Supplementary Data

Supplemental Data 1: Gene-level TPM expression levels (*Cell_RSEM*), splicing efficiencies and specificities (in WCE, Cyto and Nuc fractions), splicing efficiencies of worst intron (in WCE, Cyto and Nuc fractions), localization (Cyto/Nuc log2 ratios), and classified localization (cytoplasmic/intermediate/nuclear) for nine ENCODE cell lines.

Supplemental Data 2: Intron-level splicing efficiencies for nine ENCODE cell lines. The *Intron.pos* column contains our intron index, and coordinates of the intron in hg19.

Supplemental Data 3: Pol2 pausing indices and Pol2 occupancy over introns for six ENCODE cell lines.

Supplemental Data 4: ChIP mean signals of chromatin modifications over introns and over exon-intron junctions for seven ENCODE cell lines.

Supplemental Data 5: Gene architecture and sequence features - length parameters, splice-site scores and conservation, hexamers' enrichment and nucleotide enrichment in 3rd positions of codons (*codonPref*).

Supplemental Data 6: A list of the features used for linear regression models and random forest classifiers (Fig. 5) with a brief description and data sources. For more details, see Methods section.

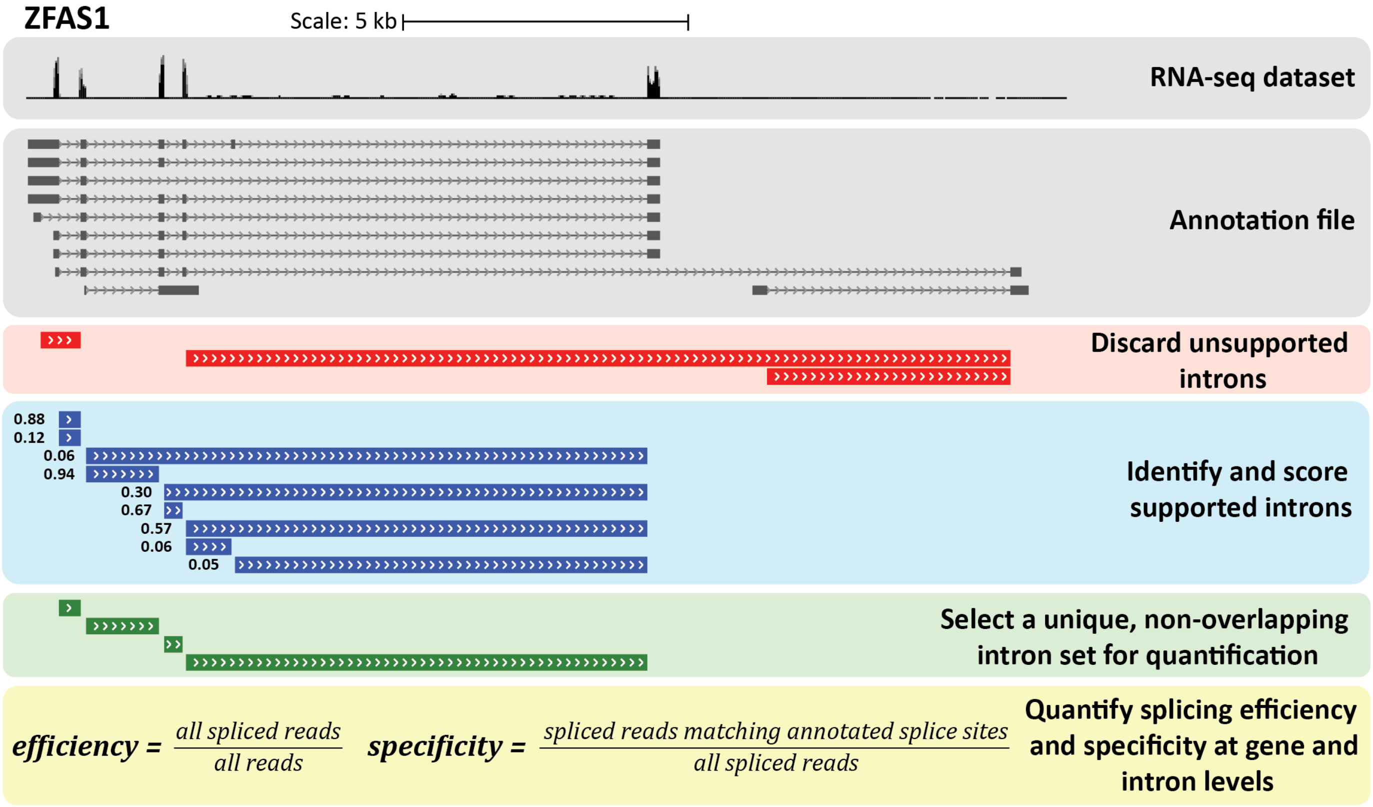
Supplemental Data 7: (**A**) AUC, precision and recall mean values for the three localization classes in indicated cell lines as calculated by 100 repeats of 10-fold cross validation. SD is indicated in parentheses. (**B**) AUC, precision and recall values of models trained and tested on pairs of cell lines. (**C**) AUC, precision and recall values of models trained on data from human HepG2 and K562 cell lines and tested on mouse liver data and vice versa.

Supplemental Data 8: Gene-level expression levels (Liver_*Cell_RSEM*), splicing efficiencies and specificities (in WCE, Cyto and Nuc fractions), splicing efficiencies of worst intron (in WCE, Cyto and Nuc fractions) and localization (Cyto/Nuc log2 ratios) for mouse liver.

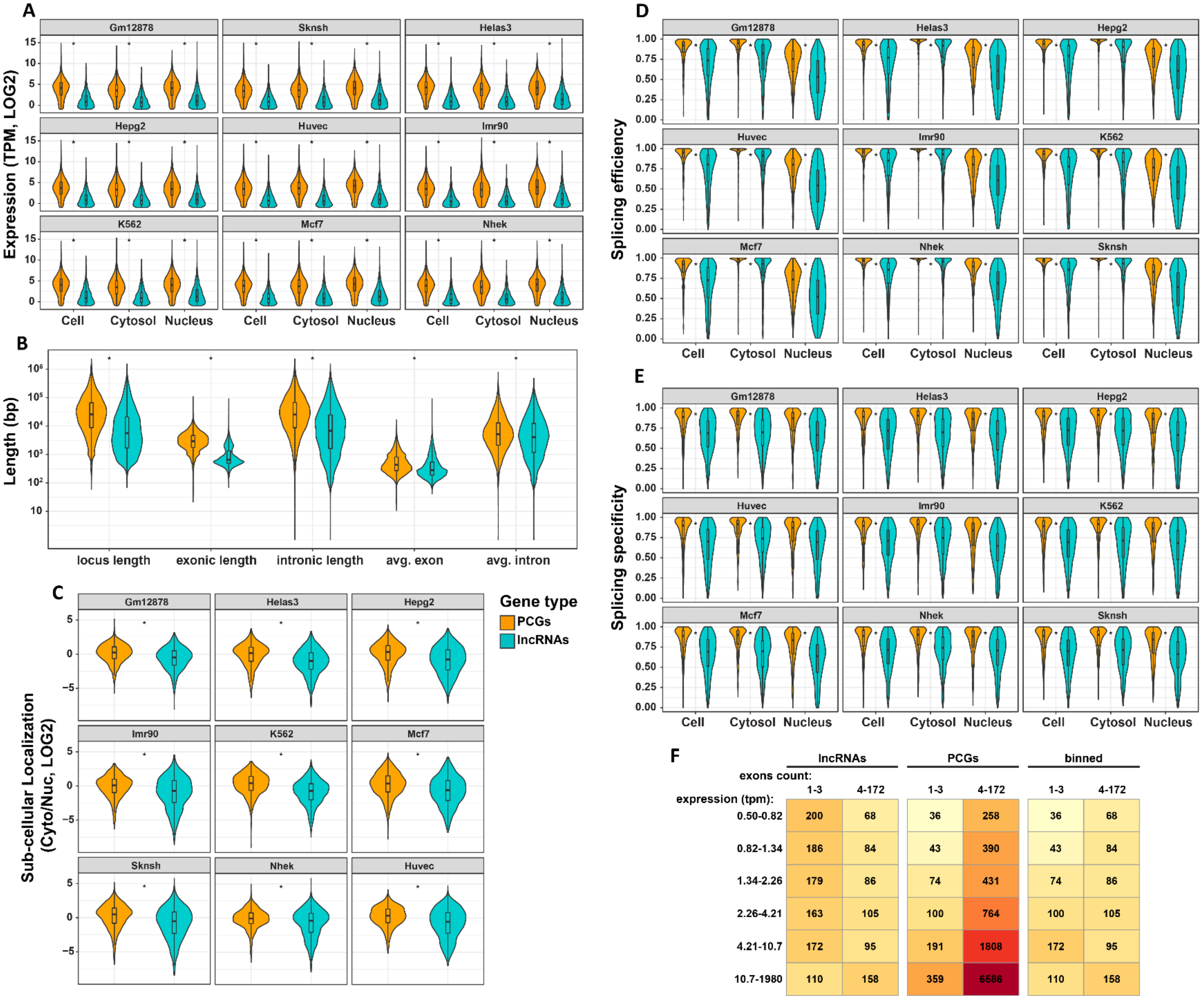
Supplemental Data 9: Human-mouse orthologs list, gene IDs refer to GENCODE v26 and M13 annotations for human and mouse, respectively.

Supplemental Data 10: GO enrichments in ranked lists of nuclearly retained PCGs for nine ENCODE cell lines.

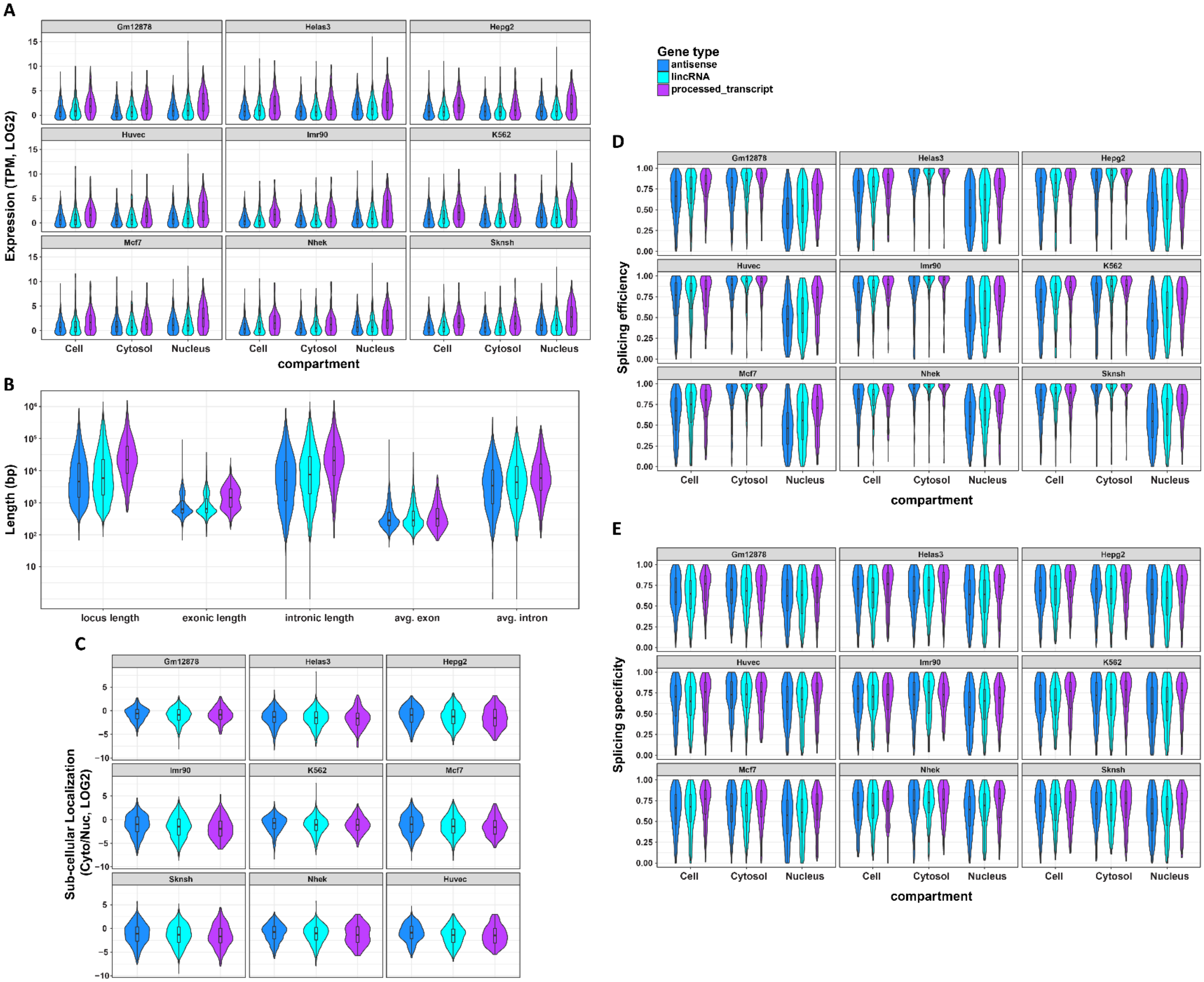
Supplementary Figures


ZFAS1
Scale: 5 kb
RNA-seq dataset
Annotation file
Discard unsupported introns
0.88
0.12
0.06
0.94
0.30
0.67
0.57
0.06
0.05
Identify and score supported introns
Select a unique, non-overlapping intron set for quantification
efficiency = all spliced reads / all reads
specificity = spliced reads matching annotated splice sites / all spliced reads
Quantify splicing efficiency and specificity at gene and intron levels

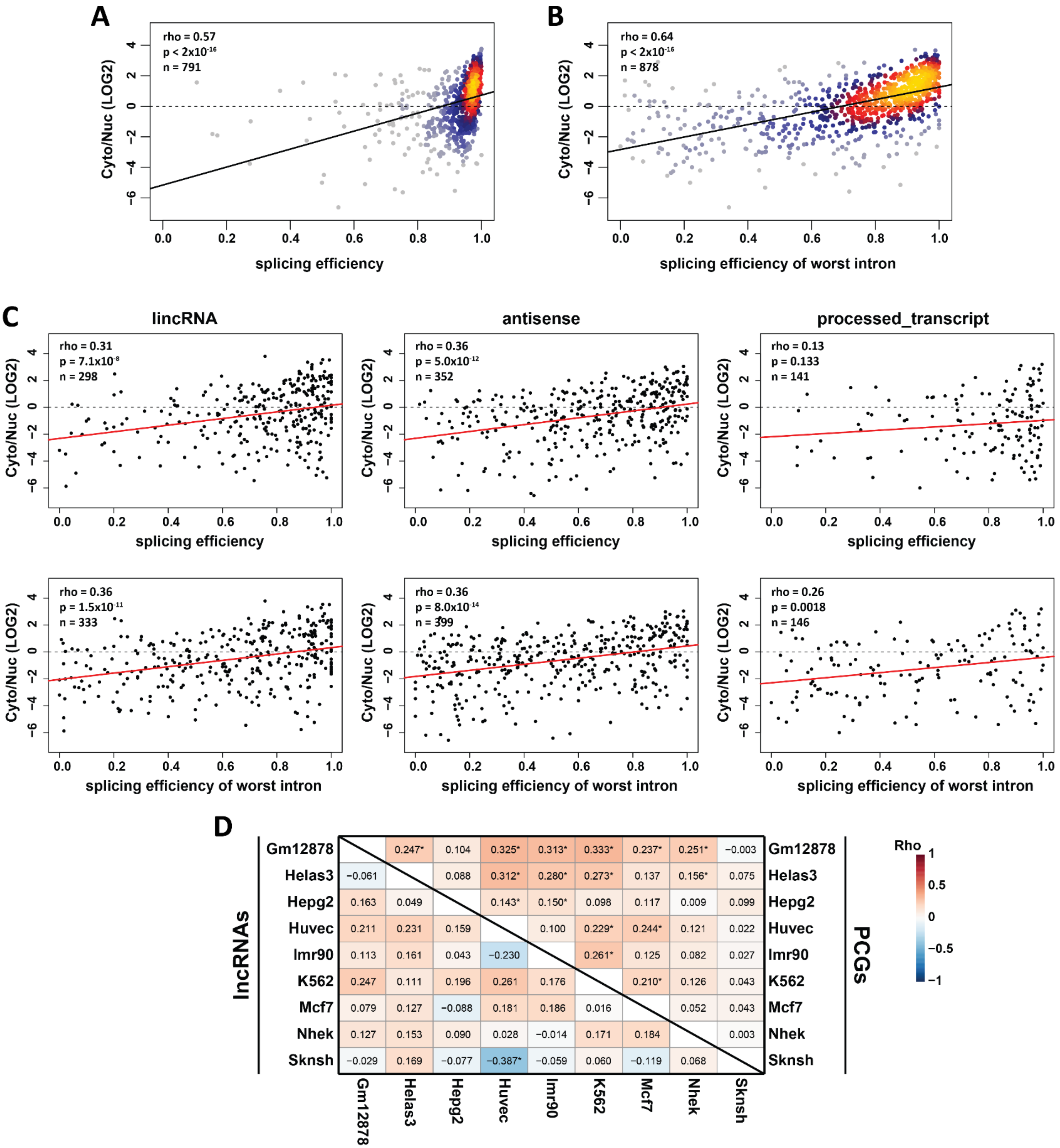
Supplemental Figure 1. An additional example of gene-level splicing quantification. As in Fig. 1, for the lncRNA *ZFAS1*.



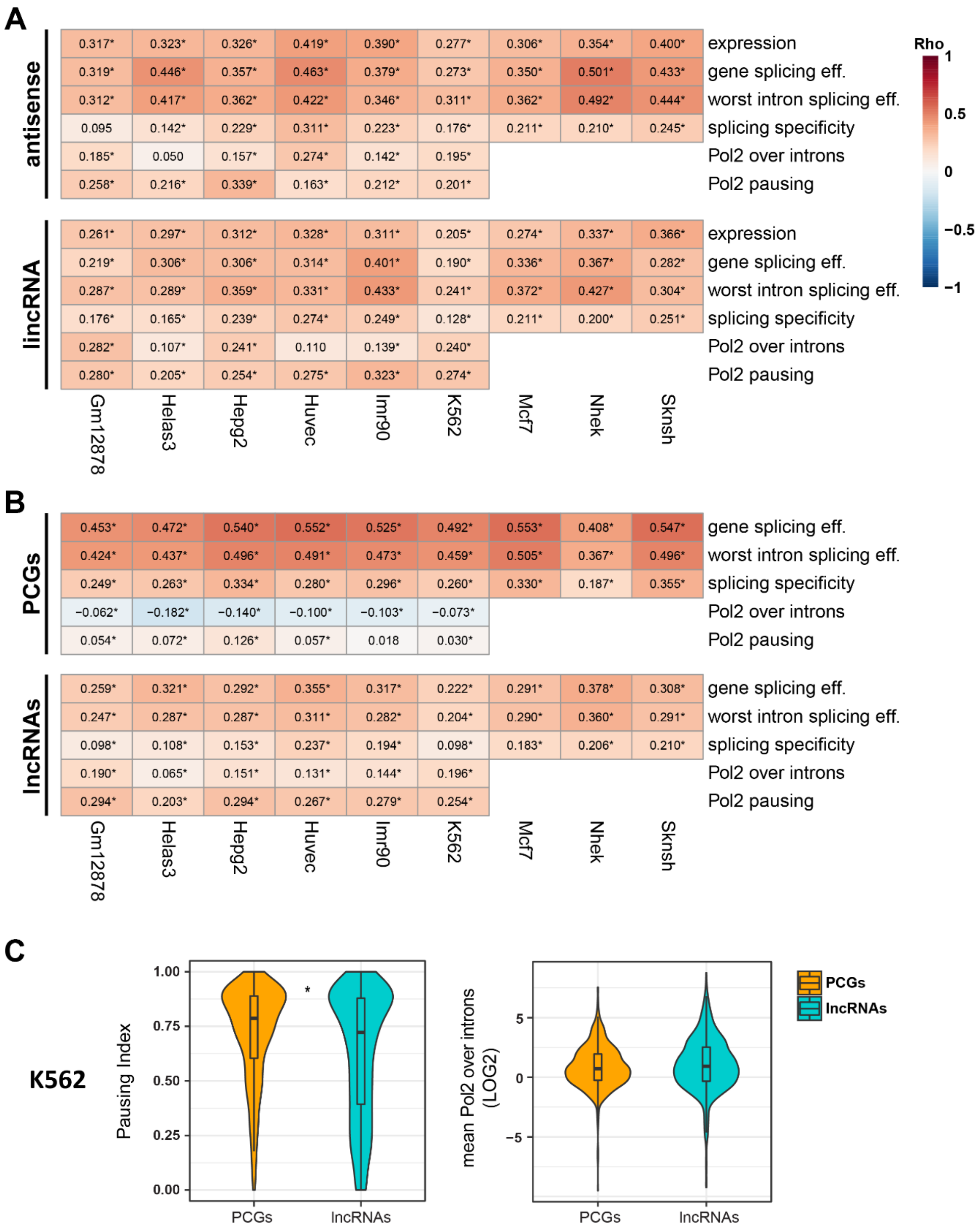
Supplemental Figure 2. Distribution of different parameters in PCGs and lncRNAs. (A) Distributions of expression levels for the nine ENCODE cell lines. (**B**) Gene-level length parameters of PCGs and lncRNAs. Maximal value among all gene isoforms were used. (**C-E**) Subcellular localization, splicing efficiency and specificity distributions for lncRNAs and a sample of PCGs matched for their expression and exon counts to the lncRNAs (see Methods). Asterisks indicate $P<10^{-6}$ for localization and $P<10^{-16}$ for the other parameters (Wilcoxon rank sum test). (**F**) Example of the binning approach used to generate gene lists for panels C-E. Binning was initially performed on lncRNAs (left), using exon counts (columns) and expression levels (rows). Numbers and color indicate amount of genes in each bin with non-missing subcellular localization values in K562 dataset. PCGs were classified using the same limits (middle). The final amount of genes per bin used for plotting localization distributions are shown in the right.



Supplemental Figure 3. (**A-B**) As Supplemental Figure 2A-B, for different classes of lncRNAs as defined by GENCODE. (**C-E**) As Fig. 2B-D, for different classes of lncRNAs as defined by GENCODE and nine ENCODE cell lines.



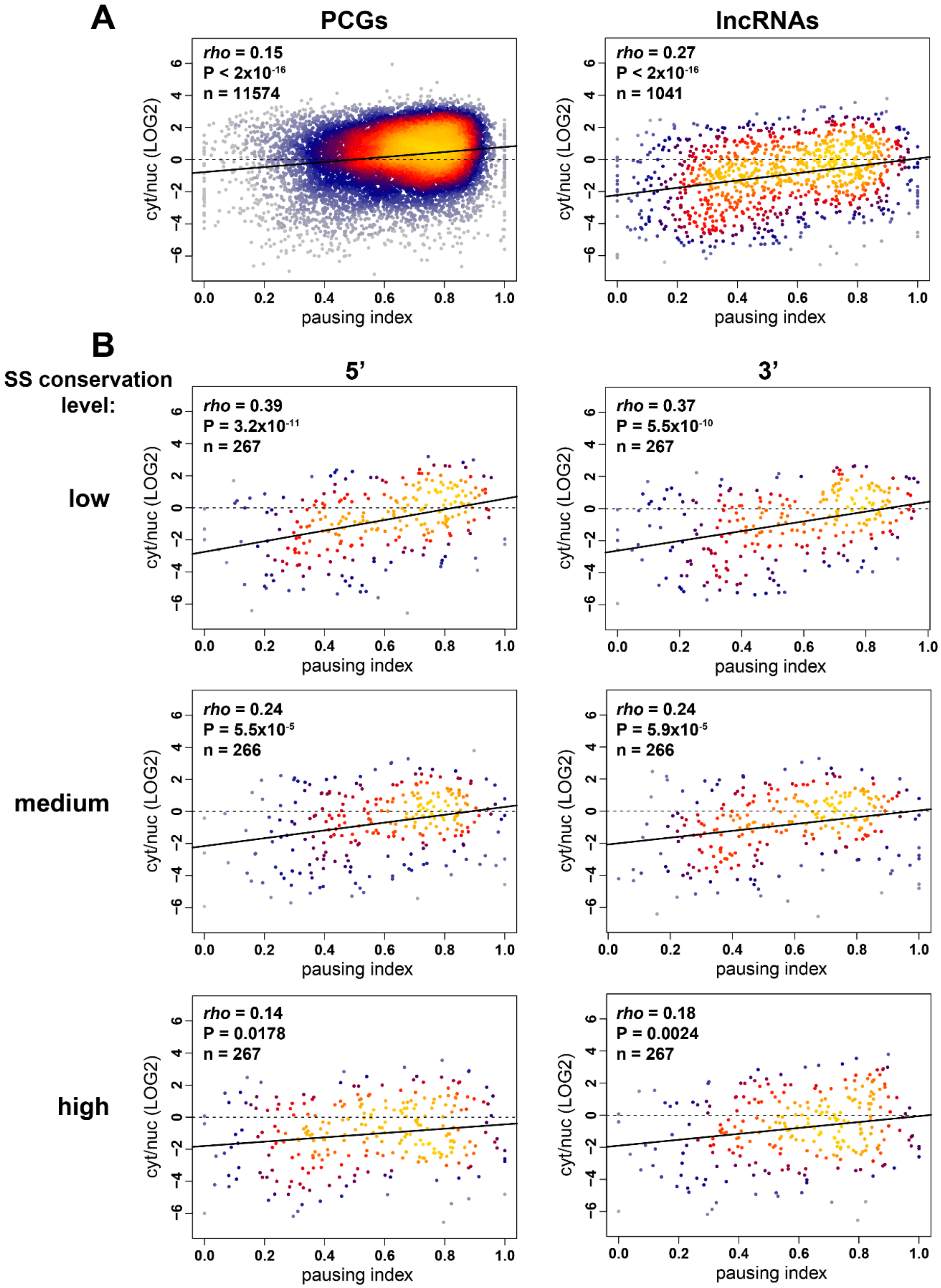
Supplemental Figure 4. (**A**) As in Fig. 3A, random sampling of PCGs to match the amount of lncRNA shown in Fig. 3B. (**B**) As in Fig. 3C, random sampling of PCGs to match the amount of lncRNA shown in Fig. 3D. (**C**) As in Fig. 3B, showing splicing-localization correlations separately for different classes of lncRNAs as defined by GENCODE. (**D**) Same as Fig. 3E, but showing correlations between differences in splicing *specificity* and localization.



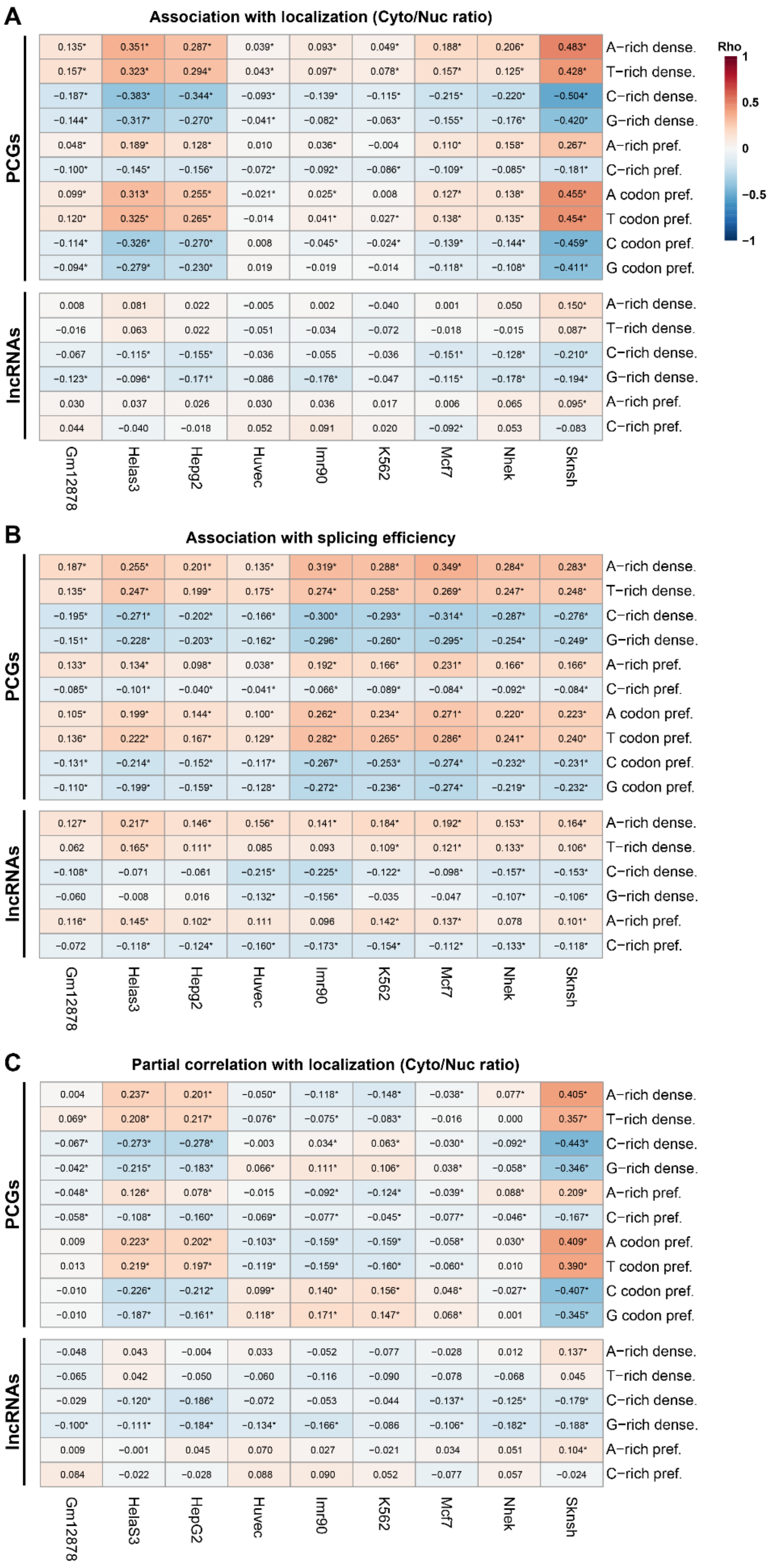
		Gm12878	Helas3	Hepg2	Huvec	Imr90	K562	Mcf7	Nhek	Sknsh
antisense	expression	0.317*	0.323*	0.326*	0.419*	0.390*	0.277*	0.306*	0.354*	0.400*
	gene splicing eff.	0.319*	0.446*	0.357*	0.463*	0.379*	0.273*	0.350*	0.501*	0.433*
	worst intron splicing eff.	0.312*	0.417*	0.362*	0.422*	0.346*	0.311*	0.362*	0.492*	0.444*
	splicing specificity	0.095	0.142*	0.229*	0.311*	0.223*	0.176*	0.211*	0.210*	0.245*
	Pol2 over introns	0.185*	0.050	0.157*	0.274*	0.142*	0.195*			
	Pol2 pausing	0.258*	0.216*	0.339*	0.163*	0.212*	0.201*			
lincRNA	expression	0.261*	0.297*	0.312*	0.328*	0.311*	0.205*	0.274*	0.337*	0.366*
	gene splicing eff.	0.219*	0.306*	0.306*	0.314*	0.401*	0.190*	0.336*	0.367*	0.282*
	worst intron splicing eff.	0.287*	0.289*	0.359*	0.331*	0.433*	0.241*	0.372*	0.427*	0.304*
	splicing specificity	0.176*	0.165*	0.239*	0.274*	0.249*	0.128*	0.211*	0.200*	0.251*
	Pol2 over introns	0.282*	0.107*	0.241*	0.110	0.139*	0.240*			
	Pol2 pausing	0.280*	0.205*	0.254*	0.275*	0.323*	0.274*			

		Gm12878	Helas3	Hepg2	Huvec	Imr90	K562	Mcf7	Nhek	Sknsh
PCGs	gene splicing eff.	0.453*	0.472*	0.540*	0.552*	0.525*	0.492*	0.553*	0.408*	0.547*
	worst intron splicing eff.	0.424*	0.437*	0.496*	0.491*	0.473*	0.459*	0.505*	0.367*	0.496*
	splicing specificity	0.249*	0.263*	0.334*	0.280*	0.296*	0.260*	0.330*	0.187*	0.355*
	Pol2 over introns	−0.062*	−0.182*	−0.140*	−0.100*	−0.103*	−0.073*			
	Pol2 pausing	0.054*	0.072*	0.126*	0.057*	0.018	0.030*			
lncRNAs	gene splicing eff.	0.259*	0.321*	0.292*	0.355*	0.317*	0.222*	0.291*	0.378*	0.308*
	worst intron splicing eff.	0.247*	0.287*	0.287*	0.311*	0.282*	0.204*	0.290*	0.360*	0.291*
	splicing specificity	0.098*	0.108*	0.153*	0.237*	0.194*	0.098*	0.183*	0.206*	0.210*
	Pol2 over introns	0.190*	0.065*	0.151*	0.131*	0.144*	0.196*			
	Pol2 pausing	0.294*	0.203*	0.294*	0.267*	0.279*	0.254*			

Supplemental Figure 5. Gene architecture and Pol2 features are associated with localization. (A) As in Fig. 4A, showing separately correlations for antisense (top) and lincRNA (bottom) gene biotypes as defined by GENCODE. **(B)** As in Fig. 4A, showing partial correlations controlling for expression levels and gene architecture (number of exons and exonic and intronic length). **(C)** Distributions of Pol2 pausing index (left) and mean Pol2 coverage over introns (right) for PCGs and lncRNAs in K562 cells. Asterisk indicates P value < 10^{-5} (Wilcoxon rank sum test).



Supplemental Figure 6. Association of promoter-proximal Pol2 pausing and subcellular localization. (**A**) Correlation between Pol2 pausing index and localization fir PCGs (left) and lncRNAs (right) in HepG2 cells. Coloring indicates local point density. Regression line is shown in bold. Fisher Z-transformation P value = 4.07e-05 (comparing PCGs and lncRNAs). (**B**) Same as A, separately for sub-groups of lncRNAs split by the conservation scores of their first intron 5' (left) or 3' (right) splice site sequences. Correlation coefficients and P-values computed using Spearman's correlation. The difference between the correlations for genes with high and low SS conservation is significant, with Fisher Z-transformation P values 0.0021 (5') and 0.022 (3').



A Association with localization (Cyto/Nuc ratio)

PCGs	0.135*	0.351*	0.287*	0.039*	0.093*	0.049*	0.188*	0.206*	0.483*	A−rich dense.
	0.157*	0.323*	0.294*	0.043*	0.097*	0.078*	0.157*	0.125*	0.428*	T−rich dense.
	−0.187*	−0.383*	−0.344*	−0.093*	−0.139*	−0.115*	−0.215*	−0.220*	−0.504*	C−rich dense.
	−0.144*	−0.317*	−0.270*	−0.041*	−0.082*	−0.063*	−0.155*	−0.176*	−0.420*	G−rich dense.
	0.048*	0.189*	0.128*	0.010	0.036*	−0.004	0.110*	0.158*	0.267*	A−rich pref.
	−0.100*	−0.145*	−0.156*	−0.072*	−0.092*	−0.086*	−0.109*	−0.085*	−0.181*	C−rich pref.
	0.099*	0.313*	0.255*	−0.021*	0.025*	0.008	0.127*	0.138*	0.455*	A codon pref.
	0.120*	0.325*	0.265*	−0.014	0.041*	0.027*	0.138*	0.135*	0.454*	T codon pref.
	−0.114*	−0.326*	−0.270*	0.008	−0.045*	−0.024*	−0.139*	−0.144*	−0.459*	C codon pref.
	−0.094*	−0.279*	−0.230*	0.019	−0.019	−0.014	−0.118*	−0.108*	−0.411*	G codon pref.
lncRNAs	0.008	0.081	0.022	−0.005	0.002	−0.040	0.001	0.050	0.150*	A−rich dense.
	−0.016	0.063	0.022	−0.051	−0.034	−0.072	−0.018	−0.015	0.087*	T−rich dense.
	−0.067	−0.115*	−0.155*	−0.036	−0.055	−0.036	−0.151*	−0.128*	−0.210*	C−rich dense.
	−0.123*	−0.096*	−0.171*	−0.086	−0.176*	−0.047	−0.115*	−0.178*	−0.194*	G−rich dense.
	0.030	0.037	0.026	0.030	0.036	0.017	0.006	0.065	0.095*	A−rich pref.
	0.044	−0.040	−0.018	0.052	0.091	0.020	−0.092*	0.053	−0.083	C−rich pref.
	Gm12878	Helas3	Hepg2	Huvec	Imr90	K562	Mcf7	Nhek	Sknsh	

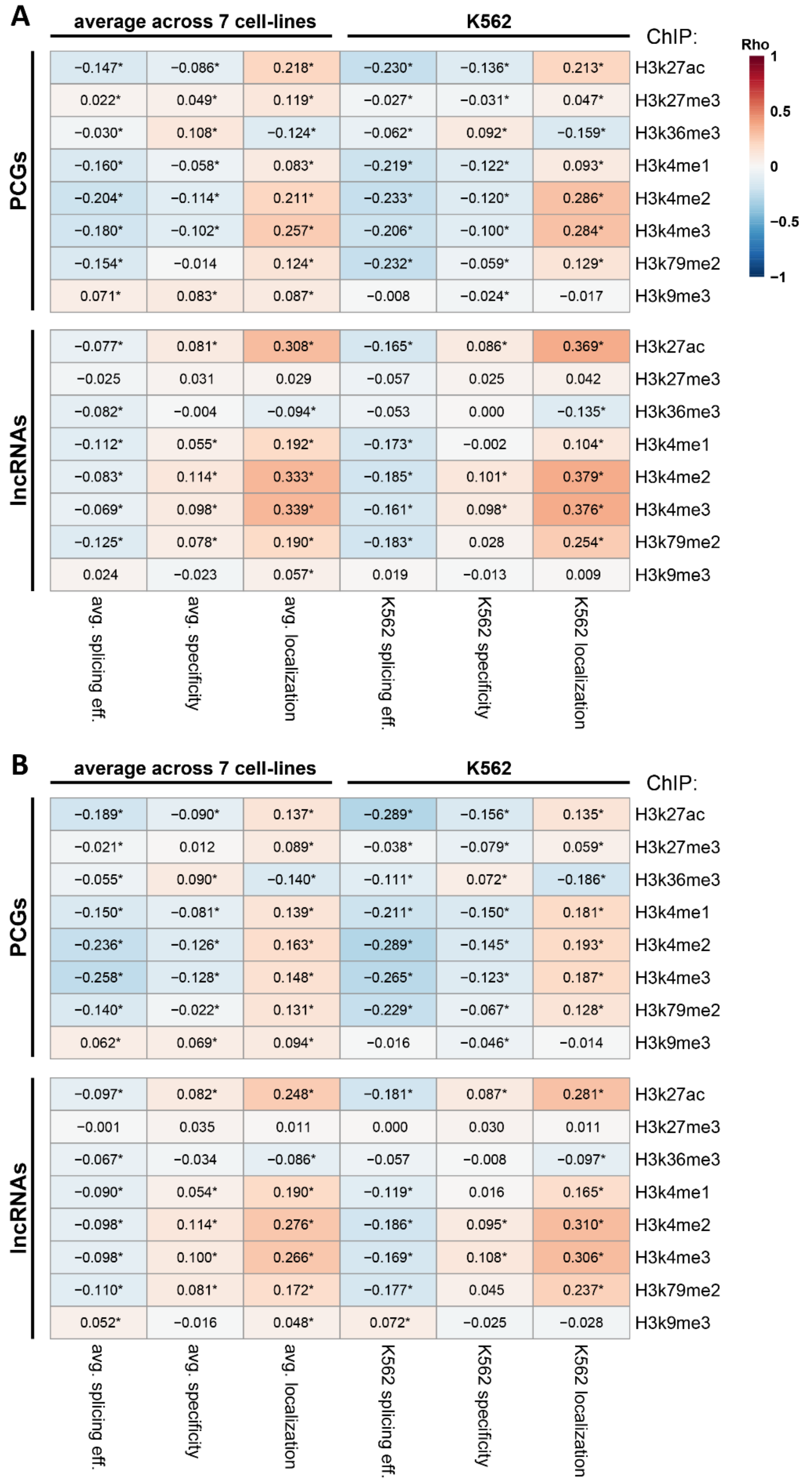
B Association with splicing efficiency

PCGs	0.187*	0.255*	0.201*	0.135*	0.319*	0.288*	0.349*	0.284*	0.283*	A−rich dense.
	0.135*	0.247*	0.199*	0.175*	0.274*	0.258*	0.269*	0.247*	0.248*	T−rich dense.
	−0.195*	−0.271*	−0.202*	−0.166*	−0.300*	−0.293*	−0.314*	−0.287*	−0.276*	C−rich dense.
	−0.151*	−0.228*	−0.203*	−0.162*	−0.296*	−0.260*	−0.295*	−0.254*	−0.249*	G−rich dense.
	0.133*	0.134*	0.098*	0.038*	0.192*	0.166*	0.231*	0.166*	0.166*	A−rich pref.
	−0.085*	−0.101*	−0.040*	−0.041*	−0.066*	−0.089*	−0.084*	−0.092*	−0.084*	C−rich pref.
	0.105*	0.199*	0.144*	0.100*	0.262*	0.234*	0.271*	0.220*	0.223*	A codon pref.
	0.136*	0.222*	0.167*	0.129*	0.282*	0.265*	0.286*	0.241*	0.240*	T codon pref.
	−0.131*	−0.214*	−0.152*	−0.117*	−0.267*	−0.253*	−0.274*	−0.232*	−0.231*	C codon pref.
	−0.110*	−0.199*	−0.159*	−0.128*	−0.272*	−0.236*	−0.274*	−0.219*	−0.232*	G codon pref.
lncRNAs	0.127*	0.217*	0.146*	0.156*	0.141*	0.184*	0.192*	0.153*	0.164*	A−rich dense.
	0.062	0.165*	0.111*	0.085	0.093	0.109*	0.121*	0.133*	0.106*	T−rich dense.
	−0.108*	−0.071	−0.061	−0.215*	−0.225*	−0.122*	−0.098*	−0.157*	−0.153*	C−rich dense.
	−0.060	−0.008	0.016	−0.132*	−0.156*	−0.035	−0.047	−0.107*	−0.106*	G−rich dense.
	0.116*	0.145*	0.102*	0.111	0.096	0.142*	0.137*	0.078	0.101*	A−rich pref.
	−0.072	−0.118*	−0.124*	−0.160*	−0.173*	−0.154*	−0.112*	−0.133*	−0.118*	C−rich pref.
	Gm12878	Helas3	Hepg2	Huvec	Imr90	K562	Mcf7	Nhek	Sknsh	

C Partial correlation with localization (Cyto/Nuc ratio)

PCGs	0.004	0.237*	0.201*	−0.050*	−0.118*	−0.148*	−0.038*	0.077*	0.405*	A−rich dense.
	0.069*	0.208*	0.217*	−0.076*	−0.075*	−0.083*	−0.016	0.000	0.357*	T−rich dense.
	−0.067*	−0.273*	−0.278*	−0.003	0.034*	0.063*	−0.030*	−0.092*	−0.443*	C−rich dense.
	−0.042*	−0.215*	−0.183*	0.066*	0.111*	0.106*	0.038*	−0.058*	−0.346*	G−rich dense.
	−0.048*	0.126*	0.078*	−0.015	−0.092*	−0.124*	−0.039*	0.088*	0.209*	A−rich pref.
	−0.058*	−0.108*	−0.160*	−0.069*	−0.077*	−0.045*	−0.077*	−0.046*	−0.167*	C−rich pref.
	0.009	0.223*	0.202*	−0.103*	−0.159*	−0.159*	−0.058*	0.030*	0.409*	A codon pref.
	0.013	0.219*	0.197*	−0.119*	−0.159*	−0.160*	−0.060*	0.010	0.390*	T codon pref.
	−0.010	−0.226*	−0.212*	0.099*	0.140*	0.156*	0.048*	−0.027*	−0.407*	C codon pref.
	−0.010	−0.187*	−0.161*	0.118*	0.171*	0.147*	0.068*	0.001	−0.345*	G codon pref.
lncRNAs	−0.048	0.043	−0.004	0.033	−0.052	−0.077	−0.028	0.012	0.137*	A−rich dense.
	−0.065	0.042	−0.050	−0.060	−0.116	−0.090	−0.078	−0.068	0.045	T−rich dense.
	−0.029	−0.120*	−0.186*	−0.072	−0.053	−0.044	−0.137*	−0.125*	−0.179*	C−rich dense.
	−0.100*	−0.111*	−0.184*	−0.134*	−0.166*	−0.086	−0.106*	−0.182*	−0.188*	G−rich dense.
	0.009	−0.001	0.045	0.070	0.027	−0.021	0.034	0.051	0.104*	A−rich pref.
	0.084	−0.022	−0.028	0.088	0.090	0.052	−0.077	0.057	−0.024	C−rich pref.
	Gm12878	HelaS3	HepG2	Huvec	Imr90	K562	Mcf7	Nhek	Sknsh	

Supplemental Figure 7. Correlation of sequence elements with splicing and localization. (A-C) Correlation between the indicated measures of nucleotide preference and Cyto/Nuc ratios (**A**), splicing efficiency (**B**), and Cyto/Nuc ratios when controlling for splicing efficiency (**C**). Correlation coefficients and FDR-adjusted P-values computed using Spearman's correlation.



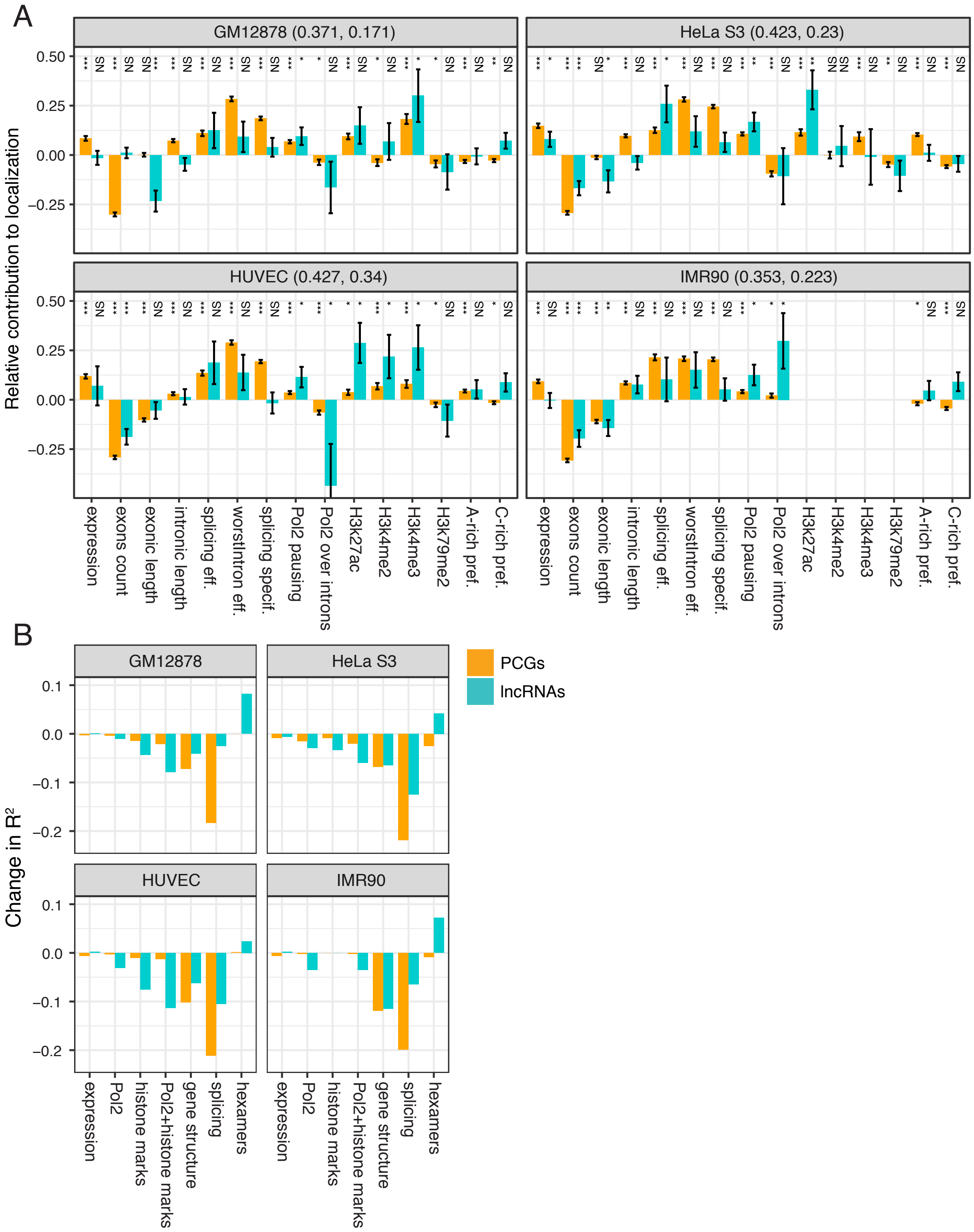
A

		average across 7 cell-lines			K562		
	ChIP:	avg. splicing eff.	avg. specificity	avg. localization	K562 splicing eff.	K562 specificity	K562 localization
PCGs	H3k27ac	−0.147*	−0.086*	0.218*	−0.230*	−0.136*	0.213*
	H3k27me3	0.022*	0.049*	0.119*	−0.027*	−0.031*	0.047*
	H3k36me3	−0.030*	0.108*	−0.124*	−0.062*	0.092*	−0.159*
	H3k4me1	−0.160*	−0.058*	0.083*	−0.219*	−0.122*	0.093*
	H3k4me2	−0.204*	−0.114*	0.211*	−0.233*	−0.120*	0.286*
	H3k4me3	−0.180*	−0.102*	0.257*	−0.206*	−0.100*	0.284*
	H3k79me2	−0.154*	−0.014	0.124*	−0.232*	−0.059*	0.129*
	H3k9me3	0.071*	0.083*	0.087*	−0.008	−0.024*	−0.017
lncRNAs	H3k27ac	−0.077*	0.081*	0.308*	−0.165*	0.086*	0.369*
	H3k27me3	−0.025	0.031	0.029	−0.057	0.025	0.042
	H3k36me3	−0.082*	−0.004	−0.094*	−0.053	0.000	−0.135*
	H3k4me1	−0.112*	0.055*	0.192*	−0.173*	−0.002	0.104*
	H3k4me2	−0.083*	0.114*	0.333*	−0.185*	0.101*	0.379*
	H3k4me3	−0.069*	0.098*	0.339*	−0.161*	0.098*	0.376*
	H3k79me2	−0.125*	0.078*	0.190*	−0.183*	0.028	0.254*
	H3k9me3	0.024	−0.023	0.057*	0.019	−0.013	0.009

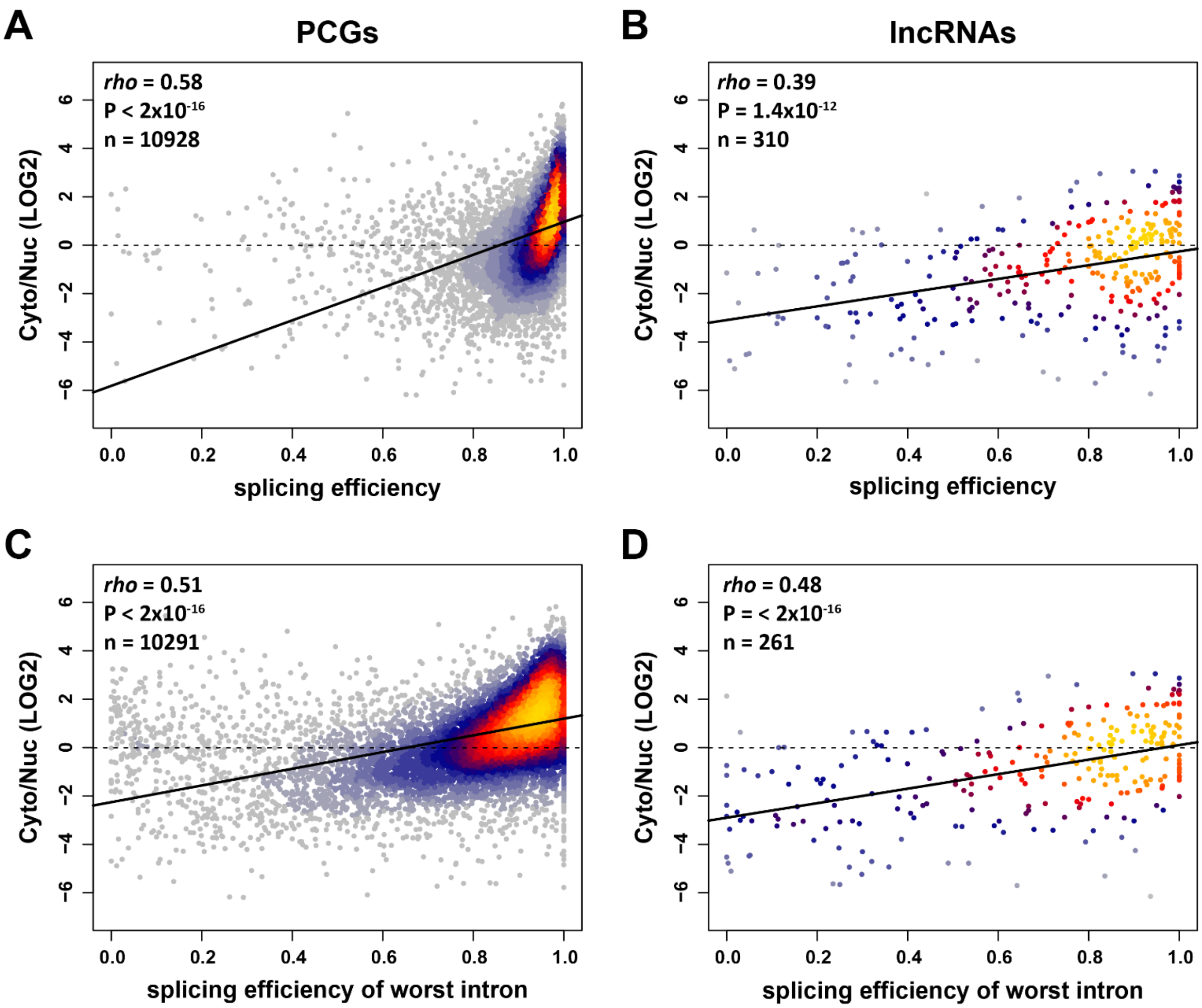
B

		average across 7 cell-lines			K562		
	ChIP:	avg. splicing eff.	avg. specificity	avg. localization	K562 splicing eff.	K562 specificity	K562 localization
PCGs	H3k27ac	−0.189*	−0.090*	0.137*	−0.289*	−0.156*	0.135*
	H3k27me3	−0.021*	0.012	0.089*	−0.038*	−0.079*	0.059*
	H3k36me3	−0.055*	0.090*	−0.140*	−0.111*	0.072*	−0.186*
	H3k4me1	−0.150*	−0.081*	0.139*	−0.211*	−0.150*	0.181*
	H3k4me2	−0.236*	−0.126*	0.163*	−0.289*	−0.145*	0.193*
	H3k4me3	−0.258*	−0.128*	0.148*	−0.265*	−0.123*	0.187*
	H3k79me2	−0.140*	−0.022*	0.131*	−0.229*	−0.067*	0.128*
	H3k9me3	0.062*	0.069*	0.094*	−0.016	−0.046*	−0.014
lncRNAs	H3k27ac	−0.097*	0.082*	0.248*	−0.181*	0.087*	0.281*
	H3k27me3	−0.001	0.035	0.011	0.000	0.030	0.011
	H3k36me3	−0.067*	−0.034	−0.086*	−0.057	−0.008	−0.097*
	H3k4me1	−0.090*	0.054*	0.190*	−0.119*	0.016	0.165*
	H3k4me2	−0.098*	0.114*	0.276*	−0.186*	0.095*	0.310*
	H3k4me3	−0.098*	0.100*	0.266*	−0.169*	0.108*	0.306*
	H3k79me2	−0.110*	0.081*	0.172*	−0.177*	0.045	0.237*
	H3k9me3	0.052*	−0.016	0.048*	0.072*	−0.025	−0.028

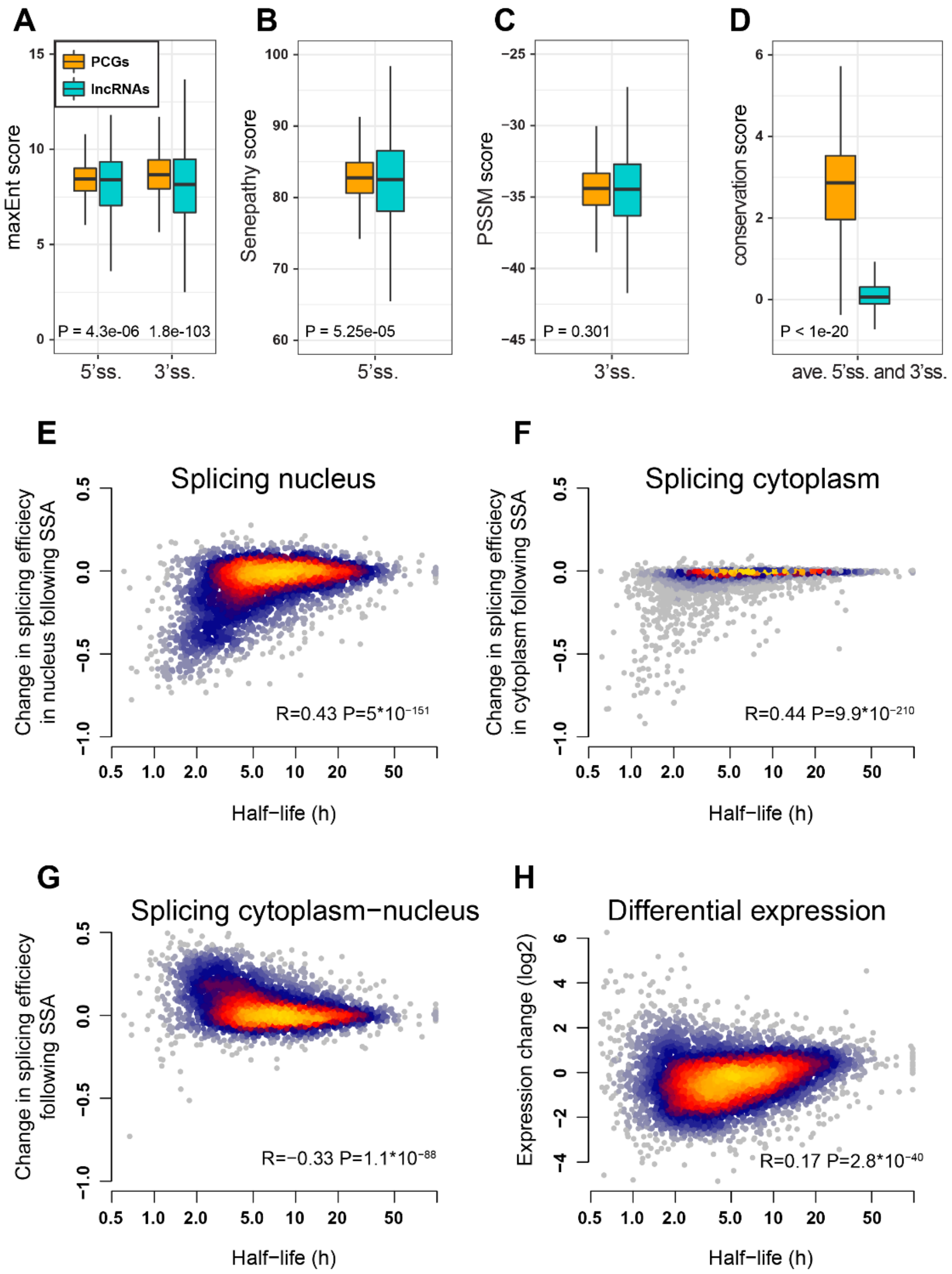
Supplemental Figure 8. Correlation of chromatin marks with splicing and localization. (A) Correlation of splicing and localization with coverage of chromatin marks over splice junctions. Correlation coefficients and FDR-adjusted P-values computed using Spearman's correlation. **(B)** Same as **A**, but for average coverage over the intron.



Supplemental Figure 9. Predictive models for localization in various cell lines. Same as Fig. 6A-B, but for additional cell lines.



Supplemental Figure 10. (**A-B**) Correlation between splicing efficiency, averaged across all introns, and localization of PCGs (A) and lncRNAs (B) in mouse liver. (**C-D**) Correlation between the splicing efficiency of the least efficient intron and localization of PCGs (C) and lncRNAs (D) in mouse liver. Coloring indicates local point density. Regression line is shown in bold. Correlation coefficients and P-values computed using Spearman's correlation.



Supplemental Figure 11. (**A-D**) Difference in splice site strength (A-C) and conservation (D) between PCGs and lncRNAs. P-values computed using Wilcoxon rank-sum test. (**E-H**) Correlation between half lives (from Schueler et al. 2014) and changes in splicing (E-G) or expression levels (H) following inhibition of splicing for six hours using spliceostatin A (from Yoshimoto et al. 2017) in HeLa S3 cells. In E, the *difference* between the splicing change in the cytoplasm and the splicing change in the nucleus is shown on the Y axis. Correlation coefficients and P-values computed using Spearman's correlation.