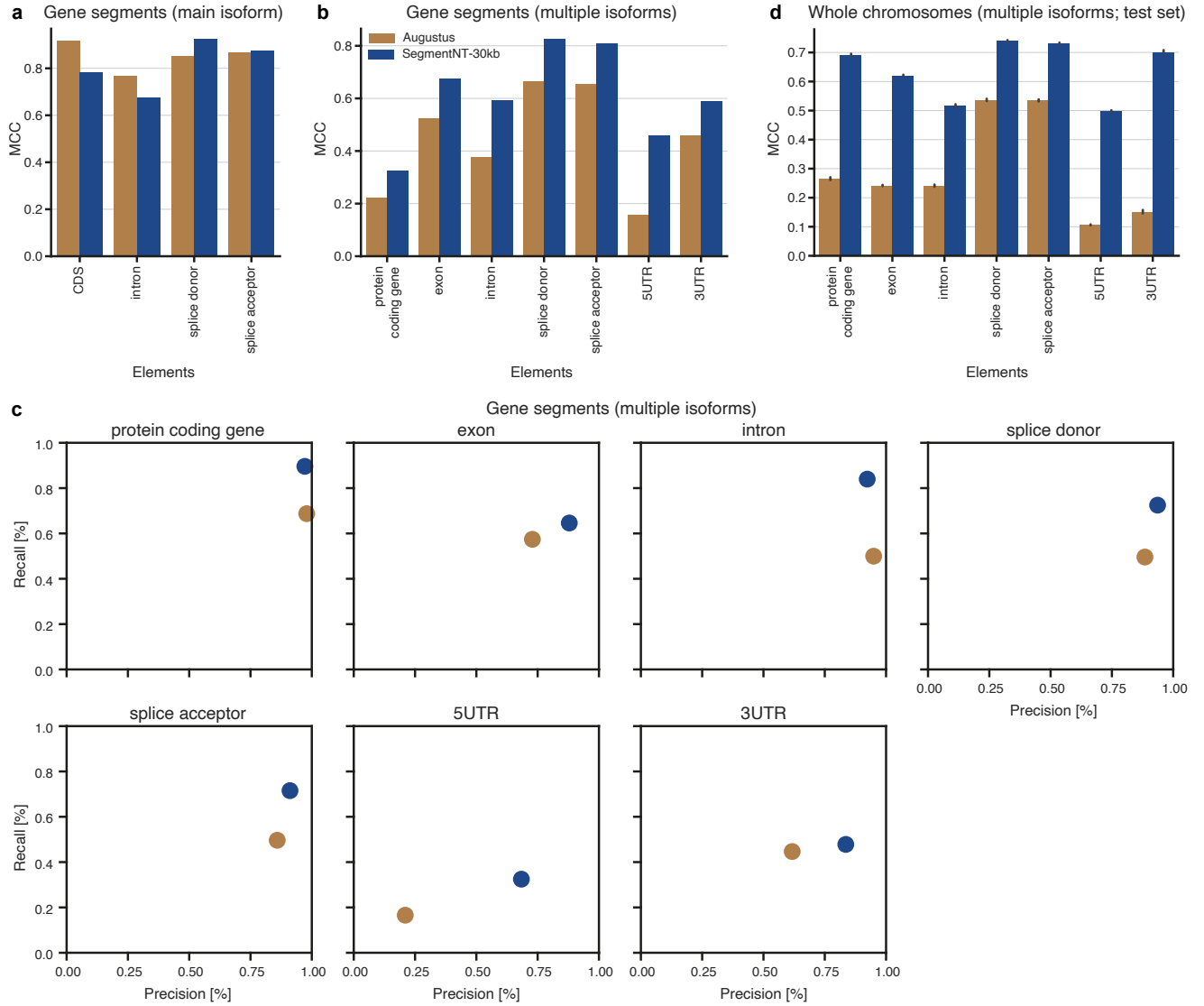
Supplementary Figure 2 | Performance of the different SegmentNT models for different metrics. a,c,e,g) Performance of SegmentNT trained on 3kb, 10kb, 20kb and 30kb sequences on 14 types of genomic elements. We used as metrics the auPRC (**a**), F1-score (**c**), Jaccard-index (**e**) and SOV score (**g**). Data are presented as mean metric values +/- 95% confidence interval from 10 different samplings of the test set. **b,d,f,h)** Average performance of the different models across the 14 elements. Data are presented as mean metric values +/- 95% confidence interval from the 14 elements.



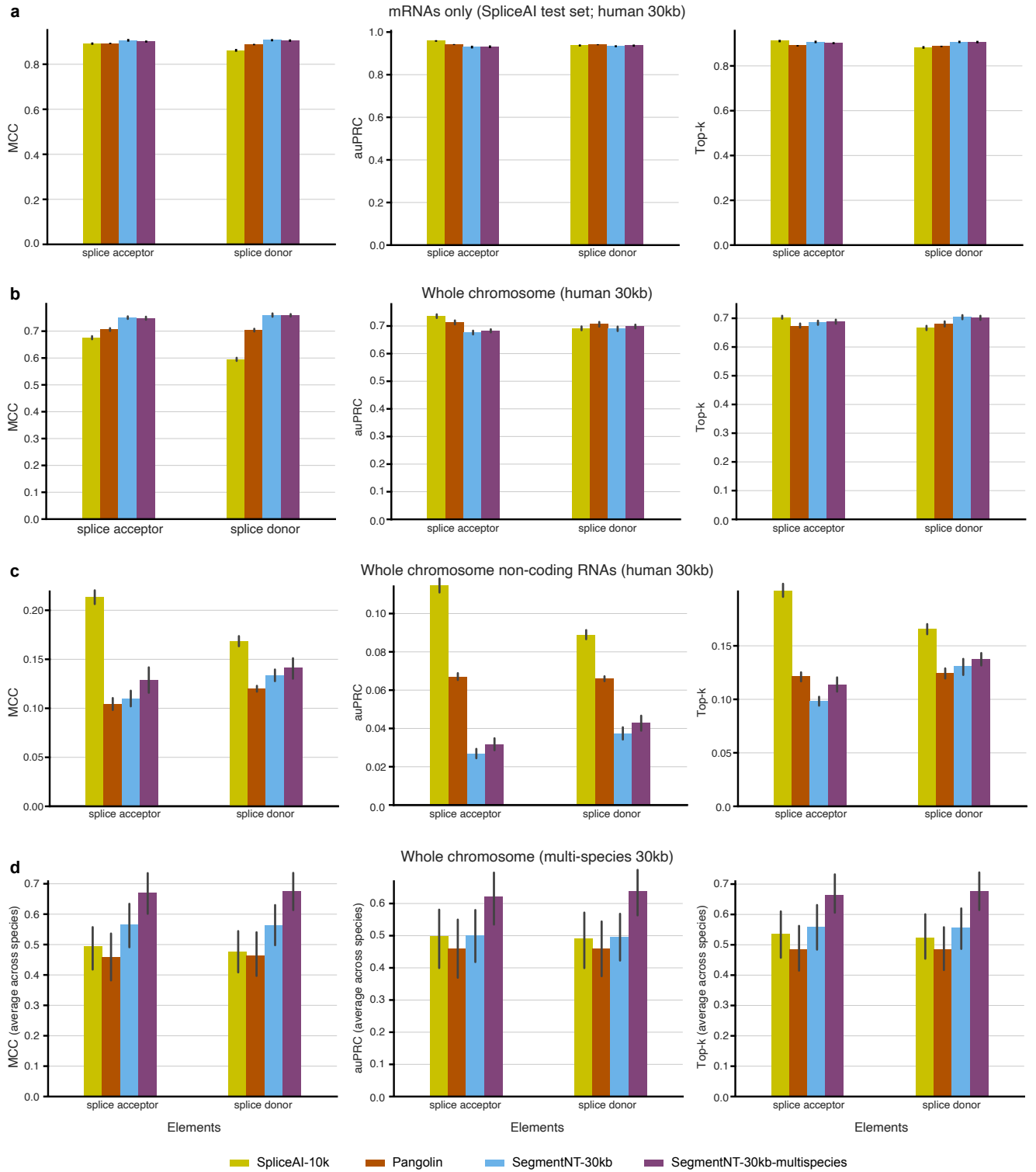
Supplementary Figure 3 | Statistical comparison between all segmentation models. Heatmap displaying the results of pairwise t-tests between different models, comparing the average MCC performance across the 14 features measured over the 10 different data splits. Each cell in the matrix represents the t-statistic, color-coded according to its magnitude and direction. Positive values (red) indicate a higher average in the y-axis label of the comparison, while negative values indicate a higher average in the x-axis label. To control for multiple comparisons, we applied the Benjamini-Hochberg False Discovery Rate (FDR) correction to the resulting p-values. Cells are annotated with significance stars indicating the adjusted p-values: *** for $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and no stars indicate $p \geq 0.05$.



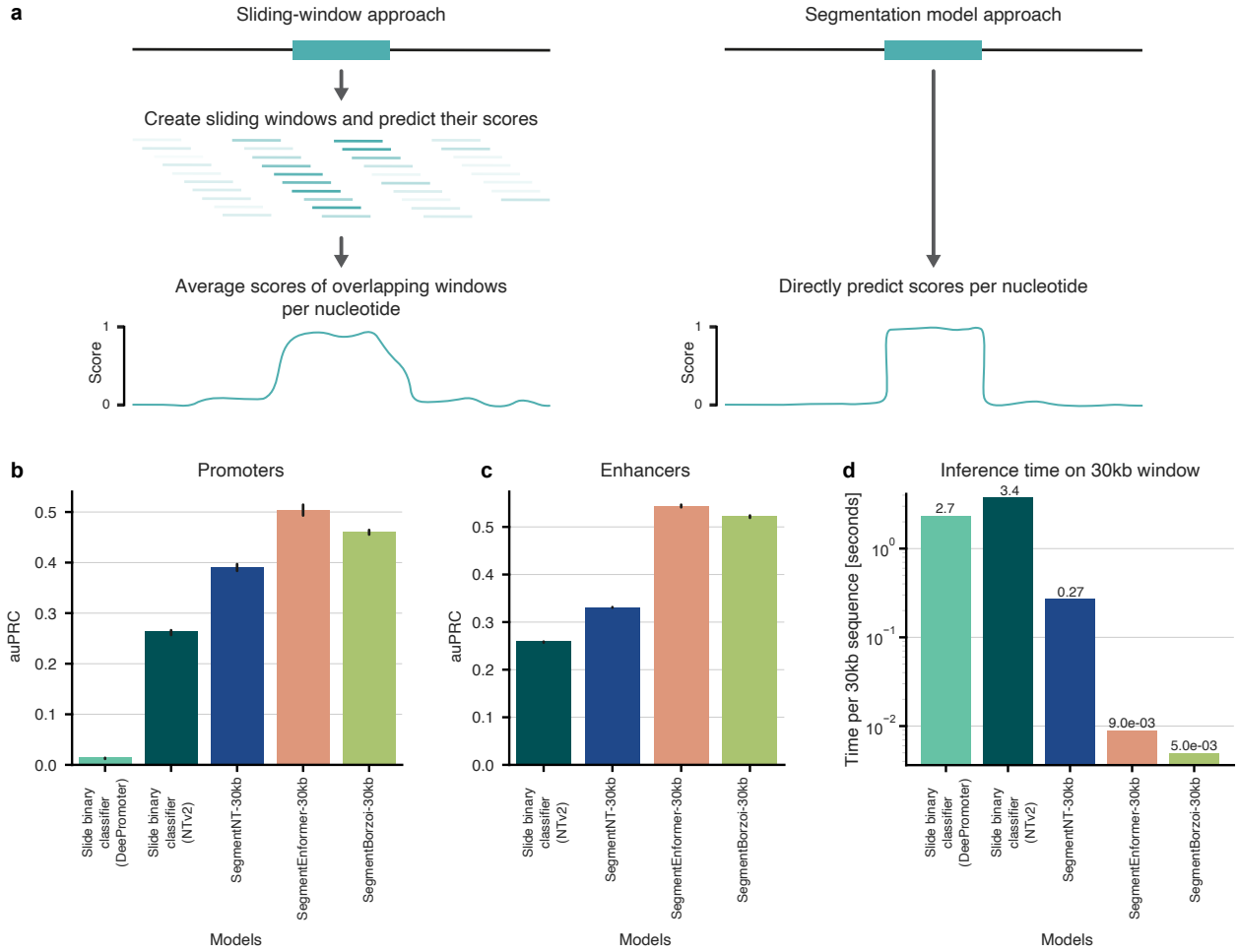
Supplementary Figure 4 | Context-length extension allows to rescale SegmentNT-10kb to 100kb sequences. a) Average MCC performance across the 14 elements for the SegmentNT-10kb model with and without context-length rescaling. b-d) Performance on different input lengths without vs with context-length extension.



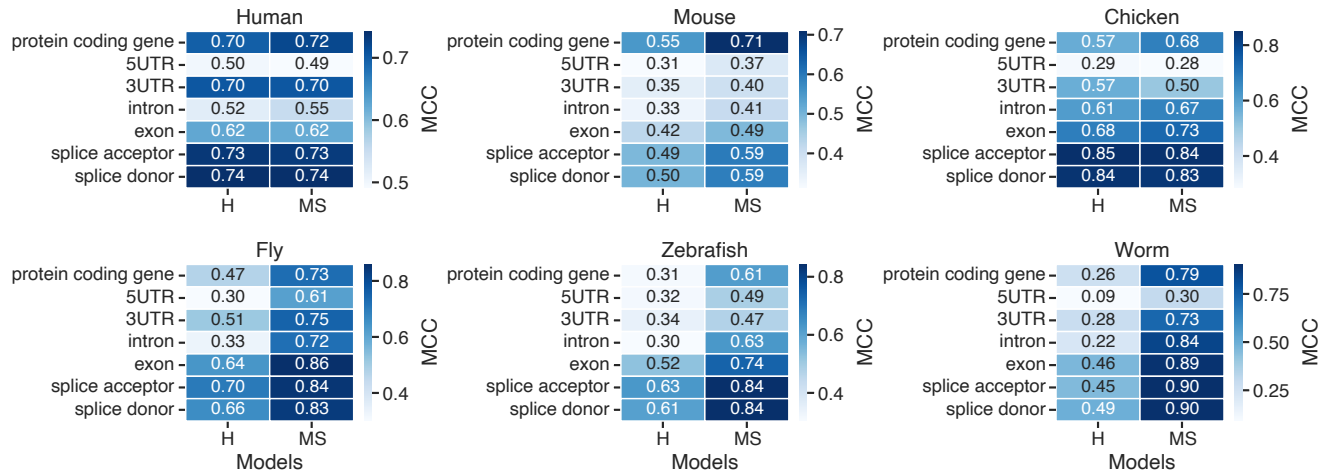
Supplementary Figure 5 | Benchmarking of gene prediction with AUGUSTUS. **a,b,d)** Performance of SegmentNT-30kb and AUGUSTUS for the different gene elements in different datasets: 30kb segments with genes using only their main isoform (**a**), 30kb segments with genes considering all annotated gene isoforms (**b**) and whole segmentation test chromosomes considering all annotated gene isoforms (**d**). The metric used was the MCC. Performance in whole-chromosome test set is presented as mean metric values +/- 95% confidence interval from 10 different data samplings. **c)** Precision and recall for SegmentNT-30kb and AUGUSTUS for the different gene elements in the dataset of 30kb segments with genes considering all annotated gene isoforms.



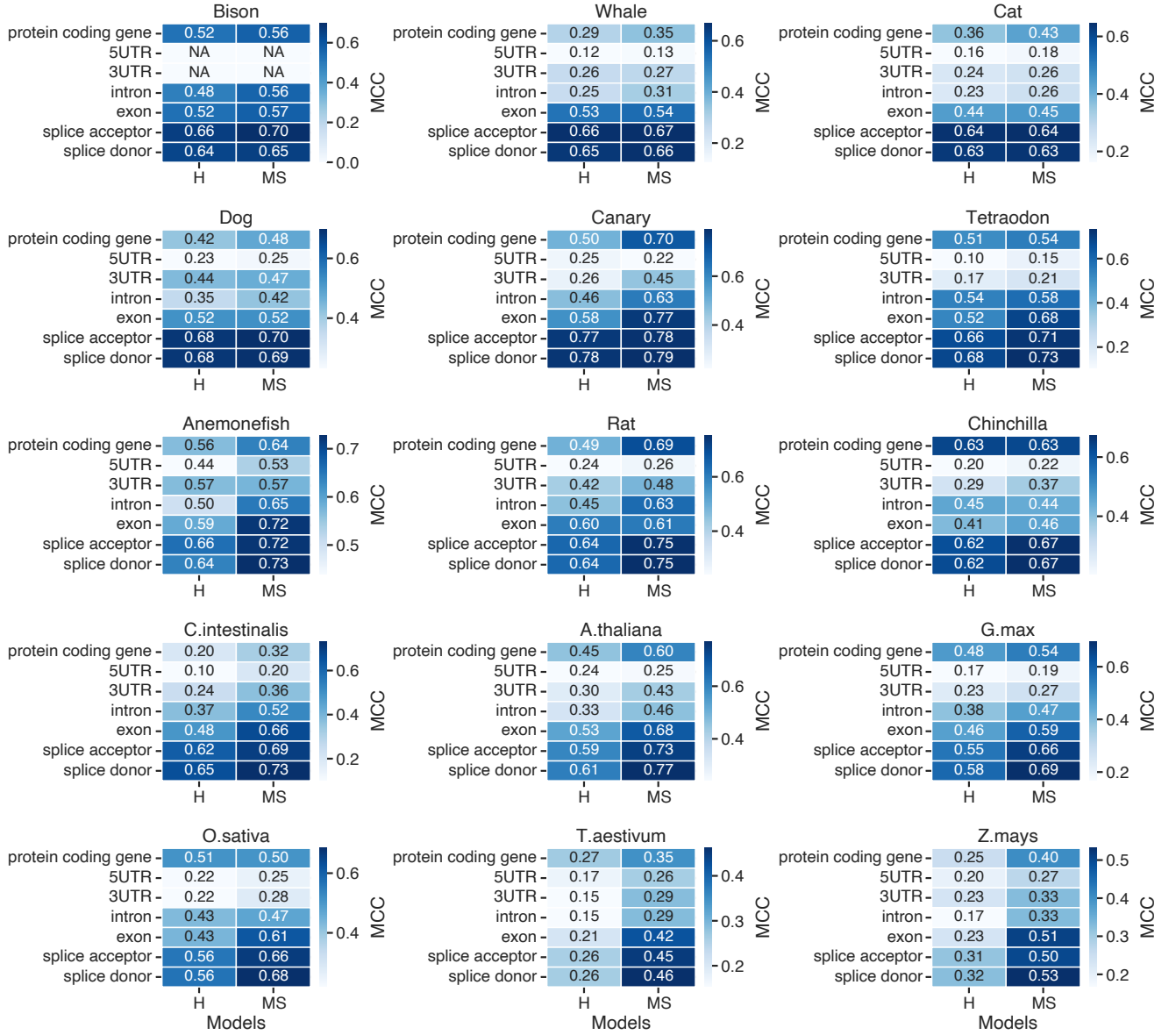
Supplementary Figure 6 | Comparison with different splicing prediction methods. Performance of the SpliceAI, Pangolin and SegmentNT-30kb and -multispecies models for splice acceptor and donor detection. We show MCC, auPRC and Top-k metrics on (a) the SpliceAI's mRNA-based test set, (b) the human SegmentNT's whole genome test sets, (c) the same but after removing protein-coding genes to focus on non-coding RNA splice sites; and (d) the multispecies SegmentNT's whole genome test sets. Error bars in mRNAs only analysis (b) represent 95% confidence intervals estimated from 100-fold bootstrapping. The performance in the multispecies dataset is based on the average across 20 species. Data are presented as mean MCC values +/- 95% confidence interval from 10 different samplings of the test set (b,c) or across species (d).



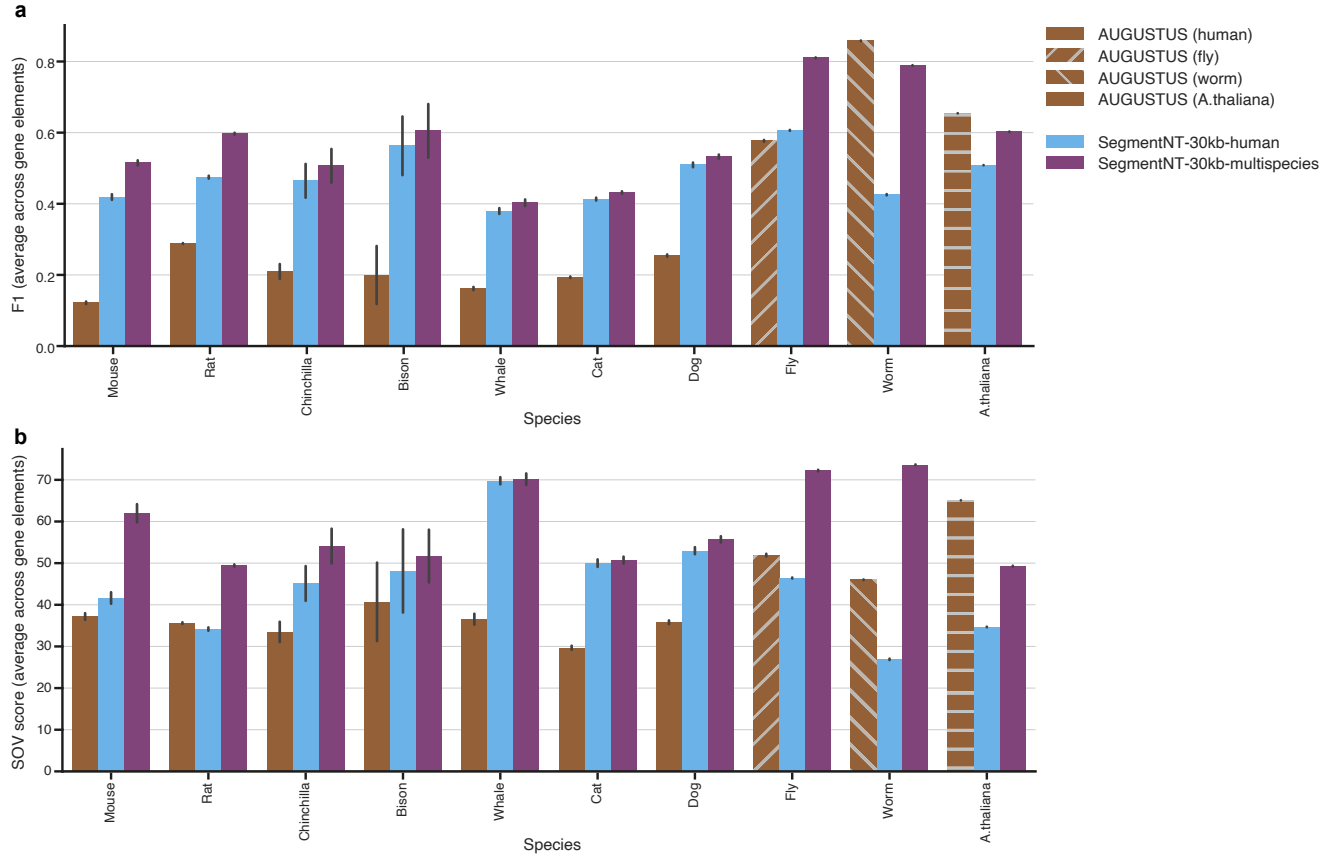
Supplementary Figure 7 | Comparison between segmentation models and alternative sliding-window approaches for promoter and enhancer predictions. **a**) Cartoon describing (left) the approach where binary classifiers are used in a sliding-window setting to make predictions at higher nucleotide resolution, contrasting with (right) our segmentation methods that directly make single-nucleotide predictions. **b**) Area under the precision-recall curve (auPRC) performance for sliding two binary promoter classifiers (DeePromoter and NTv2-promoters) or segmenting promoter elements with the different segmentation models. Data are presented as mean auPRC values $\pm$ 95% confidence interval from 10 different samplings of the test set. **c**) Area under the precision-recall curve (auPRC) performance for sliding the binary enhancer classifier NTv2 or segmenting enhancer elements with the different segmentation models. Data are presented as mean auPRC values $\pm$ 95% confidence interval from 10 different samplings of the test set. **d**) Inference times on a 30kb sequence for the different models. Inference times were calculated in a single A100 GPU.



Supplementary Figure 8 | Comparison of human (H) and multispecies (MS) SegmentNT models on training set species. Data is shown as the mean MCC values from 10 different samplings of each species' test set.



Supplementary Figure 9 | Comparison of human and multispecies SegmentNT models on test set species. Performance of the human (H) and multispecies (MS) model per element for animal and plant test set species. Data is shown as the mean MCC values from 10 different samplings of each species' test set.



Supplementary Figure 10 | Comparison with AUGUSTUS on gene annotation across species. Performance of AUGUSTUS compared with SegmentNT-30kb human and multispecies models for the annotation of genes in the genomes of the different species. We show the species for which there are AUGUSTUS models: as recommended by AUGUSTUS, we used the human model for all mammals and specialized models for the other species (fly, worm and Arabidopsis). The metric used are **(a)** the average F1-score and **(b)** the SOV score, presented as mean +/- 95% confidence interval across gene elements per species.