SUPPLEMENTARY INFORMATION

Splice Site Strength-Dependent Activity and Genetic Buffering by Poly-G Runs

Xinshu Xiao[1,2], Zefeng Wang[1,3], Minyoung Jang[1], Razvan Nutiu[1], Eric T. Wang[1,4] and Christopher B. Burge[1,5]

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge MA 02139

[2]Current address: Department of Physiological Science and the Molecular Biology Institute, University of California, Los Angeles, CA, 90095

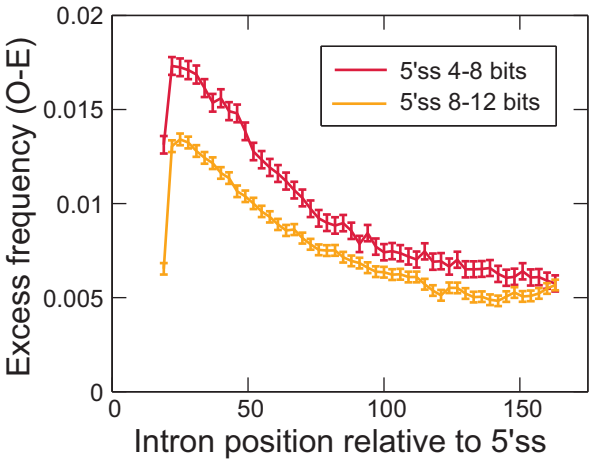
[3]Current address: Department of Pharmacology, University of North Carolina at Chapel Hill, NC 27599

[4]Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge MA 02139

[5]Correspondence should be addressed to: cburge@mit.edu. Phone: (617) 258-5997. Fax: (617) 452-2936.

Running head: Splicing Activity of G-Runs

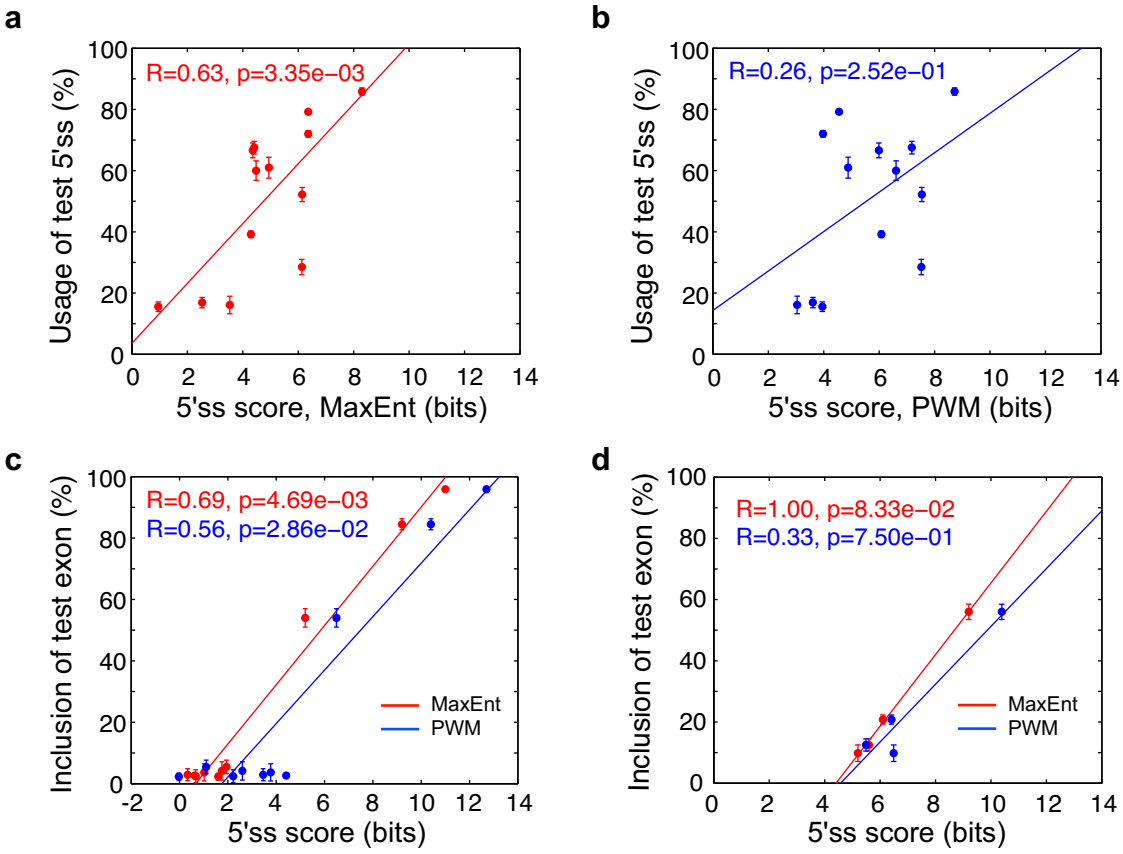
a



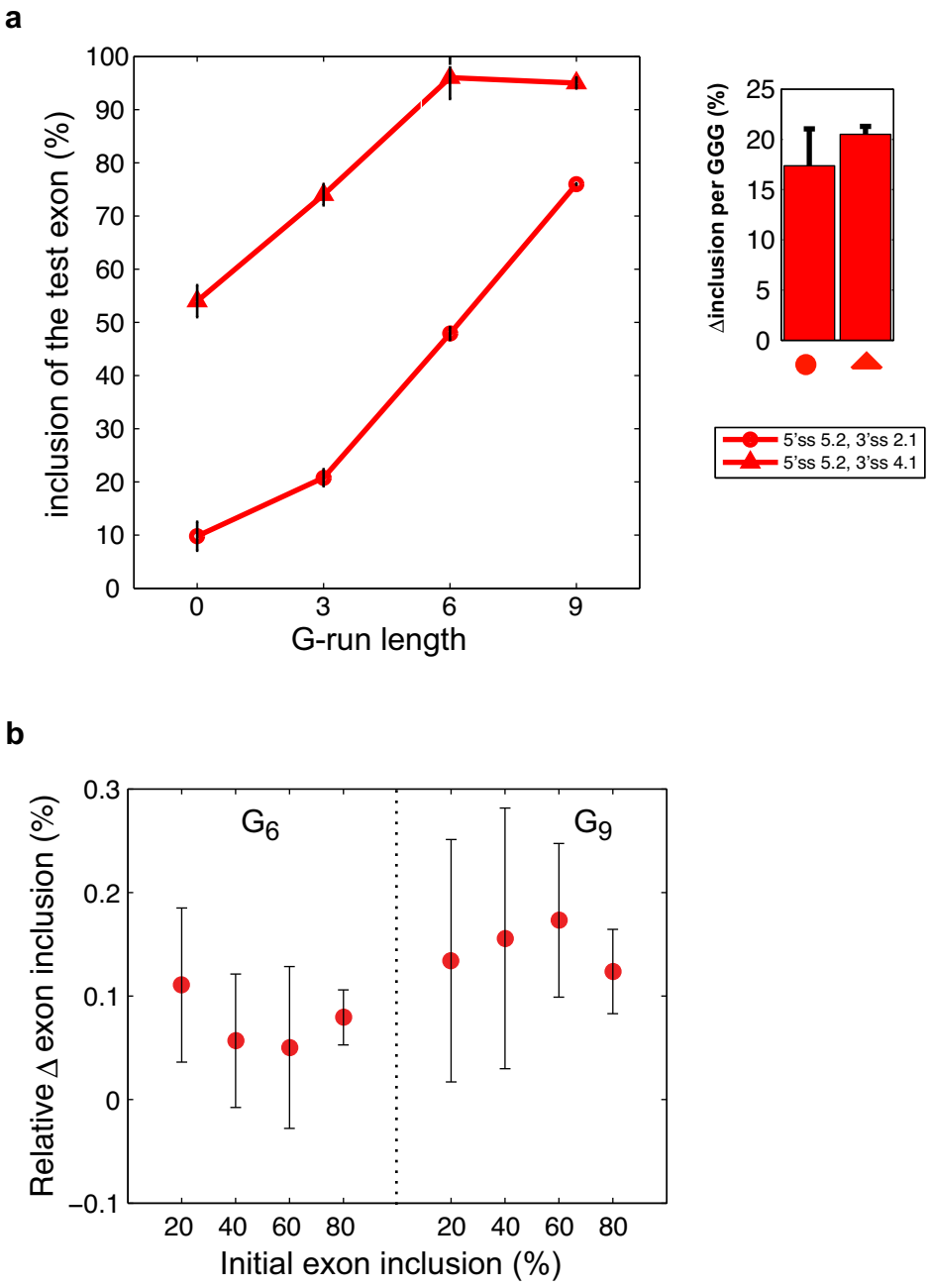
b

no. G's in G-run	[3,5]	[6,8]	[9,11]	[12,14]	[15,17]	[18,20]	≥21
No. intron	20,869	11,061	6,765	4,534	3,025	1,649	1,583
% total	22.9	12.2	7.4	5.0	3.3	1.8	1.7

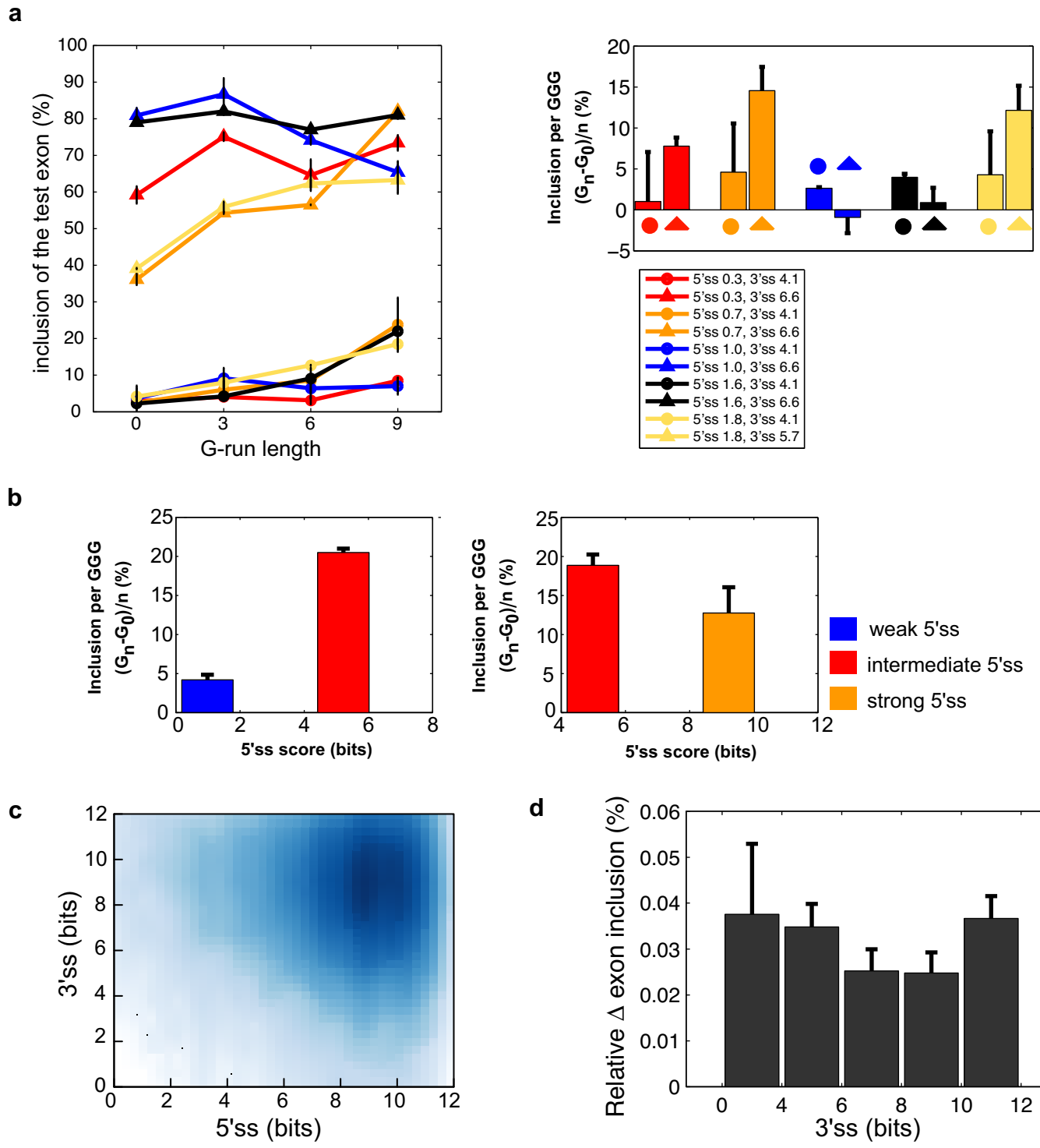
Supplementary Figure 1. Distribution of G-rich elements in human introns. (**a**) Excess frequency of GGG in downstream introns of human constitutive exons with intermediate or strong 5'ss. Excess frequency is defined as the difference between the observed total frequency of GGG among all introns, calculated in a 30-nt window, and the mean frequency of GGG in 10 random permutations of the sequence in the same window, with an offset of 3nt between successive windows (see Yeo et al, PNAS, 2004 for details). Each point represents the center of a 30-nt window. Black bars show the standard errors. (**b**) Number and percentage of human constitutive exons (N = 91,045) with different numbers of G's in runs of G_3 or longer.



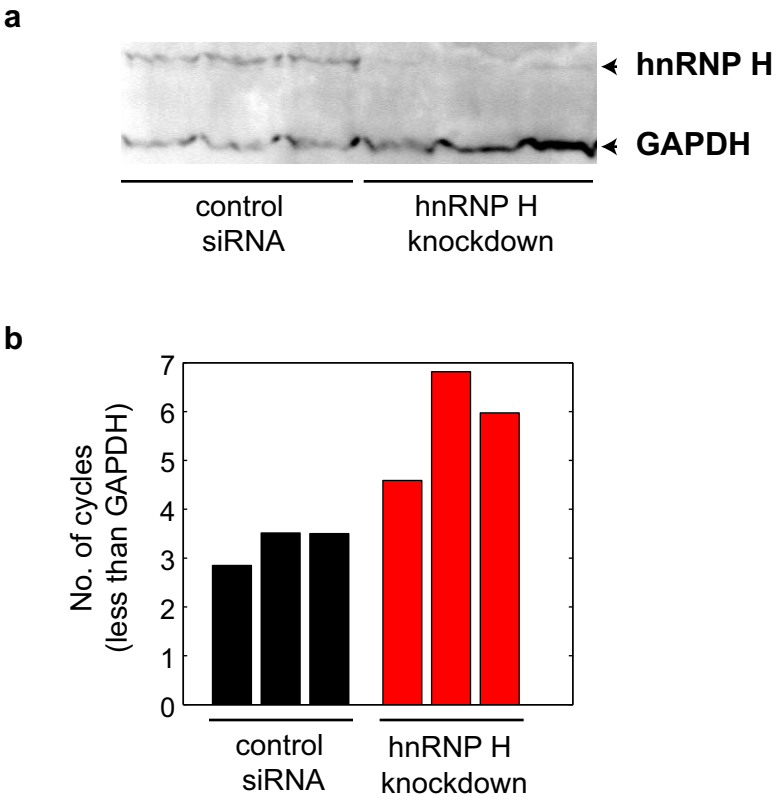
Supplementary Figure 2. Correlation (Kendall's Tau) of splicing reporter assay results with 5'ss scores calculated using MaxEnt and positional weight matrix (PWM). (**a**). Splicing reporter assays presented in Table 2 of Roca et al, RNA, 11(5): 683-98, 2005; 5'ss was scored by MaxEnt. (**b**). Same data as in (**a**), but 5'ss was scored by PWM. (**c**). Reporters presented in this paper that differ only in 5'ss sequence and with a 3'ss score of 4.1 bits. (**d**). Reporters presented in this paper that differ only in 5'ss sequence and with a 3'ss score of 2.1 bits.



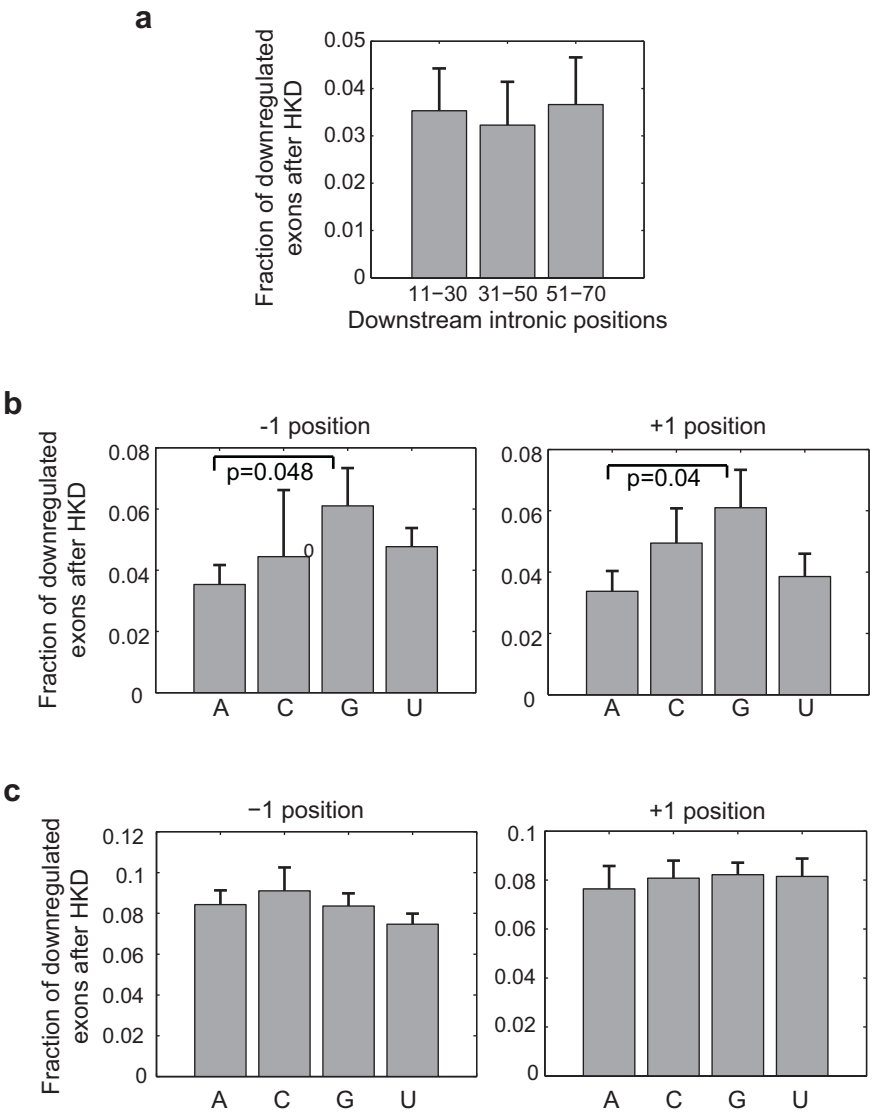
Supplementary Figure 3. Activity of GGGs independent of initial inclusion level at G_0.
(**a**) Left: Inclusion of the middle exon in reporters with the same medium 5'ss and different 3'ss. Right: change in inclusion of the middle exon normalized by the number of GGGs (data at G_9 for the construct whose 3'ss is 4.1 bits was excluded due to saturation at G_6). (**b**) Change in exon inclusion level in the mRNA-SEQ data (%, control-HKD) in exons with 6 or 9 G's in G-runs in the downstream intron relative to exons with no G-runs. Exons with exonic or upstream intronic G-runs have been excluded. Exons with or without G-runs are grouped correspondingly according to their inclusion level in the HKD experiments. Ranges of inclusion levels of the four groups are 10 to 30%, 30 to 50%, 50 to 70% and 70 to 90%. Exons with smaller than 10% or larger than 90% inclusion in the HKD experiment are excluded to allow a dynamic range for potential changes in inclusion.



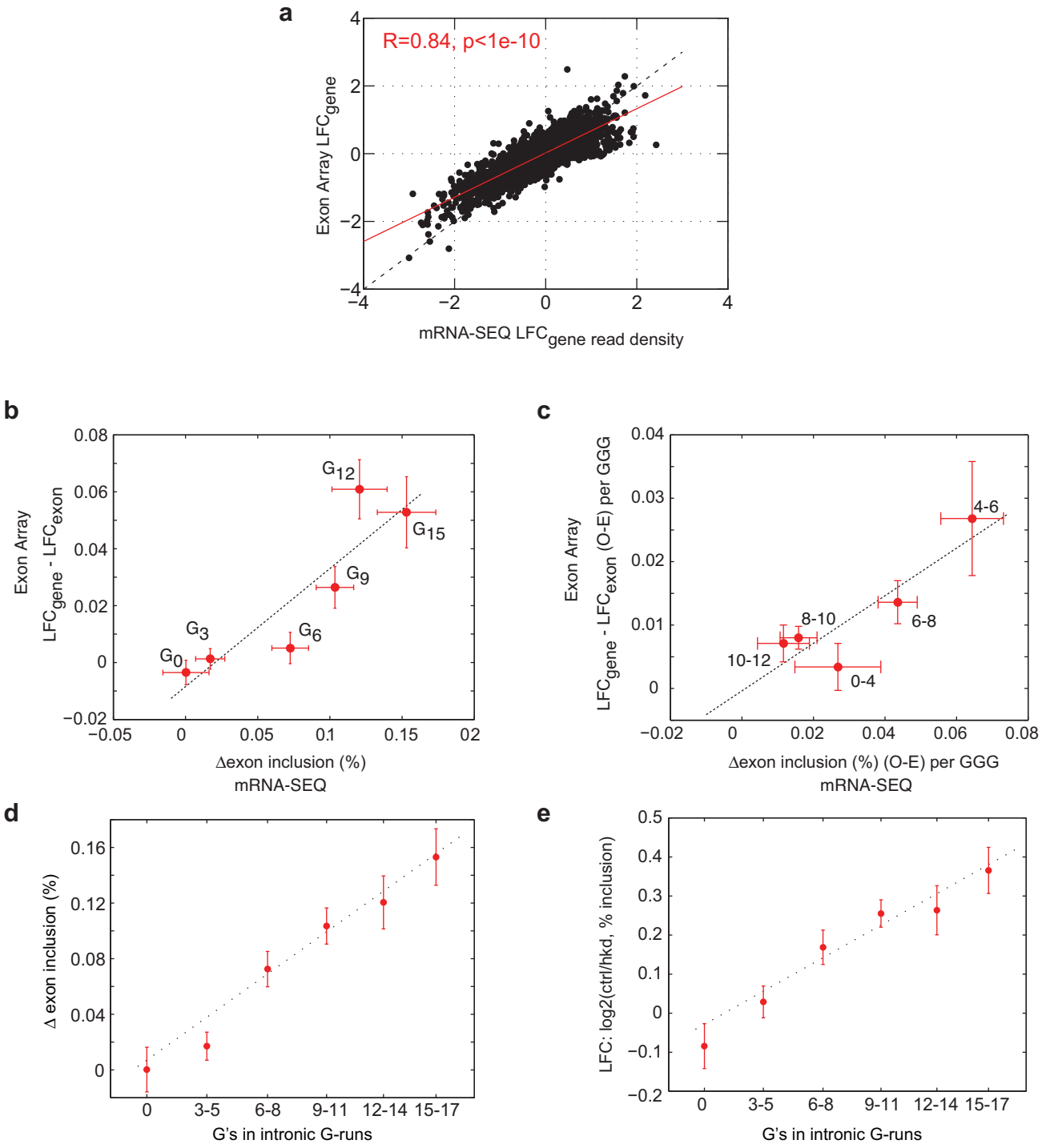
Supplementary Figure 4. Absence of consistent pattern of dependence of GGG activity on 3'ss strength. (**a**) Left: Inclusion of the middle exon in reporters with different 5'ss and 3'ss. Pairs of reporters with same 5'ss but different 3'ss are shown in the same color. Right: change in inclusion of the middle exon normalized by the number of GGGs. All 5'ss strength are in the very weak range. (**b**) Activity of GGGs dependent on 5'ss strength with fixed 3'ss. Change in inclusion of the middle exon normalized by the number of GGGs in the intron in reporters with the same 3'ss (left: 4.1 bits, right: 2.1 bits) but different 5'ss strength. (**c**) Smoothed scatter plot of the 5'ss and 3'ss scores of human constitutive exons (R=-0.029). (**d**) Change in exon inclusion level in the mRNA-SEQ data (%, control-HKD) in exons with at least 9 G's in G-runs in the downstream intron relative to exons with no G-runs, normalized by the number of G3. Exons with exonic or upstream intronic G-runs have been excluded. Exons are grouped according to the 3'ss scores. No significant difference was detected across different groups.



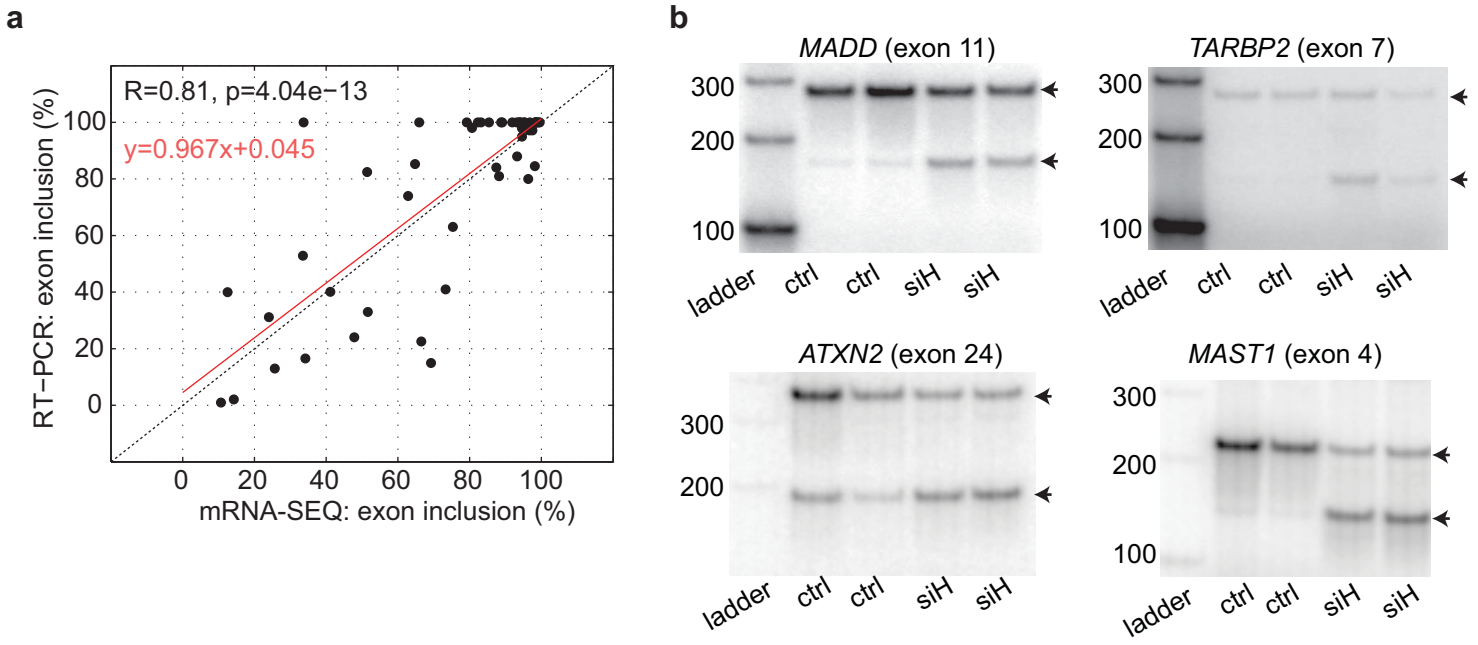
Supplementary Figure 5. hnRNP H protein and mRNA levels following control siRNA and hnRNP H siRNA transfections. (**a**) Western blot using hnRNP H specific antibody and GAPDH antibody. (**b**) qPCR of hnRNP H level.



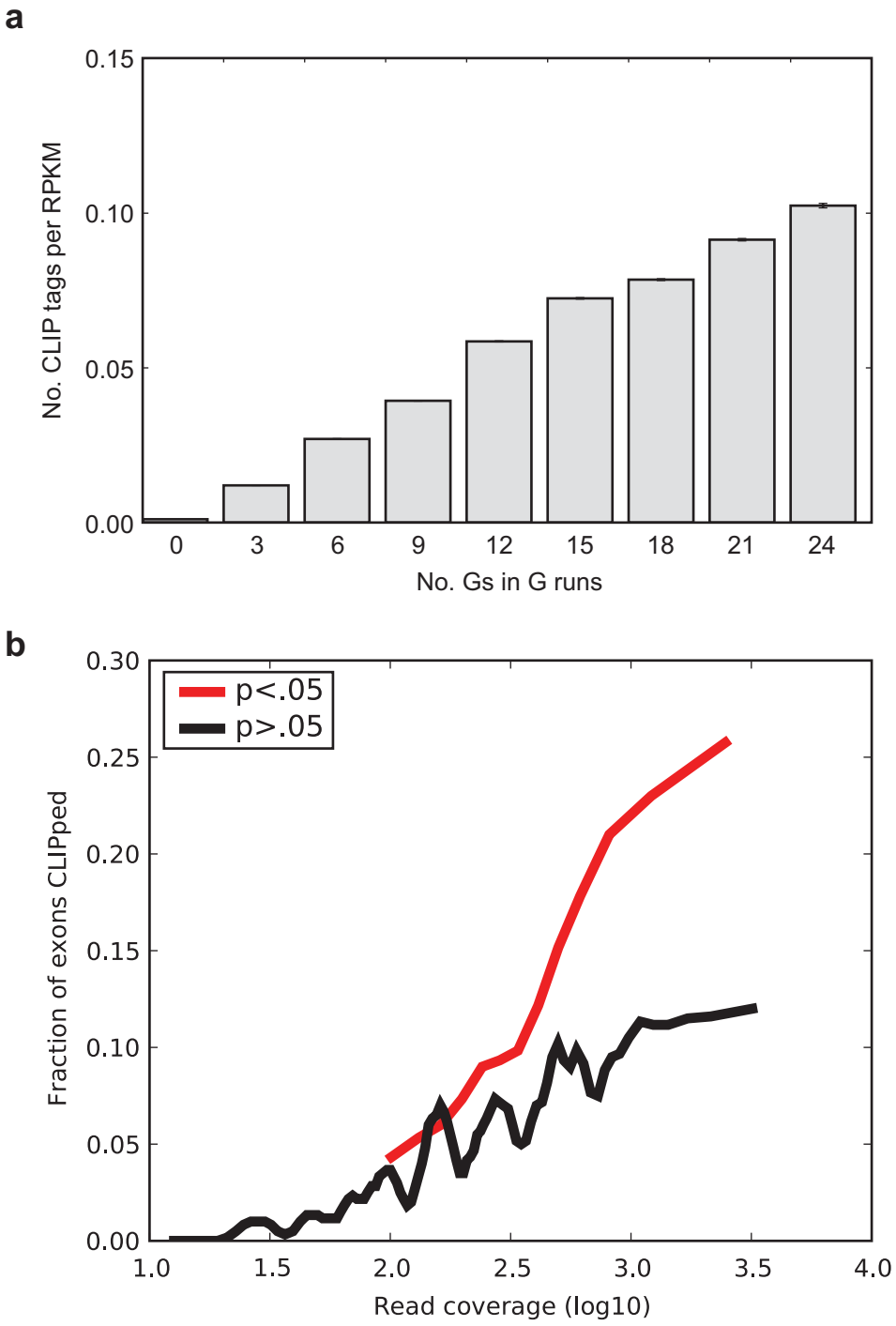
Supplementary Figure 6. Function of G-rich motifs in different intronic positions or sequence contexts evaluated using the mRNA-SEQ data. (**a**) Fraction of downregulated exons (mean and SEM) after HKD for exons with only one GGG motif in the downstream intron (11-70 nt). Exons were grouped according to the position of the intronic GGG. Data from mRNA-SEQ and exon arrays are combined together to have the largest data set. Downregulation is defined as a change of exon inclusion level of at least 20% in the mRNA-SEQ data and a change of relative exon expression (log2 based) of at least 0.5 in the exon array data. (**b**) Fraction of downregulated exons (mean and SEM) after HKD for exons with only one GGG in the downstream intronic region (11-70 nt). Exons are grouped based on the sequence at the neighboring positions of the GGG. Unlabeled comparisons have p values greater than 0.05 (Fisher's Exact test). (**c**) Same as in (**b**) for exons with one GGGGU (and no other G-runs) in the downstream intronic region (11-70 nt).



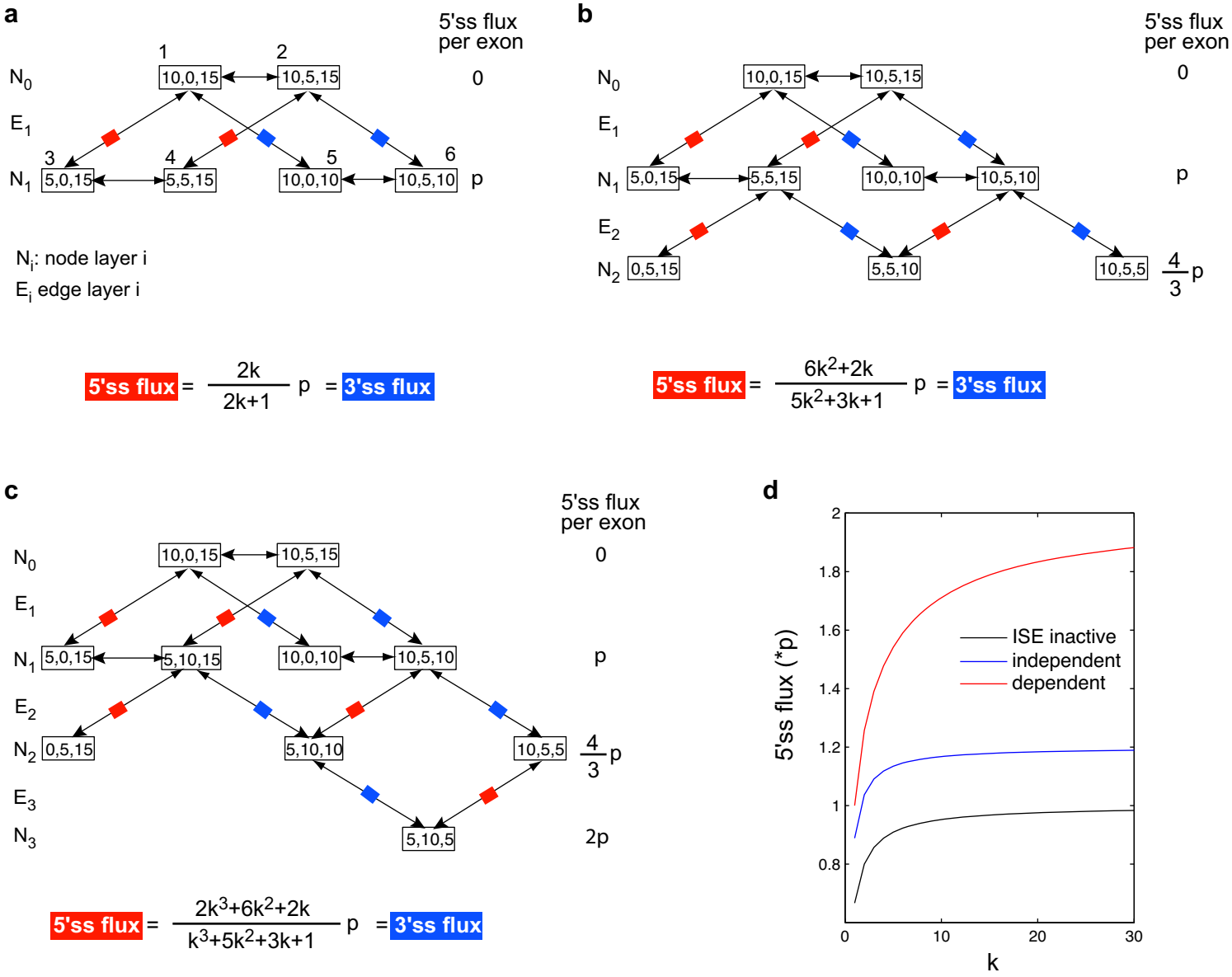
Supplementary Figure 7. Affymetrix exon array and mRNA-SEQ data after hnRNP H knockdown (HKD). (**a**) Log-fold-change (LFC, defined as Log2(HKD/Ctrl)) of gene levels calculated based on the array and mRNA-SEQ data. Pearson correlation coefficient and the correspoding p values are shown. (**b**) Relative LFC values (gene-exon) calculated from exon arrays and change in exon inclusion level (ctrl-HKD) in the mRNA-SEQ data for exons with different numbers of G's in G-runs in the downstream intron as labeled. Each point and error bars represent mean and SEM of all exons in the group respectively. (**c**) Normalized (per GGG) relative LFC values from exon arrays and normalized relative change in exon inclusion level in the mRNA-SEQ data for exons with differnet 5'ss strength. Ranges of 5'ss scores (bits) are labeled next to each point. (**d**) Change in the exon inclusion level (mean and SEM) in mRNA-SEQ defined as the difference between the values in the control and HKD experiments (mean values: R=0.99, all data: R=0.23). (**e**) Change in the exon inclusion level in mRNA-SEQ defined as log-fold-change (LFC), which is calculated as log2 exon inclusion (control/HKD) (mean values: R=0.98, all data: R=0.18).



Supplementary Figure 8. Comparison of exon inclusion levles estimated by RT-PCR, mRNA-SEQ and Exon Array. (**a**) RT-PCR vs. mRNA-SEQ for twenty-six exons each with two values from the H-knockdown and control experiments respectively. (**b**) Representative gel images of RT-PCR.

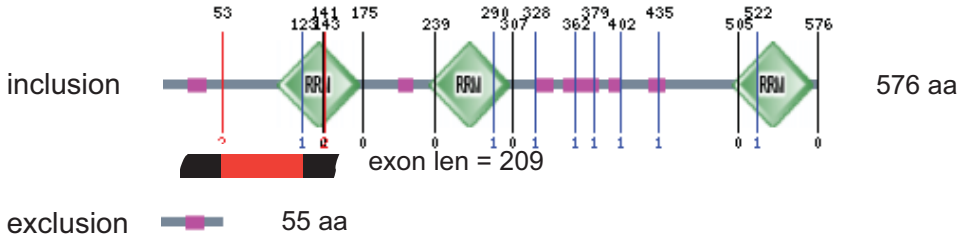


Supplementary Figure 9. CLIP-SEQ data analysis. (**a**) CLIP tag density as a function of number of G occurrences in G runs**.** CLIP tags in the region +10 to +250 downstream of all internal Ensembl exons were enumerated, along with the number of G occurrences found in runs of G3 or greater in each region. The density of CLIP tags, normalized by gene expression as inferred from RPKM values from mRNA-Seq data, is plotted as a function of the number of G occurrences. (**b**) Enrichment of CLIP tags in and around exons differentially regulated upon knockdown of hnRNP H. The number of CLIP tags in the region -250 to +250 (including the exon body) was counted for every internal Ensembl exon. Using mRNA-Seq data from control and hnRNP H knockdown cells, each exon was also evaluated for differential regulation by a Fisher's Exact test. The fraction of exons with CLIP tags is plotted as a function of read counts used in the Fisher's Exact test, for the group of exons differentially regulated ($p<0.05$) and all other exons ($p>0.05$).

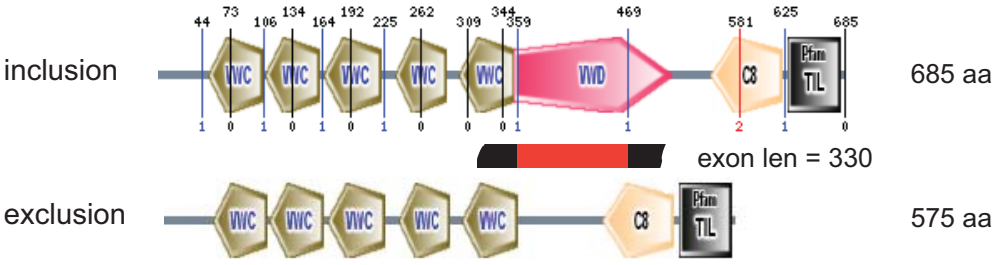


Supplementary Figure 10. Evolutionary models of splicing cis-elements. (**a**) ISE inactive. In the network, each box corresponds to a state of the exon with the 3 numbers in the box representing scores of 5'ss, ISE, 3'ss+others (e.g., ESE) respectively. Possible scores of 5'ss and 3'ss+others are assumed to be 0, 5 and 10. Possible scores of ISE are assumed to be 0 and 5. However, since the ISEs are inactive in this model, ISE scores are not counted in the total score of the exon. The summation of scores of all elements is required to be at least 20 for the exon to be spliced. The mutation rates are assumed to be p for mutations in the directions of small arrowheads, and k*p (k>1) in the directions of big arrowheads. Red bars overlap with network edges corresponding to changes in 5'ss (i.e., 5'ss flux); blue bars overlap with 3'ss flux. Numbers on the right of the network represent 5'ss flux involving the nodes in the corresponding layers normalized by the number of exons in the layer. The total flux for the 5'ss (same as 3'ss flux) is shown below the network. (**b**) ISE activity independent of 5'ss. Sames as in (A) except that the ISEs are active and the ISE scores are counted in the total score of the exon. (**c**) ISE activity dependent on 5'ss strength. Same as in (**a**) except that the scores of ISEs are dependent on the score of 5'ss. Specifically, the ISEs are scored 5 in case of a 5'ss score of 0 or 10, but 10 in case of a 5'ss score of 5. (**d**) 5'ss flux of the three models relative to different values of k.

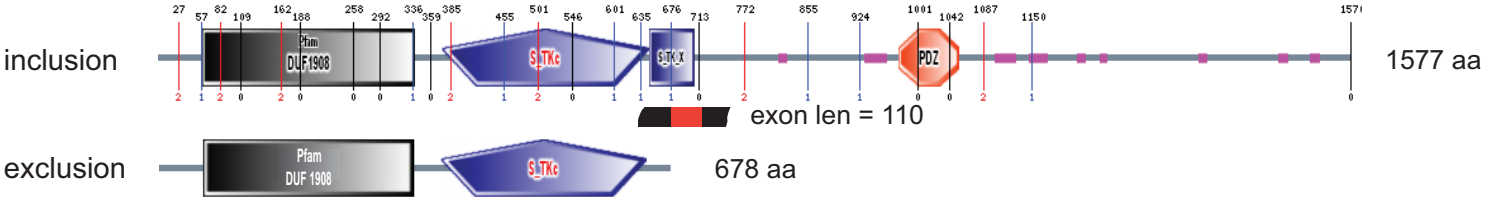
MYEF2 (Myelin expression factor 2), RRM domain



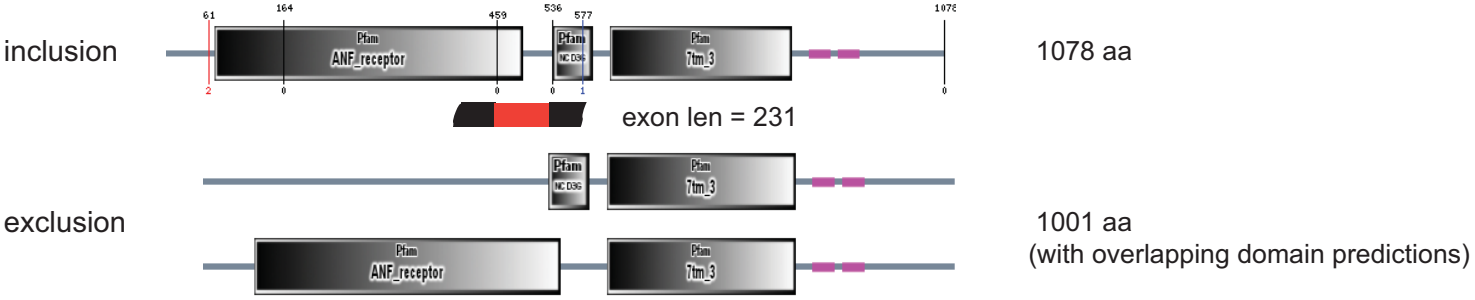
BMPER (BMP-binding endothelial regulator protein precursor), VWD domain



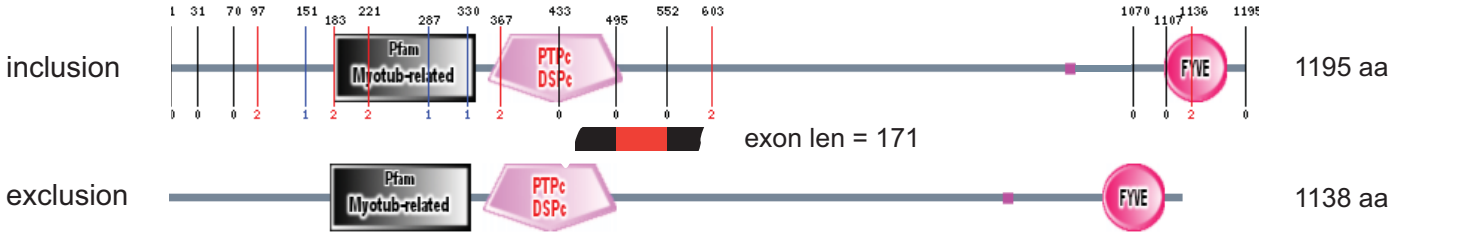
MAST1 (Microtubule-associated serine/threonine-protein kinase 1), S_TK_x domain



CASR: Extracellular calcium-sensing receptor precursor, Pfam ANF_receptor domain

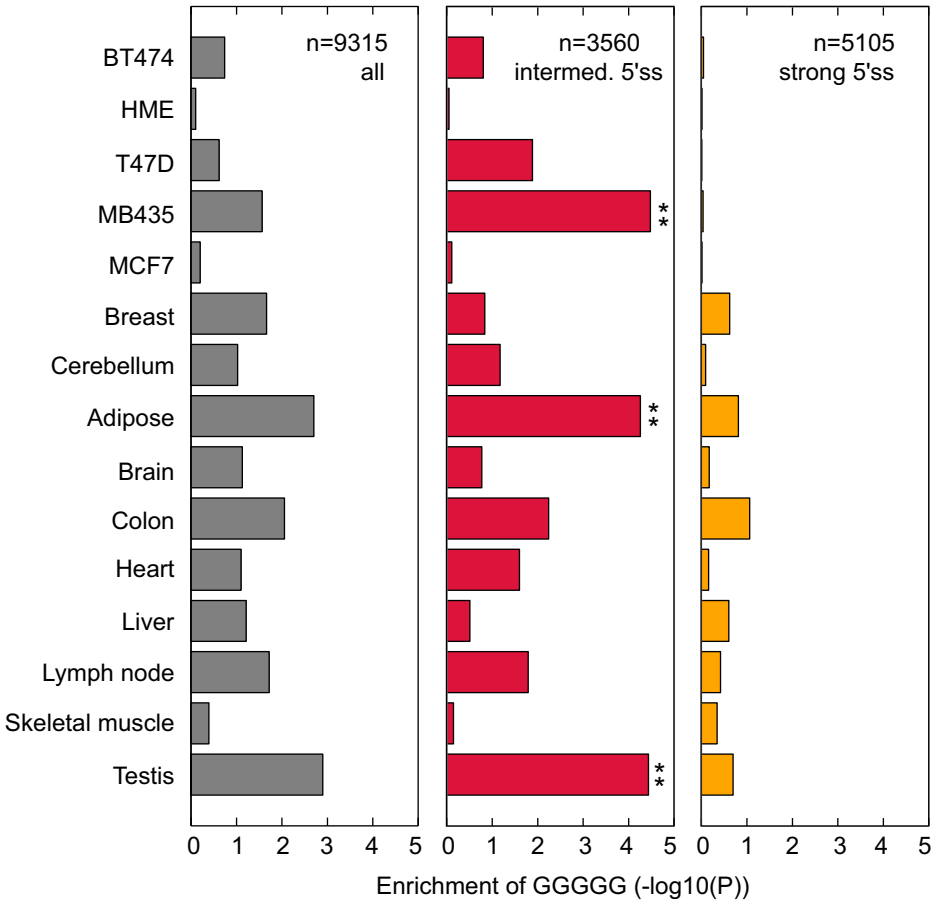


MTMR4 (myotubularin related protein 4), PTPc domain

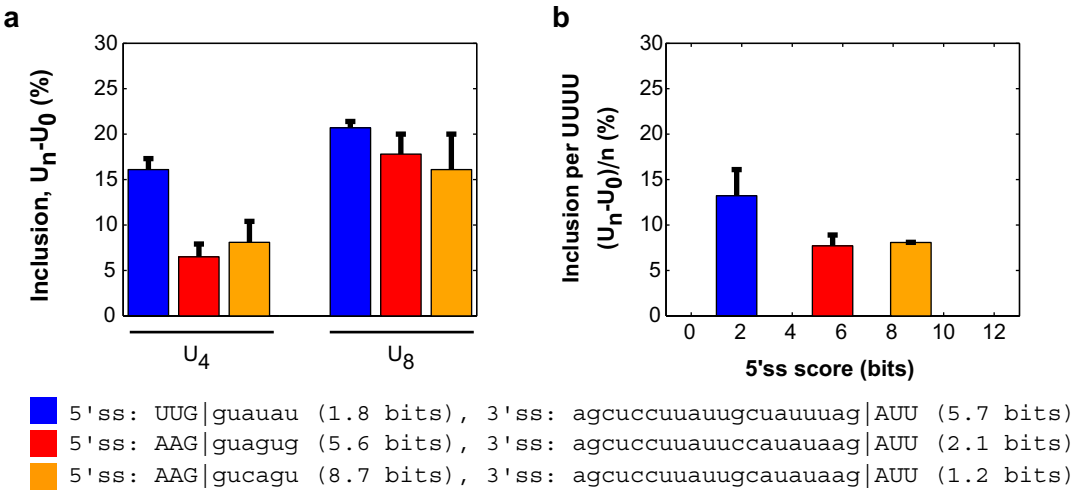


segments of low compositional complexity

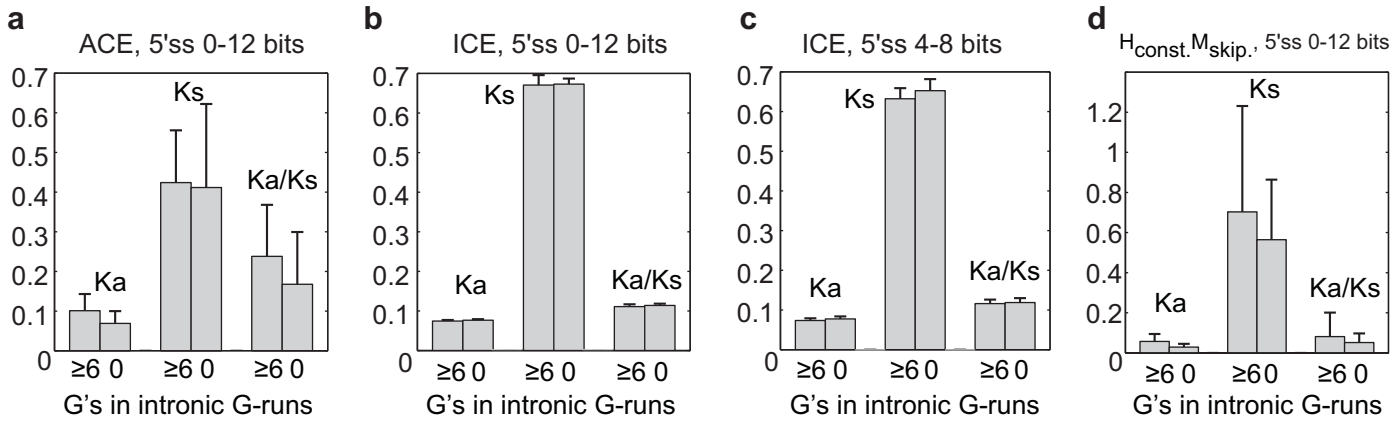
Supplementary Figure 11. Examples of protein domains overlapping human constitutive exons (red boxes) that are alternatively spliced in mouse and with G-runs (≥6 G's) in the downstream intron. Protein domain annotations were obtained from SMART and Pfam. Hypothetical exon exclusion transcripts were generated by removing the red exon from the exon inclusion isoform.



Supplementary Figure 12. Enrichment of GGGGG in the downstream introns (11-70 nt from 5'ss) of skipped exons with tissue-biased inclusion (Wang et al 2008). P values were calculated for the enrichment of GGGGG counts relative to cohorts of control pentamers in the same region. Left panel: all exons with tissue-biased inclusion; middle: exons with intermediate 5'ss (4-8 bits); right: exons with strong 5'ss (8-12 bits). P value bars are labeled * if $p < 0.01$ and ** if $p < 0.001$ after Bonferroni correction.



Supplementary Figure 13. U-runs as ISEs in the downstream intron have high activity for weak 5'ss. (**a**). Inclusion (%) of the test exon with U_4 or U_8 in the downstream intron, relative to control insert (U_0). (**b**). Change in inclusion (%) of the test exon normalized by the number of UUUU inserted in the downstream intron.



Supplementary Figure 14. Ka, Ks and Ka/Ks values of in different types of exons with or without G-runs in the downstream introns (11-70 bp). Methods of calculation of Ka, Ks and Ka/Ks are the same as in Xing et al, PNAS, 2005. (**a**) Conserved skipped exons (i.e., ACEs) with 5'ss in the range of 0-12 bits; (**b**) Conserved constitutive exons (i.e., ICEs) with 5'ss in the range of 0-12 bits; (**c**) ICEs with 5'ss in the range of 4-8 bits; (**d**) Exons that are constitutively spliced in human but alternatively skipped in mouse ($H_{const.}M_{skip.}$) with 5'ss in the range of 0-12 bits.

Supplementary Table 1. Splice sites of splicing reporters and test exon inclusion data

Name	5'ss sequence	5'ss score (bits)	3'ss score (bits)[a]	% inclusion of test exon[b]			
				Length of inserted G-run			
				G_0	G_3	G_6	G_9
pGM6 (w1)	AUG|GUAGUG	0.3	4.1	2.9±2.0	4.0±1.1	3.2±1.1	8.5±0.8
pGM5 (w2)	AAA|GUAAUC	0.6	4.1	2.7±0.5	5.1±1.7	13.5±3.9	14.0±2.8
pGM7 (w3)	UUG|GUAUCU	0.7	4.1	2.4±2.1	6.0±2.5	8.6±3.5	23.8±7.4
pGM1 (w4)	AAG|GUGGCA	1.0	4.1	3.7±2.8	9.2±2.8	6.4±2.9	7.0±2.3
pMZ1 (w5)	AAG|GUUUUG	1.6	4.1	2.3±1.1	4.2±0.7	9.1±0.6	22.0±3.0
pGM2 (w6)	UUG|GUAUAU	1.8	4.1	4.2±3.0	8.0±2.5	12.7±0.1	18.5±0.4
pGM3 (w7)	UUG|GUUUGU	1.9	4.1	5.5±2.2	13.1±2.1	16.8±2.6	46.2±2.0
pGM17	UUG|GUAUCU	0.7	5.7	36.1±1.5	54.3±0.3	56.5±0.1	82.0±0.6
pGM16	UUG|GUAUAU	1.8	5.7	39.1±0.2	55.9±1.4	62.3±1.2	63.3±3.7
pGM11	AUG|GUAGUG	0.3	6.6	59.2±2.4	75.1±1.2	64.6±4.3	73.4±2.1
pGM13	UUG|GUAUCU	0.7	6.6	83.7±1.5	92.4±0.3	89.3±0.1	96.8±0.6
pGM10	AAG|GUGGCA	1.0	6.6	80.9±2.0	86.7±4.4	74.2±1.2	65.4±3.0
pMZ3	AAG|GUUUUG	1.6	6.6	79.0±0.3	82.0±1.0	77.0±1.0	81.0±1.0
pMZ5 (i1)	AAG|GUGAUG	5.2	2.1	9.8±2.7	20.8±1.6	47.9±1.2	76.0±0.1
pMZ7 (i2)	AAG|GUAGUG	5.6	2.1	12.5±2.0	27.0±2.0	46.0±3.0	80.0±0.1
pMZ8 (i3)	AAG|GUUAGA	6.1	2.1	20.8±1.5	41.0±3.0	60.0±1.0	93.0±1.0
pMZ2	AAG|GUGAUG	5.2	4.1	54.4±2.7	73.8±1.5	95.7±4.2	94.9±0.5
pGM14 (s1)	AAG|GUCAGU	8.7	1.2	30.7±3.2	42.9±0.7	63.5±0.3	78.3±0.2
pGM9	AAG|GUCAGU	8.7	2.1	81.4±0.6	86.1±1.0	88.6±0.7	89.7±0.1
pMZ6 (s2)	AAG|GUGAGG	9.2	2.1	56.0±2.5	76.5±1.8	85.0±0.7	90.0±0.8

[a]3'ss sequences (scores):
`agcuccuuauugcauauuag/AUU` (4.1 bits)
`agcuccuuauugcuauuuag/AUU` (5.7 bits)
`agcuccuuuuugcuuauuag/AUU` (6.6 bits)
`agcuccuuauuccauauaag/AUU` (2.1 bits)
`agcuccuuauugcauauaag/AUU` (1.2 bits)

[b]Percent inclusion values represent mean and range of replicates. Values are shown in gray if the average inclusion values of corresponding constructs with fewer G-triplets were >85%. These values were not included in Fig. 1b.

Supplementary Table 2. qRT-PCR validation of reduced exon inclusion following hnRNP H knockdown and primers used.

Category	Ensembl exon ID[a]	Ensembl gene ID[b]	Gene name	Validation	Primers[c]	Ensembl exon ID	Ensembl gene ID	Gene name	Validation	Primers
candidate H-activated	1148544	139546	TARBP2	Yes	F CACAGTGACCCAGGAGTCTG R TCGTAGAGAATCCCAGGTGC	1473133	074219	TEAD2	Yes	F CAAGGGAAATCCAGTCCAAG R ATCCTCCAGACCAAAACTGG
	1316542	185359	HGS	Yes	F ACGAGCCCAAGTACAAGGTG R GTCATCACCCCGAACTGC	1472346	204681	GABBR1	Yes	F ATGTCCTCACTCTCGCCATC R GCAGCCCCTTGTAACCATAG
	1163808	144043	TEX261	Yes	F AATACACAGTGGCCACCAGC R GCTAGGTAATGATTCACCACCAC	1001187	151693	ASAP2	No	F CTGCACTACTGCTGCCTGAC R CGCCATTCATATTCAACGTG
	1186760	173020	ADRBK1	No	F GAAGTGCCTGGACAAAAAGC R AGAGGTGGTAGTGCAGGTCC	1033259	157538	DSCR3	Yes	F CGTATCATGGCGTGTTTGTC R GTAATCGTGAAGTCCACGGG
	1074445	157404	KIT	No	F GGAAGCCCTCATGTCTGAAC R ACAGATCTCCTTTTGTCGGC	1455572	119912	IDE	No	F CCTTCCAAGTCAGCTGGTTC R TGTATGCCATTAGCTCGACG
	1462548	157657	ZNF618	Yes	F ACATCCAGGAGGTGATCTGG R GATTCCGCATTCGTAAGACC	0897426	068903	SIRT2	Yes	F AGTTCTGTGCCCTATCACGG R TGTCTGCTTCTCCACCAGC
	0921221	110514	MADD	Yes	F TGATTGGTGACAAGCCAAAG R AGTGCTGCATGTGTCAGAGG	0880797	100416	TRMU	Yes	F TACAGGGACCTGCTGAGGAC R GTCAGAGCAAGGGACTCAGG
	1370700	185513	L3MBTL	Yes	F TCCAGAAGATCCCAATCAGG R CTAGCAGTGGGTTGTCCCTC	1309322	142192	APP	No	F aggaaCCCTACGAAGAAGCC R TTCTCATCCCCAGGTGTCTC
	1033563	066777	ARFGEF1	No	F TTTGCTAGTGCGGATACAGC R GGCCTCCATGAGTGCTTTAG	1474519	204842	ATXN2	Yes	F TCGCCACAGAATAGTTTCCC R AGCAGTAGAAGGGAGGAGGG
	1036207	182095	TNRC18	Yes	F GTATGCCATGAAGTCCTCCC R TCATGTTCCTCCTCAGACCC	0682741	105613	MAST1	Yes	F CACTTCTCGTTTGCCTCCTC R CACCATCCTCGTCTGTGATG
	1180438	114859	CLCN2	No	F CCTCTTCAAAACCCGATTCC R ATGAACTGTCCAAAGCCAGG	0893273	083750	RRAGB	Yes	F TGATGGGTAAAAGTGGGTCTG R CCCACAGGTTCAATACCAGG
Negative controls	1067767	066279	ASPM	Yes	F GCTAGGCGGTTAATTGTTCG R TCAGGATTCCAAAGTAGGCG					
	0915987	134762	DSC3	Yes	F CATCCAACTCAGTGTCGTGC R TGTGCCACTCCGAGTAAGTG					
	0678439	005100	DHX33	Yes	F AAGTCCCAAATGTGCTCACC R TATGAACTTCTTGCGGACCC					
	1083756	164347	GFM2	Yes	F ATGCAGGCAAAACTACCACC R CACTGCACCATCCAACACTC					

[a]Ensembl (version 45) exon IDs are "ENSE0000" followed by the numbers shown.
[b]Ensembl (version 45) gene IDs are "ENSG00000" followed by the numbers shown.
[c]F: forward primers; R: reverse primers

Supplementary Table 3. Base composition of uniquely mapped CLIP tags (3.6 million). The number of mono-, di-, and tri-nucleotides were enumerated and shown in the table, along with their percent occurrence among all mono-, di-, and tri-nucleotides observed.

	Sequence	#Occurrences	Percentage
Top 4 1mers:	G	54032175	46.3%
	A	33130053	28.4%
	T	15839678	13.6%
	C	13704408	11.7%
Top 4 2mers:	GG	23667823	20.9%
	AG	17679666	15.6%
	GA	17165640	15.2%
	AA	9872595	8.7%
Top 4 3mers:	GGG	9776419	8.9%
	GAG	9257906	8.5%
	GGA	7965353	7.3%
	AGG	7682803	7.0%

Supplementary Table 4. Differentially expressed exons[a] upon hnRNP H knockdown and the number of CLIP tags mapped to the exon, upstream (-280 to -40 nt) and downstream (+10 to +250 nt) intronic regions.

			Num of mRNA-Seq Reads									Num of CLIP tags					
			Control knockdown				hnRNP H knockdown										
Gene	Symbol	Exon coordiante	Inc	Excl	common	PSI[b]	Inc	Excl	common	PSI	P value[c]	-280 to -40 (All)	Exon (All)	+10 to +250 (All)	-280 to -40 (Unique)	Exon (Unique)	+10 to +250 (Unique)
ENSG00000141756	FKBP10	chr17:37229619:37229801:+	103	56	1724	0.281	472	10	2198	0.909	5.70E-37	0	20	0	0	4	0
ENSG00000092841	MYL6	chr12:54840677:54840721:+	26	130	1567	0.082	247	39	2114	0.738	6.30E-34	0	0	0	0	0	0
ENSG00000169592	INO80E	chr16:29920034:29923479:+	937	2	2800	0.881	359	12	2212	0.322	4.30E-28	0	1269	136	0	4	4
ENSG00000146416	AIG1	chr6:143499665:143499820:+	140	2	548	0.943	75	0	1138	1	1.10E-19	0	0	0	0	0	0
ENSG00000129351	ILF3	chr19:10659022:10659384:+	534	5	2915	0.931	732	0	2275	1	3.50E-19	15	6	10	3	3	3
ENSG00000114631	PODXL2	chr3:128869974:128870130:+	115	0	215	1	55	3	436	0.812	3.30E-16	0	0	5	0	0	2
ENSG00000204463	BAT3	chr6:31715256:31715402:-	197	13	193	0.788	97	22	291	0.52	6.50E-14	0	0	0	0	0	0
ENSG00000144043	TEX261	chr2:71072468:71072621:-	106	3	17	0.894	16	18	24	0.175	1.80E-13	0	0	0	0	0	0
ENSG00000104852	SNRP70	chr19:54297183:54298656:+	1177	65	3473	0.395	624	133	2668	0.145	3.20E-13	13	689	47	3	4	4
ENSG00000198931	APRT	chr16:87404332:87404465:-	585	12	329	0.927	269	62	270	0.531	1.90E-12	0	0	0	0	0	0
ENSG00000169592	n/a	chr16:29920236:29920352:+	99	2	2800	0.933	16	12	2212	0.274	6.50E-12	0	1	9	0	1	3
ENSG00000080603	SRCAP	chr16:30640011:30640298:+	100	0	328	1	29	13	380	0.253	8.90E-12	0	0	0	0	0	0
ENSG00000123159	GIPC1	chr19:14464669:14464724:-	39	8	48	0.666	6	11	109	0.182	2.60E-11	0	0	0	0	0	0
ENSG00000126005	EIF6	chr20:33331871:33332046:-	429	1	278	0.989	191	8	268	0.839	4.40E-11	0	0	0	0	0	0
ENSG00000173812	EIF1	chr17:37099556:37099719:+	206	2	1743	0.959	494	17	2351	0.869	5.40E-11	18	0	0	3	0	0
ENSG00000182979	MTA1	chr14:104986741:104986791:+	32	8	801	0.629	1	13	1048	0.032	6.10E-11	3	1	5	2	1	3
ENSG00000135506	n/a	chr12:56400149:56400313:+	93	23	549	0.479	362	0	1005	1	8.10E-11	0	1	0	0	1	0
ENSG00000170889	RPS9	chr19:59401956:59402142:+	140	3	816	0.907	465	47	1375	0.674	2.60E-10	0	1	0	0	1	0
ENSG00000188157	AGRN	chr1:978996:979220:+	200	0	530	1	74	1	487	0.931	3.20E-10	0	0	0	0	0	0
ENSG00000071462	WBSCR22	chr7:72745595:72745680:+	169	0	228	1	47	1	197	0.94	7.50E-10	1	0	0	1	0	0
ENSG00000142453	CARM1	chr19:10893051:10893119:+	34	15	2643	0.458	119	2	3036	0.957	9.50E-10	0	0	0	0	0	0
ENSG00000175634	RPS6KB2	chr11:66955415:66955562:+	167	0	273	1	73	2	314	0.899	1.10E-09	0	0	4	0	0	2
ENSG00000204463	BAT3	chr6:31720063:31720170:-	81	23	105	0.511	19	33	110	0.146	2.50E-09	0	0	0	0	0	0
ENSG00000126453	BCL2L12	chr19:54865285:54865557:+	174	1	875	0.965	52	7	663	0.54	3.00E-09	0	0	0	0	0	0
ENSG00000141002	TCF25	chr16:88487637:88487767:+	98	0	373	1	28	1	370	0.881	4.00E-09	0	0	0	0	0	0
ENSG00000023191	RNH1	chr11:494824:494996:-	53	0	130	1	13	6	180	0.323	4.80E-09	0	0	0	0	0	0
ENSG00000023330	ALAS1	chr3:52207728:52207904:+	16	7	255	0.332	184	3	760	0.93	6.00E-09	0	0	0	0	0	0
ENSG00000100138	NHP2L1	chr22:40408306:40408537:-	19	51	204	0.062	121	53	366	0.29	7.00E-09	0	0	0	0	0	0
ENSG00000023191	RNH1	chr11:494824:494996:-	48	2	130	0.841	11	5	180	0.327	1.30E-08	0	0	0	0	0	0
ENSG00000162910	MRPL55	chr1:226362761:226362798:-	8	11	137	0.255	53	6	136	0.806	1.60E-08	0	0	0	0	0	0
ENSG00000118267	ZNF271	chr18:31124972:31125194:+	118	5	1493	0.813	29	2	1101	0.728	2.10E-08	0	0	0	0	0	0
ENSG00000084234	APLP2	chr11:129510668:129510820:+	390	1	931	0.989	404	39	1489	0.713	3.10E-08	7	0	10	2	0	4
ENSG00000162910	MRPL55	chr1:226362761:226362798:-	9	11	137	0.278	60	25	136	0.53	3.10E-08	0	0	0	0	0	0
ENSG00000214063	TSPAN4	chr11:852550:852741:+	133	0	353	1	45	1	325	0.902	4.10E-08	2	0	0	2	0	0
ENSG00000119596	YLPM1	chr14:74365669:74365799:+	71	0	297	1	48	1	594	0.927	4.60E-08	0	1	0	0	1	0
ENSG00000102125	n/a	chrX:153293617:153293745:+	89	0	358	1	36	1	434	0.906	5.00E-08	0	0	18	0	0	3
ENSG00000065485	PDIA5	chr3:124332017:124332153:+	30	0	116	1	10	1	269	0.72	5.60E-08	0	0	0	0	0	0
ENSG00000013306	SLC25A39	chr17:39755702:39755746:-	77	1	169	0.972	17	10	152	0.43	6.00E-08	0	0	0	0	0	0
ENSG00000126351	THRA	chr17:35498021:35498279:+	76	1	785	0.926	169	0	803	1	6.60E-08	0	0	0	0	0	0
ENSG00000160688	FLAD1	chr1:153228660:153228807:+	70	3	688	0.851	24	2	798	0.746	1.00E-07	7	0	0	3	0	0
ENSG00000178802	n/a	chr15:72976405:72976613:+	239	12	2462	0.794	77	15	1565	0.498	1.40E-07	0	0	0	0	0	0

ENSG00000161939	C17orf49	chr17:6860953:6861057:+	90	0	709	1	33	7	743	0.587	1.60E-07	0	0	0	0	0	0
ENSG00000130731	C16orf13	chr16:625613:625775:-	108	4	79	0.861	59	31	105	0.304	2.40E-07	0	0	0	0	0	0
ENSG00000152904	GGPS1	chr1:233565180:233565272:+	49	1	574	0.94	36	0	1352	1	2.40E-07	0	0	0	0	0	0
ENSG00000065665	SEC61A2	chr10:12237783:12237936:+	35	2	230	0.807	224	0	569	1	2.70E-07	0	0	0	0	0	0
ENSG00000112118	n/a	chr6:52237992:52238142:-	8	61	1726	0.031	48	65	1924	0.151	4.50E-07	0	0	0	0	0	0
ENSG00000070669	ASNS	chr7:97337026:97337061:-	82	1	59	0.975	115	12	221	0.821	5.70E-07	0	0	0	0	0	0
ENSG00000173915	n/a	chr10:105145493:105145779:-	117	15	1	0.543	75	47	1	0.195	6.20E-07	0	0	0	0	0	0
ENSG00000127952	MKSTYX	chr7:75468144:75468256:-	32	1	82	0.902	15	9	192	0.325	6.30E-07	0	0	0	0	0	0
ENSG00000167302	C17orf56	chr17:76820710:76820913:-	23	1	17	0.819	5	2	48	0.329	7.00E-07	0	0	0	0	0	0
ENSG00000058600	POLR3E	chr16:22241672:22241753:+	39	0	149	1	14	4	245	0.546	9.70E-07	3	0	5	1	0	4
ENSG00000166913	YWHAB	chr20:42949703:42949797:+	16	24	980	0.175	79	33	1403	0.432	9.80E-07	0	0	0	0	0	0
ENSG00000141480	ARRB2	chr17:4568305:4568389:+	98	3	284	0.917	35	14	266	0.458	1.10E-06	0	0	0	0	0	0
ENSG00000013306	SLC25A39	chr17:39755702:39755746:-	100	2	138	0.957	23	7	104	0.594	1.10E-06	0	0	0	0	0	0
ENSG00000188917	CXorf34	chrX:100192896:100193378:-	23	1	31	0.695	92	0	23	1	1.30E-06	0	0	0	0	0	0
ENSG00000162735	PEX19	chr1:158519944:158520053:-	57	0	75	1	17	8	89	0.384	1.60E-06	0	0	0	0	0	0
ENSG00000177697	CD151	chr11:824530:824591:+	79	21	414	0.596	10	23	214	0.145	1.70E-06	0	0	0	0	0	0
ENSG00000106829	TLE4	chr9:81511858:81511953:+	3	10	252	0.087	35	7	313	0.613	1.90E-06	0	0	0	0	0	0
ENSG00000111652	COPS7A	chr12:6707353:6707428:+	61	2	309	0.916	12	17	233	0.201	2.00E-06	0	0	0	0	0	0
ENSG00000106628	POLD2	chr7:44122239:44122396:-	217	0	17	1	152	3	44	0.922	2.40E-06	0	0	0	0	0	0
ENSG00000108509	CAMTA2	chr17:4817604:4817788:-	50	0	12	1	6	2	16	0.387	2.40E-06	0	0	0	0	0	0
ENSG00000110536	PTPMT1	chr11:47547828:47548019:+	235	13	1050	0.788	192	25	1399	0.612	2.60E-06	0	0	2	0	0	2
ENSG00000182087	C19orf6	chr19:962095:962194:-	178	0	2005	1	85	3	1782	0.898	2.70E-06	0	0	0	0	0	0
ENSG00000197586	ENTPD6	chr20:25151474:25151614:+	91	0	403	1	43	1	469	0.916	2.90E-06	0	0	1	0	0	1
ENSG00000143774	GUK1	chr1:226395448:226395612:+	62	41	824	0.256	19	18	839	0.194	3.50E-06	0	0	0	0	0	0
ENSG00000116750	n/a	chr1:191263628:191263708:-	19	15	39	0.305	4	35	82	0.038	3.60E-06	0	0	0	0	0	0
ENSG00000100083	GGA1	chr22:36350922:36351029:+	69	0	299	1	16	1	241	0.826	4.70E-06	40	0	114	4	0	4
ENSG00000117448	AKR1A1	chr1:45806744:45806943:+	187	3	484	0.925	16	7	127	0.313	5.10E-06	0	0	0	0	0	0
ENSG00000072778	ACADVL	chr17:7067687:7067773:+	149	0	319	1	111	2	459	0.949	5.40E-06	0	2	0	0	1	0
ENSG00000178802	n/a	chr15:72969921:72970048:+	197	1	601	0.981	90	5	506	0.828	9.40E-06	0	0	8	0	0	4
ENSG00000154380	ENAH	chr1:223759316:223759378:-	4	2	30	0.438	28	1	19	0.916	1.00E-05	0	0	0	0	0	0
ENSG00000197586	ENTPD6	chr20:25149868:25149969:+	62	3	392	0.863	25	0	451	1	1.40E-05	0	0	0	0	0	0
ENSG00000138668	HNRNPD	chr4:83496714:83496860:-	121	95	243	0.238	246	74	313	0.449	1.80E-05	0	0	0	0	0	0
ENSG00000135473	PAN2	chr12:54998288:54998485:-	81	1	227	0.942	41	10	274	0.451	2.00E-05	0	0	0	0	0	0
ENSG00000131469	RPL27	chr17:38405476:38405645:+	932	24	3523	0.897	1360	73	4143	0.806	2.10E-05	0	2	0	0	2	0
ENSG00000169241	RAG1AP1	chr1:153376661:153376822:+	107	2	405	0.925	34	0	307	1	2.10E-05	0	0	0	0	0	0
ENSG00000162910	MRPL55	chr1:226363279:226363345:-	66	11	137	0.694	24	25	136	0.266	2.20E-05	0	0	0	0	0	0
ENSG00000010270	n/a	chr7:38185217:38185272:+	11	6	398	0.428	0	15	640	0	2.80E-05	0	0	0	0	0	0
ENSG00000115226	FNDC4	chr2:27569037:27569161:-	26	0	68	1	6	3	96	0.352	2.90E-05	0	0	0	0	0	0
ENSG00000169689	STRA13	chr17:77570806:77570859:-	67	33	175	0.457	41	96	218	0.15	2.90E-05	0	0	0	0	0	0
ENSG00000151500	THYN1	chr11:133623913:133624063:-	86	1	106	0.954	32	11	100	0.413	2.90E-05	0	0	0	0	0	0
ENSG00000171163	ZNF692	chr1:247111504:247111603:-	43	0	173	1	14	8	196	0.351	3.00E-05	0	0	0	0	0	0
ENSG00000214021	TTLL3	chr3:9832758:9832886:+	38	1	186	0.91	6	0	159	1	3.20E-05	0	0	0	0	0	0
ENSG00000141556	TBCD	chr17:78458926:78458986:+	52	0	160	1	17	1	175	0.87	3.20E-05	4	0	18	2	0	4
ENSG00000169241	RAG1AP1	chr1:153375928:153376051:+	95	11	351	0.702	104	0	196	1	3.30E-05	0	1	0	0	1	0
ENSG00000100151	PICK1	chr22:36799726:36799776:+	42	0	206	1	25	5	360	0.68	3.40E-05	0	0	0	0	0	0
ENSG00000169241	RAG1AP1	chr1:153376661:153376822:+	111	0	321	1	42	10	261	0.492	3.40E-05	0	0	0	0	0	0
ENSG00000099991	CABIN1	chr22:22902079:22902243:+	84	0	616	1	65	2	979	0.881	3.50E-05	0	0	34	0	0	4
ENSG00000177830	CHID1	chr11:883427:883519:-	132	3	115	0.934	77	26	122	0.488	4.70E-05	0	0	0	0	0	0

ENSG00000119688	PMP69	chr14:73829610:73829704:-	29	2	41	0.822	8	2	63	0.56	5.30E-05	0	0	0	0	0	0
ENSG00000076928	ARHGEF1	chr19:47101167:47101253:+	34	0	58	1	20	4	122	0.625	5.50E-05	0	0	0	0	0	0
ENSG00000123159	GIPC1	chr19:14454501:14454818:-	245	2	232	0.945	210	7	328	0.808	5.50E-05	0	0	0	0	0	0
ENSG00000125695	LYK5	chr17:59138339:59138510:-	61	1	12	0.931	19	4	19	0.513	5.50E-05	0	0	0	0	0	0
ENSG00000089154	GCN1L1	chr12:119053372:119053520:-	168	0	99	1	44	5	61	0.682	5.90E-05	0	0	0	0	0	0
ENSG00000175061	C17orf45	chr17:16283699:16283742:+	203	250	3603	0.267	144	359	3896	0.152	7.40E-05	0	0	0	0	0	0
ENSG00000110514	MADD	chr11:47304865:47304934:+	19	0	129	1	8	1	282	0.748	7.60E-05	1	0	43	1	0	4
ENSG00000168724	DNAJC21	chr5:34985405:34985539:+	5	4	111	0.245	34	3	133	0.746	8.00E-05	0	44	0	0	4	0
ENSG00000120913	PDLIM2	chr8:22505009:22505126:+	27	0	254	1	3	5	204	0.144	8.60E-05	0	0	0	0	0	0
ENSG00000198715	C1orf85	chr1:154531174:154531431:-	139	0	143	1	32	5	78	0.514	8.80E-05	0	0	0	0	0	0
ENSG00000070047	KIAA1542	chr11:595605:595724:+	20	0	79	1	6	4	133	0.295	1.00E-04	0	0	0	0	0	0
ENSG00000101294	HM13	chr20:29606210:29606293:+	171	2	598	0.967	159	7	890	0.885	1.00E-04	1	0	0	1	0	0
ENSG00000132819	RBM38	chr20:55401741:55401795:+	41	0	1956	1	19	11	2610	0.416	1.00E-04	22	0	0	3	0	0
ENSG00000123374	CDK2	chr12:54649526:54649627:+	89	1	582	0.965	67	12	836	0.631	1.10E-04	4	5	0	3	2	0
ENSG00000093167	LRRFIP2	chr3:37107962:37108033:-	2	4	43	0.155	18	2	32	0.767	1.10E-04	0	0	0	0	0	0
ENSG00000089159	n/a	chr12:119137746:119137847:-	5	3	22	0.338	29	1	16	0.899	1.20E-04	0	0	0	0	0	0
ENSG00000177697	CD151	chr11:824530:824591:+	65	2	415	0.927	10	4	214	0.495	1.30E-04	0	0	0	0	0	0
ENSG00000183010	PIG45	chr17:77486504:77486683:-	536	0	591	1	409	7	623	0.926	1.30E-04	0	0	0	0	0	0
ENSG00000165948	FAM14B	chr14:93637913:93638074:+	80	0	431	1	34	1	411	0.887	1.30E-04	0	0	0	0	0	0
ENSG00000102710	P38IP	chr13:36482689:36482792:-	6	8	70	0.185	41	7	101	0.639	1.40E-04	0	0	0	0	0	0
ENSG00000100416	MTU1	chr22:45130551:45130633:+	38	0	518	1	12	10	538	0.291	1.40E-04	0	0	1	0	0	1
ENSG00000116898	MRPS15	chr1:36700276:36700351:-	141	1	23	0.981	75	13	26	0.673	1.40E-04	0	0	0	0	0	0
ENSG00000143537	ADAM15	chr1:153300515:153300589:+	27	9	316	0.519	4	17	271	0.078	1.40E-04	0	0	0	0	0	0
ENSG00000187741	FANCA	chr16:88352486:88352614:-	59	0	6	1	19	2	13	0.717	1.50E-04	0	0	0	0	0	0
ENSG00000198917	C9orf114	chr9:130625847:130625994:-	67	1	598	0.942	28	2	593	0.774	1.60E-04	0	0	0	0	0	0
ENSG00000175279	APITD1	chr1:10416486:10416609:+	53	0	148	1	33	2	232	0.818	1.60E-04	0	0	0	0	0	0
ENSG00000010219	DYRK4	chr12:4586754:4586814:+	1	6	241	0.062	17	6	258	0.528	1.80E-04	0	0	7	0	0	2
ENSG00000136295	TTYH3	chr7:2662252:2662344:+	66	0	239	1	20	5	191	0.563	1.80E-04	0	0	0	0	0	0
ENSG00000085733	CTTN	chr11:69945224:69945334:+	67	4	498	0.83	116	3	467	0.919	2.00E-04	0	2	0	0	1	0
ENSG00000169683	LRRC45	chr17:77580708:77580800:+	61	0	238	1	42	3	365	0.818	2.20E-04	0	0	0	0	0	0
ENSG00000160767	C1orf2	chr1:153486691:153487236:-	451	0	26	1	202	4	29	0.819	2.30E-04	0	0	0	0	0	0
ENSG00000103275	UBE2I	chr16:1302349:1302605:+	37	28	437	0.18	123	27	742	0.43	2.30E-04	1	0	0	1	0	0
ENSG00000159352	PSMD4	chr1:149501261:149501401:+	22	1	249	0.847	11	1	497	0.735	2.40E-04	0	0	0	0	0	0
ENSG00000180964	TCEAL8	chrX:102396189:102396261:-	56	1	676	0.953	21	6	639	0.56	2.40E-04	0	0	0	0	0	0
ENSG00000125482	TTF1	chr9:134266663:134268036:-	191	0	34	1	224	7	83	0.552	2.60E-04	0	0	0	0	0	0
ENSG00000196476	C20orf96	chr20:218200:218317:-	20	1	13	0.849	11	7	37	0.307	2.60E-04	0	0	0	0	0	0
ENSG00000108840	HDAC5	chr17:39517916:39518049:-	72	0	54	1	14	2	36	0.646	2.70E-04	0	0	0	0	0	0
ENSG00000197694	SPTAN1	chr9:130395083:130395142:+	11	7	266	0.384	56	1	434	0.957	2.80E-04	0	0	0	0	0	0
ENSG00000064490	RFXANK	chr19:19170466:19170532:+	103	0	281	1	34	1	199	0.928	3.20E-04	0	0	0	0	0	0
ENSG00000142327	RNPEPL1	chr2:241161865:241161981:+	53	0	294	1	17	2	256	0.706	3.20E-04	0	0	0	0	0	0
ENSG00000198715	C1orf85	chr1:154529735:154529991:-	207	1	354	0.972	68	1	209	0.918	3.20E-04	0	0	0	0	0	0
ENSG00000108272	MGC4172	chr17:32030213:32030305:+	44	1	200	0.934	6	0	119	1	3.20E-04	0	0	0	0	0	0
ENSG00000115694	STK25	chr2:242096094:242096223:-	132	0	51	1	25	10	21	0.399	3.30E-04	0	0	0	0	0	0
ENSG00000102030	ARD1A	chrX:152850718:152850762:-	192	0	253	1	40	4	107	0.816	3.30E-04	0	0	0	0	0	0
ENSG00000117335	CD46	chr1:206030221:206030313:+	10	19	1853	0.145	55	16	3271	0.525	3.60E-04	0	0	0	0	0	0
ENSG00000083457	ITGAE	chr17:3573331:3573426:-	4	8	30	0.137	26	7	30	0.54	3.70E-04	0	0	0	0	0	0
ENSG00000119139	TJP2	chr9:71054111:71054221:+	24	0	136	1	16	1	300	0.824	3.80E-04	0	0	0	0	0	0
ENSG00000137817	PARP6	chr15:70330240:70330349:-	37	1	23	0.916	5	2	20	0.423	4.00E-04	0	0	0	0	0	0

ENSG00000167130	DOLPP1	chr9:130888239:130888367:+	55	0	105	1	30	4	141	0.667	4.00E-04	0	0	0	0	0	0
ENSG00000122406	RPL5	chr1:93071534:93071603:+	30	3	479	0.788	82	4	604	0.884	4.00E-04	1	0	0	1	0	0
ENSG00000137817	DKFZp566D244	chr15:70330240:70330352:-	36	1	23	0.912	5	2	20	0.419	4.20E-04	0	0	0	0	0	0
ENSG00000141480	ARRB2	chr17:4565057:4565087:+	27	10	363	0.574	8	5	420	0.444	4.30E-04	13	0	2	3	0	2
ENSG00000125648	SLC25A23	chr19:6404992:6405099:-	80	1	44	0.96	77	8	94	0.74	4.50E-04	0	0	0	0	0	0
ENSG00000141551	CSNK1D	chr17:77797141:77797204:-	58	5	434	0.818	36	14	579	0.498	4.60E-04	0	0	0	0	0	0
ENSG00000173653	RCE1	chr11:66369547:66369609:+	31	2	541	0.858	12	14	675	0.25	4.60E-04	0	0	0	0	0	0
ENSG00000100023	PPIL2	chr22:20378105:20378161:+	28	0	283	1	19	1	555	0.885	4.70E-04	2	0	0	1	0	0
ENSG00000069188	SDK2	chr17:68860207:68860371:-	28	0	7	1	17	2	24	0.659	4.80E-04	0	0	0	0	0	0
ENSG00000087460	GNASL	chr20:56907391:56907435:+	245	54	1199	0.668	418	64	1505	0.744	4.90E-04	7	0	10	3	0	2
ENSG00000084234	APLP2	chr11:129498717:129498884:+	107	26	1284	0.481	253	43	2004	0.57	4.90E-04	0	0	0	0	0	0
ENSG00000100207	TCF20	chr22:40894559:40894686:-	17	3	191	0.603	75	0	335	1	5.10E-04	0	0	0	0	0	0
ENSG00000063854	HAGH	chr16:1809117:1809225:-	92	0	60	1	33	9	47	0.519	5.10E-04	0	0	0	0	0	0
ENSG00000143569	n/a	chr1:152508462:152508512:+	4	8	277	0.175	20	6	239	0.586	5.20E-04	0	0	0	0	0	0
ENSG00000063601	MTMR1	chrX:149674819:149674995:+	65	0	370	1	49	1	563	0.914	5.70E-04	3	0	0	1	0	0
ENSG00000089351	GRAMD1A	chr19:40196000:40196112:+	65	0	331	1	17	1	224	0.831	5.70E-04	3	0	0	2	0	0
ENSG00000165280	FANCG	chr9:35068138:35068340:-	100	0	41	1	34	4	36	0.626	6.00E-04	0	0	0	0	0	0
ENSG00000099875	MKNK2	chr19:1991133:1991176:-	33	0	171	1	152	6	382	0.919	6.10E-04	0	0	0	0	0	0
ENSG00000071626	DAZAP1	chr19:1376877:1376959:+	219	5	2237	0.937	294	1	2182	0.99	6.50E-04	0	0	4	0	0	2
ENSG00000184277	BLP2	chr15:100009422:100009499:-	28	19	173	0.342	71	10	204	0.714	6.60E-04	0	0	0	0	0	0
ENSG00000102098	SCML2	chrX:18170480:18170631:-	11	2	19	0.569	4	3	57	0.243	7.00E-04	0	0	0	0	0	0
ENSG00000169241	RAG1AP1	chr1:153375928:153376051:+	80	2	405	0.916	106	0	307	1	7.00E-04	0	1	0	0	1	0
ENSG00000108671	PSMD11	chr17:27830982:27831017:+	27	0	95	1	32	19	288	0.446	7.40E-04	0	0	0	0	0	0
ENSG00000139546	TARBP2	chr12:52185186:52185313:+	60	1	220	0.941	26	2	221	0.777	8.70E-04	0	0	0	0	0	0
ENSG00000129484	PARP2	chr14:19895395:19895519:+	56	1	325	0.938	186	0	626	1	8.80E-04	0	0	0	0	0	0
ENSG00000005486	RHBDD2	chr7:75348619:75348740:+	23	14	612	0.312	15	41	1184	0.092	9.20E-04	0	0	0	0	0	0
ENSG00000102858	MGRN1	chr16:4667465:4667574:+	37	0	204	1	19	1	281	0.848	9.40E-04	0	0	3	0	0	1
ENSG00000108591	DRG2	chr17:17943672:17943762:+	69	0	228	1	39	1	265	0.927	9.50E-04	2	0	9	1	0	3
ENSG00000162910	MRPL55	chr1:226363279:226363345:-	64	11	137	0.688	26	6	136	0.621	9.80E-04	0	0	0	0	0	0
ENSG00000169188	APEX2	chrX:55046120:55046266:+	72	0	134	1	52	1	197	0.927	1.00E-03	0	0	0	0	0	0
ENSG00000105771	C19orf61	chr19:48940764:48940876:-	49	2	21	0.876	16	2	27	0.698	1.00E-03	0	0	0	0	0	0
ENSG00000105397	TYK2	chr19:10325204:10325322:-	72	2	74	0.91	25	0	68	1	1.10E-03	0	0	0	0	0	0
ENSG00000178188	SH2B1	chr16:28791992:28792091:+	88	0	773	1	75	2	1119	0.921	1.20E-03	0	0	0	0	0	0
ENSG00000125503	PPP1R12C	chr19:60299442:60299515:-	28	0	26	1	8	3	31	0.491	1.30E-03	0	0	0	0	0	0
ENSG00000171603	CLSTN1	chr1:9720143:9720199:-	4	6	95	0.213	26	5	126	0.678	1.30E-03	0	0	0	0	0	0
ENSG00000178188	SH2B1	chr16:28791992:28792091:+	90	0	773	1	77	4	1119	0.856	1.40E-03	0	0	0	0	0	0
ENSG00000100207	n/a	chr22:40894559:40894653:-	15	3	44	0.614	58	0	60	1	1.40E-03	0	0	0	0	0	0
ENSG00000100379	KCTD17	chr22:35787525:35787615:+	45	0	1043	1	19	1	1047	0.861	1.40E-03	0	0	8	0	0	3
ENSG00000115486	GGCX	chr2:85641449:85641619:-	52	1	61	0.92	35	12	88	0.393	1.40E-03	0	0	0	0	0	0
ENSG00000105698	USF2	chr19:40452546:40452746:+	154	7	638	0.814	146	16	906	0.644	1.40E-03	0	1	0	0	1	0
ENSG00000083845	RPS5	chr19:63596155:63596364:+	181	3	1543	0.921	381	15	2382	0.83	1.40E-03	0	0	0	0	0	0
ENSG00000082014	SMARCD3	chr7:150573479:150573690:-	38	0	1	1	27	9	2	0.364	1.50E-03	0	0	0	0	0	0
ENSG00000100350	FOXRED2	chr22:35230091:35230360:-	136	2	103	0.916	63	2	93	0.834	1.50E-03	0	0	0	0	0	0
ENSG00000115204	MPV17	chr2:27388586:27388618:-	21	0	452	1	6	1	514	0.747	1.50E-03	0	0	0	0	0	0
ENSG00000072071	LPHN1	chr19:14132405:14132508:-	34	0	34	1	9	4	32	0.405	1.50E-03	0	0	0	0	0	0
ENSG00000123562	MORF4L2	chrX:102826265:102826313:-	43	22	146	0.457	34	69	226	0.175	1.50E-03	0	0	0	0	0	0
ENSG00000132716	WDR42A	chr1:158461476:158461533:-	33	1	11	0.93	18	3	25	0.707	1.50E-03	0	0	0	0	0	0
ENSG00000171311	EXOSC1	chr10:99192984:99193058:-	73	0	28	1	44	1	44	0.94	1.60E-03	0	0	0	0	0	0

ENSG00000129250	KIF1C	chr17:4844239:4844371:+	98	0	168	1	30	3	108	0.724	1.60E-03	0	0	0	0	0	0
ENSG00000124357	NAGK	chr2:71151140:71151224:+	46	0	184	1	30	4	264	0.717	1.70E-03	0	0	0	0	0	0
ENSG00000144524	COPS7B	chr2:232364687:232364762:+	25	0	192	1	91	2	331	0.942	1.70E-03	4	0	0	1	0	0
ENSG00000112983	BRD8	chr5:137532058:137532276:-	82	1	73	0.939	22	4	47	0.507	1.70E-03	0	0	0	0	0	0
ENSG00000100281	HMG2L1	chr22:33989975:33990089:+	19	3	482	0.644	66	0	747	1	1.70E-03	0	2	0	0	1	0
ENSG00000171720	HDAC3	chr5:140985436:140985515:-	84	0	125	1	18	5	63	0.556	1.80E-03	0	0	0	0	0	0
ENSG00000197111	PCBP2	chr12:52145983:52146024:+	105	0	424	1	206	5	538	0.949	1.80E-03	0	0	0	0	0	0
ENSG00000164405	UQCRQ	chr5:132230460:132230626:+	1312	2	2066	0.993	957	15	1768	0.935	1.80E-03	0	4	0	0	2	0
ENSG00000015475	n/a	chr22:16612871:16612940:-	33	0	59	1	25	2	117	0.823	1.80E-03	0	0	0	0	0	0
ENSG00000180900	SCRIB	chr8:144961710:144961772:-	73	1	40	0.966	26	5	35	0.669	1.80E-03	0	0	0	0	0	0
ENSG00000163719	MTMR14	chr3:9706648:9706827:+	40	1	591	0.896	18	7	635	0.356	1.90E-03	0	0	1	0	0	1
ENSG00000159079	n/a	chr21:32896453:32896548:-	152	0	138	1	89	2	140	0.934	1.90E-03	0	0	0	0	0	0
ENSG00000149923	PPP4C	chr16:30000081:30000132:+	160	3	783	0.957	167	8	1183	0.898	2.00E-03	3	0	0	1	0	0
ENSG00000161091	C19orf28	chr19:3493774:3493975:-	34	3	210	0.692	144	3	475	0.905	2.00E-03	0	0	0	0	0	0
ENSG00000186111	PIP5K1C	chr19:3584435:3584518:-	9	2	470	0.604	1	5	683	0.064	2.00E-03	0	0	0	0	0	0
ENSG00000141556	TBCD	chr17:78489214:78489301:+	101	1	469	0.971	23	1	225	0.884	2.00E-03	0	0	0	0	0	0
ENSG00000119777	TMEM214	chr2:27116113:27116209:+	74	0	294	1	20	6	175	0.512	2.00E-03	0	0	0	0	0	0
ENSG00000060069	CTDP1	chr18:75589895:75590057:+	13	1	117	0.749	22	0	647	1	2.00E-03	0	0	0	0	0	0
ENSG00000105613	MAST1	chr19:12812260:12812348:+	16	0	33	1	7	3	62	0.435	2.10E-03	0	0	0	0	0	0
ENSG00000197111	PCBP2	chr12:52147856:52147894:+	169	1	442	0.987	255	9	452	0.93	2.10E-03	0	0	1	0	0	1
ENSG00000173511	VEGFB	chr11:63759861:63760057:+	146	0	243	1	44	4	133	0.689	2.20E-03	0	0	0	0	0	0
ENSG00000186815	TPCN1	chr12:112210351:112210422:+	14	0	91	1	4	1	139	0.594	2.20E-03	0	0	0	0	0	0
ENSG00000128272	n/a	chr22:38247305:38247622:+	41	0	2214	1	122	1	3814	0.945	2.20E-03	267	6	0	4	4	0
ENSG00000197448	GSTK1	chr7:142671763:142671891:+	114	0	468	1	38	1	286	0.91	2.20E-03	0	0	0	0	0	0
ENSG00000169020	ATP5I	chr4:657701:657755:-	472	4	610	0.98	316	10	538	0.929	2.20E-03	0	0	0	0	0	0
ENSG00000073060	SCARB1	chr12:123833182:123833310:-	32	10	469	0.46	55	0	410	1	2.40E-03	0	0	0	0	0	0
ENSG00000107164	FUBP3	chr9:132499875:132499946:+	55	1	344	0.953	56	4	651	0.837	2.40E-03	4	0	0	1	0	0
ENSG00000151445	C14orf133	chr14:76993191:76993461:-	10	11	6	0.126	63	7	18	0.589	2.50E-03	0	0	0	0	0	0
ENSG00000117114	KIAA0786	chr1:82191259:82191297:+	0	10	128	0	13	4	212	0.603	2.50E-03	0	0	0	0	0	0

[a]Displayed are exons whose PSI values change by at least 0.05 and are significantly differentially expressed according to the Fisher's Exact Test (FDR<5%)

[b]PSI: 'percent spliced in' (PSI or): the fraction of mRNAs that include a given exon as a proportion of mRNAs that either included or exclude the exon, based on analysis of mRNA-Seq read density

[c]P value: Fisher's Exact Test

Supplementary Table 5. Protein domains[a] overlapping human exons[b] with G-runs (≥6 G's) in the downstream intron that are constitutively spliced in human but alternatively skipped in mouse

Ensembl Exon ID	Gene Name	Domain Name	Domain ID	Domain Definition
ENSE00001276536	SLC2A4	transmembrane_domain		
ENSE00001322387	AIRE	PHD	SM00249	PHD zinc finger
ENSE00000776613	CASR	Pfam:ANF_receptor	PF01094	Receptor family ligand binding region
ENSE00001095037	SLC18A2	transmembrane_domain		
ENSE00000900964	SLC22A2	transmembrane_domain		
ENSE00000901278	NR2C2	ZnF_C4	SM00399	c4 zinc finger in nuclear hormone receptors
ENSE00001062271	NOXO1	SH3	SM00326	Src homology 3 domains
ENSE00000858358	PLXNA1	PSI	SM00423	domain found in Plexins, Semaphorins and Integrins
ENSE00000725328	SLC22A8	transmembrane_domain		
ENSE00001184800	SLCO2A1	transmembrane_domain		
ENSE00001160576	RAD51L3	Pfam:HhH-GPD	PF00730	HhH-GPD superfamily base excision DNA repair protein
ENSE00000860409	JARID1C	Pfam:PLU-1	PF08429	PLU-1-like protein
ENSE00000904839	ARHGEF11	PDZ	SM00228	Domain present in PSD-95, Dlg, and ZO-1/2.
ENSE00000925052	MYST2	Pfam:MOZ_SAS	PF01853	MOZ/SAS family
ENSE00000969365	LNX1	PDZ	SM00228	Domain present in PSD-95, Dlg, and ZO-1/2.
ENSE00000928809	KIAA0319	FN3	SM00060	Fibronectin type 3 domain
ENSE00000772412	IL1R2	IG	SM00409	Immunoglobulin
ENSE00001328854	TRIM37	RING	SM00184	Ring finger
ENSE00001078588	SLC6A20	transmembrane_domain		
ENSE00001147940	CHRNB1	Pfam:Neur_chan_LBD	PF02931	Neurotransmitter-gated ion-channel ligand binding domain
ENSE00001054369	LIMD1	LIM	SM00132	Zinc-binding domain present in Lin-11, Isl-1, Mec-3.
ENSE00001027250	PRG3	CLECT	SM00034	C-type lectin (CTL) or carbohydrate-recognition domain (CRD)
ENSE00000902126	SLC2A9	transmembrane_domain		
ENSE00001144910	IL1F5	Pfam:IL1	PF00340	Interleukin-1 / 18
ENSE00000719866	SLC13A1	transmembrane_domain		
ENSE00000898574	SLC4A9	Pfam:HCO3_cotransp	PF00955	HCO3- transporter family
ENSE00000898574	SLC4A9	transmembrane_domain		
ENSE00001468480	RSC1A1	Pfam:Asp_protease	PF09668	Aspartyl protease
ENSE00001468480	RSC1A1	Pfam:RVP	PF00077	Retroviral aspartyl protease
ENSE00000682715	MAST1	S_TK_X	SM00133	Extension to Ser/Thr-type protein kinases
ENSE00000682715	MAST1	Pfam:Pkinase_C	PF00433	Protein kinase C terminal domain
ENSE00000904964	IGSF9	FN3	SM00060	Fibronectin type 3 domain
ENSE00000653067	PLA2G3	Pfam:Phospholip_A2_2	PF05826	Phospholipase A2

ENSE00001085998	BMPER	Pfam:VWD	PF00094	von Willebrand factor type D domain
ENSE00001078343	SLC4A1AP	DSRM	SM00358	Double-stranded RNA binding motif
ENSE00000925173	MTMR4	PTPc_DSPc	SM00012	Protein tyrosine phosphatase, catalytic domain, undefined specificity
ENSE00001319813	DSCAML1	FN3	SM00060	Fibronectin type 3 domain
ENSE00001227480	SLC6A18	transmembrane_domain		
ENSE00001243192	SMC1A	coiled_coil_region		
ENSE00000730179	GLE1L	coiled_coil_region		
ENSE00001097895	GYLTL1B	Pfam:Glyco_transf_8	PF01501	Glycosyl transferase family 8
ENSE00001317121	CHD6	Pfam:BRK	PF07533	BRK domain
ENSE00000710810	CADM4	Pfam:C2-set_2	PF08205	CD80-like C2-set immunoglobulin domain
ENSE00000710810	CADM4	IG	SM00409	Immunoglobulin
ENSE00001260528	SYCE2	coiled_coil_region		
ENSE00001176633	MYEF2	RRM	SM00360	RNA recognition motif
ENSE00000770211	RRP9	WD40	SM00320	WD40 repeats
ENSE00000947730	RHBDL3	transmembrane_domain		
ENSE00001108386	C21orf63	Pfam:Gal_Lectin	PF02140	Galactose binding lectin domain
ENSE00001056243	FCHO1	Pfam:SAFF	PF10291	SAFF domain
ENSE00000894651	NCAN	IG	SM00409	Immunoglobulin
ENSE00000894651	NCAN	LINK	SM00445	Link (Hyaluronan-binding)
ENSE00000725428	TRIM24	BBOX	SM00336	B-Box-type zinc finger
ENSE00000815783	SLC10A7	transmembrane_domain		
ENSE00001379933	MS4A10	transmembrane_domain		
ENSE00001379933	MS4A10	Pfam:CD20	PF04103	CD20/IgE Fc receptor beta subunit family
ENSE00001310615	DDX51	HELICc	SM00490	helicase superfamily c-terminal domain
ENSE00000349542	TMEM16H	transmembrane_domain		
ENSE00001463691	TXNDC8	Pfam:Thioredoxin	PF00085	Thioredoxin
ENSE00001417683	SPNS3	transmembrane_domain		
ENSE00001302237	MBOAT1	transmembrane_domain		
ENSE00001302237	MBOAT1	Pfam:MBOAT	PF03062	MBOAT family
ENSE00001119065	VPS37C	coiled_coil_region		
ENSE00001090368	SRCRB4D	SR	SM00202	Scavenger receptor Cys-rich
ENSE00001175399	DCST1	transmembrane_domain		
ENSE00001310371	SLC43A2	transmembrane_domain		
ENSE00001310371	SLC43A2	Pfam:MFS_1	PF07690	Major Facilitator Superfamily
ENSE00000960585	ATP8B2	transmembrane_domain		
ENSE00001119112	VWCE	VWC	SM00214	von Willebrand factor (vWF) type C domain
ENSE00001465528	NP_001013765.1	transmembrane_domain		
ENSE00000575433	PPCDC	Pfam:Flavoprotein	PF02441	Flavoprotein

ENSE00001208051	AIM1L	XTALbg	SM00247	Beta/gamma crystallins
ENSE00001175371	NP_116231.2	Pfam:CytochromB561_N	PF09786	Cytochrome B561, N terminal
ENSE00001175371	NP_116231.2	transmembrane_domain		
ENSE00000667398	FAM86A	Pfam:Methyltransf_16	PF10294	Putative methyltransferase
ENSE00000927140	GARNL3	Pfam:Rap_GAP	PF02145	Rap/ran-GAP
ENSE00001171153	MYT1_HUMAN	Pfam:zf-C2HC	PF01530	Zinc finger, C2HC type
ENSE00001363796	RPTOR_HUMAN	WD40	SM00320	WD40 repeats
ENSE00001272781	TRSPAP1	RRM	SM00360	RNA recognition motif
ENSE00001067177	FNDC7	FN3	SM00060	Fibronectin type 3 domain
ENSE00001280805	SLC39A11	Pfam:Zip	PF02535	ZIP Zinc transporter
ENSE00001280805	SLC39A11	transmembrane_domain		
ENSE00001239305	ASAH3L	transmembrane_domain		
ENSE00001476044	ZNF783	KRAB	SM00349	krueppel associated box
ENSE00001365697	NP_001010866.1	transmembrane_domain		
ENSE00001260286	ZNF545	KRAB	SM00349	krueppel associated box
ENSE00000798452	LRRC40	LRR	SM00370	Leucine-rich repeats, outliers
ENSE00000798452	LRRC40	LRR_TYP	SM00369	Leucine-rich repeats, typical (most populated) subfamily
ENSE00001302858	SLC35F3	transmembrane_domain		
ENSE00001035708	SLC5A6	transmembrane_domain		
ENSE00000655119	C20orf54	transmembrane_domain		
ENSE00000655119	C20orf54	Pfam:DUF1011	PF06237	Protein of unknown function (DUF1011)
ENSE00000807601	ZCCHC9	Pfam:zf-CCHC	PF00098	Zinc knuckle
ENSE00000807601	ZCCHC9	ZnF_C2HC	SM00343	zinc finger
ENSE00001260327	ZNF545	KRAB	SM00349	krueppel associated box
ENSE00001260327	ZNF545	transmembrane_domain		
ENSE00001450929	FAM40A	Pfam:N1221	PF07923	N1221-like protein
ENSE00000829291	NP_112203.1	RRM	SM00360	RNA recognition motif
ENSE00001230281	NP_612637.1	transmembrane_domain		
ENSE00000683831	ARHGEF15	RhoGEF	SM00325	Guanine nucleotide exchange factor for Rho/Rac/Cdc42-like GTPases
ENSE00001100288	PAQR5	Pfam:HlyIII	PF03006	Haemolysin-III related
ENSE00001100288	PAQR5	transmembrane_domain		
ENSE00001113510	KLHL26	BTB	SM00225	Broad-Complex, Tramtrack and Bric a brac

[a]Protein domains as annotated in SMART or Pfam

[b]Exons that do not overlap protein domains annotated in SMART of Pfam are not listed.

Supplementary Table 6. GO categories enriched in group 1: genes with 9 or more G's in G-runs within +11 to +70 nt from 5'ss in at least 50% of introns, or group 2: genes containing mRNA-SEQ and array-predicted candidate hnRNP H-activated exons

GO id	P values[a] in group 1	P values[b] in group 2	GO description	GO id	P values[a] in group 1	P values[b] in group 2	GO description
0004252	**1.90E-07**	NS[c]	**serine-type endopeptidase activity**	0006816	5.18E-04	NS	calcium ion transport
0005576	**3.44E-07**	NS	**extracellular region**	**0032502**	5.36E-04	**4.00E-04**	**developmental process**
0032501	**3.98E-07**	**1.00E-04**	**multicellular organismal process**	0006066	5.84E-04	NS	alcohol metabolic process
0005509	**5.11E-07**	NS	**calcium ion binding**	0005891	6.01E-04	NS	voltage-gated calcium channel
0017171	**1.05E-06**	NS	**serine hydrolase activity**	0005272	6.08E-04	NS	sodium channel activity
0008236	**1.05E-06**	NS	**serine-type peptidase activity**	0005887	6.78E-04	NS	integral to plasma membrane
0005886	**3.19E-06**	**1.00E-04**	**plasma membrane**	0004175	7.76E-04	NS	endopeptidase activity
0048856	**8.72E-06**	NS	**anatomical structure development**	**0044425**	8.96E-04	**1.00E-04**	**membrane part**
0007275	**9.61E-06**	NS	**multicellular organismal development**	0005625	9.23E-04	NS	soluble fraction
0016020	**1.05E-05**	**1.00E-04**	**membrane**	0003924	9.28E-04	NS	GTPase activity
0044459	**1.41E-05**	**1.00E-04**	**plasma membrane part**	0007267	9.66E-04	NS	cell-cell signaling
0051179	**1.62E-05**	NS	**localization**	0009581	9.71E-04	NS	detection of external stimulus
0048731	**1.71E-05**	8.00E-04	**system development**	**0008092**	NS	**1.00E-04**	**cytoskeletal protein binding**
0007154	**3.34E-05**	**1.00E-04**	**cell communication**	**0060089**	NS	**1.00E-04**	**molecular transducer activity**
0007264	**3.63E-05**	NS	**small GTPase mediated signal transduction**	**0007166**	NS	**1.00E-04**	**cell surface receptor linked signal**
0005179	**8.82E-05**	NS	**hormone activity**	**0009966**	NS	**1.00E-04**	**regulation of signal transduction**
0006810	**9.06E-05**	NS	**transport**	**0046872**	NS	**1.00E-04**	**metal ion binding**
0007265	**9.57E-05**	NS	**Ras protein signal transduction**	**0005488**	NS	**1.00E-04**	**binding**
0005102	**1.28E-04**	NS	**receptor binding**	**0004871**	NS	**1.00E-04**	**signal transducer activity**
0005215	**1.62E-04**	NS	**transporter activity**	**0043169**	NS	**1.00E-04**	**cation binding**
0051234	**1.66E-04**	NS	**establishment of localization**	**0043167**	NS	**1.00E-04**	**ion binding**
0031224	**2.00E-04**	**1.00E-04**	**intrinsic to membrane**	**0005509**	NS	**1.00E-04**	**calcium ion binding**
0030182	2.68E-04	NS	neuron differentiation	**0004872**	NS	**2.00E-04**	**receptor activity**
0016021	2.70E-04	**1.00E-04**	**integral to membrane**	**0046914**	NS	**2.00E-04**	**transition metal ion binding**
0007242	3.06E-04	**1.00E-04**	**intracellular signaling cascade**	**0006629**	NS	**2.00E-04**	**lipid metabolic process**
0007399	3.33E-04	NS	nervous system development	**0030695**	NS	**2.00E-04**	**GTPase regulator activity**
0007165	3.61E-04	**1.00E-04**	**signal transduction**	**0008270**	NS	**2.00E-04**	**zinc ion binding**
0048699	4.19E-04	NS	generation of neurons	**0030234**	NS	**4.00E-04**	**enzyme regulator activity**
0015674	4.23E-04	NS	di-, tri-valent inorganic cation transport	**0016491**	NS	**4.00E-04**	**oxidoreductase activity**
0031226	4.75E-04	NS	intrinsic to plasma membrane	0005856	NS	8.00E-04	cytoskeleton
0022008	5.10E-04	NS	neurogenesis				

[a]P values below Bonferroni cutoff of 2.1E-04 are shown in bold.

[b]P-values below Bonferroni cutoff of 5.2E-04 are shown in bold.

[c]NS: not significant

Supplementary Table 7. Motifs with significantly different conservation levels in the downstream intronic regions of ACEs and ICEs with different 5'ss strength

Exons:	ACEs			ACEs			ICEs						ICEs					
intron L	70			200			70						200					
5'ss (bits)	[0,4]	[4,8]	[8,12]	[0,4]	[4,8]	[8,12]	[0,2]	[2,4]	[4,6]	[6,8]	[8,10]	[10,12]	[0,2]	[2,4]	[4,6]	[6,8]	[8,10]	[10,12]
AAAU										*								
AACA								*										
AAGC									*						*			
AAUC										*				*				
ACAG								*										
ACCG		*																
ACGC								*										
AGAA							+			*								
AGAG															*			
AGCA										*	+							
AGCC							*				+	+						
AGCU									*	*						*		
AUCG		*	*															
AUGC					*													
AUUC	*												*					
CAAA							*											
CACA	*																	
CACG		*						*										
CAGC							*											
CAGG	*																	
CCCC									*									
CCGA								*										
CCGU	*																	
CCUA							*											
CCUG												*						
CGCU		*																
CGGG				*														
CUAA	*	+	+									*				*		*
GAAC																		*
GAAG				*														
GACA				*														
GACC									*		+	+			*	*	*	*
GACG													*					
GAGG							*											
GAUC					*													
GCAU				+	*	+												
GCCA											*							
GCCC											*	*						
GCGA												*						
GCUG					*	+												
GGCC												*						
GGGC							+	+	+	+	*	+						
GGGG								+	+	*	+	+						
GUAG								*										
GUCA							*		*	*	+							
GUCG									*									
GUUA																*		+
UAAA											*	*			*			
UAAC								*			+							
UAGA								*	+	*		+						
UAGC								*										
UAGG				*														
UCAA							*											
UCCG				*														
UGAA										*								
UGAC									*		+	+			*	+	+	+
UGGG								+	+	*	+	+						

UGUC								*	*									
UUGC													*					
UUUU							*											
AAACA												*						
AAACC								*		*								
AAAGC									*									
AAAUA									*			+						
AACCC		*																
ACAAC			*															
ACACA															+	*	*	+
ACACU							*											
ACAUA				*														
ACCAC									*									
ACCCA											*							
ACCGU														*				
ACGAC									*									
ACGCU		*								*								
ACGGA					*													
ACUGU				*														
AGAGA									*									
AGAGC					*													
AGCGC											*						*	
AGCUU	*									*								
AGUAA															*		+	
AGUAG					*													
AUCAG							*											
AUCCG				*								*						
AUCGU		*																
AUGUU												*						
AUUCC	*			*														
CACAC																*	*	
CACAU													*					
CACCG		*																
CACGC								*						*				
CAGCC							+			*	+							
CAGGA										*								
CAUCG		*																
CAUGU						*			*				*					
CCAGA									*	*								
CCAGC														*				
CCCAG											*							
CCCCG									*									
CCGAA												*						
CCGAG								*						*				
CCGGC			*	*														
CCUUU	*																	
CGAGC															*			
CGAGG														*				
CGCAU												*						
CGCCU								*						*				
CGCUA															*			
CGCUG												*						
CGCUU					*	+												
CGGAG									*									
CGGUU																*		
CGUCU			*															
CUAUU				*														
CUGAC										*						*	*	*
CUGCU			*						*						*	+		+
CUGUG												*						
CUUUC														*				
GAAGA				*														
GAAUG																*		

GAAUU																	*	
GACAC				*														
GAGCA		*																
GAGUU							*											
GCACG		*																
GCAGA								*										
GCCCA		*																
GCGAC												*						
GCGCU												*						
GCUGG			*															
GCUUC																*		
GCUUU					*													
GGACC				*														
GGCAG						*												
GGCCC														*	+		+	+
GGCGA												*						
GGUAA				*		+												
GGUGU								*										
GGUUU				*														
GUAGA															*			
GUCAC															*	+		+
GUCAG													+	+		*		+
GUGAA	*																	
GUGAC								*	+		+							
GUGGU									+	*								
GUGUA								*										
GUGUG		*						+	+	*	+	+						
GUUCG				*														
UAACC		*			*	+												
UAAGC										*								
UAAUC					*													
UAAUG												*						
UAGAA								+		*								
UAGAC		*																
UAGAG							*											
UAGGC													*					
UAUCU													*					
UAUUU	*																	
UCACA															*	+		
UCCCA															*			
UCGUU									*									
UCUGC													+		+	*		+
UGAAA	*																	
UGACC					*				*		+				*	+	+	+
UGACG															*			
UGAGU											*							
UGCAU	+	*	+															
UGCCA							+				*							
UGCCU		*																
UGCUC				*	*													
UGGUU									*									
UGUCG								*										
UUAAU													*					
UUAGA																	*	
UUCAC								*						*				
UUCAU												*						
UUCCG	*			*														
UUGCC		*																
UUUCA								*										
UUUUC	*																	

*: more conserved in the corresponding 5'ss group than at least one other group
+: significantly conserved in the corresponding 5'ss group

SUPPLEMENTARY METHODS

Real Time RT-PCR of hnRNP H mRNA

The amount of hnRNP H mRNA was measured in all samples relative to the housekeeping gene glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) by real-time RT-PCR using the primers:

hnRNP H forward: 5' GAGGGATTCGTGGTGAAGG 3'

hnRNP H reverse: 5' TCCACCGGCAATGTTATCC 3'

GAPDH forward: 5' TCGTGGAGTCCACTGGCGTCTT 3'

GAPDH reverse: 5' TGGCAGTGATGGCATGGACTG 3'

Western analysis

Total cell protein extract was analyzed by PAGE, transferred to a PVDF membrane and probed with primary antibodies to hnRNP H (polyclonal from rabbit, Bethyl Laboratories) and GAPDH (monoclonal from mouse, Abcam). The secondary antibodies were linked to alkaline phosphatase and visualization was done using ECF substrate (Amersham Biosciences).

qRT-PCR validation of hnRNP H-responsive exons

To validate the splicing changes following H-KD, 22 exons were selected from the set of exons with a change in exon inclusion level ($\Delta\Psi$) of 20% or more (N=136) in the mRNA-SEQ data or with a relative reduction in expression level (LFC_{gene}-LFC_{exon}) of 0.5 or more (N=205) in the array data (for computational analysis, see below). Test exons were chosen so that there was no significant difference in the mean estimated exon expression change between the tested exons and the complete set of exons that passed the criteria. The test exons selected (Supplementary Table 2) were among those flanked by constitutive exons according to cDNA and EST data or whose

flanking exons had small (($\Delta\Psi$< 5% or |LFC_{gene}-LFC_{exon}| < 0.1) relative changes in exon expression level based on the mRNA-SEQ or array experiments. As negative controls, four exons were selected from the "non-target" set that had a small (($\Delta\Psi$ < 5% or | LFC_{gene}-LFC_{exon}| < 0.1) relative change in exon expression level. Primers were designed to target the flanking exons to amplify both exon inclusion and exon skipping isoforms (Supplementary Table 2).

Total RNAs used for the second batch of array experiments (H-KD arrays 2 and 3, control arrays 2 and 3) were used to perform the RT-PCR experiments. The reverse transcription (RT) reaction was carried out using SuperScript III (Invitrogen), followed by radioactive PCR (25 or 30 cycles of amplification depending on the gene expression level of the tested exon, with trace amounts of α-^{32}P-dCTP in addition to non-radioactive dNTPs). All RT-PCR reactions were run in duplicate. Quantitation of spliced isoforms was conducted as described previously[1].

Supplementary Fig. 8a shows the comparison of the Ψ values based on RT-PCR and those estimated from the mRNA-SEQ. We defined positively validated H-responsive exons as those that 1) had at least one isoform with the expected size visible in each experiment; 2) showed increased relative abundance of exon skipping isoform in both biological H-KD experiments compared to the control experiments and 3) the mean relative abundance of the exon skipping isoform of the two H-KD experiments was at least 10% higher compared to that of the two controls. Among the 22 tested candidate H-responsive exons, 15 were validated by the RT-PCR experiments. Supplementary Table 2 shows the list of the 22 exons and their associated genes with the 15 validated ones noted. No significant change in exon inclusion was detected for the four negative control exons.

Genomic Sequence Analysis

Genomic sequences of human (genome release hg18) and mouse (mm8) and alignments of human and mouse cDNA and EST sequences to the human genome and the mouse genome were obtained from the University of California Santa Cruz (UCSC) Genome browser (http://genome.ucsc.edu)[2]. Human and mouse skipped exons (SEs) and constitutive exons (CEs) were identified independently by using cDNA or EST transcript data specific to each organism. An exon present in at least one transcript was designated as an SE if it was excluded in one or more other transcripts. Exons with evidence of both skipping and alternative splice site usage were excluded. Exons that were not associated with evidence of skipping, alternative splice site usage, or intron retention were identified as CEs. Only exons flanked by canonical AG/, /GT terminal dinucleotides were included. Exons flanked by putative U12-type introns[3] were excluded. A minimum exon length of 30 bases and a maximum exon length of 300 bases were required. In addition, the flanking introns were required to be at least 50 bases long. The CEs and SEs were further filtered for internal exons based on human and mouse Ensembl gene annotations (release 40, August 2006) [4]. Orthologous exons were defined as those that are present in orthologous gene pairs obtained from Ensembl and whose splice sites were aligned according to the UCSC whole-genome alignment [5]. Alternative-conserved exons (ACEs) were defined as orthologous SEs with transcripts supporting exon skipping and inclusion in both human and mouse. Included-conserved exons (ICEs) were defined as orthologous CEs with transcripts supporting exon inclusion only in both human and mouse data. Overall, 15,256 human SEs, 13,162 mouse SEs, 91,045 human CEs, 90,449 mouse CEs, 2964 ACEs and 44,368 ICEs were identified.

Human SNP data were downloaded from dbSNP Build 128 (ftp://ftp.ncbi.nih.gov/snp/). As a quality filter, SNPs that mapped to multiple genomic regions or known repetitive elements were excluded. The remaining SNPs were mapped to the 5'ss region (-3, -2, …, +6 positions

relative to the exon-intron boundary) of the 91,045 human CEs. Overall, 1482 human exons were identified to have 5'ss SNPs.

Conservation of G triplets

The conservation of G triplets was analyzed in the downstream intronic region excluding the 5'ss (i.e. bases 11 to 70) of ICEs (Fig. 1a). The cutoff of 70 was used since the density of G-triplets, which peaks at about 20-30 nt downstream of the exon boundary (Supplementary Fig. 1a), is significantly reduced beyond this position compared to the more proximal regions to the 5'ss [6]. If the intron length was less than 140, only the region from base 11 up to half of the intron length was considered to avoid including sequences in the vicinity of the branch point and 3'ss. Scores of the 5'ss of each intron were calculated by MaxEnt log-odds scoring[7]. The introns were stratified according to their 5'ss scores in the human orthologs into 5 bins with the ranges of (0-4, 4-6, 6-8, 8-10, 10-12) bits. The numbers of introns in each bin (from lowest to highest 5'ss scores) were, respectively, 1106, 3153, 9295, 18532, 9961. (The small numbers of exons with 5'ss scores below 0 or above 12 bits were excluded.) Within each bin, the conservation rate (CR) of G-triplets was evaluated as follows. First, a conserved G-triplet occurrence was defined as an instance of the motif in the human genome that was aligned to the mouse genome and perfectly conserved in sequence in mouse in the UCSC whole-genome alignments. The CR of G-triplets was then calculated as the number of conserved occurrences divided by the total number of occurrences of the motif in the human intron regions considered. Next, an expected CR of G-triplets was calculated using a first-order Markov model. The expected CRs of all other trinucleotides were calculated similarly. Subsequently, two control 3-mers were chosen according to the following criteria: 1) same G+C frequency (100%) as in G-triplets; 2) same number (0) of CG dinucleotide as in G-triplets; 3) the least different expected CRs compared to that of G-triplets in each 5'ss bin;

4) absent from the RESCUE-ISE motifs (GGG and CCC) as previously identified[6]. The control CR for G-triplets was then calculated as the average CR of the two control 3mers. The ratio between the actual CR and the control CR for G-triplets (*i.e.*, observed/control) was then computed (Fig. 1a). The standard deviation of the CR ratio of G-triplets was calculated based on a binomial distribution with parameters N, the number of G triplets considered (in human), and p, the CR of G-triplets.

To evaluate the significance of G-triplet conservation in specific 5'ss bins, a bootstrapping approach was used to construct background sets of introns. Specifically, introns were randomly picked from each bin to reach a total of 500 occurrences of G-triplets in the human orthologs. Introns picked from all 5 bins were then pooled together to form one background set. The CR of G-triplet in the background set was calculated as described above. The bootstrapping procedure was repeated 1000 times. A P-value for high conservation of G-triplets in each bin was calculated as the fraction of times the actual CR ratio in the bin was lower than or equal to the background CR ratio and a P-value for low conservation of G-triplets was calculated as the fraction of times the background CR ratio was lower than or equal to the actual CR ratio. The P-values were then Bonferroni-corrected for multiple comparisons (i.e., multiplied by 10).

mRNA-Seq data analysis

mRNA-Seq reads (32 bases long) were mapped to the human genome (hg18) and spliced exon-exon junctions annotated in Ensembl genes (release 40, August 2006)[4]. The mapping was carried out using the ELAND (Efficient Alignment of Nucleotide Databases) module within the Illumina Genome Analyzer Pipeline software. Up to two mismatches were allowed in all mappings. Reads mapped to the splice junctions were required to match to at least 4 bases on each side of the junction, which we found is sufficient to ensure unique mapping in the vast majority of cases. To

be retained for downstream analyses, reads were required to map uniquely to the union of the genome and the transcriptome (i.e. including splice junctions). We obtained 22,594,112 and 30,385,240 reads in the control and H-KD experiments, respectively, of which 67% (15,092,536) and 65% (19,849,170) mapped uniquely to the genome, 5% (1,191,074 and 1,539,861) mapped uniquely to the splice junctions, and 21% (4,746,766) and 20% (6,086,971) mapped non-uniquely to the union of the genome and the transcriptome.

The genomic reads were then mapped to Ensembl exons (termed exon body reads). Exon body reads and exon junction reads were filtered to remove those that were mapped to overlapping sense and antisense Ensembl exons. Among the 273,858 Ensembl exons, about 65% (69%) were found to have exon body reads in the control (H-KD) data. Approximately 52% and 55% of all exons had exon junction reads in the control and H-KD data respectively.

In the study of the function of hnRNP H, we focused on exons that were either constitutive or alternatively skipped in the control or H-KD experiment. For each exon, we estimated the exon inclusion level (i.e. Ψ) based on the number of mRNA-Seq reads. The value of Ψ was defined as (inclusion density / inclusion density + exclusion density), where inclusion density was computed as the read density of the exon and both exon inclusion junctions, and exclusion density was computed as the read density of the exclusion junction. Read density of each region was defined as the ratio of the number of reads mapped to the region divided by the total number of unique positions in the region, i.e. positions in the region that can generate 32mers which are unique in the genome. The rationale for this is that only those positions that are unique can have reads mapping uniquely to them. Following Laplace's Rule of Succession, we added a single pseudocount to the total number of reads, which was divided proportionally to the number of unique positions corresponding to exclusion and inclusion, respectively, e.g., the fraction $N_{\text{exclusion}}/(N_{\text{exclusion}}+N_{\text{inclusion}})$ was added to the count of exclusion reads, where $N_{exclusion}$ and $N_{inclusion}$ are

the numbers of unique positions corresponding to exclusion and inclusion. For the functional studies of hnRNP H (Fig. 2) described in the following sections, we focused on "skippable" exons by requiring the exons to have at least one junction read supporting exon exclusion in at least one experiment.

Expression level of each gene/transcript (Supplementary Fig. 7a) was evaluated by calculating the gene read density, i.e. the total read counts for each Ensembl transcript normalized by the total number of unique positions in the transcript. In the comparison of gene expression levels of the control and H-KD experiments, the gene read densities were further scaled using the total number of reads obtained in each experiment.

Function of G-runs as ISEs

In this section, we focus on the splicing activator function of hnRNP H, which has been associated with binding to G-runs in the intron downstream of the regulated exon[8]. One prediction of this function, as supported by the splicing reporter assays (Fig. 1c, Supplementary Figs. 3 and 4), is that splicing of an exon should be more strongly enhanced in the presence of longer (or more) G-runs in its downstream intron. Thus, in the mRNA-Seq data, exons with longer (or more) intronic G-runs are expected to have larger reductions in their inclusion levels following knockdown of hnRNP H. To investigate this aspect in our data, subsets of exons were extracted with different numbers of G's in runs G_3 or longer (G-runs) within bases +11 to +70 downstream of the 5'ss. Since exonic G-runs can function as hnRNP H-dependent silencers of splicing (see below), exons with more than 3 internal Gs in G-runs were excluded. Exons with more than 3 G's in G-runs in the upstream intronic regions (up to 150 nt from the 3'ss) were also excluded. The change in Ψ (control – H-KD) was then calculated for exons in the subsets. Figure 2a shows the average and standard error of this value for exons with 3-5, 6-8, 9-11, 12-14 and 15-17 Gs in G-

runs. As controls, exons with no intronic (within 70 bases from splice sites) or exonic G-runs were also analyzed (such exons are termed "non-targets"). Larger changes in Ψ were observed for exons with flanking G-runs conserved between human and mouse than for exons flanked by comparably sized G-runs not conserved in mouse (data not shown), suggesting that conserved G-runs are more often functional as ISEs.

To study the ISE activity of G-runs as a function of 5'ss strength, we focused on the set of exons (558 in total) with at least 6 G's in G-runs in the downstream intron (+11 to +70 bases) and no more than 3 exonic or upstream intronic G's in G-runs. We refer to these exons as "candidate H-activated exons". Among these exons, 136 had a ≥ 20% reduction in Ψ after H-KD and 28 had a ≥ 20% increase in Ψ, indicating a strong bias ($p < 1e-10$, Fisher's exact test) towards H-activated exons in this set, as expected.

Using the 558 candidate H-activated exons, the function of hnRNP H relative to 5'ss strength was investigated. Similar to the motif conservation analysis for G-triplets, the exons were stratified into 5 bins with the 5'ss score in the ranges of (0-4, 4-6, 6-8, 8-10, 10-12) bits. The change in Ψ following H-KD (control – H-KD) was calculated for exons in each bin. The corresponding change in Ψ was also calculated for the "non-target" exons in each bin. Figure 2b shows the exon inclusion change (mean and SE) of the H-activated exons relative to that of the non-target exons (observed – expected, O-E) in each bin.

To evaluate whether certain bins have a statistically higher or lower change in exon inclusion than other bins, a sampling procedure was carried out to control for 5'ss strength and enrichment of G's in G-runs. For example, if there were *N* exons in bin *X*, then *N*/4 exons were randomly picked from each of the other 4 bins to form a control set of exons. To control for effects of G-run length, the random sampling algorithm was designed so that the *N* control exons

and the *N* bin-*X* exons were matched in the total number of downstream intronic Gs in G-runs. In this analysis, exons in bin *X* whose number of G's did not match at least 4 exons in the other bins were excluded. The sampling procedure was repeated 1000 times and the average relative exon inclusion levels of the bin-*X* exons were compared with that of the control exons to obtain a P value. Exons in each bin were analyzed similarly. The P-values were Bonferroni corrected for multiple comparisons (i.e. multiplied by 10).

No consistent dependence of G-run ISE activity on 3'ss strength was seen based on the mRNA-Seq data (Supplementary Fig. 4d) or the reporter data (Supplementary Fig. 4a), and no correlation was observed between the 5'ss and 3'ss strength of constitutive exons (R = -0.029, Supplementary Fig. 4c). This absence of correlation indicates that the activity of downstream intronic G-run is not expected to be dependent on the 3'ss strength as a result of its dependence on the 5'ss strength.

The function of G-run ISEs was also largely unaffected by initial inclusion levels of exons. In the splicing reporter analyses, the 5'ss strength dependence of G-run ISE activity was robust to changes in "initial" exon inclusion level – i.e. prior to G-run insertion – except when the initial level was so high (e.g., ≥ 95%) as to preclude additional enhancement (Supplementary Fig. 3a). For example, pairing of an intermediate 5'ss with 2.1 bit and 4.1 bit 3'ss yielded initial inclusion levels of 10% and 54%, respectively. However, the mean change in exon inclusion per inserted G triplet for these two reporters was not significantly different, at 17% and 20%, respectively, indicating additive contribution to exon inclusion independent of initial inclusion level (Supplementary Fig. 3a). Based on the mRNA-Seq data, the initial inclusion level also did not significantly influence inclusion level changes following H-KD (Supplementary Fig. 3b).

Function of G-runs as ESSs

In this section, we focus on the splicing silencer function of hnRNP H associated with binding to G-runs in the exons[9]. To investigate the ESS function relative to G-run lengths, we extracted subsets of exons with different minimum number of G's in G-runs in the exons and no more than 3 G's in G-runs in the flanking intronic regions defined above. The change in Ψ (control – H-KD) was then calculated for exons in the subsets. Figure 2d shows the average and standard error of this value for exons with 3-5, 6-8, 9-11 and 12-14 Gs in G-runs.

To study the function of G-runs as ESSs, we focused on the set of exons (167 in total) with at least 6 G's in G-runs in exon. We refer to these exons as "candidate H-repressed exons". Among these exons, 44 had a ≥ 20% increase in Ψ after H-KD and 21 had a ≥ 20% reduction in Ψ. Thus, there was a bias ($p = 0.002$, Fisher's Exact test) towards derepression following H-KD, as expected. These 167 exons were grouped by their 5'ss scores and their relative changes in exon inclusion level (Fig. 2d) were evaluated similarly as for the candidate H-activated exons.

Exon array data analysis

The raw intensity data (CEL files) of exon arrays obtained from Affymetrix GeneArray Scanner were processed using the Affymetrix Power Tools (APT, see http://www.affymetrix.com/support/developer/powertools/index.affx). The apt-cel-transformer program was used to carry out RMA background correction and quantile normalization of the probe intensities. The exon array probes were mapped to Ensembl genes and exons using the custom CDF files[10] (version 9). The R package "affy"[11] (version 1.14.1) was then used to obtain probe intensity values from the processed CEL files. The gene-level expression values were calculated by summarizing the RMA and quantile-processed probe intensities of core probes

mapped to each gene in the custom CDF file using the median polish method[12] implemented in "affy".

Since two slightly different RNAi protocols were used, the array data were analyzed in a pairwise manner by pairing H-KD and control arrays completed with the same protocols. To detect exons differentially expressed in the control arrays and H-KD arrays, the probe intensity values of each gene were processed as follows. First, probes were discarded if associated with probe or gene log intensity values of less than 6 on the control arrays. This cutoff was chosen based on the poor correlation of probes with log intensities below 6 between the control arrays (data not shown). About 80% antigenomic and genomic background probes (negative controls) had log intensities below 6. Exons with 2 or more probes passing this filter were retained for further analysis. Next, the difference of probe intensity values ($\Delta_{probe} = \log_2(\text{H-KD})-\log_2(\text{control})$, i.e., log-fold change or LFC) between the H-KD and control arrays was calculated for each probe and each control/H-KD array pair. The median Δ_{probe} value of all probes mapped to an exon in the 3 pairs of array experiments was used to calculate log-fold change in exon expression (LFC_{exon}). Similarly, the median of the differences of gene intensity values between H-KD and control arrays was used to calculate LFC_{gene}. Thus, exons with a differential expression in the H-KD and control experiments could be identified as those with a large difference between LFC_{exon} and LFC_{gene} values (see below).

Similar to the analysis of the mRNA-Seq data, the exons were grouped according to the number of G's in G-runs in the downstream introns. The relative change in exon expression levels LFC_{gene}-LFC_{exon} obtained from the array experiments were then calculated for the subsets of exons and the non-target exons. The results are shown in Supplementary Fig. 7 together with the mRNA-Seq results.

To analyze the function of intronic G-runs relative to 5'ss strength with the array data, we used the set of exons (4200 in total) with at least 9 G's in G-runs. (G_9 was required instead of G_6 as for the mRNA-Seq data since G_6 showed only slightly functional increase than G_3 in the array data in Supplementary Fig. 7.) In this dataset, 315 exons had a relative reduction in expression level (LFC_{gene}-LFC_{exon}) of 0.5 on a $\log_2$ scale (i.e., ~30%, illustrated by the lower red dotted line) or more. In contrast, 117 exons were detected with an increase in relative expression level of at least 0.5 ($P < 1e-10$). These exons were analyzed similarly in different 5'ss bins as for the mRNA-Seq data (Supplementary Fig. 7).

The function of exonic G-runs was not analyzed using the array data due to smaller data sets and reduced data quality (e.g., a "dose-response" in the function of splicing silencing was not observed for exonic G-runs based on the array data).

Genetic variation in splice sites

In Fig. 3b, the fraction of human CEs with SNPs in the 5'ss (last 3 bases of exon and first 6 bases of intron) was analyzed. The data were grouped into subsets (test exons) associated with different numbers of G-runs (within 70 bases downstream from the 5'ss). As controls, exons with no G-runs were randomly sampled to match the following features: (i) number of CpGs in the 5'ss sequences; (ii) C+G content of 5'ss sequences (differing by ≤ 1 out of 9 bases); (iii) 5'ss score (differing by ≤ 2 bits); (iv) intronic C+G frequency (differing by ≤ 20%). These quite stringent controls were imposed to ensure that such background features do not bias SNP density. Fig. 3b shows the ratio between the fraction of exons with 5'ss SNPs in test exons and that in control exons. Error bars were calculated by assuming a binomial distribution with parameters *N*, the number of exons analyzed and *p*, the fraction of exons with 5'ss SNPs in the test exons. In-frame and out-of-frame exons were analyzed separately for those with at least 6 G's in G-runs.

In Fig. 3c, the change in 5'ss score was analyzed for subsets of ICEs (test exons) with different numbers of G-runs (between bases +11 to +70 downstream from the 5'ss). In each subset, the number of G-runs was required to be conserved between the human and mouse orthologs. As controls, exons with no G-runs were used. Control exons were randomly sampled to match the following sequence features of the test exons (average values of all features in human and mouse orthologs were used): (i) number of CpG in the 5'ss sequences; (ii) intronic C+G frequency (differing by ≤0.1); (iii) MaxEnt 5'ss score (differing by ≤ 2 bits); (iv) intronic sequence conservation (PhastCons scores averaged for the 11-70 base positions of an intron, differing by ≤ 0.1); (v) 5'ss C+G frequency (within the same range, either ≥0.35 or <0.35). These stringent controls we used to ensure that such background features do not bias the difference in 5'ss scores between human and mouse orthologs. Fig. 3c shows the ratio between the mean 5'ss score difference in test exons and that in control exons. Error bars represent the standard errors of the change in 5'ss scores associated with the set of the test exons.

Evolution of splicing phenotype of exons with G-run ISEs

Orthologous human/mouse exons were analyzed for the relationship between presence of conserved G-runs and evolutionary change in splicing pattern (constitutive splicing in human and alternative splicing in mouse). To ensure reliable identification of constitutive splicing, the set of candidate human CEs were filtered further based on Illumina high-throughput transcriptome sequencing data from 10 diverse human tissues[13], and exons flanked by (or overlapping with) alternative splice junctions in any tissue in the Solexa data were excluded. Among the original 91,045 candidate human CEs, 71,235 exons passed this filtering step. These human CEs were aligned to mouse exons based on the UCSC whole-genome alignment. Overall, 36,240 exons were aligned to mouse CEs or SEs that have EST or cDNA support. In this data set, 1,796 exons were

constitutively spliced in human but alternatively skipped in mouse. Exons categorized as CEs in mouse but as SEs in human were not analyzed because the somewhat lower number of EST sequences available from mouse (and the absence of Solexa transcript data) would make it difficult to reliably identify a set of mouse CEs not contaminated by unrecognized SEs.

Exon boundaries of the human/mouse orthologous exons (36,240 in total) were aligned to the genomes of 16 other placental mammals included in the UCSC 28-way whole genome alignment[5]. The presence of G-runs in the downstream introns (within 70 nt from the 5'ss) of aligned exons was calculated in each genome. If the G-run was present (not necessarily in aligned positions) in at least 10 of the 16 genomes, then it was designated as 'present' in the last common ancestor (LCA) of the organisms considered. On the other hand, if the G-run was absent from at least 10 of the 16 genomes, then it was considered as 'absent' from the LCA. In total, 15,609 (8957) test exons were identified as having at least 3 (6) G's in G-runs, and 22,205 control exons were identified.

To study the evolution of splicing phenotypes related to G-run ISEs, we asked the question of whether exons with ancestral G-runs in downstream introns were more likely to have different splicing phenotypes in human and mouse (i.e., constitutive in human, alternative in mouse) compared to those lacking ancestral G-runs. In Fig. 3d, the fraction of exons that were constitutively spliced in human but alternatively spliced in mouse among all exons with ancestral G-runs was analyzed for subsets of data (test exons) with different numbers of ancestral G-runs, compared to the set of exons with no ancestral G-runs (control exons). To account for potential biases in EST/cDNA coverage of mouse genes, the control exons were sampled randomly to match the transcript coverage of the test and control genes by requiring the number of transcripts per kbp of mRNAs to differ by less than 0.2. In addition, the C+G frequency and conservation level of the introns (70 nt downstream from the 5'ss, averaged among all genomes for each intron)

were controlled for. The C+G frequency was required to differ by less than 5%. Since the conservation of G-run ISEs was determined in this instance based on presence of the G_3 motif in the 70 nt intronic region without requiring alignments, the overall conservation level of introns was calculated analogously, by averaging the 'conserved occurrence rate' (COR) of each 3-mer[1]. The control data and test data were matched by requiring their average COR values to differ by less than 0.1. Figure 3e shows the ratio between the fraction of test exons that are constitutively spliced in human but alternatively skipped in mouse and this fraction in control exons in subsets with different number of ancestral G-runs. Error bars were calculated by assuming a binomial distribution with parameters *N*, the number of test exons, and *p* the fraction of test exons that were CEs in human and alternatively spliced in mouse. In-frame and out-of-frame exons were analyzed separately for those with at least 6 G's in G-runs.

Exons with internal G-runs were analyzed analogously (Fig. 3e) by controlling for the EST/cDNA coverage, exonic G+C frequency and conservation level.

Gene ontology analysis

Genes containing exons detected as downregulated following H-KD (404 exons with at least 9 G's in G-run in the downstream intron and with either a ≥20% reduction in exon inclusion in the mRNA-Seq or a LFC of 0.5 or more in the array data) were analyzed for enrichment of GO categories. Genes containing the "non-target" exons (N = 4184) defined previously were used as controls. A total of 350 test genes and 1099 control genes were identified, excluding genes that belonged to both groups. GO identifiers (IDs) for each gene were obtained from EnsMart (Ensembl release 40, August 2006, http://www.ensembl.org). Function and process ontologies for all genes were obtained from http://www.geneontology.org. In addition to the GO IDs annotated by Ensembl, all general "parent" GO categories within the "Molecular Function" and "Biological

Process" ontologies were assigned to each gene. Since conservation of G-triplets was used in defining the set of candidate H-activated exons, we controlled for gene conservation level in the test genes and control genes. The candidate H-activated exons (genes) have, essentially by definition, high C+G content. Thus, the C+G frequency of the test and control genes was not controlled. However, the number of exons in each gene was controlled considering the higher probability to identify an exon as a H-activated candidate in genes with more exons. Specifically, for each GO category, the fraction of genes in the test set (F_{test}) associated with this category was calculated. Then, a total of 350 genes were selected randomly from the control set to achieve a one-to-one match with the test genes in terms of average phastCons conservation score [14] and number of exons (allowing phastCons scores to differ between the test and control gene pairs by no more than 5% of the range across all genes, and requiring exact match of number of exons). Then, the fraction of genes in this randomly selected control set ($F_{control}$) associated with the current GO category was calculated. This random sampling process was repeated 10,000 times. The P value for enrichment of each GO category in the set of test genes was then calculated as the fraction of times that F_{test} was lower than or equal to $F_{control}$. Supplementary Table 6 lists the GO categories identified in this analysis ($P < 0.001$), with those passing the Bonferroni cutoff (1/total number of GO categories considered) shown in bold.

Genes rich in intronic G-runs were selected from Ensembl-annotated genes using the criterion that ≥50% of introns contain at least 9 Gs in G-runs within 70 bp of the 5'ss. A total of 3220 genes passed this criterion. The set of genes (12,504 in total) with no intronic G-runs within 70 bp of the 5'ss were used as controls. GO categories enriched in the set of genes with G-runs were identified by comparing the number of genes in this data set and that in the set of controls using the Chi-square test. Supplementary Table 6 lists the GO categories identified in this analysis

($P < 0.001$), with those passing the Bonferroni cutoff (1/total number of GO categories considered) shown in bold.

Sequence motif analyses

For the analyses described in Fig. 4, the conservation rate (CR) of 4-mers and 5-mers in the downstream intronic regions of ICEs was calculated similarly as for G-triplets. The expected CR of each motif was also calculated using first-order Markov models. A control CR was then computed for each motif (test kmer) in each 5'ss group as the average CR of 10 control kmers with the same length. The control k-mers were chosen according to the following criteria: 1) same G+C frequency as in the test k-mer; 2) same number of CG dinucleotide; 3) the least different expected CRs compared to that of the test k-mer. For motif discovery, since we sought to identify motifs with function in introns, and the subset with functional difference in different 5'ss ranges, a two-step procedure was used to evaluate the statistical significance. First, we required a minimum level of conservation above background (represented by the control CR) in at least one 5'ss group using a motif conservation score (MCS) similar to that described by [15]. The MCS is essentially a z-score calculated using the actual CR and the control CR of each kmer. Second, to test whether a motif was more conserved in one 5'ss group than another, a t-statistic for the difference between the excess conservation rate (actual relative to control) of the motif in two different 5'ss groups was determined. Statistical significance was assessed using shuffled datasets generated by randomly permuting the 5'ss scores between exons. Specifically, a MCS cutoff was determined to allow 5% of the motifs to pass the first step of the test in the shuffled data and a t-statistic cutoff was determined to allow 1% of the motifs to pass the second step of the test independent of the first test in the shuffled data.

The above analyses (including data randomization) were conducted separately using intronic windows 11-70 and 11-200, for motif lengths 4 and 5, and for ACEs and ICEs. ACEs were grouped into 5'ss bins in the ranges of 0-4, 4-8 and 8-12 bits. The number of introns in each bin (from lowest to highest 5'ss score) was, respectively, 219, 996, and 1314. In consideration of the larger data size of ICEs and the smaller difference in intron conservation profiles across 5'ss groups (Fig. 4a), ICEs were grouped into 5'ss bins with a step of 2 bits (0-2, 2-4, …, 8-12). The number of introns in each bin (from lowest to highest 5'ss score) was, respectively, 243, 863, 3153, 9295, 18532 and 9961. To reduce the number of comparisons, each 2-bit bin was compared to a 4-bit bin in a different 5'ss strength category, i.e., 0-4 defined as weak 5'ss, 4-8 as intermediate 5'ss and 8-12 as strong 5'ss, in the second step of the statistical procedure. The shuffled data were used to estimate the false discovery rate for motifs passing the statistical tests as ~20%. Representative 5mer analyses for ICEs (with window 11-70) and for ACEs (with window 11-200) are shown in Fig. 4, with full results shown in Supplementary Table 7.

To assess whether other exonic motifs may have 5'ss strength-dependent activity, analyses of sequence conservation were performed, using the improved method that was used for intronic motifs (Fig. 4), as described below. In brief, we searched for 4mer and 5mer motifs whose conservation level differed significantly (compared to randomly shuffled data) between two 5'ss groups in at least 2 out of the 3 reading frames (based on Ensembl annotation) at $P < 0.01$. Only one 5mer (UACAG) passed this test. UACAG was observed to be most significantly conserved in the very weak (0-2 bits) 5'ss group of ICE exons. This 5mer is contained in one of the RESCUE-ESE 6mers[16], UACAGA, suggesting possible ESE activity. Therefore, activity of at least one exonic SRE may also depend on 5'ss strength.

CLIP-Seq analysis

We have adapted standard CLIP protocols[17] to isolate RNA bound by hnRNP H in vivo. HEK 293T cells were grown to confluence and UV irradiated at 400 mJ/cm2. Total protein was isolated using RIPA buffer (50 mM Tris-HCl pH 7.4, 0.1% Na-deoxycholate, 1% NP-40). Lysate was treated with DNase I and dilute RNase A (Fermentas). Immunoprecipitation of hnRNP H was performed using rabbit polyclonal anti-hnRNP H antibodies (Bethyl labs, A300-511A) that were pre-conjugated to protein A beads (Invitrogen). IP'd protein-RNA complexes were washed with RIPA buffer, treated with alkaline phosphatase, and ligated to the Illumina 3' end small RNA adapter. The complexes were subsequently radiolabeled with P32-ATP using T4 PNK, and electrophoresed on a NuPage Bis-Tris SDS-PAGE gel (Invitrogen). Protein-RNA complexes were transferred to nitrocellulose and exposed to film. Protein-RNA complexes from bands excised between 50 and 80 kD, in increments of ~7 kD, were proteinase K treated. RNA was eluted and precipitated, and RNA of length 40-60 nucleotides was ligated to the Illumina 5' end small RNA adapter. Reverse transcription was performed using a primer complementary to the 3' adapter, and 35 cycles of PCR was performed using the Illumina small RNA PCR primers. PCR products of length ~125-175 nucleotides were gel purified and sequenced on an Illumina 1G Genome Analyzer.

Thirty-two base-pair long sequences were aligned to the human genome build hg18 and transcript junction database[13] using Bowtie[18], with parameter settings allowing for up to 2 mismatches in the seed region (first 28 nucleotides). Other analyses were performed using Python scripts.

Equilibrium model of splicing element evolutionary dynamics

To understand the evolutionary implications of splicing enhancement by ISEs generally and of the type of 5'ss strength-dependent activity observed for G-run ISEs specifically, we developed

models of the evolutionary dynamics of splicing cis-elements in the evolution of constitutive exons, which capture essential features of this system. Supplementary Fig. 10 illustrates specific instance of the model, including cases in which: (a) ISEs have no activity; (b) ISEs have activity independent of 5'ss strength; and (c) ISE activity is higher in the context of intermediate 5'ss strength. In each model, three types of splicing *cis*-elements are included: 5'ss, ISE and 3'ss. In some cases we can also consider the third variable as encompassing the 3'ss and other positive-acting cis-elements not explicitly listed (e.g., exonic splicing enhancers). In these examples, we assume that the 5'ss and 3'ss elements each have three possible discrete values (0, 5 and 10) and ISEs have two states (absent, present) which are associated with two (0 and 5) or three (0, 5, and 10) distinct values. Each box (node) in the network corresponds to exons which have a specific combination of scores of the three elements (members in the box are, from left to right, scores of 5'ss, ISE, and 3'ss). A minimum cutoff of 20 was used for the total score of one state.

Up to the relevant minima/maxima, elements in each state may acquire strengthening or weakening mutations. The ratio of motif-weakening to motif-strengthening mutations is represented in the model by the parameter k. The value of k is expected to always exceed one since random mutations more often reduce information (increase entropy) by disrupting/weakening motifs than the reverse. The rate of strengthening mutations is p, so that the rate of motif weakening mutations is kp. For simplicity, the same values of k and p are assumed for all types of score changes (i.e. $10\leftrightarrow5$, $5\leftrightarrow0$). This assumption does not qualitatively affect the conclusions given below. Because the absolute rate of mutations per position per generation is generally very small, multiple mutations, *i.e.*, transitions between states that differ in more than one element, or by a score increment of more than 5, were not considered. Therefore, only 'point mutations' – changes in the score of a single element by 5 units – were considered.

The above model is essentially Markov in character. As in general Markov models, the following assumptions were made: (i) neighbor independence: each splicing element mutates randomly and independently of each other; (ii) history independence: the probability of mutation at each state only depends on the present state (and not on its history); and (iii) element independence: the probability of mutating from one state to another depends only on the information content (score) of the two states, and not on the specific splicing element involved (note that we have assumed uniform k and p values). For such a Markov model, as long as any state can be reached from any other through a finite series of steps with positive probability, the existence and uniqueness of the equilibrium ('steady-state' or stationary) distribution is assured. The equilibrium frequency of each state in the networks can be solved using N (where N is the total number of states) equations representing the balance of the outgoing and incoming mutational flux at each state. Here, 5'ss flux is defined as the probability of changes in the 5'ss between two states (i.e., current state frequency mutation rate of 5'ss from current state to the next).For example, the equations describing the flux equilibrium at the states (labeled as 1, 2, …, 6) in Supplementary Fig. 10a are:

State	Incoming flux		Outgoing flux
1	$bkp+cp+ep$	=	$ap+2akp$
2	$ap+dp+fp$	=	$3bkp$
3	$akp+dkp$	=	$2cp$
4	$bkp+cp$	=	$dkp+dp$
5	$akp+fkp$	=	$2ep$
6	$bkp+ep$	=	$fp+fkp$

where a, b, …, f are the equilibrium frequencies of states 1, 2, …, 6, respectively.

The values of a, b, …, f can be solved from the above equations as: $a = kD$, $b = D$, $c = k^2D$, $d = kD$, $e = k^2D$ and $f = kD$, where $D = 1/(2k^2+3k+1)$. Indeed, in a Markov model satisfying conditions (i)-(iii) above, if the equilibrium frequency of the node with highest total score (the root

node, state 2 in Supplementary Fig. 10a) in the network is designated q (in Supplementary Fig. 10a, $D=q$), then nodes that are one mutation away from the root node have an equilibrium frequency of kq, nodes that are two mutations away from the root node have an equilibrium frequency of k^2q, and so on. This property guarantees that the network flux in the two directions along any network edge are balanced (also known as 'detailed balance'), i.e., the mutation flux from node x to y is equal to the mutation flux from y to x.

Next, the overall network flux associated with each type of splicing element can be calculated as the total network flow (sum of state frequency × mutation rate values) associated with this element in the entire network. For example, the 5'ss flux (edges marked by red blocks) in Supplementary Fig. 10a is calculated as: $akp+cp+bkp+dp = 2kp/(2k+1)$.

To compare the 5'ss fluxes in the three models, we introduce a variable of 'flux per exon' (FPE) for each node layer (node layers and edge layers are illustrated in Supplementary Fig. 10). 5'ss FPE of node layer N_i is defined as the 5'ss flux between node layer N_i and N_{i-1} normalized by the total number of exons in node layer N_i. In Supplementary Fig. 10, the 5'ss FPE of each node layer is shown, where FPE of node layer N_i is essentially p*(number of edges in either direction in E_i with 5'ss flux between N_i and N_{i-1}/number of nodes in layer N_i). Importantly, the 5'ss FPE for any node layer with non-zero probability is invariant across the three models in Supplementary Fig. 10. Comparing Supplementary Figs. 10a and b, it can be appreciated that due to the existence of an extra node layer in Supplementary Fig. 10b, the distribution of exons in 10a is expanded to populate the extra node layer. Because the 5'ss FPE associated with the new layer is larger ($4p/3$) vs. p, the 5'ss flux is larger in Supplementary 10b than in 10a. Similarly, comparing Supplementary Figs. 10b and c, the 5'ss flux is larger in 10c because of the shift of exons to the new node layer (N_3) and a larger 5'ss FPE value of $2p$ associated with this layer.

Supplementary Fig. 10 shows the formulas for total 5'ss flux and 3'ss flux associated with the three models. Due to the symmetry of the 5'ss and 3'ss scores, the fluxes of these two variables are the same for all three models. Supplementary Fig. 10d illustrates the flux values calculated for different values of *k*. A higher 5'ss (3'ss) flux value was observed for the ISE-independent model compared with the ISE-inactive model, while the flux is the highest in the ISE-dependent (i.e. 5'ss strength-dependent) model compared with the other two models. In these models, the 3'ss flux changes in the same pattern as the 5'ss flux, even in the ISE-dependent model where the ISE activity is dependent on 5'ss strength but not 3'ss strength. This is an indirect effect as a result of tolerance to 3'ss mutations when the overall splicing strength of the exon is higher due to the increased activity of ISEs in the context of certain 5'ss scores (for example, refer to the network flow between [5, 10, 10] and [5, 10, 5] in Supplementary Fig. 10c). Asymptotically (in the value of *k*), the 5'ss flux of the ISE-inactive model is O(*p*), while that of the ISE-independent and ISE-dependent models are O(1.2*p*) and O(2*p*), respectively.

Analyses of more complicated models (e.g., containing more discrete values for splicing element scores) invariably yielded the same conclusion that in general presence of ISEs leads to increased splice site flux and models with higher activity for ISEs associated with intermediate 5'ss strength give higher 5'ss flux than those with 5'ss strength-independent activity. In general, if the network nodes are arranged so that all nodes that differ by one single point mutation from the highest-scored node are in layer 1, and all those that differ by two point mutations are in layer 2, etc., and if the layers being considered are complete (i.e., if all combinations of *cis*-element scores in the layers exceed the minimum cutoff), then the FPE of layer *I* can be expressed as:

$2\sum_{j=1}^{I-1}\binom{n}{j}\binom{I-2}{j-1}\Big/\sum_{j=1}^{I}\binom{n}{j}\binom{I-1}{j-1}$, where *n* is the number of types of cis-elements involved in the

model and terms $\binom{n}{j}$, etc. have their the standard combinatorial definitions. Solved numerically, as I increases, the FPE value also increases for n in the range of 2-500. We expect this conclusion will hold generally, i.e. that FPE increases for successively lower layers. As discussed above, the equilibrium probability of nodes in each layer also increases as I increases (since k increases). Thus, the total 5'ss or 3'ss flux should also increase as more layers are included in the model.

Though we have not yet determined the precise set of conditions, in cases when the node layers are not complete (i.e., some nodes have total score below the threshold), higher splice site flux tends to occur in the ISE-active model than the no-ISE model whenever the ISE-active model contains extra states associated with splice site flux, i.e. whenever ISEs afford a greater diversity in the strength of other splicing elements. Similar considerations appear to be sufficient for having greater splice site flux in the 5'ss strength-dependent versus 5'ss strength-independent model as well.

The above models and discussions also apply to networks involving exonic splicing silencers (ESSs). Opposite to the ISE-containing models, the no-ESS model is generally associated with higher splice site flux than the ESS-independent model (where the ESS activity is independent of the 5'ss) whenever the no-ESS model contains extra states associated with splice site flux, or the ESS-independent model contains extra states un-associated with splice site flux. Similarly, ESSs with splice site strength dependent activities as discussed for G-runs in the exons, i.e., higher ESS activity for exons with strong 5'ss, but less ESS activity in the context of intermediate 5'ss, tend to enable more splice site flux than ESSs independent of 5'ss since they often afford extra states associated with splice site flux. The difference between the splice site flux in the no-ESS model and the ESS-dependent model depends on the specific extent of the 5'ss

strength-dependent function of the ESS which determines the gain or loss of states associated with splice site flux.

Supplementary References

1. Wang, Z., Xiao, X., Van Nostrand, E. & Burge, C.B. General and specific functions of exonic splicing silencers in splicing control. *Mol Cell* **23**, 61-70 (2006).
2. Karolchik, D. et al. The UCSC Genome Browser Database. *Nucleic Acids Res* **31**, 51-4 (2003).
3. Burge, C.B., Padgett, R.A. & Sharp, P.A. Evolutionary fates and origins of U12-type introns. *Mol Cell* **2**, 773-85 (1998).
4. Birney, E. et al. Ensembl 2006. *Nucleic Acids Res* **34**, D556-61 (2006).
5. Miller, W. et al. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.*, gr.6761107 (2007).
6. Yeo, G., Hoon, S., Venkatesh, B. & Burge, C.B. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A* **101**, 15700-5 (2004).
7. Yeo, G. & Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-94 (2004).
8. Marcucci, R., Baralle, F.E. & Romano, M. Complex splicing control of the human Thrombopoietin gene by intronic G runs. *Nucl. Acids Res.* **35**, 132-142 (2007).
9. Chen, C.D., Kobayashi, R. & Helfman, D.M. Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. *Genes Dev* **13**, 593-606 (1999).
10. Dai, M. et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* **33**, e175 (2005).
11. Gautier, L., Cope, L., Bolstad, B.M. & Irizarry, R.A. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-15 (2004).
12. Irizarry, R.A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-64 (2003).
13. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-6 (2008).
14. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).
15. Xie, X. et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338-45 (2005).
16. Fairbrother, W.G., Yeh, R.F., Sharp, P.A. & Burge, C.B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007-13 (2002).
17. Ule, J., Jensen, K., Mele, A. & Darnell, R.B. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* **37**, 376-86 (2005).
18. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).