Supplementary information

Study of the RNA splicing kinetics via *in vivo* 5-EU labelling

Anastasiia K. Bolikhova[1,2#], Andrey I. Buyan[3,4#], Sofia S. Mariasina[2,4,5], Alexander Y. Rudenko[2], Daria S. Chekh[7], Alexander M. Mazur[6], Egor B. Prokhortchouk[6], Olga A. Dontsova[1,2,4,8], Petr V. Sergiev[1,2,4,9*]

[1] Center of Life Sciences, Skolkovo Institute of Science and Technology, Skolkovo, 121205, Russia;

[2] A.N. Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, 119991, Moscow, Russia.

[3] Institute of Protein Research, Russian Academy of Sciences, Pushchino, 142290, Russia.

[4] Department of Chemistry, Lomonosov Moscow State University, Moscow, 119991, Russia

[5] Faculty of Fundamental Medicine, Lomonosov Moscow State University Moscow, 119991, Russia.

[6] Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, 119071, Russia

[7] Faculty of Biology, Lomonosov Moscow State University Moscow, 119991, Russia.

[8] Department of Functioning of Living Systems, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, 117997, Russia;

[9] Institute of Functional Genomics, Lomonosov Moscow State University, Moscow, 119991, Russia.

[#] Equal contribution

* To whom correspondence should be addressed. Email: petya@genebee.msu.ru (P.V.S.)

SUPPLEMENTARY FIGURES

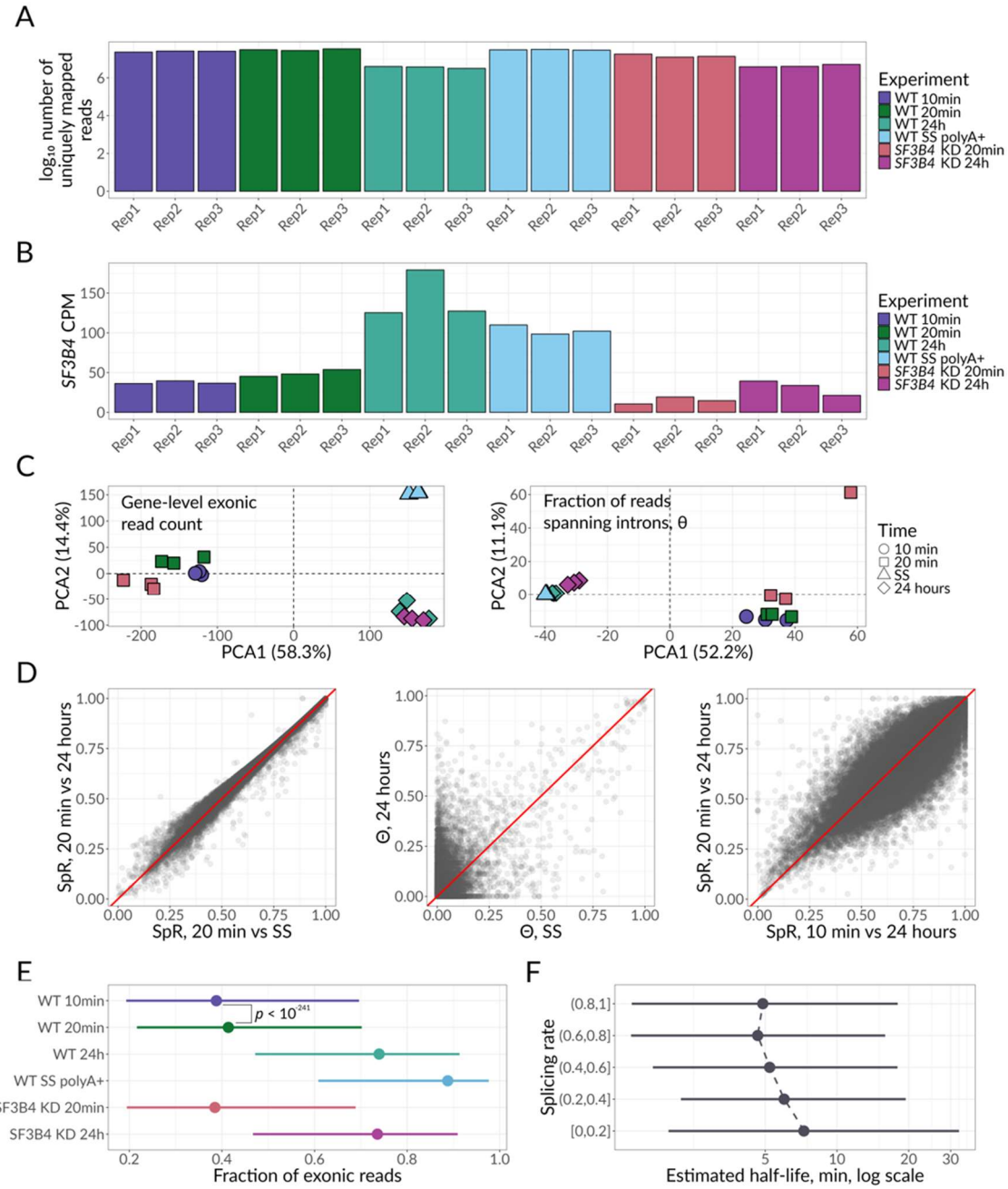

Figure S1. The results of EU-RNA-Seq quality check. **(A)** The number of uniquely mapped reads. Colour corresponds to cell line and kinetic point. **(B)** Relative SF3B4 expression in different cell lines. Y-axis: counts per million mapped reads. Colour scheme is the same as in (A). **(C)** First and second PCA principal components for samples based on gene counts (left) and Θ (right). Colour scheme is the same as in (A). **(D)** Correlation between SpR calculated using 20 minutes compared to steady state or 24 hours (left), Θ estimates for steady state and 24 hour kinetic point (middle) and SpR obtained using 24 hour and 10 or 20 minute kinetic points (right). Red line corresponds to y = x. **(E)** Gene-wise fractions of exonic reads for different kinetic points and cell lines. Dots correspond to medians with lower and upper error bars showing 25% and 75% quantiles. Colour scheme is the same as in (A). *P*-value was calculated using a one-sided Wilcoxon signed rank test. **(F)** Bond half-life

estimations from (1) for different SpR bins illustrated in the same manner as in (E), ART ANOVA *P*-value < 10^{-15}.

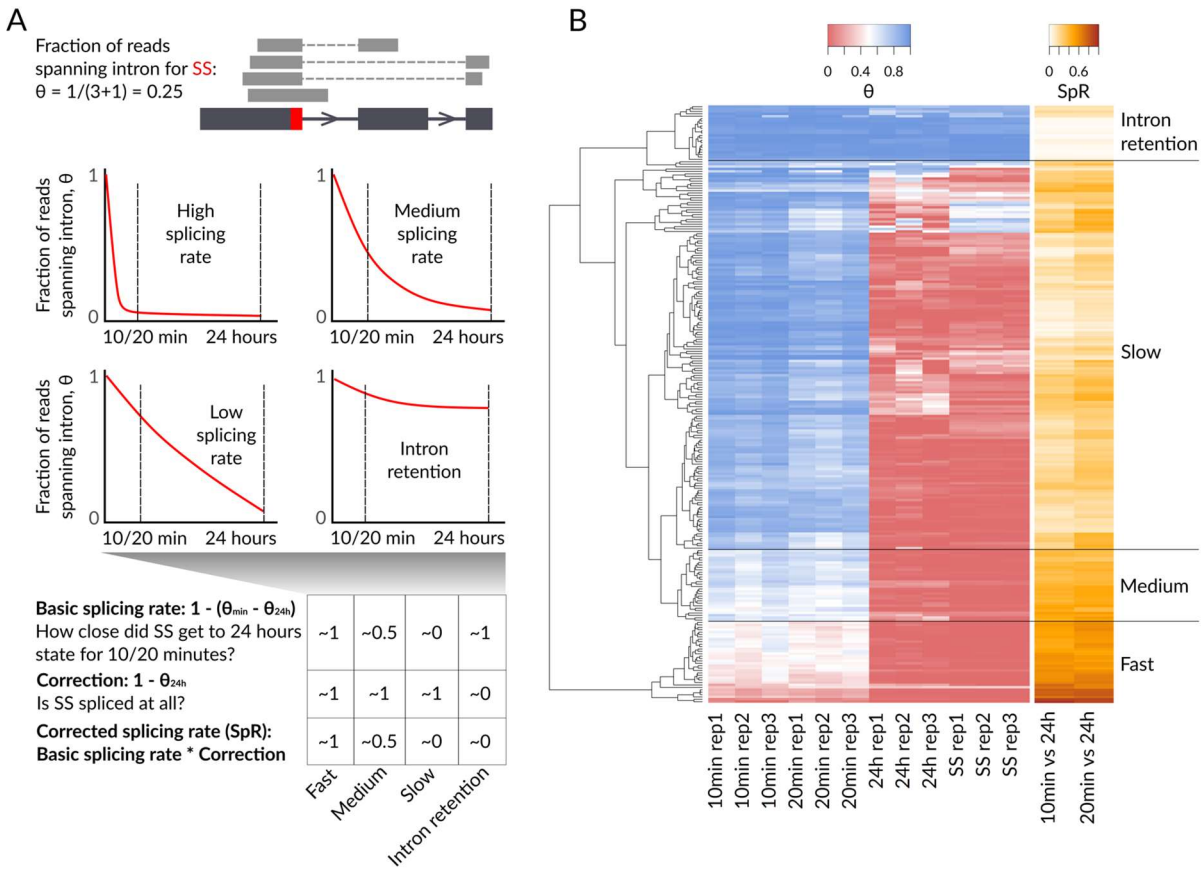

Figure S2. Splicing rate calculation. **(A)** Top: scheme for obtaining Θ values. Middle: examples of the splice sites with different splicing kinetics. Bottom: formulas to calculate corrected SpR with the approximate values corresponding to fast, medium and slow splice sites as well as sites of the retained introns. **(B)** EU-RNA-Seq based SpR calculation results: Θ and SpR estimates for a subset of 249 splice sites with at least 10 counts per million mapped non-split reads covering splice sites in at least 3 samples. Hierarchical clusterization was performed with the complete linkage method using euclidean distances. Splice site groups were manually annotated as slow, medium, fast and those exhibiting intron retention.

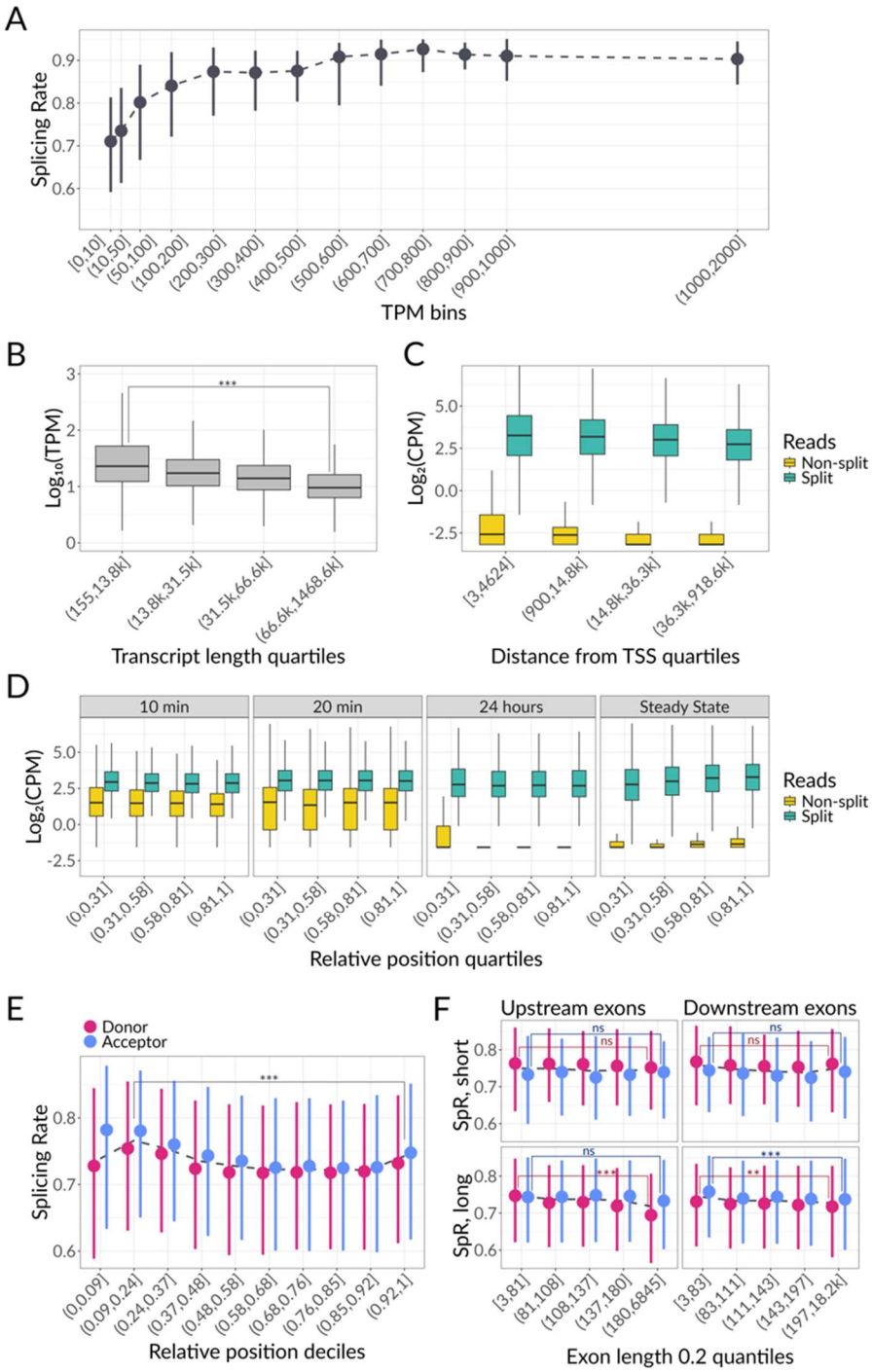

Figure S3. EU-RNA-Seq reads distribution in transcripts. **(A)** SpR association with the expression level bins. Dots correspond to medians with lower and upper error bars showing 25% and 75% quantiles. Dashed line represents the overall median tendency. **(B)** The dependency of the transcript abundance on the transcript length in steady state samples, ART ANOVA *P*-value < 10^{-15}. **(C)** Normalised number of split (green) and non-split (yellow) reads for different distance from TSS quartiles in steady state samples. **(D)** Normalised number of split and non-split reads depending on the relative position (the distance from TSS to SS relative to the overall transcript length). Colour scheme is the same as in (B). The increase in the number of split and non-split reads near the transcript end for steady state may reflect the effect of polyA selection. **(E)** Dependency of the SpR on the SS relative position illustrated in the same manner as in (A). ART ANOVA *P*-value < 10^{-15} for both

Position and SS type * Position interaction terms. Colour corresponds to donor and acceptor SS. **(F)** Dependency of the SpR on the adjacent exon length for short (≤ 134 nt, top) and long (> 134 nt, bottom) introns illustrated in the same manner as in (A). Colour corresponds to donor and acceptor SS. ART ANOVA *P*-value is significant only for long introns, being < 10^{-15} and 10^{-12} for upstream exon length and SS type * length interaction terms, respectively, and 10^{-10} for downstream exon length. The significance of the Tukey test between groups is depicted as * for *P*-value < 0.05, ** for < 0.01 and *** for < 10^{-4}, ns stands for non-significant.

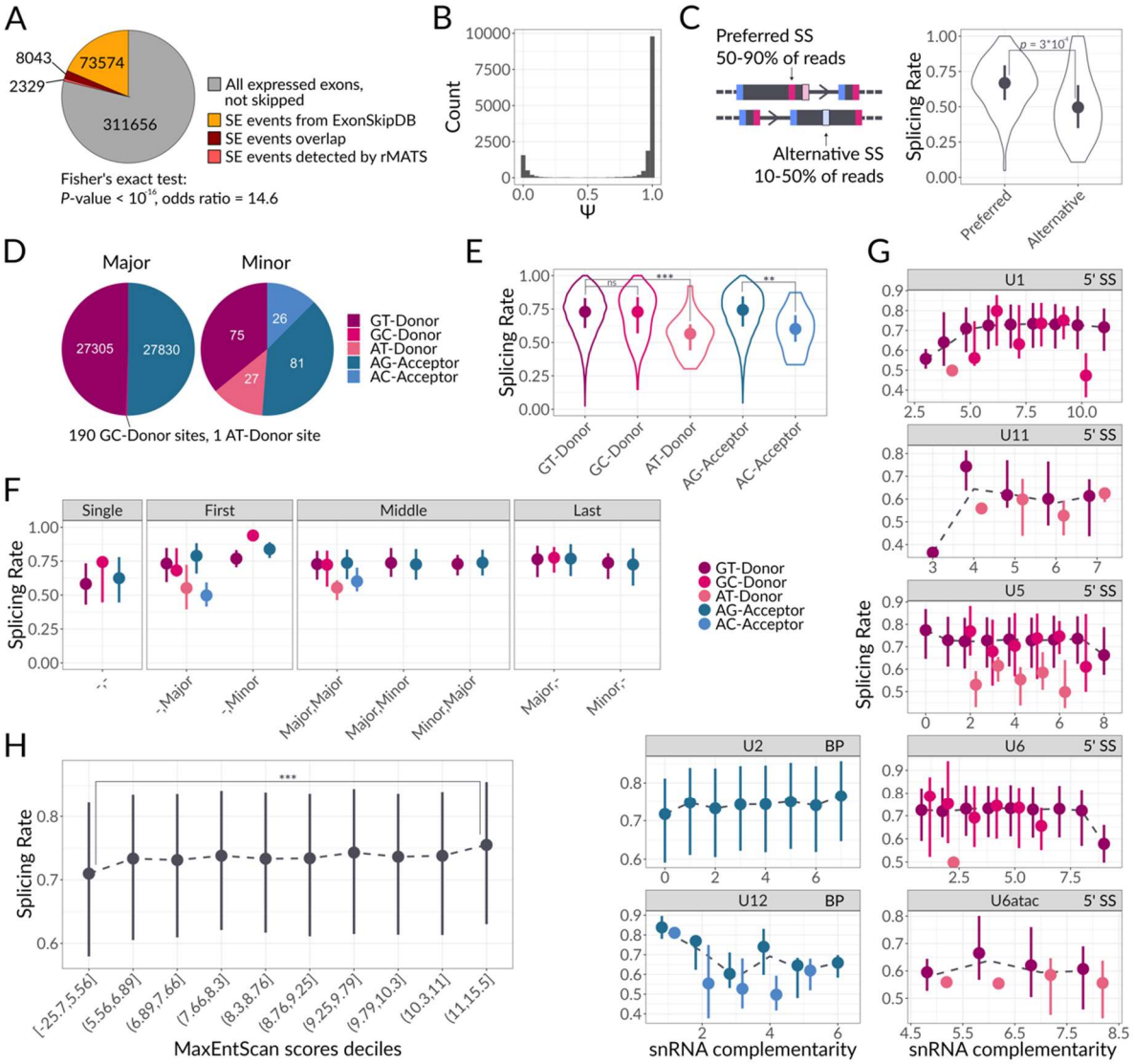

Figure S4. Additional splice site properties influencing SpR. **(A)** Pie chart demonstrating number of all expressed exons that are not skipped according to both ExonSkipDB and our data, number of exon skipping events identified solely by rMATS using our data, number of events annotated in ExonSkipDB that were not found by rMATS, and the intersection of these two. Results of one-sided Fisher's exact test are shown below. **(B)** Distribution of the percent spliced in (PSI, Ψ) value for the exon skipping events identified by rMATS. **(C)** Left: scheme illustrating how the preferred and alternative SS were identified. Right: SpR of the preferred and alternative SS involved in alternative splicing, one-sided Wilcoxon rank sum test *P*-value = $3*10^{-4}$ for preferred vs alternative SS. Dots correspond to medians with lower and upper error bars showing 25% and 75% quantiles. **(D)** The number of different types of donor and acceptor splice sites recognized by major and minor spliceosomes. **(E)** SpR for different SS types illustrated in the same manner as in (C). ART ANOVA *P*-value < 10^{-15}. **(F)** SS splicing rate dependency on the presence of the neighbour splice junctions. Colour scheme is the same as in (D). **(G)** Dependency of the donor and acceptor SS splicing rate on the complementarity to various snRNAs, colour scheme is the same as in (D). Most significant ART ANOVA *P*-values included = 0.008 for U1 complementarity to GT-donors and 0.0002 for U2 complementarity to AG-acceptors. **(H)** Splicing rate dependency on MaxEntScan scores, depicted in the same way as in (C), ART ANOVA *P*-value < 10^{-15}. The

significance of the Tukey test between groups is depicted as * for *P*-value < 0.05, ** for < 0.01 and *** for < 10^{-4}, ns stands for non-significant.

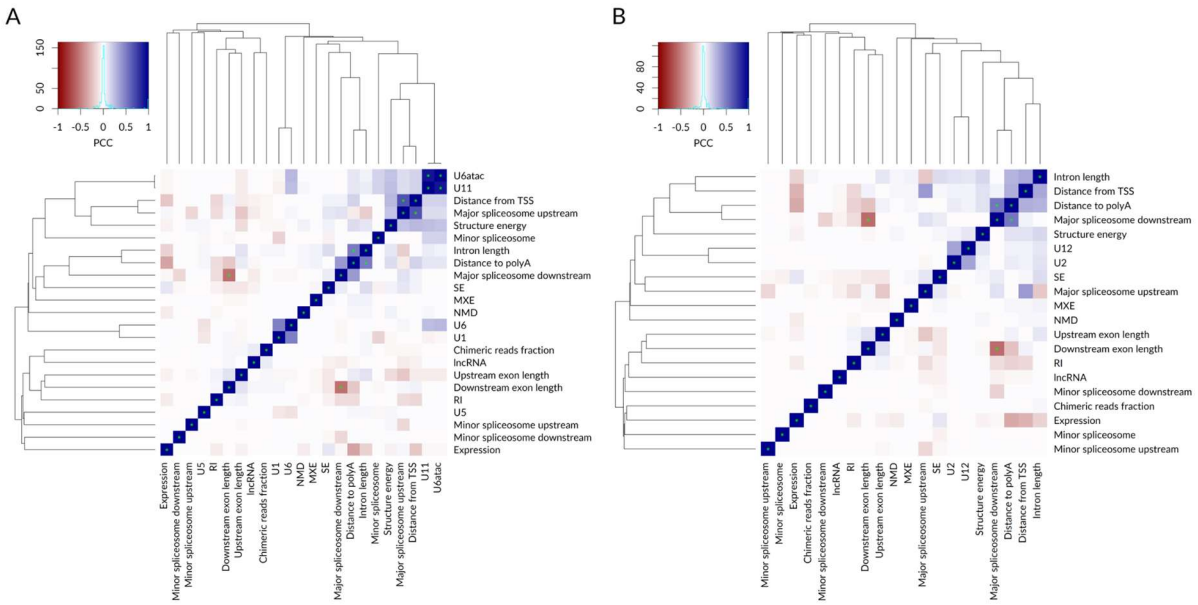

Figure S5. Pearson correlation coefficients (PCCs) for the SS features. **(A)** PCCs between different independent variables used to build the CDF-Quantile regression model for donor SS splicing rates. Colour corresponds to PCC value, asterisk shows strong correlations with FDR < 0.05 and the absolute PCC value > 0.5. Hierarchical clusterization was performed with the average-linkage method (UPGMA) using cosine distances. **(B)** PCCs between different independent variables used to build the CDF-Quantile regression model for acceptor SS splicing rates, built in the same way as in (A).

SUPPLEMENTARY REFERENCES

1. Wachutka,L., Caizzi,L., Gagneur,J. and Cramer,P. (2019) Global donor and acceptor splicing site kinetics in human cells. *eLife*, **8**, e45056.