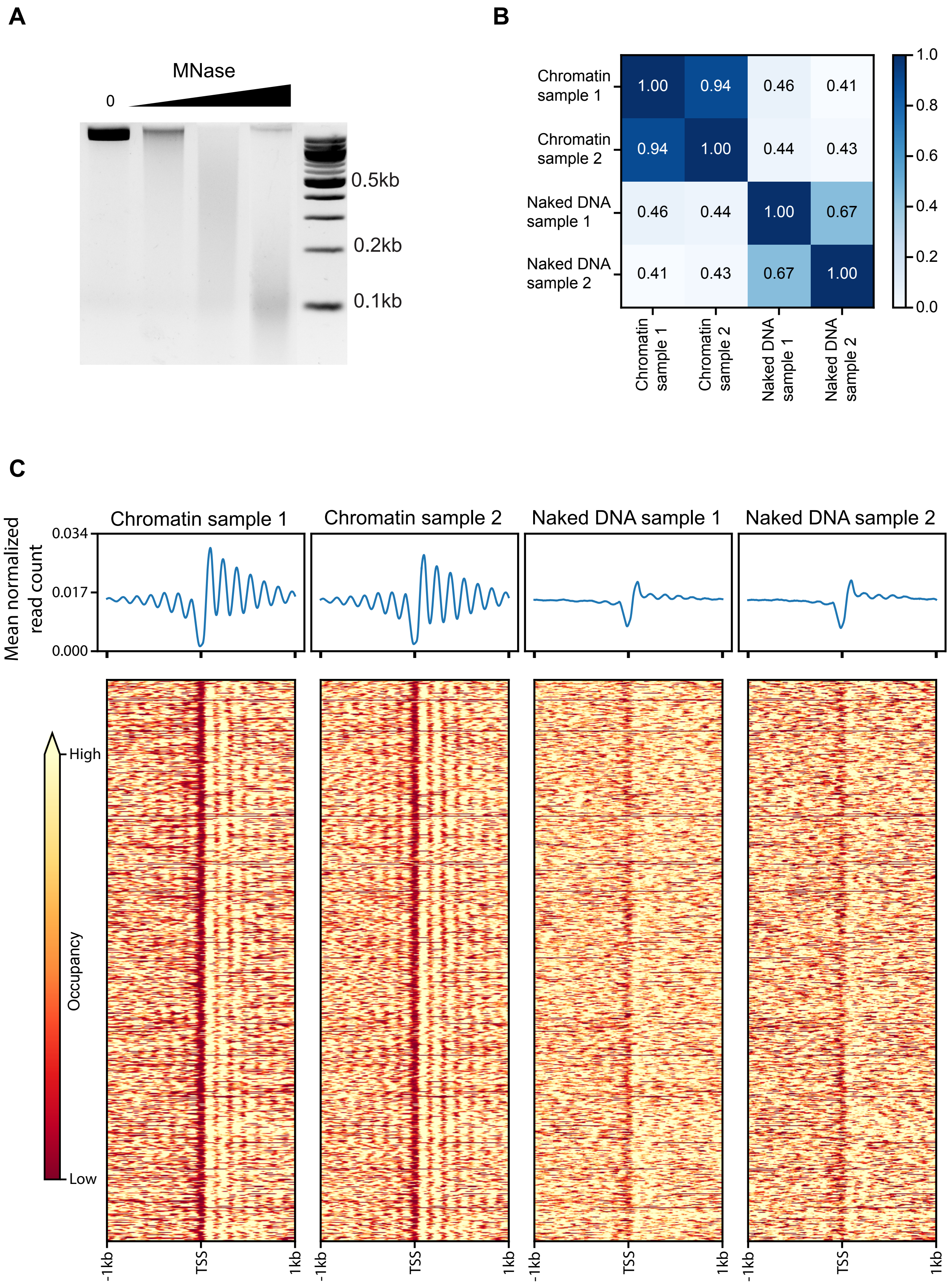
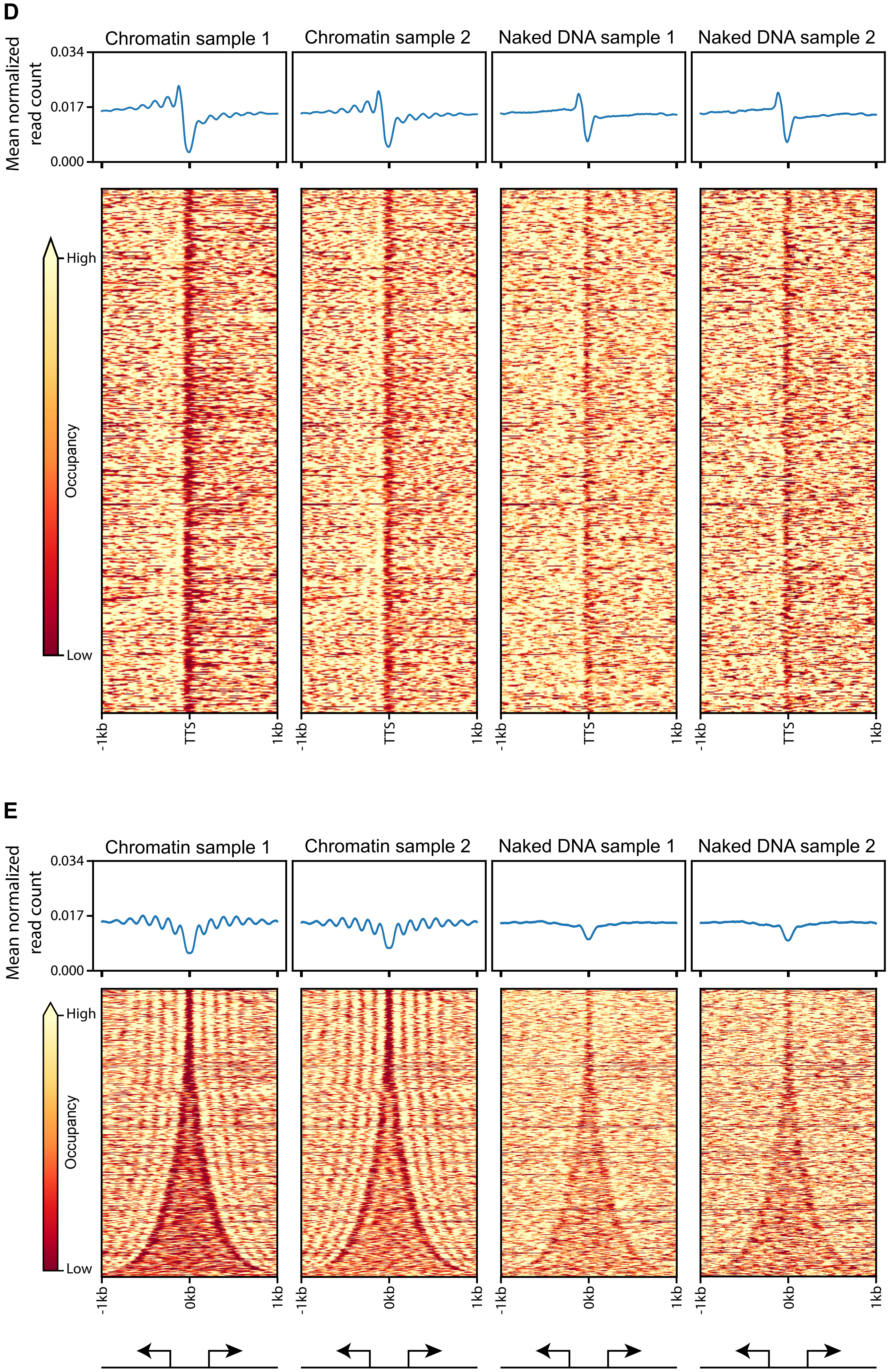
Supplemental Material

GC content but not nucleosome positioning directly contributes to intron-splicing efficiency in *Paramecium*

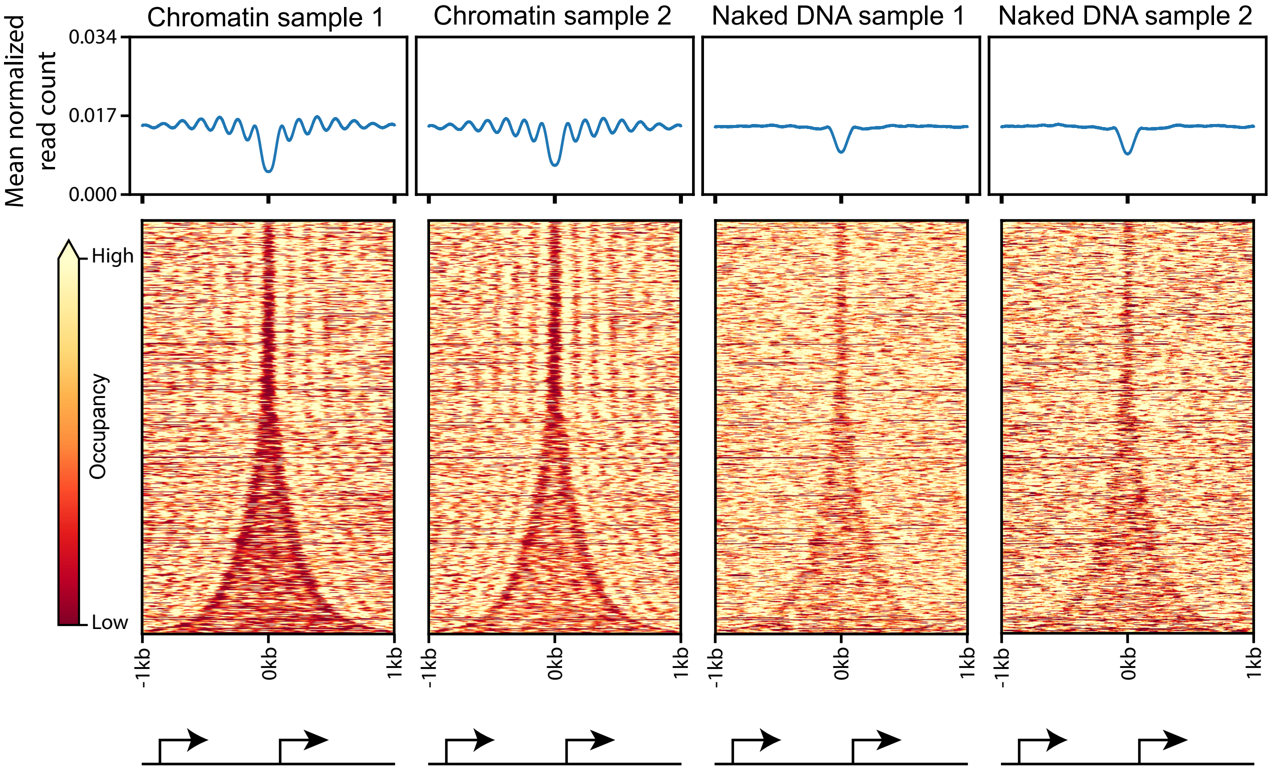
Stefano Gnan, Mélody Matelot, Marion Weiman, Olivier Arnaiz, Frédéric Guérin, Linda Sperling, Mireille Bétermier, Claude Thermes, Chun-Long Chen and Sandra Duharcourt

Table of contents

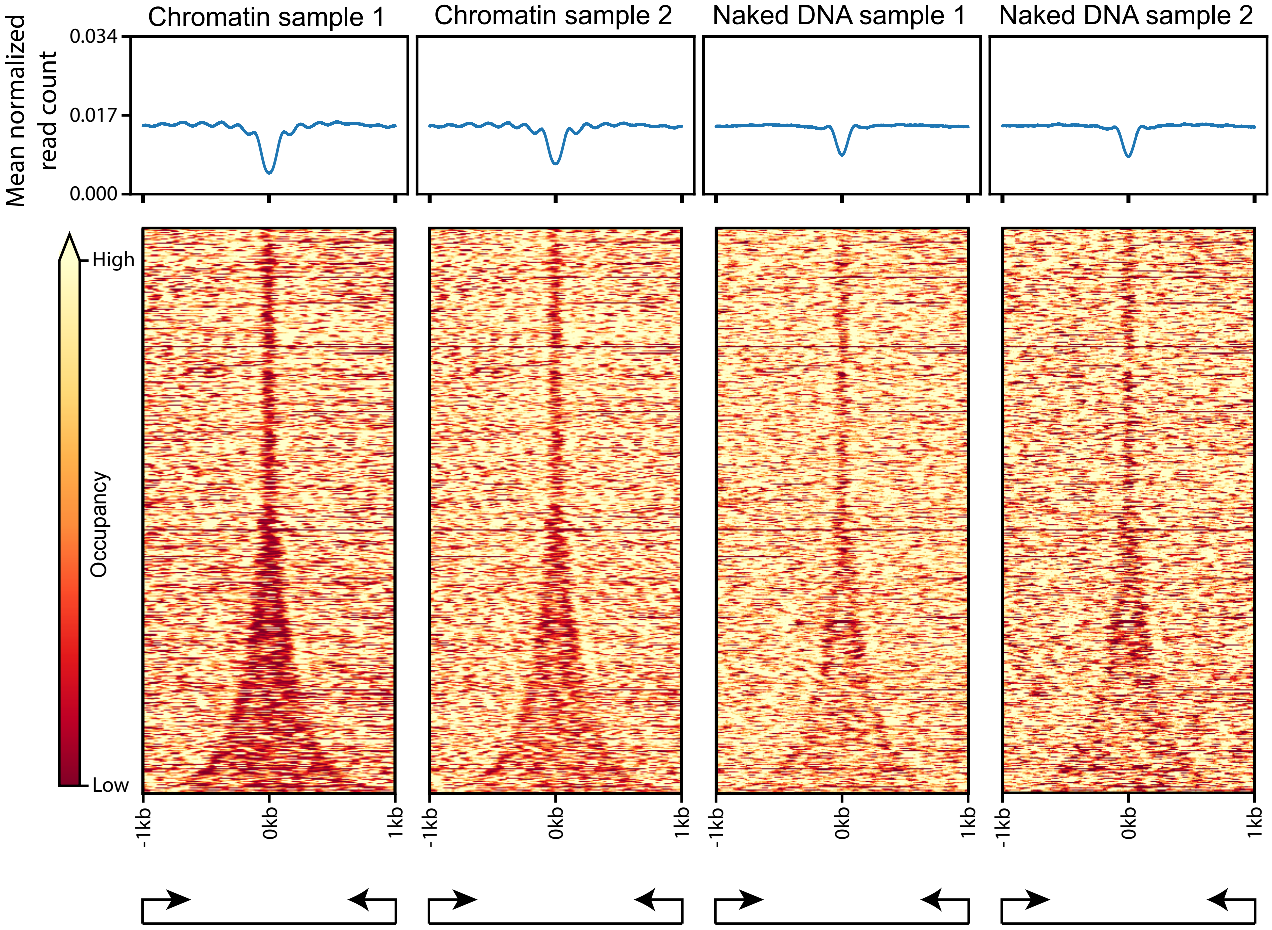

A
MNase
0
0.5kb
0.2kb
0.1kb
B
Chromatin sample 1
Chromatin sample 2
Naked DNA sample 1
Naked DNA sample 2
1.00 0.94 0.46 0.41
0.94 1.00 0.44 0.43
0.46 0.44 1.00 0.67
0.41 0.43 0.67 1.00
1.0
0.8
0.6
0.4
0.2
0.0
C
Chromatin sample 1
Chromatin sample 2
Naked DNA sample 1
Naked DNA sample 2
Mean normalized read count
0.034
0.017
0.000
High
Occupancy
Low
-1kb
TSS
1kb

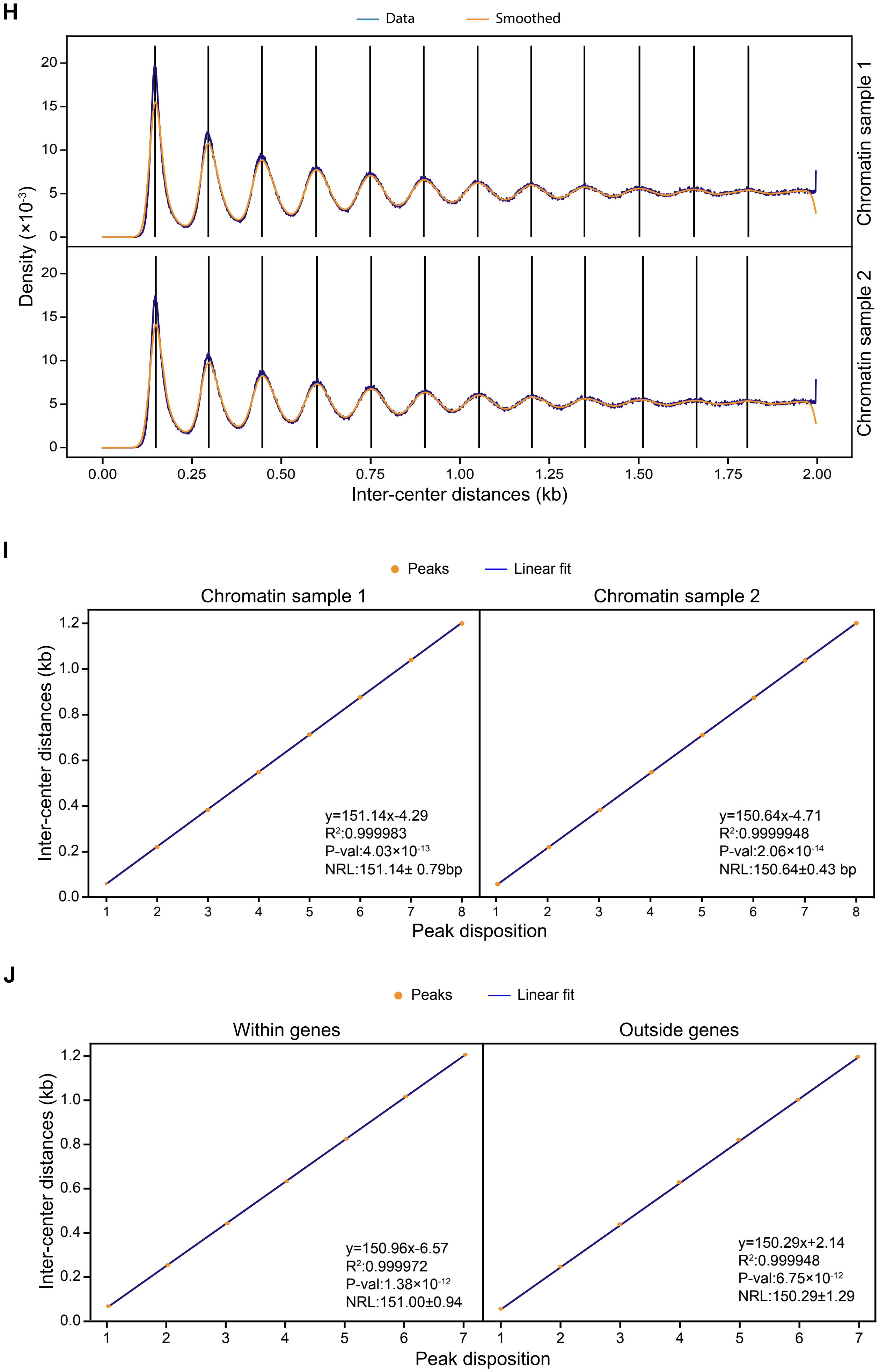

D
Chromatin sample 1
Chromatin sample 2
Naked DNA sample 1
Naked DNA sample 2
Mean normalized read count
0.034
0.017
0.000
High
Occupancy
Low
-1kb
TTS
1kb
E
Chromatin sample 1
Chromatin sample 2
Naked DNA sample 1
Naked DNA sample 2
Mean normalized read count
0.034
0.017
0.000
High
Occupancy
Low
-1kb
0kb
1kb

F

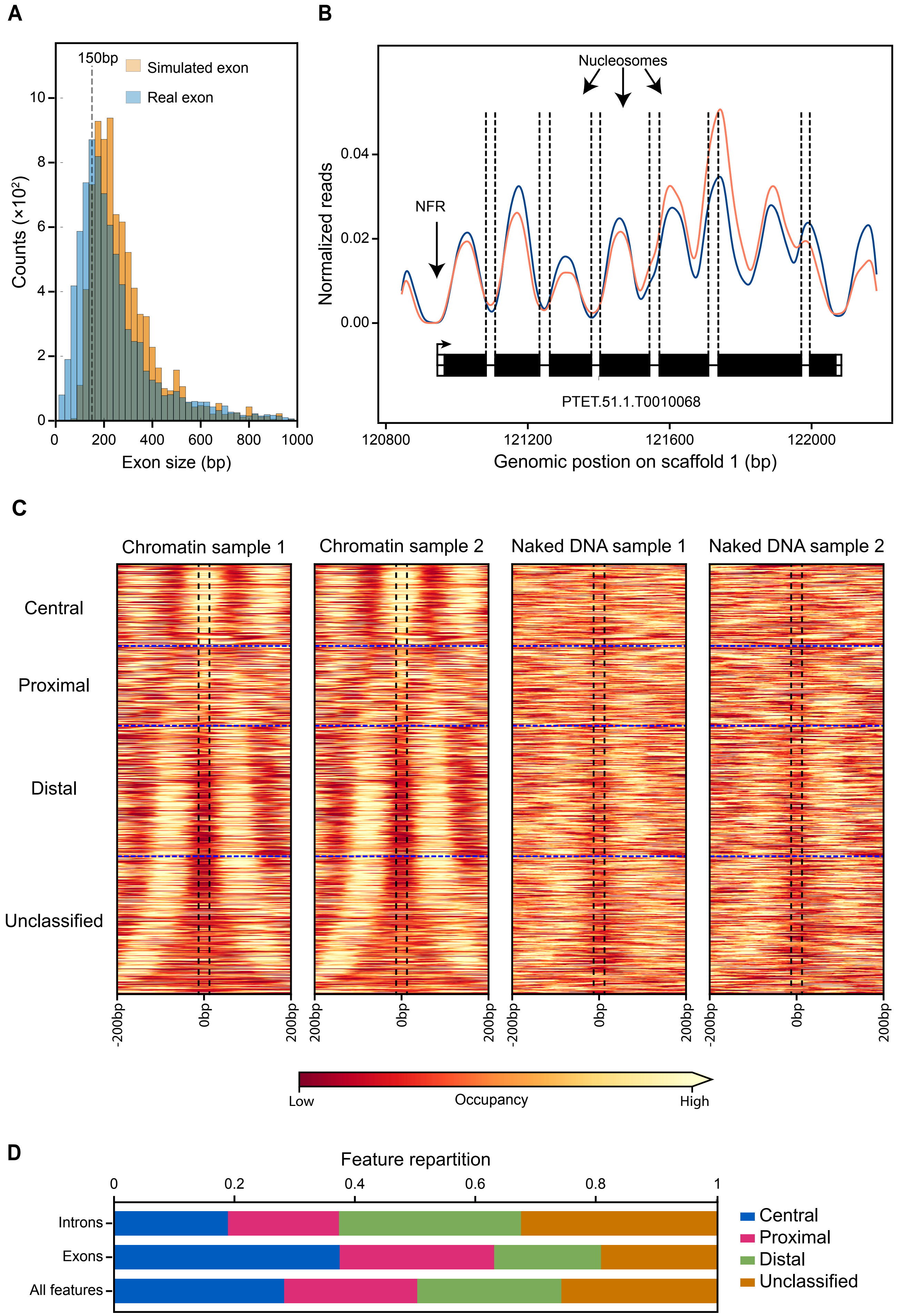

Chromatin sample 1
Chromatin sample 2
Naked DNA sample 1
Naked DNA sample 2
Mean normalized read count
0.034
0.017
0.000
High
Occupancy
Low
-1kb
0kb
1kb

G

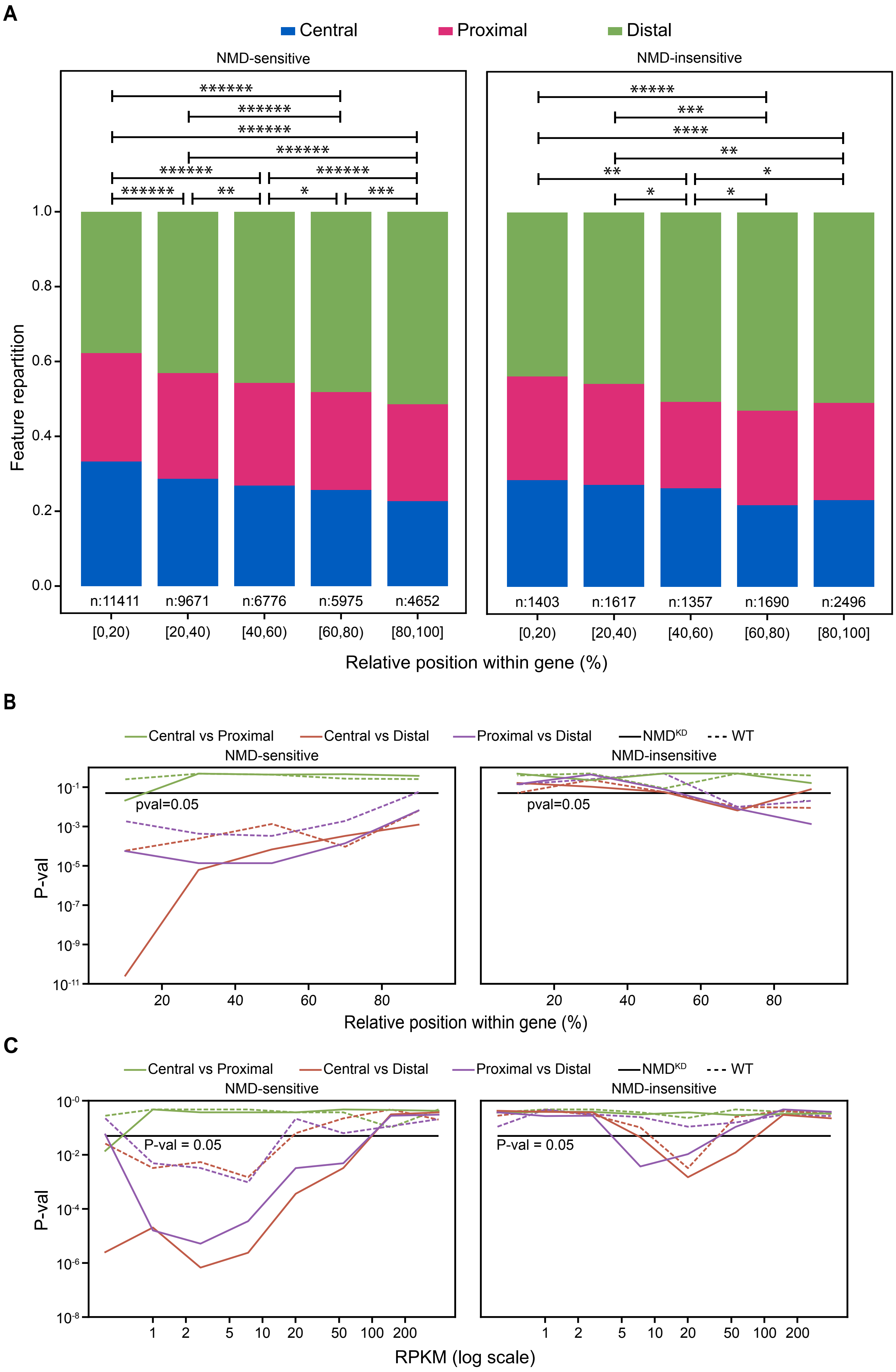

Chromatin sample 1
Chromatin sample 2
Naked DNA sample 1
Naked DNA sample 2
Mean normalized read count
0.034
0.017
0.000
High
Occupancy
Low
-1kb
0kb
1kb

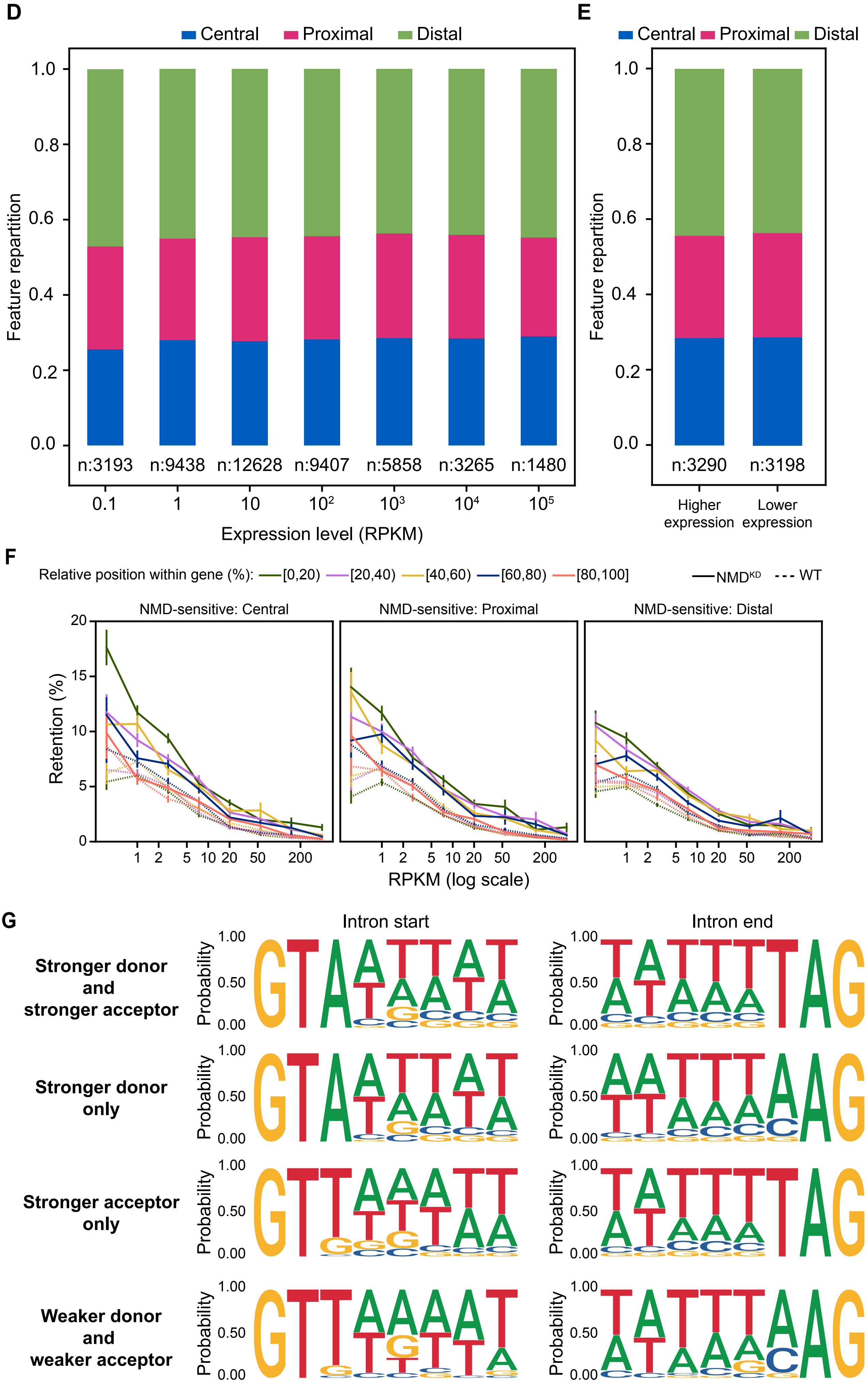

H
Data
Smoothed
Density (×10-3)
20
15
10
5
0
Chromatin sample 1
Chromatin sample 2
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Inter-center distances (kb)
I
Peaks
Linear fit
Chromatin sample 1
Chromatin sample 2
Inter-center distances (kb)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
y=151.14x-4.29
R2:0.999983
P-val:4.03×10-13
NRL:151.14± 0.79bp
y=150.64x-4.71
R2:0.9999948
P-val:2.06×10-14
NRL:150.64±0.43 bp
1
2
3
4
5
6
7
8
Peak disposition
J
Peaks
Linear fit
Within genes
Outside genes
Inter-center distances (kb)
y=150.96x-6.57
R2:0.999972
P-val:1.38×10-12
NRL:151.00±0.94
y=150.29x+2.14
R2:0.999948
P-val:6.75×10-12
NRL:150.29±1.29
Peak disposition

Supplemental Figure S1. Nucleosome occupancy along the MAC *Paramecium* genome. (A) MNase digestion of naked DNA with increasing MNase enzyme concentration. **(B)** Heatmap showing pairwise Pearson's correlation between chromatin and naked DNA samples (two biological replicates for each). It should be noted that the naked DNA samples show a lower correlation than the chromatin samples. **(C)** Average profiles and heatmaps showing nucleosome occupancy around Transcription Start Sites (TSSs) in individual samples. **(D)** Average profiles and heatmaps showing nucleosome occupancy around Transcription Termination Sites (TTSs) in individual samples. **(E-G)** Average profiles and heatmaps showing nucleosome occupancy ±1 kb around the center of intergenic regions. Pairs of genes have been divided into three groups based on their relative orientation: divergent (E), tandem (F) and convergent (G). Genes are sorted based on their intergenic distance. **(H)** Inter-center distance between nucleosomes on the same scaffold for the two chromatin samples. In blue, distance distributions from actual data (from 1 bp to 2 kb, binning=1 bp); and in orange, the gaussian smoothed signal. Black dashed lines identify the local maxima (peak centers) of the smoothed data (Methods). **(I)** In orange, the local maxima from Supplemental Fig S1H ordered by increasing distance; in blue, the linear fitted model. On the bottom right of each panel, the information about the linear fitting and the estimated NRL for both chromatin replicates. P-values are calculated using a two-sided Z-test. **(J)** The same analysis as in Fig 1G by separating into the nucleosome centers overlapping with gene bodies (left panel) or outside of them (right panel). As in Fig 1G, local maxima are ordered by increasing distance. The relative linear fitting model is reported in blue. The bottom right gives the information about the linear fitting and the estimated NRL (Mean±SD). P-value is calculated using a two-sided Z-test.

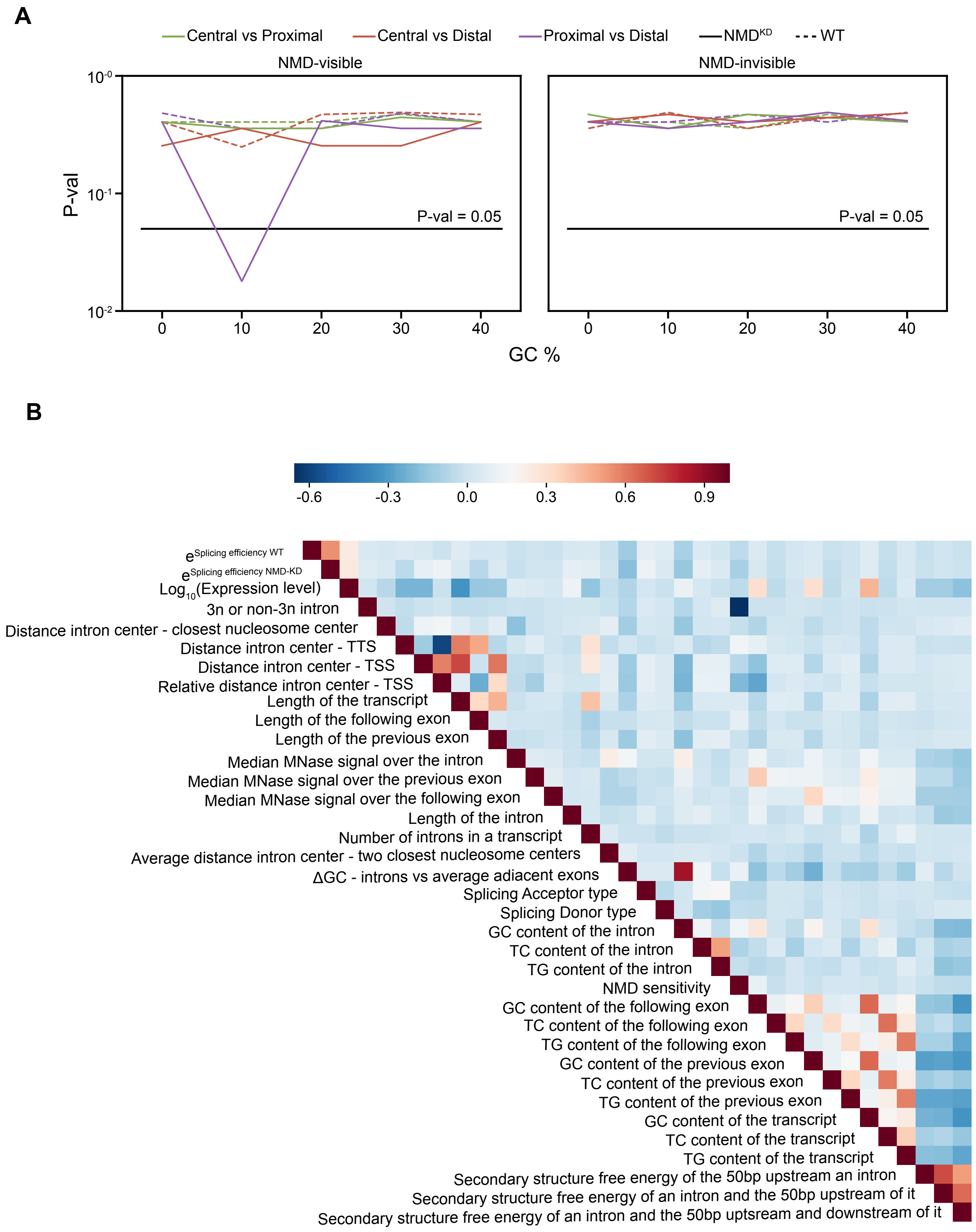

A
150bp
Simulated exon
Real exon
Counts (×10²)
0
2
4
6
8
10
0
200
400
600
800
1000
Exon size (bp)
B
Nucleosomes
NFR
Normalized reads
0.00
0.02
0.04
PTET.51.1.T0010068
120800
121200
121600
122000
Genomic postion on scaffold 1 (bp)
C
Chromatin sample 1
Chromatin sample 2
Naked DNA sample 1
Naked DNA sample 2
Central
Proximal
Distal
Unclassified
-200bp
0bp
200bp
Low
Occupancy
High
D
Feature repartition
0
0.2
0.4
0.6
0.8
1
Introns
Exons
All features
Central
Proximal
Distal
Unclassified

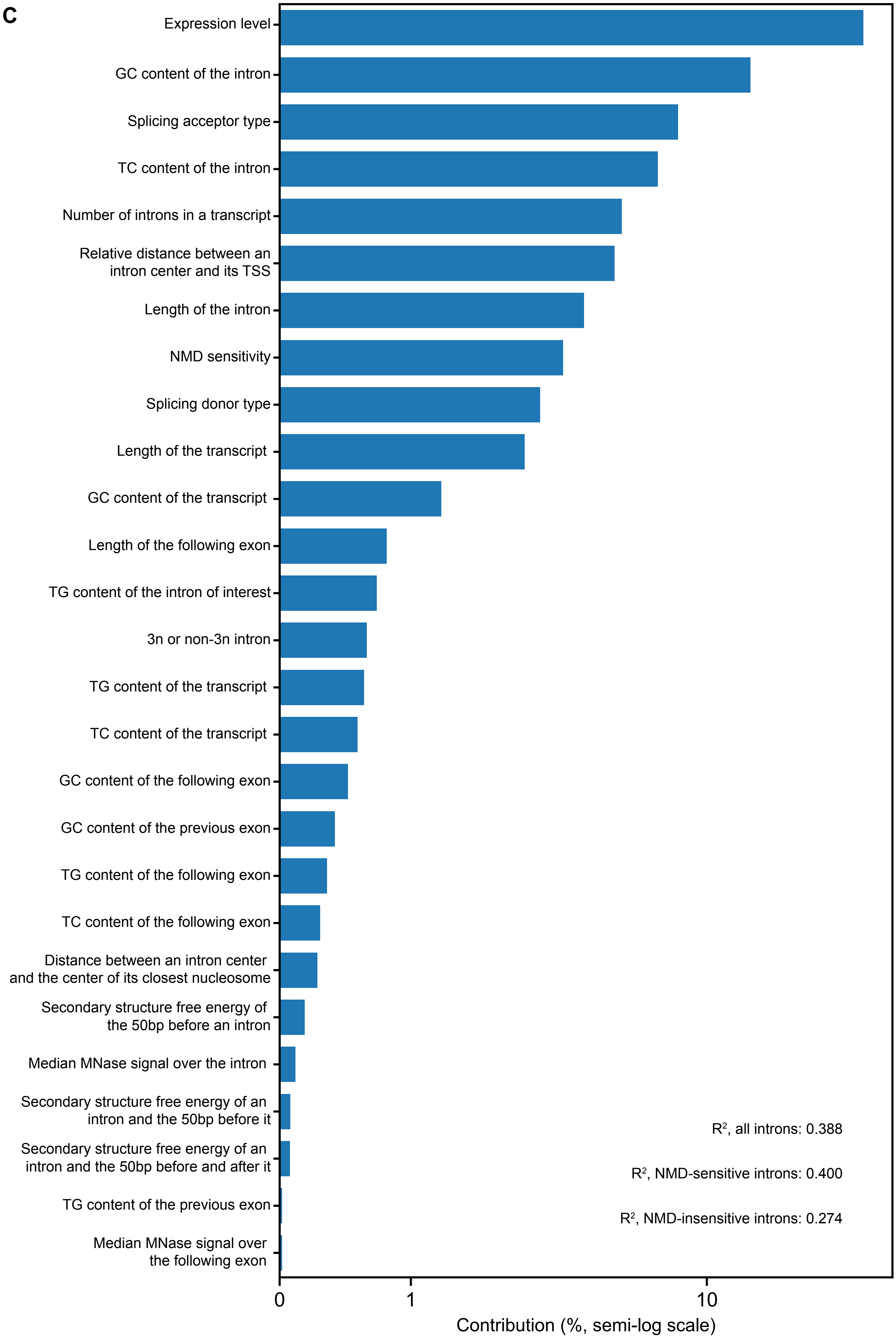
Supplemental Figure S2. Inter-nucleosomal DNA is frequently associated with intron position. (A) Histogram showing exon size distribution (bin size = 25 bp) of transcription units whose extremities have been identified by 5' CAP-seq and poly(A) detection: in blue, real exons; and in orange, simulated exons created assuming uniform exon sizes within each transcript. **(B)** The same genomic region as shown in Fig 2B. Tracks of two independent chromatin treated samples reporting nucleosome occupancy over a representative gene. Intron locations are indicated by dashed vertical lines. Nucleosome free regions (NFRs) are found around the gene promoters and introns are frequently associated with inter-nucleosomal DNA. **(C)** Heatmap showing nucleosome occupancy ± 200 bp around intron centers in individual samples. Introns are ordered as in Fig 2D. Horizontal blue dashed lines define the boundaries that identify central, proximal, distal and unclassified introns, while vertical black dashed lines delineate the average size of an intron (25 bp). **(D)** Same figure as Fig 2G including features with distance >75 bp. “Unclassified” corresponds to introns and/or exons that do not belong to the three defined categories (central, proximal, distal).


A
Central
Proximal
Distal
NMD-sensitive
NMD-insensitive
******
******
******
******
******
******
******
**
*
***
*****
***
****
**
**
*
*
*
Feature repartition
1.0
0.8
0.6
0.4
0.2
0.0
n:11411
n:9671
n:6776
n:5975
n:4652
n:1403
n:1617
n:1357
n:1690
n:2496
[0,20)
[20,40)
[40,60)
[60,80)
[80,100]
Relative position within gene (%)
B
Central vs Proximal
Central vs Distal
Proximal vs Distal
NMD^KD
WT
NMD-sensitive
NMD-insensitive
pval=0.05
P-val
10^-1
10^-3
10^-5
10^-7
10^-9
10^-11
20
40
60
80
Relative position within gene (%)
C
Central vs Proximal
Central vs Distal
Proximal vs Distal
NMD^KD
WT
NMD-sensitive
NMD-insensitive
P-val = 0.05
P-val
10^-0
10^-2
10^-4
10^-6
10^-8
1
2
5
10
20
50
100
200
RPKM (log scale)


D
Central
Proximal
Distal
1.0
0.8
0.6
0.4
0.2
0.0
Feature repartition
n:3193
n:9438
n:12628
n:9407
n:5858
n:3265
n:1480
0.1
1
10
10²
10³
10⁴
10⁵
Expression level (RPKM)
E
Central
Proximal
Distal
n:3290
n:3198
Higher expression
Lower expression
F
Relative position within gene (%): [0,20) [20,40) [40,60) [60,80) [80,100]
NMDKD
WT
NMD-sensitive: Central
NMD-sensitive: Proximal
NMD-sensitive: Distal
Retention (%)
20
15
10
5
0
1
2
5
10
20
50
200
RPKM (log scale)
G
Intron start
Intron end
Stronger donor and stronger acceptor
Stronger donor only
Stronger acceptor only
Weaker donor and weaker acceptor
Probability
1.00
0.50
0.00

Supplemental Figure S3. Nucleosome positioning is associated with intron-splicing efficiency. (A) Barplot representation of Fig 3B. P-values are calculated using the χ^2 test and adjusted using false discovery rate (5%). Only the significant ones are indicated. **(B)** P-values relative to Fig 3C were calculated using the Mann–Whitney *U* test and adjusted using false discovery rate (5%). Comparisons of intron retention rates were evaluated in WT (dashed lines) and NMD-depleted cells (NMD^{KD}, solid line), respectively, for NMD-sensitive introns (left) and NMD-insensitive introns (right). The black line indicates P-value = 0.05**. (C)** P-values relative to Fig 3D were calculated using the Mann–Whitney *U* test and adjusted using false discovery rate (5%). Color and line style codes are the same as in panel B. **(D)** Introns analyzed in Fig 2G were further divided accordingly to their expression levels. Expression ranges were selected to have maintain a minimal statistical power. **(E)** Intron classification for genes duplicates from the last whole genome duplication. Only duplicates with the same number of identified introns were used. Each pair of genes was assigned to the “higher” or “lower” expression groups based on their expression levels. (D-E) P-values were calculated using a χ^2 test and were adjusted using the false discovery rate (5%) when needed. No significant difference was identified. **(F)** Retention level of NMD-sensitive introns as a function of expression levels in NMD-depleted (NMD^{KD}) and WT cells. Introns are grouped according to their relative position within a gene and to their relative position to the closest nucleosome: central (left), proximal (center) and distal (right). **(G)** Sequence logo plots showing the nucleotide composition of the splicing donors and acceptors sites relative to Fig 3E.


A
Central vs Proximal
Central vs Distal
Proximal vs Distal
NMD^{KD}
WT
NMD-visible
NMD-invisible
P-val
10^{-0}
10^{-1}
10^{-2}
P-val = 0.05
P-val = 0.05
0
10
20
30
40
GC %
B
-0.6
-0.3
0.0
0.3
0.6
0.9
$e^{\text{Splicing efficiency WT}}$
$e^{\text{Splicing efficiency NMD-KD}}$
Log_{10}(Expression level)
3n or non-3n intron
Distance intron center - closest nucleosome center
Distance intron center - TTS
Distance intron center - TSS
Relative distance intron center - TSS
Length of the transcript
Length of the following exon
Length of the previous exon
Median MNase signal over the intron
Median MNase signal over the previous exon
Median MNase signal over the following exon
Length of the intron
Number of introns in a transcript
Average distance intron center - two closest nucleosome centers
ΔGC - introns vs average adiacent exons
Splicing Acceptor type
Splicing Donor type
GC content of the intron
TC content of the intron
TG content of the intron
NMD sensitivity
GC content of the following exon
TC content of the following exon
TG content of the following exon
GC content of the previous exon
TC content of the previous exon
TG content of the previous exon
GC content of the transcript
TC content of the transcript
TG content of the transcript
Secondary structure free energy of the 50bp upstream an intron
Secondary structure free energy of an intron and the 50bp upstream of it
Secondary structure free energy of an intron and the 50bp uptsream and downstream of it


C
Expression level
GC content of the intron
Splicing acceptor type
TC content of the intron
Number of introns in a transcript
Relative distance between an intron center and its TSS
Length of the intron
NMD sensitivity
Splicing donor type
Length of the transcript
GC content of the transcript
Length of the following exon
TG content of the intron of interest
3n or non-3n intron
TG content of the transcript
TC content of the transcript
GC content of the following exon
GC content of the previous exon
TG content of the following exon
TC content of the following exon
Distance between an intron center and the center of its closest nucleosome
Secondary structure free energy of the 50bp before an intron
Median MNase signal over the intron
Secondary structure free energy of an intron and the 50bp before it
Secondary structure free energy of an intron and the 50bp before and after it
TG content of the previous exon
Median MNase signal over the following exon
R², all introns: 0.388
R², NMD-sensitive introns: 0.400
R², NMD-insensitive introns: 0.274
0
1
10
Contribution (%, semi-log scale)

Supplemental Figure S4. GC content related to nucleosome positioning contributes to intron-splicing efficiency. (A) P-values relative to Fig 4B were calculated using the Mann–Whitney *U* test and adjusted using false discovery rate (5%). Comparisons of intron retention rates were evaluated in WT (dashed lines) and NMD-depleted (NMD^{KD}, solid line) cells, respectively. Left: NMD-sensitive introns. Right: NMD-insensitive introns. The black line indicates the position of P-value = 0.05. **(B)** Pairwise Pearson's correlation heatmap of all parameters used to model Splicing Efficiency (SE) in NMD-depleted cells. **(C)** Histogram reporting the contribution of the final parameters used in our model. P-values are displayed in Supplemental Table S1. Pearson's correlation between model and real data for all introns, NMD-sensitive and NMD-insensitive introns are displayed at the bottom right corner.

Supplemental Table S1. Parameters and their contribution in modelling the splicing efficiency in the multilinear regression model.

Group	Parameter	Abbreviation	Contribution	p-value	Group contribution
Expression level	Expression level	EL	46.09%	$< 10^{-16}$	46.09%
Base composition of an Intron	GC content of the intron of interest	$\mathrm{GC_I}$	15.30%	$< 10^{-16}$	22.23%
	TC content of the intron of interest	$\mathrm{TC_I}$	6.20%	$< 10^{-16}$	
	TG content of the intron of interest	$\mathrm{TG_I}$	0.74%	4.95×10^{-7}	
Splicing signals	Splicing acceptor type*	SA	7.56%	$< 10^{-16}$	9.55%
	Splicing donor type*	SD	1.99%	$< 10^{-16}$	
Transcript size and base composition	Number of introns in a transcript	N	4.37%	$< 10^{-16}$	8.71%
	Length of the transcript	$\mathrm{L_T}$	1.87%	$< 10^{-16}$	
	GC content of the transcript	$\mathrm{GC_T}$	1.23%	9.73×10^{-6}	
	TG content of the transcript	$\mathrm{TG_T}$	0.64%	9.45×10^{-8}	
	TC content of the transcript	$\mathrm{TC_T}$	0.59%	9.71×10^{-10}	
Intron size and position	Relative distance between an intron center and its TSS	$\mathrm{D_{TSS}}$	4.06%	$< 10^{-16}$	7.08%
	Length of the intron	$\mathrm{L_I}$	3.02%	$< 10^{-16}$	
NMD sensitivity	NMD sensitivity*	NMD	2.46%	9.60×10^{-13}	2.46%
Base composition and size of the flanking exons	Length of the following exon	$\mathrm{L_{FE}}$	0.82%	2.53×10^{-7}	2.45%
	GC content of the following exon	$\mathrm{GC_{FE}}$	0.52%	4.00×10^{-6}	
	GC content of the previous exon	$\mathrm{GC_{PE}}$	0.42%	1.20×10^{-2}	
	TG content of the following exon	$\mathrm{TG_{FE}}$	0.36%	9.73×10^{-6}	
	TC content of the following exon	$\mathrm{TC_{FE}}$	0.31%	1.66×10^{-4}	
	TG content of the previous exon	$\mathrm{TG_{PE}}$	0.02%	1.45×10^{-6}	
3n or non-3n introns	3n or non-3n introns*	3N	0.66%	8.84×10^{-4}	0.66%

Nucleosome position	Distance between an intron center and the center of its closest nucleosome	D_N	0.29%	2.69×10^{-3}	0.43%
	Median MNase signal over the intron	M_I	0.12%	1.18×10^{-2}	
	Median MNase signal over the following exon	M_{FE}	0.02%	2.72×10^{-2}	
Secondary structure	Secondary structure free energy of the 50 nt upstream of an intron	ΔG_{p50}	0.19%	1.66×10^{-4}	0.35%
	Secondary structure free energy of an intron and the 50 nt upstream	$\Delta G_{p50\text{-}I}$	0.08%	4.22×10^{-3}	
	Secondary structure free energy of an intron and the 50 nt upstream and downstream	$\Delta G_{fp50\text{-}I}$	0.08%	5.57×10^{-5}	
	Average distance between intron center and the center of the two closest nucleosomes[$]		NA	0.21	
	DeltaGC (Intron-GC content - average flanking-exons-GC contents) [&]		NA	NA	
	Distance between an intron center and its TSS[&]		NA	NA	
	Distance between an intron center and its TTS[&]		NA	NA	
	Length of the previous exon[$]		NA	0.29	
	Median MNase signal over the previous exon[$]		NA	0.30	
	TC content of the previous exon[$]		NA	0.45	

* Binary variables

& Parameter that was excluded after trying to lower the variance inflation factor below 5

$ Not statistically significant parameter