SUPPLEMENTARY INFORMATION: RNA *in situ* conformation sequencing reveals novel long-range RNA structures with impact on splicing

Sergey Margasyuk[1], Marina Kalinina[1], Marina Petrova[1], Dmitry Skvortsov[1,2], Changchang Cao[3], and Dmitri D. Pervouchine[1,*]

[1]Skolkovo Institute of Science and Technology, Moscow 143026, Russia

[2]Moscow State University, Faculty of Chemistry, Moscow 119991, Russia

[3]Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

[*]e-mail: d.pervouchine@skoltech.ru; tel/fax +7 (495) 280 14 81

May 26, 2023

List of Figures

List of Tables

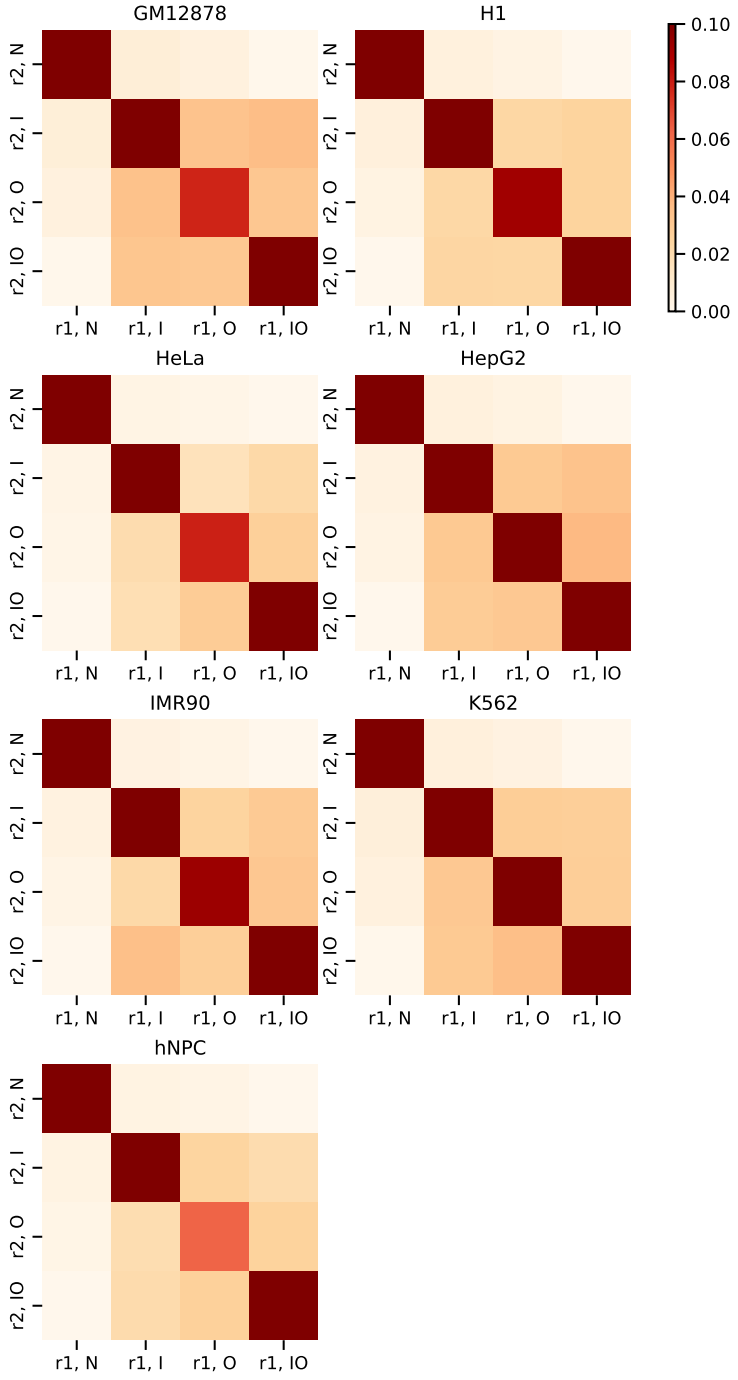


Figure S1: The Jaccard distance $J(A, B) = |A \cap B|/|A \cup B|$ between PCCR sets N, I, O, and IO in human cell lines: GM12878, H1, HeLa, HepG2, IMR90, K562, and hNPC. The values of $J(A, B)$ are indicated by the color map (top right).

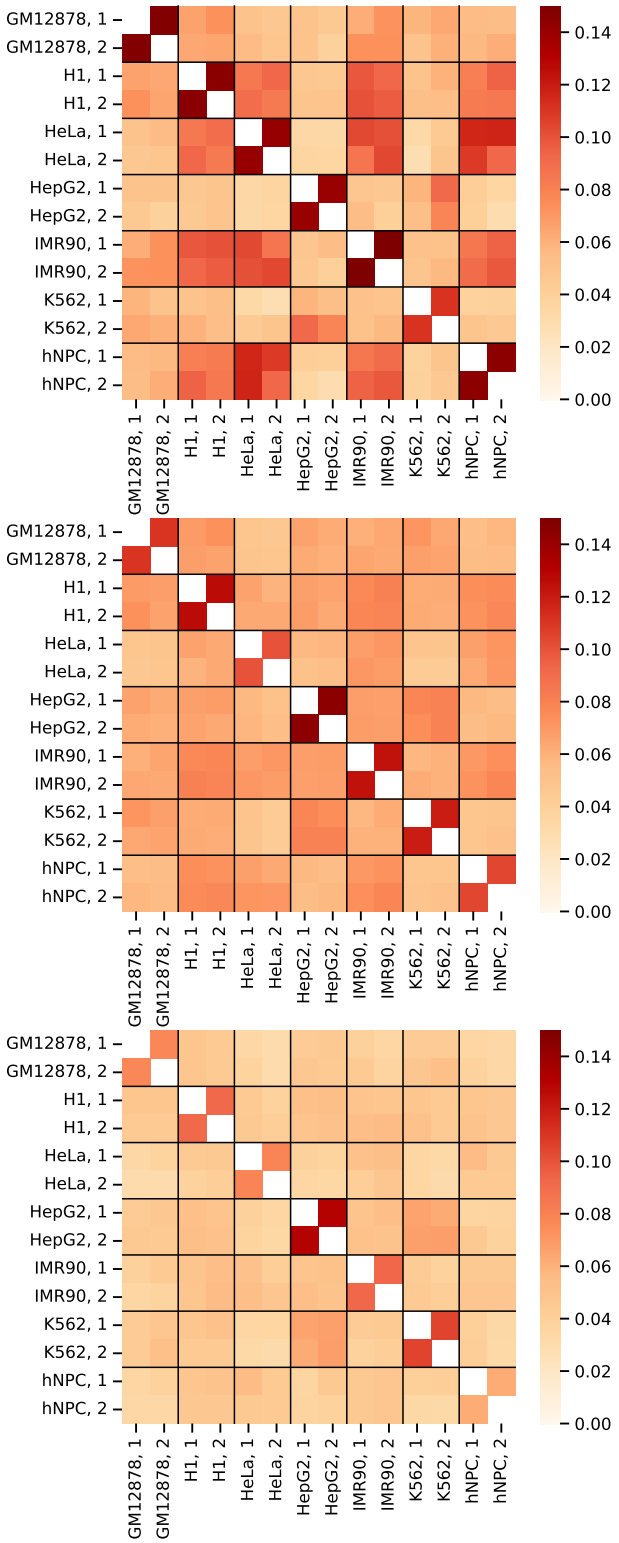


Figure S2: The Jaccard distance ($J(A, B)$) between bioreplicates and cell lines for three PCCR groups: IO (top), I (middle), and O (bottom). The values of $J(A, B)$ are indicated by the color maps (right).

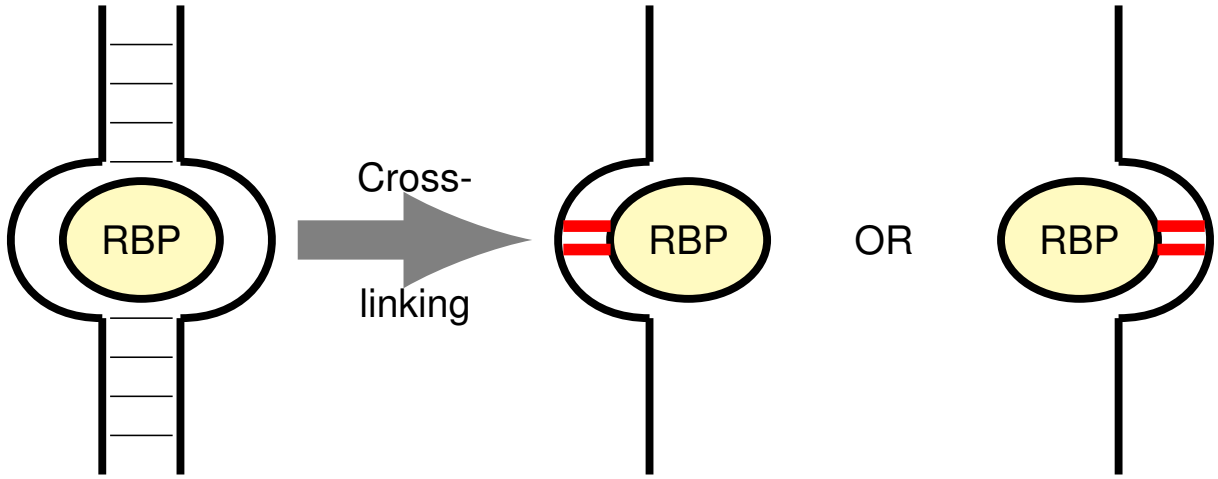


Figure S3: Forked eCLIP peaks occur as a result of RBP crosslinking to one or the other RNA strand near complementary regions. After peak calling, it produces forked RBP footprints near each of the two complementary sequences.

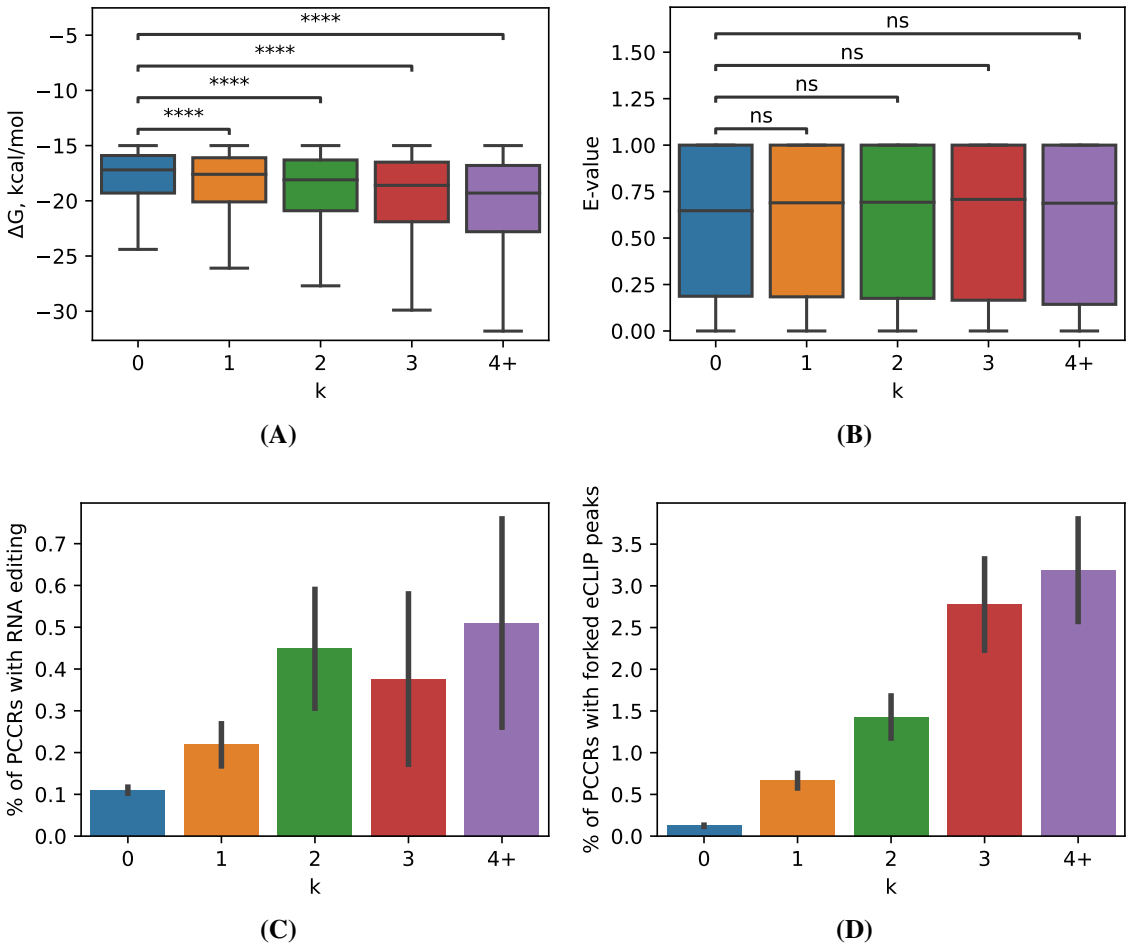


Figure S4: The legend is as in Figure 2, but for PCCRs supported inside (I category) in at least k cell lines.

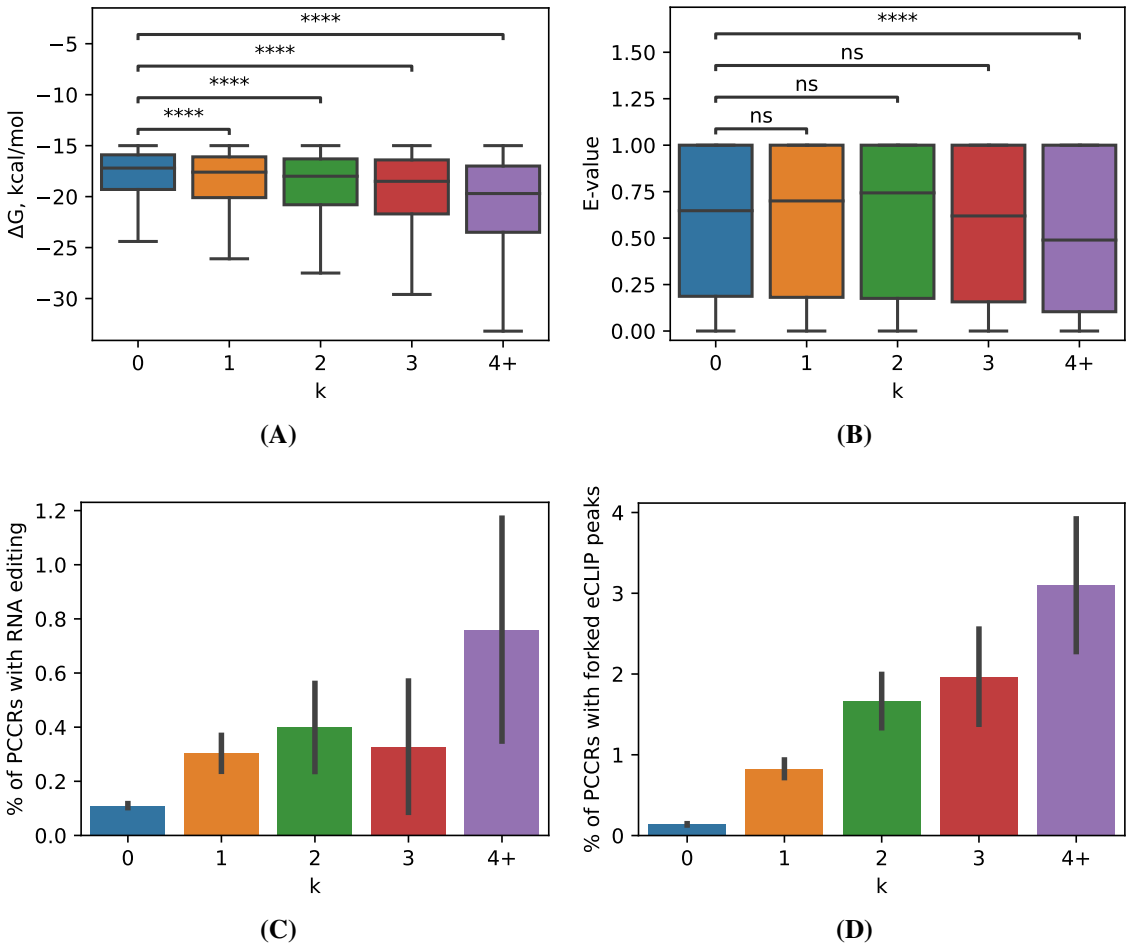


Figure S5: The legend is as in Figure 2, but for PCCRs supported outside (O category) in at least k cell lines.

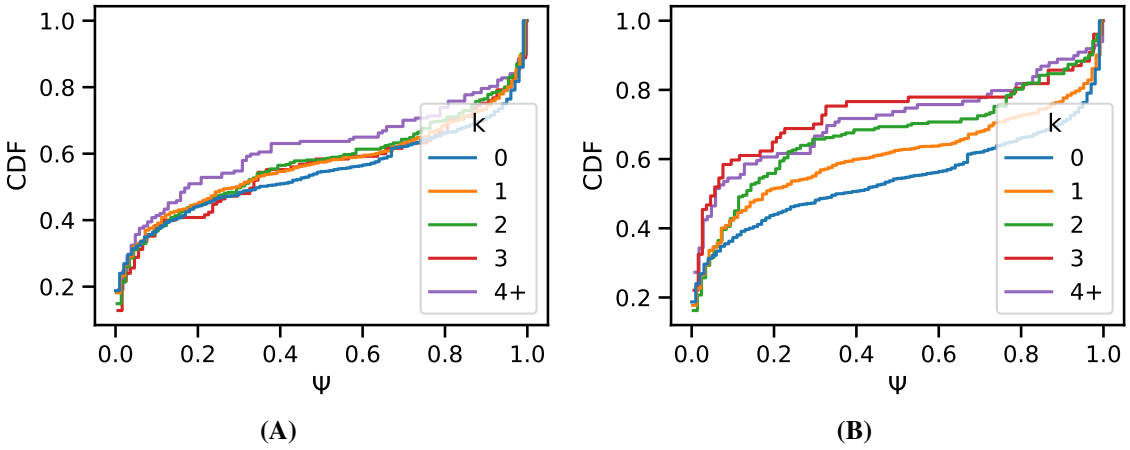


Figure S6: The distribution of average exon inclusion rate (Ψ) of exons looped out by PCCRs that are supported inside (I category, left) and outside (O category, right) in at least k cell lines.

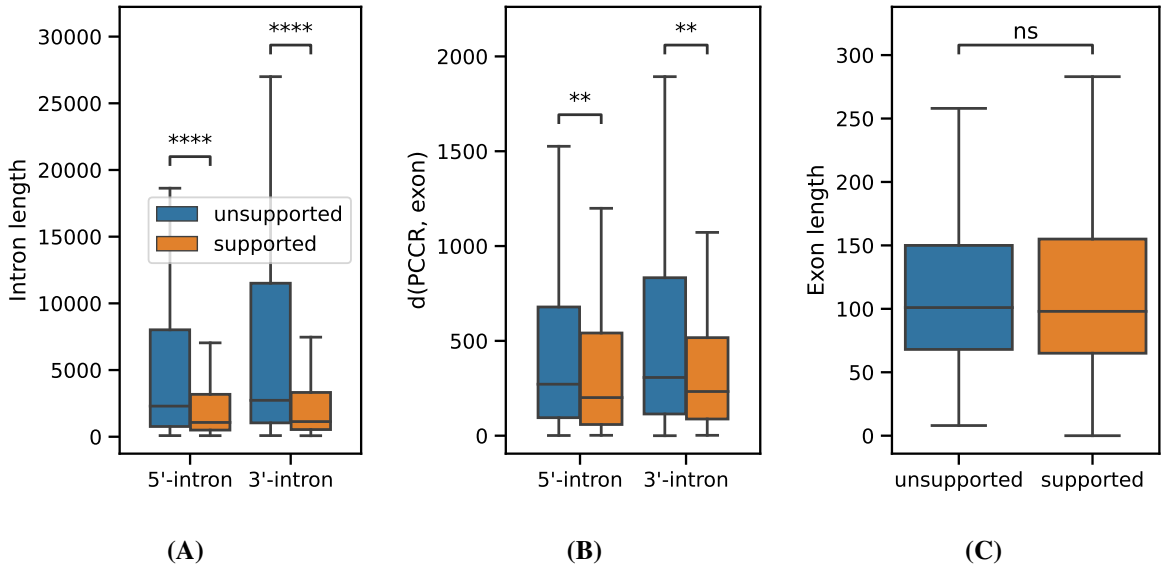


Figure S7: The characteristics of introns and exons looped out by PCCRs with and without RIC-seq support. (A) Lengths of flanking introns (5'-intron and 3'-intron) adjacent to exons looped out by PCCRs. (B) The distance from PCCR to the looped out exon. (C) Lengths of exons looped out by PCCRs.

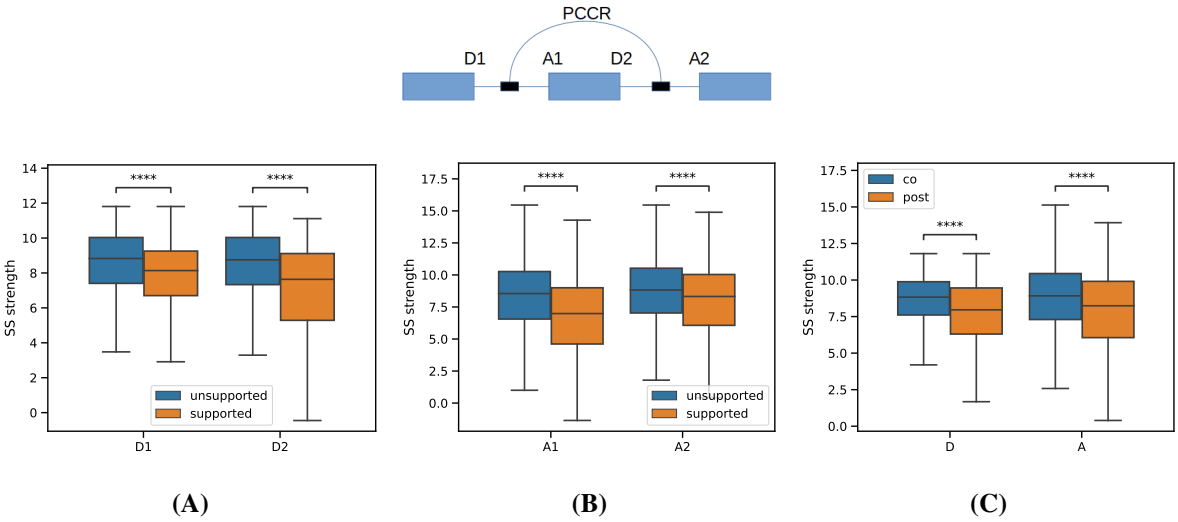
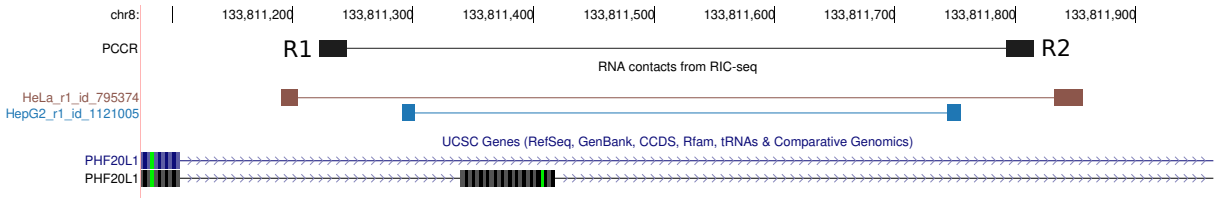
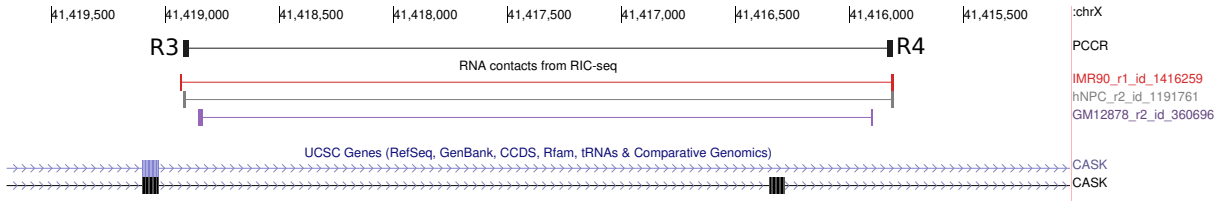


Figure S8: Strengths of donor (panel A) and acceptor (panel B) splice sites around exons looped out by PCCRs. D1 and A1 are the donor and acceptor splice sites of the 5'-intron; D2 and A3 are the donor and acceptor splice sites of the 3'-intron (see ideogram on the top). Strengths of donor (D) and acceptor (A) splice sites in co-transcriptoinal (co) and post-transcriptional (post) introns.



(A)

Figure S9: UCSC genome browser diagram of RIC-seq contacts (clusters of contacts) with respect to R1 and R2 sequences in *PHF20L1*.



(A)

Figure S10: UCSC genome browser diagram of RIC-seq contacts (clusters of contacts) with respect to R3 and R4 sequences in *CASK*.

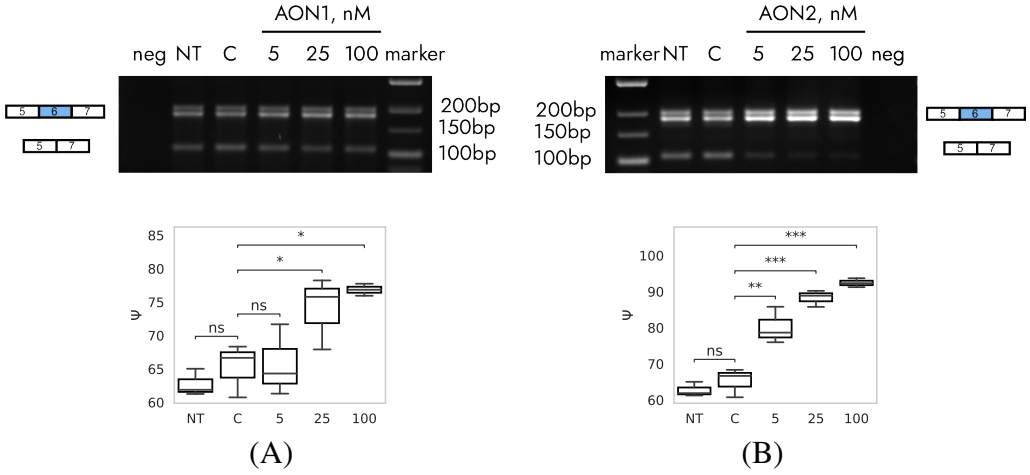


Figure S11: The response of *mPHF20L1* exon 6 splicing to the addition of AON1 (A) and AON2 (B). NT (non-treated control); C (control AON); 5, 25, and 100 are AON1 and AON2 concentrations (nM). Ψ is the exon inclusion rate.

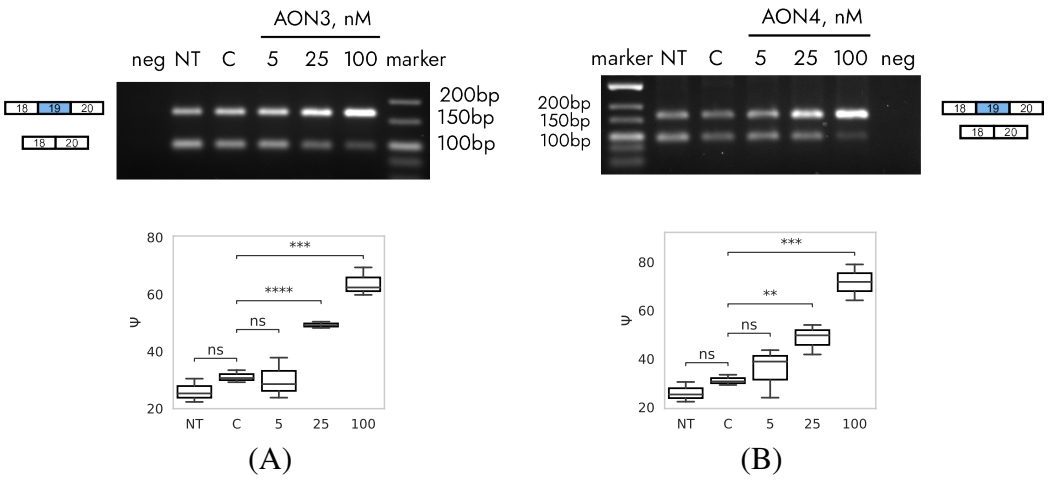
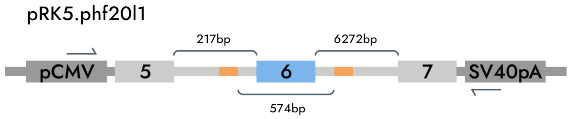
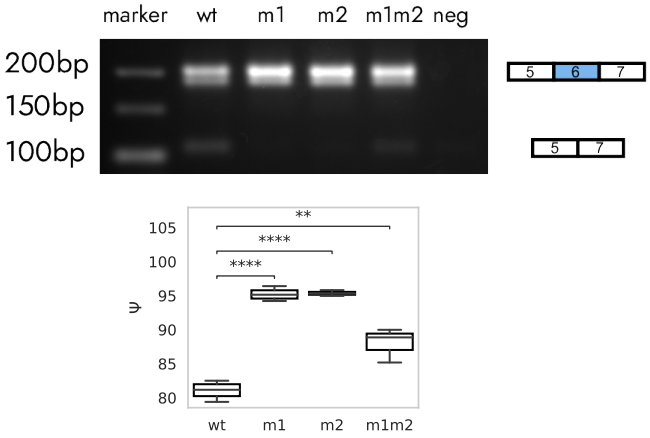


Figure S12: The response of *mCASK* exon 19 splicing to the addition of AON3 (A) and AON4 (B). NT (non-treated control); C (control AON); 5, 25, and 100 are AON3 and AON4 concentrations (nM). Ψ is the exon inclusion rate.



(A)



(B)

Figure S13: *mPHF20L1* mutagenesis. **(A)** A scheme of the *mPHF20L1* minigene. **(B)** Qualitative (top) and quantitative (bottom) assessment of m1, m2, and m1m2 mutants using RT-PCR and qPCR, respectively (see mutagenesis scheme in Figure 5D); neg (negative control).

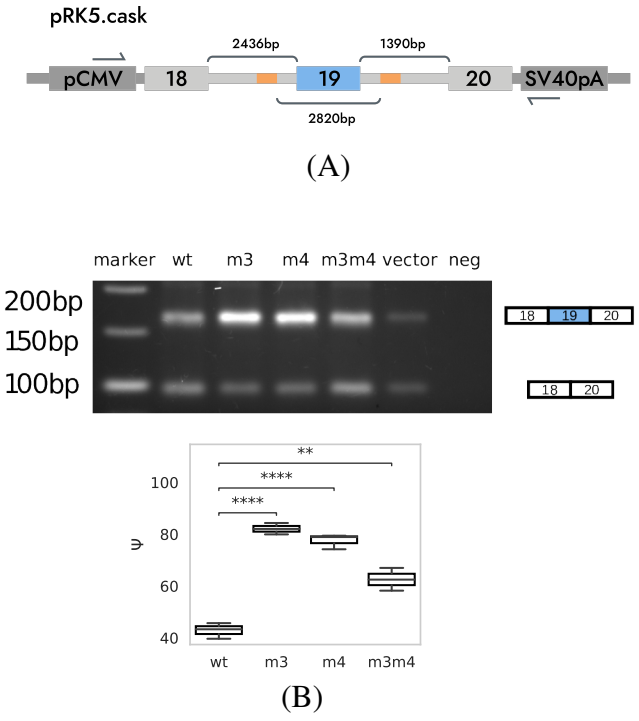


Figure S14: *mCASK* mutagenesis. (A) A scheme of the *mCASK* minigene. **(B)** Qualitative (top) and quantitative (bottom) assessment of m3, m4, and m3m4 mutants using RT-PCR and qPCR, respectively (see mutagenesis scheme in Figure 6D); vector (empty vector-transfected A549 cells); neg (negative control).

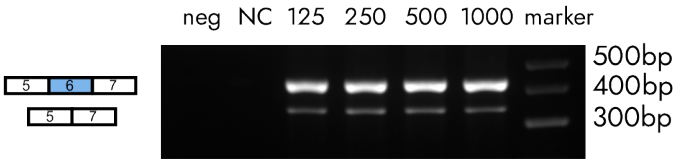


Figure S15: The exon inclusion rate in the *PHF20L1* minigene does not depend on the plasmid concentration. Neg is a negative control (PCR reaction without DNA); NC is transfection that lacks the vector; 125, 250, 500, 1000 are the amounts of the transfected vector (ng).

Experiment	Replicate	n. contacts	n. reads
GM12878	1	3,753,435	7,215,220
GM12878	2	3,671,150	7,231,758
H1	1	2,102,731	4,204,569
H1	2	2,226,229	4,450,153
HeLa	1	861,788	1,701,182
HeLa	2	780,148	1,522,313
HepG2	1	1,223,628	2,418,003
HepG2	2	1,124,851	2,220,530
hNPC	1	1,292,268	2,481,788
hNPC	2	1,214,394	2,323,734
IMR90	1	1,437,933	2,999,298
IMR90	2	1,497,632	3,115,577
K562	1	1,926,823	3,578,811
K562	2	2,110,703	387,927
Total		25,223,713	45,850,863

Table S1: The number of RNA contacts and their supporting split reads.

I\O	0	1	2	3	4	5+
0	0	27,347	4,664	1,357	543	344
1	31,530	7,787	2,340	873	358	204
2	6,035	2,910	1,323	557	266	172
3	1,858	1,286	735	400	232	167
4	745	591	400	234	149	148
5+	436	374	279	257	184	313

Table S2: The number of PCCRs supported by inner contacts in k cell lines (rows) and outer contacts in l cell lines (column).

I\O	0	1	2	3	4	5+
0	NaN	(-18.6, -18.5)	(-18.9, -18.7)	(-19.3, -18.9)	(-19.9, -19.1)	(-21.1, -20.0)
1	(-18.7, -18.7)	(-19.0, -18.9)	(-19.3, -19.0)	(-19.9, -19.3)	(-20.4, -19.4)	(-20.9, -19.6)
2	(-19.4, -19.2)	(-19.5, -19.2)	(-20.0, -19.4)	(-20.1, -19.4)	(-20.7, -19.6)	(-22.0, -20.5)
3	(-19.9, -19.5)	(-20.3, -19.8)	(-20.6, -19.9)	(-21.5, -20.3)	(-22.0, -20.6)	(-21.0, -19.7)
4	(-20.3, -19.6)	(-20.5, -19.8)	(-20.9, -19.9)	(-21.1, -20.0)	(-21.9, -20.2)	(-22.9, -20.8)
5+	(-20.4, -19.6)	(-21.3, -20.2)	(-21.7, -20.6)	(-21.8, -20.5)	(-23.1, -21.6)	(-23.9, -22.5)

Table S3: The equilibrium free energy ranges (95% confidence interval) in PCCRs supported by inner contacts in k cell lines (rows) and outer contacts in l cell lines (column).

Cell line	Accession number
HeLa	ENCSR000CPR
HepG2	ENCSR000CPE
GM12878	ENCSR000AED
H1	ENCSR000COU
IMR90	ENCSR000CTQ
hNPC	ENCSR244ISQ
K562	ENCSR000CPH

Table S4: Accession numbers of RNA-seq experiments used as a matched set control.

AON1	T*+T*G*+C*T*+G*C*+T*A*+T*T*+T*G*+G*G*+G*C*+T
AON2	A*+A*T*+C*C*+C*A*+A*A*+T*A*+A*C*+A*G*+C*A*+G
AON3	+A*A*+A*T*+T*G*+G*T*+G*T*+G*C*+A
AON4	+G*C*+A*C*+A*C*+C*A*+A*T*+T*C*+G

Table S5: AON sequences. DNA nucleotides: G,A,T,C. LNA nucleotides (red): +G, +A, +T, +C. Phosphorothioated DNA nucleotides: G*, A*, T*, C*. Phosphorothioated LNA nucleotides (red): +G*, +A*, +T*, +C*.

hcask_pRK5_fwd	GGGGATCCTCTAGAGTCGACCTGC
hcask_pRK5_rev	GGGGAATTCAATCGATAGAACCGAGG
hcask_cask_fwd	GGGAAATGCGGGGGAGTATT
hcask_cask_rev	TGTCGTCCTTTTGGTTGGGT
mcask_pRK5_fwd	GGGGATCCTCTAGAGTCGACCTGC
mcask_pRK5_rev	GGGGAATTCAATCGATAGAACCGAGG
mcask_cask_clai_fw	TAAGCAATCGATGGGAAATGCGAGGGAGTATTAC
mcask_cask_sali_rv	TGCTTAGTCGACCCTTTTGGTTGGGTGGTTGA
pRK5_phf_rf_fwd	GCACCTCGGTTCTATCGATTGAATTCCCCGC CATGCCCGAGGATGC
pRK5_phf_rf_rev	CTGCAGGTCGACTCTAGAGGATCCCCTCTCT ACGTCTGGTGTGGCTA
phf20l1_mm_fw	ttctatcgattgaattccccACAGTTCAGTTTTATGATGG
phf20l1_mm_rv	gtcgactctagaggatccccCAAATGTTTCACTGGATG

Table S6: Cloning primers.

hcask_m1_fwd	CGTGTGGTTAAATACAACAAAGATAACGGAACAG
hcask_m1_rev	AGTGATCTAACAGCACAGG
hcask_m2_fwd	CTTAACCACACGCCTCTGTTATTGTGGTACGTGCAC
hcask_m2_rev	GTAAACTTAGATAGCGGAAACTAAA
mcask_m1_fwd	CGTGTGGTTAAATACAACAAAGATAACGGAACAG
mcask_m1_rev	AGTGATCTAACAGCACAGG
mcask_m2_fwd	CTTAACCACACGCCTCTGTTATTGTGGTACGTGCAC
mcask_m2_fw_mm	CTTAACCACACGCCTCTTGTTATTGTGGTACGTGCAC
phf_m1_rear_f	GGGTTTATCGTCCAAGAAAAGGAAGCTTAGACTGTAAA
phf_m1_rear_rev	GCTATCAGCACATAACCCTACAATTA
phf_m2_rear_fwd	GACAATAAACCCATTTTCAGCACCTGGACCTACAAT
phf_m2_rear_rev	CAGAATAGAAAAGAAAATATTTTAAATTTCTAAT

Table S7: Mutagenesis primers.

hcask_rtpcr_cask_fwd	GGGAAATGCGGGGGAGTATT
hcask_rtpcr_cask_rev	TGTCGTCCTTTTGGTTGGGT
mcask_rtpcr_cask_fw	GGGAAATGCGAGGGAGTATTAC
mcask_rtpcr_cask_rv	CCTTTTGGTTGGGTGGTTGA
rtpcr_phf20l1_fwd	GCCATGCCCGAGGATGCTAA
rtpcr_phf20l1_rev	TCTCTACGTCTGGTGTGGCTATAC

Table S8: RT-PCR primers.

qpcr_mcask_fw_mm	GGGAAATGCGAGGGAGTATTAC
qpcr_mcask_rv_mm_in	TGATGGCAAGTCCGAAACAG
qpcr_mcask_rv_mm_ex	GTTGATGGCAAGTCCTCACA
qpcr_gapdh_mm_fw	TTCAACAGCAACTCCCACTCTT
qpcr_gapdh_mm_rv	GCCGTATTCATTGTCATACCAG

Table S9: Mouse Cask qPCR primers.

SupplementaryDataFile 1: The list of 134 PCCRs with alternative RIC-seq support and alternative splicing across cell lines.

SupplementaryDataFile 2: The list of 115,162 PCCRs with RIC-seq support in at least one cell line. Class/I and class/O columns contain the number of cell lines with RIC-seq support from inside and outside, respectively.