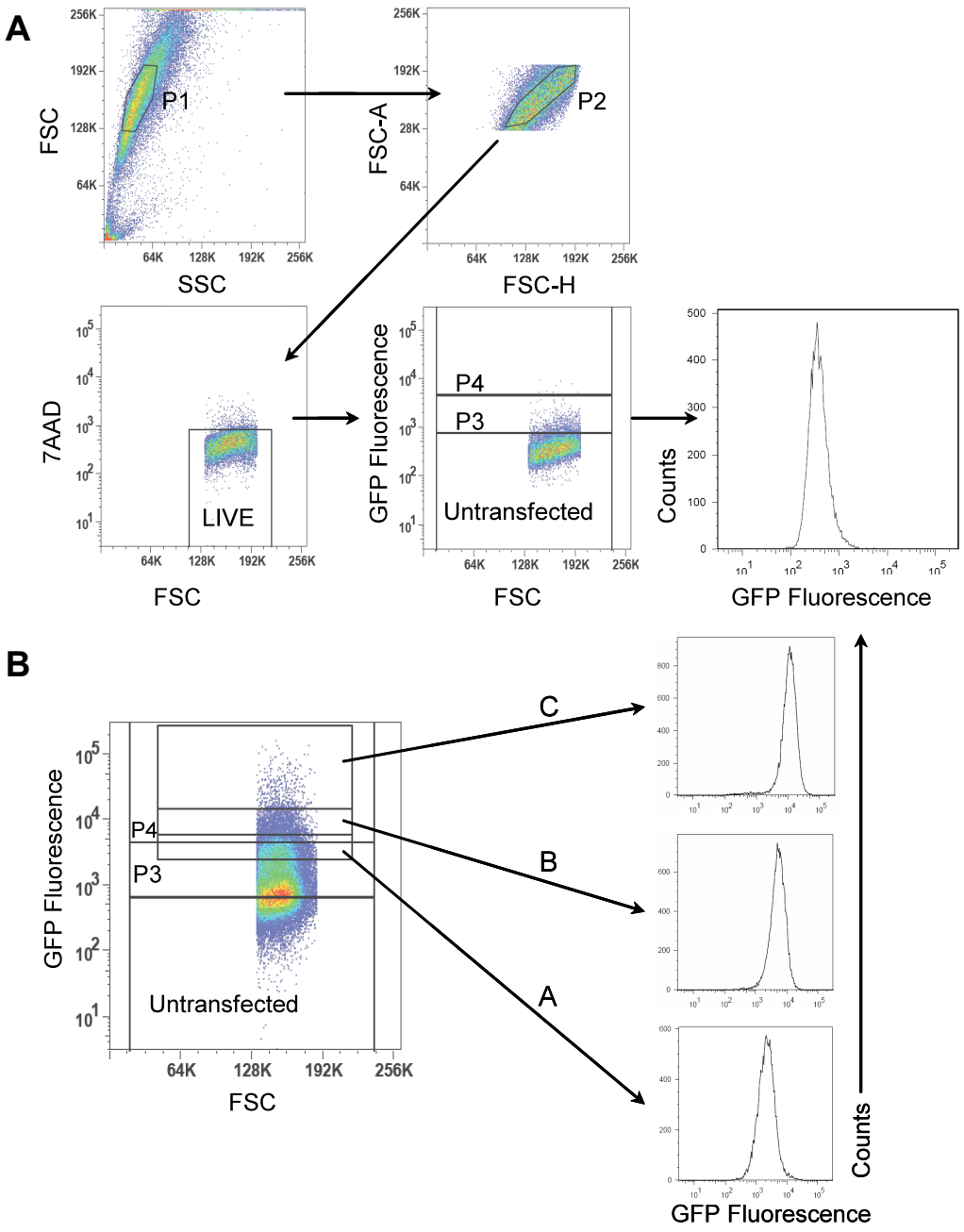



**Figure S1. FACS analysis and ISRE library sorting scheme**

(A) FACS analysis and gating procedure for all HEK-293 FLP-In cells. As an example, flow cytometry data from the stable NMD control is presented. Dot plots show initial gating of stable cells (P1), followed by P2 gating for cell uniformity (i.e., to remove cell aggregates) and finally the selection of live cells using 7-Amino-Actinomycin D (7AAD) staining. The P3 gate reflects the GFP positive cells and the P4 gate is drawn to indicate the upper GFP fluorescence limit of the NMD control population. P4 was used as the gate for the selection of ISS positive cells. The histogram reports the intensity of GFP fluorescence in the NMD control population. (B) FACS analysis of ISS positive stable cells after one round of sorting. Cells from gates A, B and C were sorted and the resulting histograms indicate the intensity of GFP fluorescence after 1 week in culture.

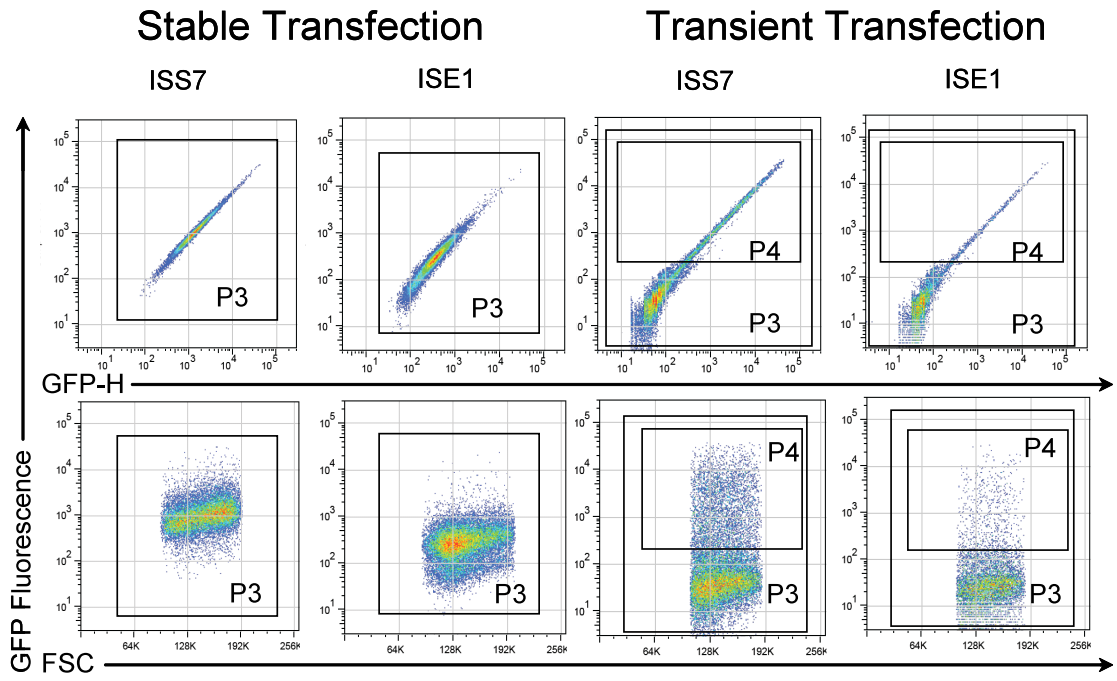
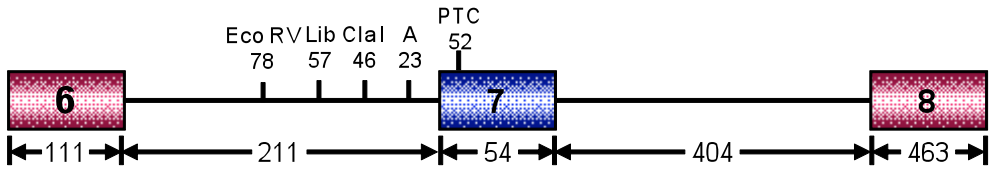



**Figure S2. Assessment of splicing regulatory activity through stable and transient transfection assays**

Sixteen recovered ISS sequences (ISS1-ISS16) and 1 recovered ISE sequence (ISE1) were examined for regulatory activity in both transient and stable transfection assays. Examples of assay results for two recovered sequences (ISS7, ISE1) are shown. For the stable cell line assays, mean GFP fluorescence levels were determined using gate P3. For the transient transfection assays, the P3 gate represents the untransfected cell population and the P4 gate represents the GFP-positive cells. The results of an ANOVA analysis applied to data from the transient and stable assays indicate that the two methods are not statistically similar ($P = 0.27$).

**A**




**B**

GFP

…ACTACCTGAG CACCCAGTCC GCCCTGAGCA AAGACCCCAA CGAGAAGCGC GATCACATGG TCCTGCTGGA

GTTCGTGACC GCCGCCGGGA TCACTCTCGG CATGGACGAG CTGTACTAAC ATAATTCCCC CACCACCTCC

CATATGTCCA GATTCTCTTG ATGATGCTGA TGCTTTGGGA AGTATGTTAA TTTCATGGTA CATGAGTGGC
TATCATACTG GCTATTATAT GGTAAGTAAT CACTCAGCAT CTTTTCCTGA CAATTTTTTT GTAGTTATGT

GACTTTGTTT GGCTGATCAT ATTTTGTTGA ATAAAATAAG TAAAATGTCT TGTGAAACAA AATGCTTTTT

GGTACCAACA TCCATATAAA GCTATAGATA TCGATCAGTN NNNNNNNNNN NNNNGCATCA TCGATGTCTA

TATAGCTATT TTTTTTAACT TCCTTTATTT TCCTTACAGT AATTCAGACA AAATCAAAAA GAAGGAAGGT
GCTCACATTC CTTAAATTAA GGAGTAAGTC TGCCAGCATT ATGAAAGTGA ATCTTACTTT TGTAAAACTT
TATGGTTTGT GGAAAACAAA TGTTTTTGAA CAGTTAAAAA GTTCAGATGT TAAAAAGTTG AAAGGTTAAT
GTAAAACAAT CAATATTAAA GAATTTTGAT GCCAAAACTA TTAGATAAAA GGTTAATCTA CATCCCTACT
AGAATTCTCA TACTTAACTG GTTGGTTATG TGGAAGAAAC ATACTTTCAC AATAAAGAGC TTTAGGATAT

GATGCCATTT TATATCACTA GTAGGCAGAC CAGCAGACTT TTTTTTATTG TGATATGGGA TAACCTAGGC
ATACTGCACT GTACACTCTG ACATATGAAG TGCTCTAGTC AAGTTTAACT GGTGTCCACA GAGGACATGG

TTTAACTGGA ATTCGTCAAG CCTCTGGTTC TAATTTCTCA TTTGCAGGAA ATGCTGGCAT AGAGCAGCAC

TAAATGACAC CACTAAAGAA ACGATCAGAC AGATCTGGAA TGTGAAGCGT TATAGAAGAT AACTGGCCTC
ATTTCTTCAA AATATCAAGT GTTGGGAAAG AAAAAAGGAA GTGGAATGGG TAACTCTTCT TGATTAAAAG

TTATGTAATA ACCAAATGCA ATGTGAAATA TTTTACTGGA CTCTTTTGAA AAACCATCTA GTAAAAGACT
GGGGTGGGGG TGGGAGGCCA GCACGGTGGT GAGGCAGTTG AGAAAATTTG AATGTGGATT AGATTTTGAA
TGATATTGGA TAATTATTGG TAATTTTATG GCCTGTGAGA AGGGTGTTGT AGTTTATAAA AGACTGTCTT
AATTTGCATA CTTAAGCATT TAGGAATGAA GTGTTAGAGT GTCTTAAAAT GTTTCAAATG GTTTAACAAA
ATGTATGTGA GGCGTATGTG

pair 1,5 (5’–3’)

pair 2 (3’)

pair 2 (3’–5’)

Eco RV

15-mer library

Cla I

Exon6 (5’)

Exon7

PTC

BP

pair 3 (5’–3’)

Exon8

pair 3 (3’–5’)

pair 4 (3’–5’)

**C**

**Exon included isoform:**

Exon7 Exon8

TAATTCAGACAAAATCAAAAAGAAGGAAGGTGCTCACATTCCTTAAATTAAGGAGAAATGCTGGCATAGAGCAGC…

5’–3’ pair 4

**Exon excluded isoform:**

Exon6 Exon8

…TTTCATGGTACATGAGTGGCTATCATACTGGCTATTATATG | GAAATGCTGGCATAGAGCAGCAC…

3’–5’ pair 5

**Figure S3. Schematic representation of SPLICE and primer set binding sites**
(A) Schematic representation of the SMN1 mini-gene. Shown below each exon and intron are their respective lengths (bp). The positions (relative to the 3' ss of exon 7) of restriction sites Eco RV and Cla I used to insert the 15-mer library are indicated. The PTC was inserted 51-nt upstream of the 5' ss of exon 7. (B) Mapping of primer set binding sites on the SMN1 mini-gene sequence. Schematic representing the locations of primer set binding for transcript isoform analysis by qRT-PCR. The locations of the branch point (BP), restrictions sites Eco RV and Cla I, the 15-mer library (plus flanking regions) and PTC are shown. (C) Schematic representing the exact locations of primer sequences spanning exon-exon junctions.

**A**

| | |
|---|---|
| ISS17 | GCAAGGTCCCTCTAG |
| ISS18 | GACGGAGCCGTCTGG |
| ISS19 | AGAGTGGCGGTGGAG |
| ISS20 | GATATGGCGAGGGTG |
| ISS21 | GGTGGCAGACACGAT |
| ISS22 | AAATAGAGGCCCCAG |
| ISS23 | TTATGGAGTTCCTAG |
| ISS24 | GAGGGCAGTCCGTGG |
| ISS25 | TGGACACGTCAGTCA |
| ISS26 | TCTGACTCAATAGTA |
| ISS27 | AATTGGGTTTGGGGG |
| ISS28 | TATGACATGTGGGGA |

**B**

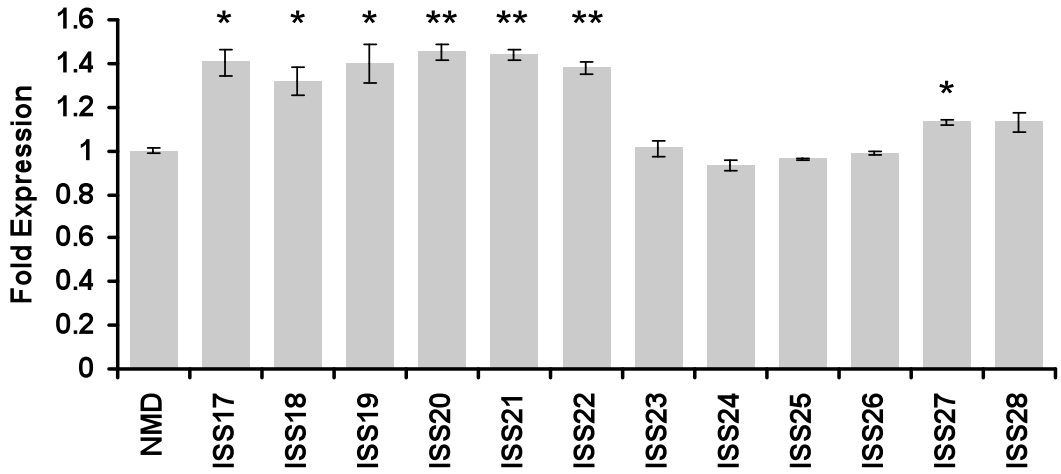



**Figure S4. The activity of additional recovered ISRE sequences is validated by stable cell line assays**

(A) Additional recovered ISRE sequences examined for regulatory activity. (B) Flow cytometry analysis of HEK-293 FLP-In stable cell lines generated for each recovered ISRE sequence and control construct. Mean GFP levels from two independent experiments were determined and normalized to the NMD control. The fold expression of each sample relative to NMD and average error are reported. Resulting *P*-values in comparison to the NMD control: * $P < 0.05$ and ** $P < 0.01$.

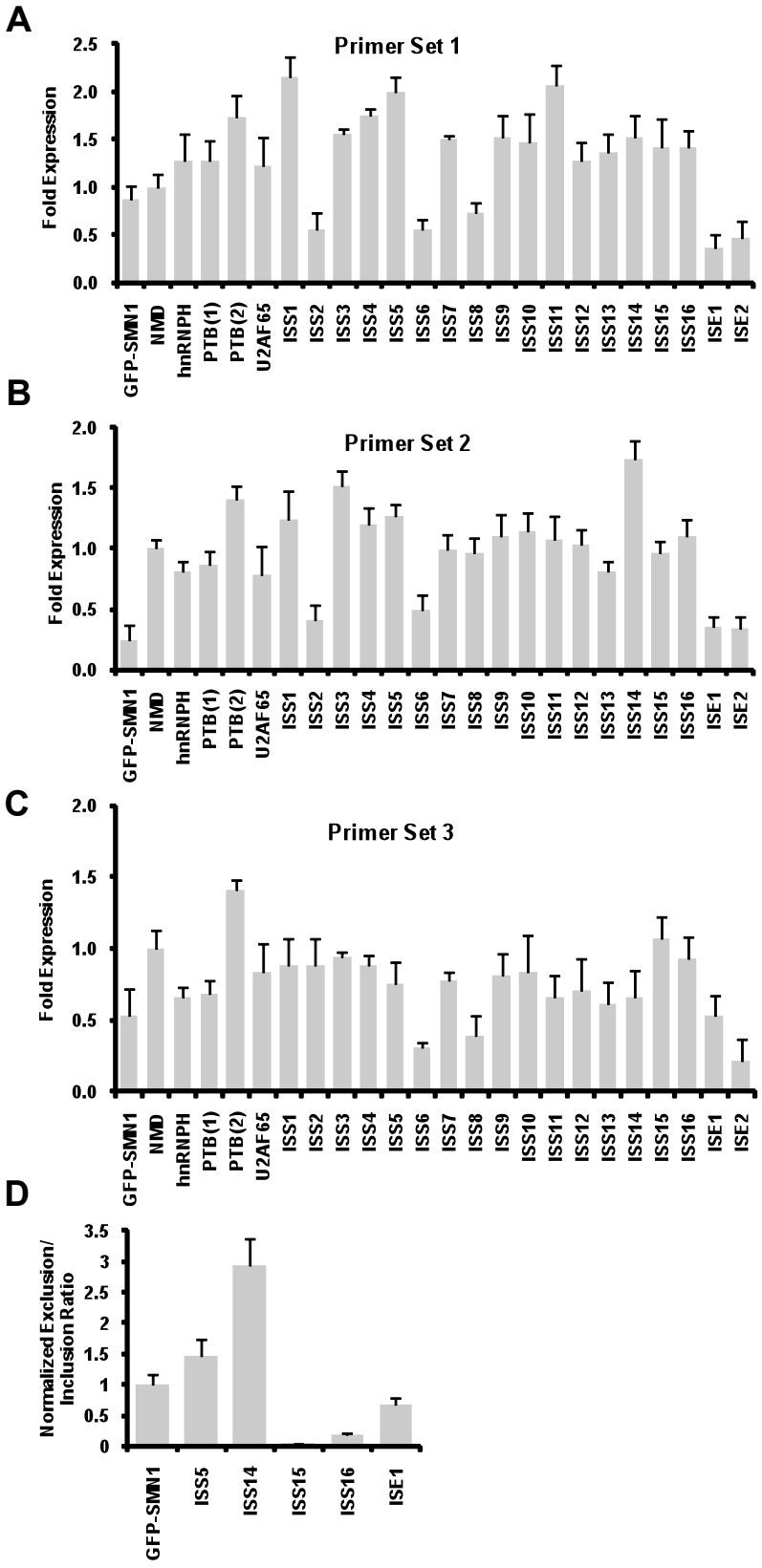

A
Primer Set 1
Fold Expression
2.5
2.0
1.5
1.0
0.5
0.0
GFP-SMN1
NMD
hnRNPH
PTB(1)
PTB(2)
U2AF65
ISS1
ISS2
ISS3
ISS4
ISS5
ISS6
ISS7
ISS8
ISS9
ISS10
ISS11
ISS12
ISS13
ISS14
ISS15
ISS16
ISE1
ISE2
B
Primer Set 2
Fold Expression
2.0
1.5
1.0
0.5
0.0
GFP-SMN1
NMD
hnRNPH
PTB(1)
PTB(2)
U2AF65
ISS1
ISS2
ISS3
ISS4
ISS5
ISS6
ISS7
ISS8
ISS9
ISS10
ISS11
ISS12
ISS13
ISS14
ISS15
ISS16
ISE1
ISE2
C
Primer Set 3
Fold Expression
2.0
1.5
1.0
0.5
0.0
GFP-SMN1
NMD
hnRNPH
PTB(1)
PTB(2)
U2AF65
ISS1
ISS2
ISS3
ISS4
ISS5
ISS6
ISS7
ISS8
ISS9
ISS10
ISS11
ISS12
ISS13
ISS14
ISS15
ISS16
ISE1
ISE2
D
Normalized Exclusion/ Inclusion Ratio
3.5
3
2.5
2
1.5
1
0.5
0
GFP-SMN1
ISS5
ISS14
ISS15
ISS16
ISE1

**Figure S5. Additional qRT-PCR isoform analysis of recovered ISREs and control constructs**

(A) qRT-PCR analysis with primer set 1 (Figure 1, Supplementary Figure S1 and Supplementary Table S1). Results demonstrate that overall transcript levels for the GFP-SMN1, ISS controls, ISSs and ISEs did not significantly differ from the NMD control ($P = 0.2$). For all subsequent analyses, expression levels of duplicate PCR samples were normalized to the levels of *HPRT*. Fold expression data is reported as the mean expression for each sample divided by the mean NMD expression value $\pm$ the average error. (B) qRT-PCR analysis with primer set 2. The levels of intron 6 retained in transcripts containing the selected and control ISS sequences are similar to the NMD control ($P = 0.65$). In contrast, intron 6 retention in ISE transcripts are similar to the GFP-SMN1 control ($P = 0.74$) and different from the NMD control ($P < 0.05$), suggesting that intron 6 in the GFP-SMN1 control and ISEs are processed similarly by the general splicing machinery. The retention level of intron 6 for the GFP-SMN1 control is statistically different from the NMD control ($P < 0.05$). (C) qRT-PCR analysis with primer set 3. The levels of intron 7 retention for the recovered and control ISS sequences and the GFP-SMN1 are similar to the NMD control ($P = 0.33$). The intron 7 retention levels in ISE transcripts are significantly different from the NMD control ($P < 0.05$). (D) qRT-PCR analysis with primer sets 4 and 5 on ISS5, ISS14-ISS16 and ISE1 inserted in the non-NMD-based GFP-SMN1 control construct. The transcript isoform analysis of stable cell lines demonstrates that the tested sequences maintain their regulatory activities (Figure 2B and C) in the non-NMD-based reporter. However, the transcript isoform levels of ISS15 and ISS16 displayed significant enhancer activity ($P < 0.05$), and do not correlate with measured fluorescence levels from the NMD-based reporter. The results suggest that ISS15 and ISS16 may exhibit enhanced fluorescence levels in the context of the NMD reporter due to the evasion of the NMD process. Data is reported as the expression ratio of the mean expression of the exon excluded isoform to the exon included isoform normalized to the ratio for the GFP-SMN1 control $\pm$ the average error.

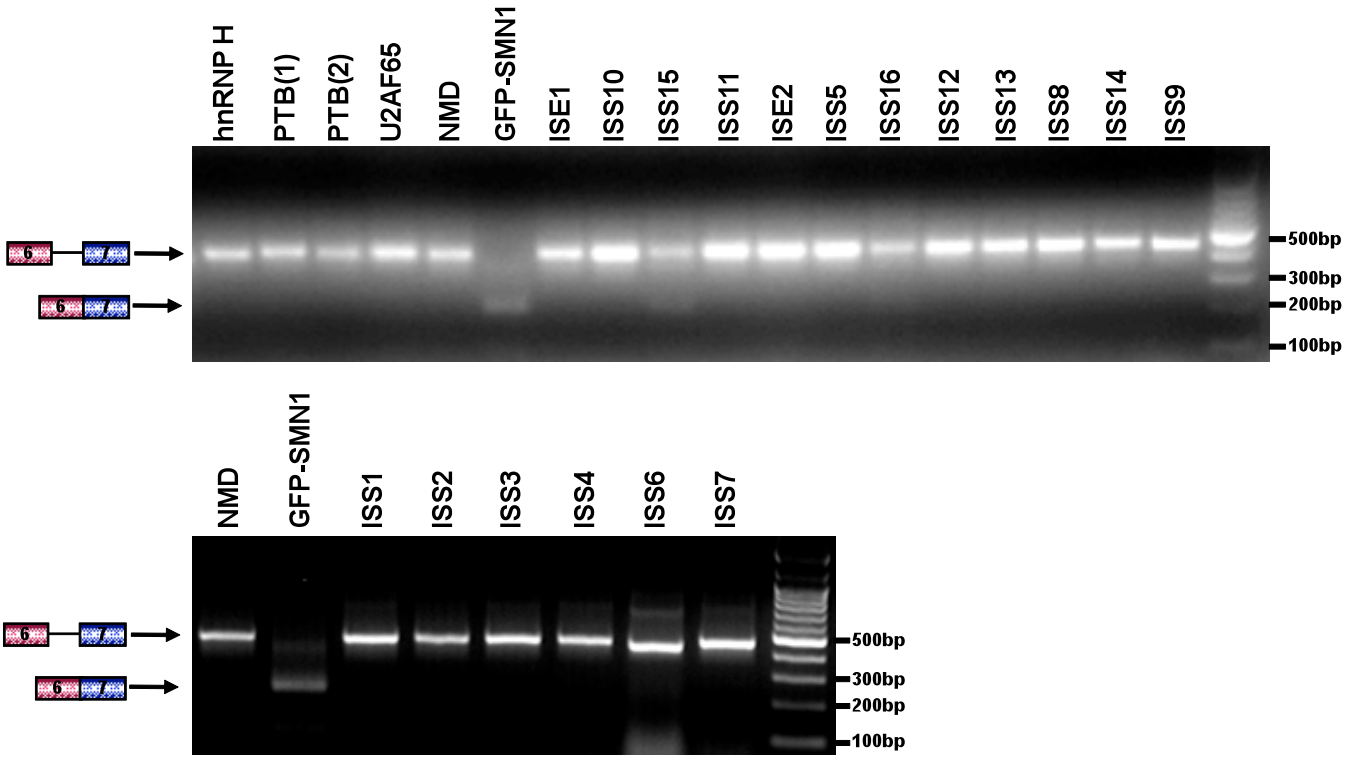



**Figure S6. Examination of possible alternative 3' ss by qRT-PCR analysis of recovered ISREs and control constructs**

We examined the possibility that selected ISREs and control constructs may include an alternative 3' ss by qRT-PCR analysis using the forward primer of primer set 1 (Figure 1 and Table S3) and a unique reverse primer for exon 7 (primer ex7, Table S1). The position of PCR products corresponding to the intron 6 retained and exon 7 included isoforms are indicated on the left. Using the above primer set, the expected sizes of the intron 6 retained and exon 7 included isoforms are 462 bp and 251 bp, respectively. Given the placement and length of our 15-mer library cassette (39-nt, Supplementary Figure S1), an ISRE with an alternative 3' ss would display the alternative 3' ss included isoform at a length between 297-329 bps. None of the recovered ISREs and control constructs display a PCR product within this range and therefore rule out the possibility that selected ISRE sequences may contain an alternative 3' ss. As shown above, the exon 7 included isoform was also detected in cell lines ISS15 and 16 as previously observed in our qRT-PCR analysis with primer set 4 (Figure 2C). While this data suggests that selected ISREs do not lead to alternative 3' ss processing, we cannot rule out the possibility of a minor change at the 3' ss due to aberrant splicing that would alter the reading frame of the PTC.

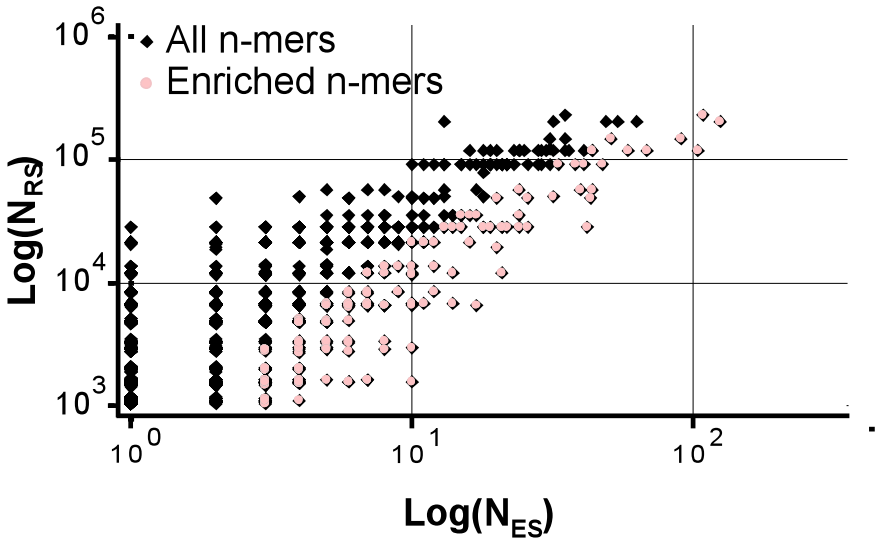



**Figure S7. Scatter-plot for the occurrence frequency of all 4-6-nt n-mers in the ISRE sample set**

Scatter-plot for the occurrence frequency of all 4-6-nt n-mers in the enriched sample set ($N_{ES}$) vs. a corresponding random sample set ($N_{RS}$) (black). A similar scatter-plot based on n-mers determined to be significantly enriched in the recovered ISREs is overlaid (pink).

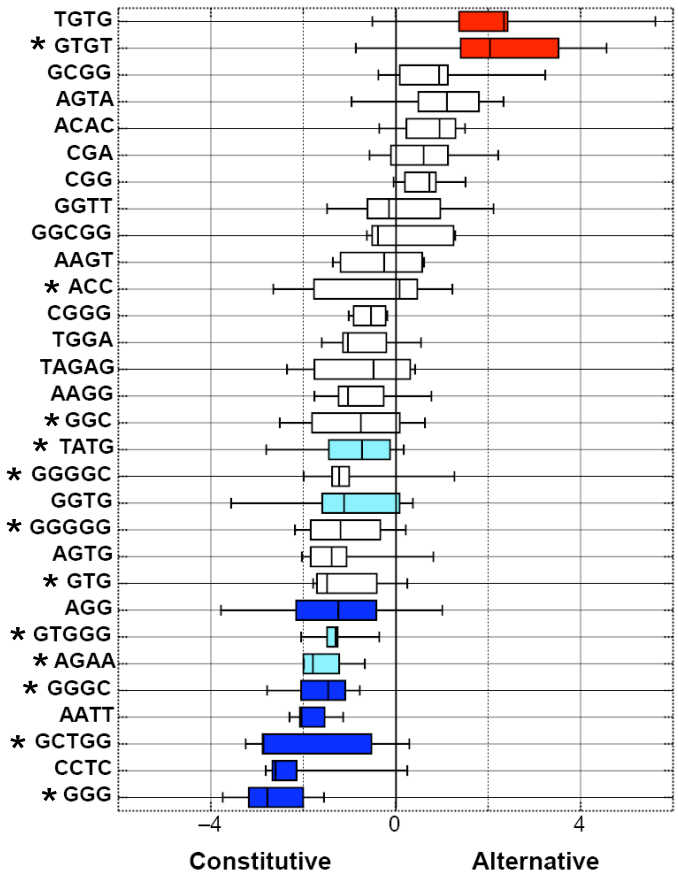



**Figure S8. Enriched n-mers associate with constitutive and alternative splicing**

Box-plots revealing the distribution of TA-scores for GCCS derived ISREs. The GCCS consensus motifs that are significantly associated with alternative splicing are shown in red ($P_{\text{t-test}} < 0.01$) and those that are significantly associated with constitutive splicing are shown in shades of blue (dark blue, $P_{\text{t-test}} < 0.01$; light blue, $P_{\text{t-test}} < 0.05$). In total, 9 consensus motifs are biased toward alternative splicing and 21 consensus motifs display a bias towards constitutive splicing. Elements exhibiting no significant association with either category are not shaded. Starred motifs are present in hexamers subjected to RNAi silencing studies to examine regulated splicing. The entire population of consensus n-mers significantly associates with constitutive splicing ($P_{\text{t-test}} = 1.8e^{-8}$). The stronger association with constitutively spliced exons may be a result of the selected ISREs functioning as ISSs, which have been shown to be enriched in the intronic flanks of constitutively spliced exons (1). A previous analysis of conserved intronic sequences revealed that a large number of motifs strongly associate with constitutive splicing and are more abundant than those associated with alternative splicing (2). Additional studies have also demonstrated that splice silencing may be a mechanism that represses pseudoexon inclusion (3) and that intronic sequences which repress splicing might have a fundamental role in defining real exons by silencing nearby decoy sites (4). In addition, several of our enriched motifs that associate with constitutive splicing also overlap with elements that have been previously identified upstream of constitutively spliced exons (5). Taken together, these observations are in line with results from our genome-wide analysis and suggest the utility of future computational investigations to determine the association between selected ISREs and pseudoexons.

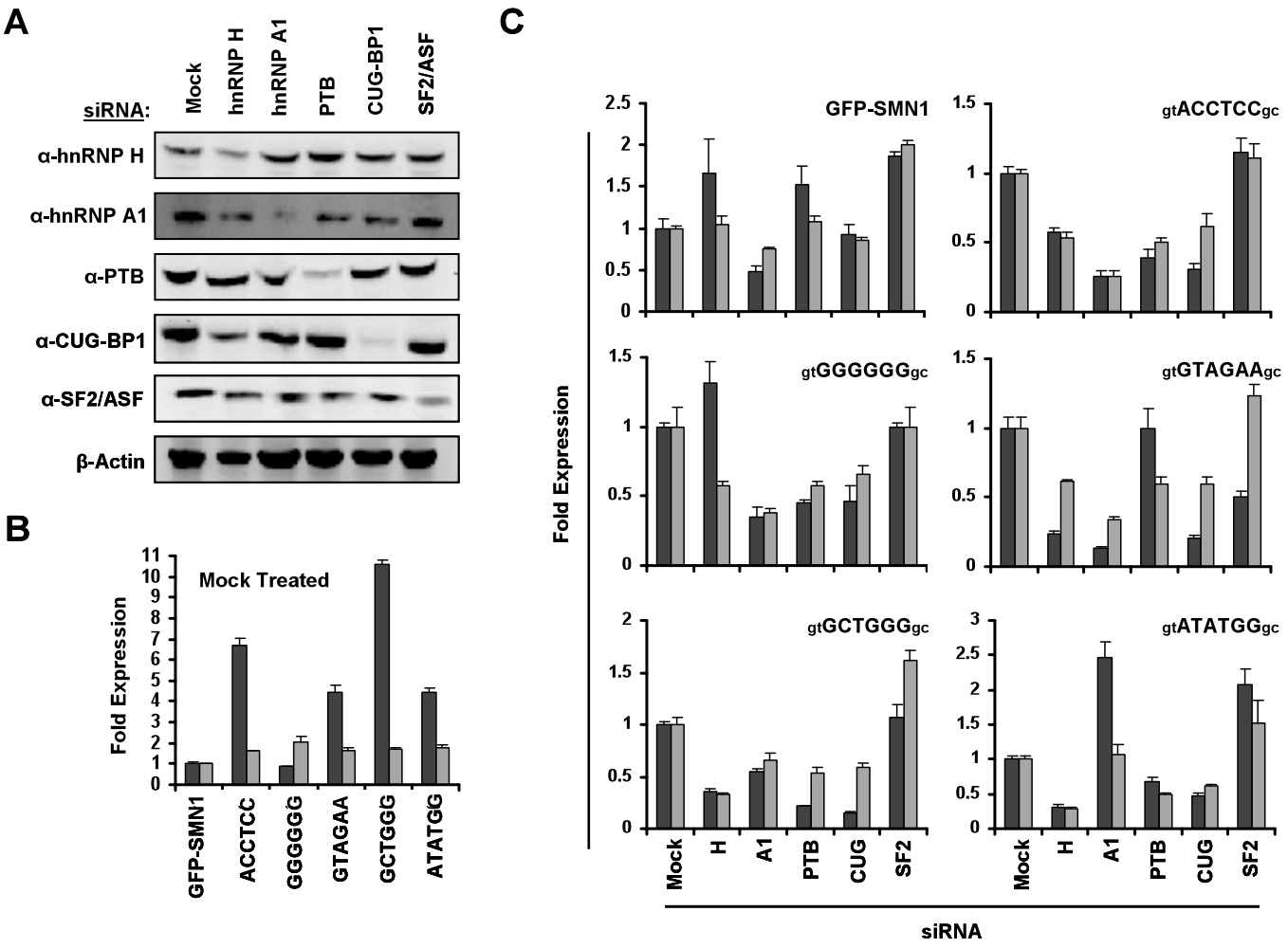



**Figure S9. The effects of *in vivo* depletion of splicing factors on ISRE regulated splicing patterns**

(A) Western blot analysis of total cell lysates prepared from the GFP-SMN1 control cell line treated with siRNAs targeted to trans-acting splicing factors and a mock siRNA negative control. The results demonstrate that individual siRNAs have minimal to no off-target affects. (B) qRT-PCR analysis of the mock treated ISRE hexamer and GFP-SMN1 control cell lines with primer sets specific for exon 7 excluded (black bars) and included (gray bars) products. Expression levels of duplicate PCR samples were normalized to the levels of *HPRT*. Fold expression data is reported as the mean expression for each sample divided by the mean GFP-SMN1 control expression value $\pm$ the average error. (C) qRT-PCR analysis of the siRNA treated ISRE hexamer and GFP-SMN1 control cell lines with primer sets specific for exon 7 excluded (black bars) and included (gray bars) products. Fold expression data is reported as the mean expression for each sample divided by the mean mock siRNA treated cell line control expression value $\pm$ the average error.

**Table S1. Primer and oligonucleotide sequences**

| Name | Primer Sequence (5 ’- 3’) |
|---|---|
| Ex6 | CATGGACGAGCTGTACGTTAACATAATTCCCCCACCACCTC |
| Ex8 | CGCTCG AGCACATACGCCTCACATACATTTTG |
| GFP1 | GCGGTACCATGGTGAGCAAGGGCG |
| GFP2 | GGTGGTGGGGGAATTATGTTAACGTACAGCTCGTCCATGCC |
| ECmutF | CTTTTTAACATCCATATAAAGCTATCGATATCTAGCTATCGAT GTCTATATAGCTATTTTTTTTAACT |
| ECmutR | AGTTAAAAAAAAATAGCTATATAGACATCGATAGCTAGATATCG ATAGCTTTATATGGATGTTAAAAAG |
| ISStemp | GCGCGATATCGATCAGT ($N_{15}$) GCATCATCGATGCGC |
| Lib1 | GCGCGATATCGATCAGT |
| Lib2 | GCGCATCGATGATGC |
| Lib3 | GAAACAAAATGCTTTTTAACATCCATA |
| Lib4 | GGAAAATAAAAGGAAGTTAAAAAAAATAGC |
| SMN1cDNA | TAGAAGGCACAGTCGAGG |
| Ex7 | AAGGAATGTGAGCACCTT |

**Table S2. Plasmid constructs used in this work**

| Name | Description |
|---|---|
| pCS238 | GFP-SMN1. Contains the wild-type SMN1 mini-gene fused to the N-terminus of GFP. Positive control used for all flow cytometry analysis. |
| pCS516 | SMN1 NMD-based reporter construct. Contains the SMN1 mini-gene with a PTC in exon 7 fused to the N-terminus of GFP. Recovered ISREs as well as control ISS were inserted into this construct. |
| pCS517 | SMN1 NMD-based containing random 15-mer. Negative control used for all flow cytometry analysis. |
| pCS668 | U2AF65 binding site inserted into pCS516. |
| pCS669 | hnRNP H binding site inserted into pCS516. |
| pCS670 | PTB (1) binding site inserted into pCS516. |
| pCS667 | PTB (2) binding site inserted into pCS516. |

**Table S3. Primer sequences for SMN1 transcript isoform analysis through qRT-PCR**

| Name | Forward Primer (5' - 3') | Reverse Primer (5' - 3') | Isoform |
|---|---|---|---|
| Pair 1 | TGAGCAAAGACCCCAA | TGATAGCCACTCATGTACC | GFP and Ex 6 |
| Pair 2 | CTCCCATATGTCCAGATTCT | AGCATTTTGTTTCACAAGACA | Ex 6 and Int 6 |
| Pair 3 | CACTAGTAGGCAGACCAG | CAGTTATCTTCTATAACGCTTCAC | Int 7 and Ex 8 |
| Pair 4 | TAAATTAAGGAGAAATGCT | GGTTTTTCAAAAGAGTCCAGTAA | Ex 7/8 and Ex 8 |
| Pair 5 | TGAGCAAAGACCCCAA | CCAGCATTTCCATATAATAG | GFP and Ex 6/8 |
| Pair 6 | CAAAGATGGTCAAGGTCGCAAG | GGCGATGTCAATAGGACTCC | HPRT |

**Table S4. Primer sequences for endogenous transcript isoform analysis through qRT-PCR**

| Name | Gene | Hexamer | Sequence (5' – 3') | Isoform | Type of Alternative splicing |
|---|---|---|---|---|---|
| Fw.ADD3ex15_16 | ADD3 | ACCTCC | TGAAAAATTAGAAGAA AACCATGAGC | Exon15/16 | cassette |
| Fw.ADD3ex14_16 | ADD3 | ACCTCC | GGCC TAG AAGAAA ACCATG AGC | Exon14/16 | cassette |
| Rv.ADD3ex16 | ADD3 | ACCTCC | CTTCGATTTTCTCTGGA GACT | ADD3 cDNA, Exon15/16, Exon 14/16 | cassette |
| Fw.hnRNPCex1_3 | HNRNPC | ACCTCC | CCC CTT CTT GTT TTC GGC TTT | Exon1/3 | cassette |
| Fw.hnRNPCex2_3 | HNRNPC | ACCTCC | CTT CAGCTACATTTT C GGCTTT | Exon2/3 | cassette |
| Rv.hnRNPCex3 | HNRNPC | ACCTCC | CGAAAAGATTGCCTCC ACAT | hnRNPC cDNA and Exon1/3, Exon 2/3 | cassette |
| Fw.CLK3ex4 | CLK3 | GGGGGG | CCGTGACAGCGATACA TAC | Exon 4/5, Exon4/6 | cassette |
| Rv.CLK3ex4_5 | CLK3 | GGGGGG | GTTGGCTTCTCGAGGA GG | Exon 4/5 | cassette |
| Rv.CLK3ex4_6 | CLK3 | GGGGGG | CCACAATCTCATCGAG GAGG | Exon 4/6 | cassette |
| Rv.CLK3cDNA | CLK3 | GGGGGG | CAAGCACTCCACCACC T | CLK3 cDNA | cassette |
| Fw.CADPSex16 | CADPS | GGGGGG | GAAAGATATTGTTACC CCAGT | Exon 16/19, Exon16/18, Exon16/17 | mutually exclusive |
| Rv.CADPSex16_18 | CADPS | GGGGGG | CCTTTTGATTCTCTTCG ATTTTG | Exon16/18, | mutually exclusive |
| Rv.CADPSex16_19 | CADPS | GGGGGG | GGCCTACATTTTCTTCG ATTTTG | Exon 16/19 | mutually exclusive |
| Rv.CADPSex16_17 | CADPS | GGGGGG | CTCTCTTTTTCCCTTCG ATTTTG | Exon16/17 | mutually exclusive |
| Rv.CADPScDNA | CADPS | GGGGGG | AAG CTT TTT GGC AGG AGT GA | CADPS cDNA | mutually exclusive |
| Fw.c6orf60ex15_16 | C6orf60 | GTAGAA | CTTTACAAGTGTCATTA GAAGAAATG | Exon 15/16 | cassette |
| Fw.c6orf60ex14_16 | C6orf60 | GTAGAA | CCA ACA GAT AAG ATT AGA AGA AAT GG | Exon 14/16 | cassette |
| Rv.c6orf60ex16 | C6orf60 | GTAGAA | GATCTGGTCTCTTTCTG TAAGC | C6orf60 cDNA, Exon 15/16, Exon 14/16 | cassette |
| Fw.RREB1ex11_12 | RREB1 | GTAGAA | GATAGCACAGACAGTC AGTCG | Exon11/12 | cassette |
| Fw.RREB1ex10_12 | RREB1 | GTAGAA | ACA CAC ACT GAC AGT CAG TCG | Exon10/12 | cassette |
| Rv.RREB1ex12 | RREB1 | GTAGAA | CTCCTCCTCCGGCTCAT | RREB1 cDNA, Exon11/12, Exon10/12 | cassette |

| | | | | | |
|---|---|---|---|---|---|
| Fw.MADDex35 | MADD | GCTGGG | AGTTCCCTGTGCGAC | Exon35/36, Exon35/37 | cassette |
| Rv.MADDex35_36 | MADD | GCTGGG | TCTATGAAAACCTGATTGTGCA | Exon35/36 | cassette |
| Rv.MADDex35_37 | MADD | GCTGGG | TAATTTCAGGAACTGATTGTGCA | Exon35/37 | cassette |
| Rv.MADDcDNA | MADD | GCTGGG | TAGTACAGCTCCCGACACTT | MADD cDNA | cassette |
| Fw.CAMK2Gex13_14 | CAMK2G | GCTGGG | CGGGCAAGCTGCCAAAAG | Exon13/14 | cassette |
| Fw.CAMK2Gex12_14 | CAMK2G | GCTGGG | GAA CTT CTC AGC TGC CAA AAG | Exon12/14 | cassette |
| Rv.CAMK2Gex14 | CAMK2G | GCTGGG | TTGACACCGCCATCCG | CAMK2G cDNA, Exon13/14, Exon12/14 | cassette |
| Fw.A2BP1ex15_17 | A2BP1 | ATATGG | GCAGACATTTATGGTGGTTATG | Exon15/17 | mutually exclusive |
| Fw.A2BP1ex16_17 | A2BP1 | ATATGG | TAA ATT GCT GCA GGG TGG TTA TG | Exon16/17 | mutually exclusive |
| Rv.A2BP1ex17 | A2BP1 | ATATGG | CTGTCACTGTAGGCAGCG | A2BP1 cDNA, Exon15/17, Exon16/17 | mutually exclusive |
| Fw.HNRNPA2B1ex1 | HNRNPA2B1 | ATATGG | CTCTAGCGGCAGTAGCA | Exon1/2, Exon1/3 | cassette |
| Rv. HNRNPA2B1ex1_2 | HNRNPA2B1 | ATATGG | GTTTCTAAAGTTTTCTCCATCGCG | Exon1/2 | cassette |
| Rv. HNRNPA2B1ex1_3 | HNRNPA2B1 | ATATGG | GTTCCTTTTCTCTCTCCATCGC | Exon1/3 | cassette |
| Rv. HNRNPA2B1cDNA | HNRNPA2B1 | ATATGG | CCTCAAACTTTCTTCTGTGG | HNRNPA2B1 cDNA, Exon1/2, Exon1/3 | cassette |

**Table S5. Identified ISRE regulatory sequences**

| ISS sequences | Name | Tested stably | Tested transiently | *Occurrence |
|---|---|---|---|---|
| GACGTGTGTCTCGGG | ISE1 | Y | Y | 5 |
| ATAGTGGCGGTGGAG | ISE2 | Y | N | 1 |
| TACATCCCTCGGTTG | ISS1 | Y | Y | 2 |
| AGAATAAGTGGGGTG | ISS2 | Y | Y | 1 |
| AGTATATGGTGAGGA | ISS3 | Y | Y | 1 |
| TGTTTTGCGTCCAAG | ISS4 | Y | Y | 2 |
| AGAATAAGTGAGGTG | ISS5 | Y | Y | 2 |
| CCGAGTGCGACGGTG | ISS6 | Y | Y | 3 |
| ACAGGCCAAGGGGGG | ISS7 | Y | Y | 1 |
| CAAACACCTCCGATG | ISS8 | Y | Y | 20 |
| GGTCGAGTCGCAAGG | ISS9 | Y | Y | 2 |
| TAGGTGTGTCTCGGG | ISS10 | Y | Y | 1 |
| ACAGTGCTAAGTAGG | ISS11 | Y | Y | 4 |
| AAAGACCGGGATATG | ISS12 | Y | Y | 7 |
| AGTCACCTATTATAG | ISS13 | Y | Y | 5 |
| TTGTAAGGTGCTGGG | ISS14 | Y | Y | 2 |
| GGGGCGCGCGGGGGG | ISS15 | Y | Y | 1 |
| AGAGTGGGGCGGGTG | ISS16 | Y | Y | 1 |
| GCAAGGTCCCTCTAG | ISS17 | Y | N | 4 |
| GACGGAGCCGTCTGG | ISS18 | Y | N | 1 |
| AGAGTGGCGGTGGAG | ISS19 | Y | N | 1 |
| GATATGGCGAGGGTG | ISS20 | Y | N | 1 |
| GGTGGCAGACACGAT | ISS21 | Y | N | 2 |
| AAATAGAGGCCCCAG | ISS22 | Y | N | 1 |
| TTATGGAGTTCCTAG | ISS23 | Y | N | 8 |
| GAGGGCAGTCCGTGG | ISS24 | Y | N | 1 |
| TGGACACGTCAGTCA | ISS25 | Y | N | 3 |
| TCTGACTCAATAGTA | ISS26 | Y | N | 1 |
| AATTGGGTTTGGGGG | ISS27 | Y | N | 1 |
| TATGACATGTGGGGA | ISS28 | Y | N | 1 |
| CCGAGGAACCATAGG | | | | 11 |
| CCCTATGGTTCCTCG | | | | 1 |
| GACGGGTGCCTCGGG | | | | 2 |
| GGCTGGAAGACCTGC | | | | 2 |
| GGAGTGGCTGGTTCG | | | | 1 |
| GGCTGGGCTAGGATG | | | | 1 |
| ACCTCAGGCTCTGAA | | | | 2 |
| GACTGTGTTAGGCGG | | | | 1 |
| AAAGAACGGGATATG | | | | 1 |
| TCGAATCTCTCCAGT | | | | 1 |
| CCTACGCTCATTATT | | | | 4 |
| TCTTCTCTTCTCTTC | | | | 1 |
| TGTTCGCACCGCTGG | | | | 1 |
| TGTTCGCACCACTGA | | | | 1 |

| | |
|---|---|
| GTTAACCAACGATGG | 1 |
| GGTATCGAAAGTTGT | 1 |
| TACATCCAGAAGTCG | 2 |
| TGGACCAGGCGTACG | 1 |
| CACACGTGAGAGAGA | 3 |
| GAAGGGCGACAGATA | 1 |
| AGAACGCTGGATTAA | 1 |
| TTACTTTAAGGATAA | 2 |
| ATACGGAAAGGCCTT | 1 |
| GTGCTTATATGGGTT | 1 |
| TTAGTCCCATTCCGA | 2 |
| CCACTTCGGTTGCCT | 1 |
| ACGTCCGTCGTGGAT | 1 |
| ACCTCGAGGTCTGAA | 1 |
| AAGGCTAGTTTAGTA | 1 |
| AAGGCTAGATTAGTA | 3 |
| AGAGGAGTCGTGTCA | 1 |
| AGTGGAATCGTATCA | 1 |
| ATTCCAGCTGGAGCT | 3 |
| GCCGAGTAAAGTGTA | 1 |
| CTTGAGTACCCCCGA | 1 |
| CATGCACCGACCAAG | 1 |
| AATTGTGTTTGTGAT | 1 |
| AATTGTGTTTGGCGG | 7 |
| TATGACGTGTGGGGG | 1 |
| TATGACATGTGGGGG | 4 |
| CAATTGAGTTGGTGT | 1 |
| CGATGGGGCAGGGGA | 1 |
| CAGTGAACTTTGCGA | 1 |
| CCTTGGTCCTGACAT | 1 |
| GAGTGGCCTAGGGAG | 1 |
| AAGTGGGCACGGTTG | 1 |
| AGGTAGCCACCGTTG | 1 |
| GGGGGGGTCACTTAG | 1 |
| TGGTTGGACCCGTAG | 4 |
| CTAGTAACCAGCCAG | 1 |
| CTAAGCACCACTGAG | 1 |
| CATGTCAGGACCAAG | 2 |
| CATGGACCGACCAAG | 1 |
| TATGCCTCCCCGATA | 1 |
| CGAAGAACCCCAAGG | 1 |
| CGGAGAAACCGGAGG | 1 |
| CTATCTCCTTCTATG | 1 |
| TTAACACCTCCCAAG | 1 |
| CAAAGACCTGCGATG | 1 |
| CAAACACGTCCGATG | 1 |

| | |
|---|---|
| CTAACACCTCCGATG | 2 |
| GTGGCTAAGAATTGG | 1 |
| GTAAAGGGTGTCAGT | 1 |
| ATTAATAATACTGGG | 1 |
| GTTAATAGCGCGGGA | 2 |
| TGTGGTCGCGACCTG | 1 |
| GGCGGTCGAGTACAG | 1 |
| GTTGTGAAAGAGGAG | 1 |
| GCGGTTTGCGGGCGG | 1 |
| GCATGGCCCCGCTGG | 1 |
| GCACTAGAATCTGAG | 1 |
| GCAGTACGGGCTTAG | 1 |
| CGAGCGGCTTTAGAG | 1 |
| AGAATGGACCGTGAG | 1 |
| GTACAGCGGAGAGGG | 1 |
| GTACGGTGCAGAGGG | 2 |
| GTAGTGTAGGGAGGG | 1 |
| GAAGTGTAGGGAGGG | 1 |
| ATACCGTTCAGTGGG | 1 |
| ATACCGTTCAGTGAG | 3 |
| AAAGGGGCAAGGTGG | 1 |
| AGAGTGCGAAGCGGG | 1 |
| GTAAATCGGCGGGTG | 1 |
| GGAAATCGGCGGATG | 1 |
| GGCAATCGGCGGGTG | 1 |
| CAGAGGAGTCTCTAG | 1 |
| CAAGACCGGGATATG | 1 |
| AATTATTAGTCGATG | 2 |
| GCTTAGTGAGTGATG | 1 |
| AGAAGACAAGTGGTG | 1 |
| GGTTGAAGGGGGGCG | 1 |
| ACATTATGAGGGTCG | 2 |
| AGAGTAAGTGAGGTG | 1 |
| AATTGTGTTCGGTGG | 1 |
| GTGGCTATGAATTTG | 1 |

The splicing activity of the first 30 sequences was assessed by stable transfection assays. Additional sequences validated through transiently transfection assays are indicated (Y= Tested and N=Not tested). The occurrence of each sequence from the sequencing of 226 clones is also noted.*~30% of the recovered 15-mers were recovered more than once, which is likely due to assay conditions where enriched cell populations were grown for several weeks and then examined for sequence content.

## Table S6. Significantly enriched ISRE n-mers

| Field | Description |
|---|---|
| n-mer | The n-mer |
| Length | N-mer length |
| Count(ISS) | Counts observed in ISS sample |
| Count(RS) | Counts observed in RS sample |
| N(ISS) | Total counts performed in ISS sample |
| N(RS) | Total counts performed in RS sample |
| P(ISS) | Probability of n-mer in ISS sample |
| P(RS) | Probability of n-mer in RS sample |
| CI(low) | "Lower cutoff for confidence interval (alpha= 0.02, two tailed)" |
| CI(high) | "Upper cutoff for confidence interval (alpha= 0.02, two tailed)" |
| Z | Z-score |
| P(Z) | P-value based on Z-score |

| n-mer | length | count(ISS) | count(RS) | N(ISS) | N(RS) | P(ISS) | P(RS) | CI(low) | CI(high) | Z | P(Z) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AAGG | 4 | 19 | 28158 | 500 | 1799620 | 0.038 | 0.015647 | 0.009956 | 0.024508 | 4.02621 | 5.67E-05 |
| AAGT | 4 | 10 | 20912 | 500 | 1799620 | 0.02 | 0.01162 | 0.006882 | 0.019556 | 1.74801 | 0.080462 |
| AGAA | 4 | 11 | 20909 | 750 | 2699430 | 0.014667 | 0.007746 | 0.004585 | 0.013058 | 2.16143 | 0.030662 |
| AGAG | 4 | 14 | 28151 | 1000 | 3599240 | 0.014 | 0.007821 | 0.004971 | 0.012286 | 2.21742 | 0.026594 |
| AGGC | 4 | 32 | 49223 | 500 | 1799620 | 0.064 | 0.027352 | 0.019439 | 0.03836 | 5.02257 | 5.10E-07 |
| AGGG | 4 | 21 | 28200 | 500 | 1799620 | 0.042 | 0.01567 | 0.009975 | 0.024537 | 4.73884 | 2.15E-06 |
| AGTA | 4 | 11 | 20932 | 750 | 2699430 | 0.014667 | 0.007754 | 0.004591 | 0.013069 | 2.15759 | 0.03096 |
| AGTG | 4 | 24 | 28201 | 500 | 1799620 | 0.048 | 0.015671 | 0.009975 | 0.024537 | 5.8182 | 5.95E-09 |
| ATGG | 4 | 22 | 28223 | 500 | 1799620 | 0.044 | 0.015683 | 0.009985 | 0.024552 | 5.09436 | 3.50E-07 |
| CCGA | 4 | 11 | 21079 | 500 | 1799620 | 0.022 | 0.011713 | 0.006951 | 0.019672 | 2.13739 | 0.032566 |
| CCTC | 4 | 10 | 21415 | 750 | 2699430 | 0.013333 | 0.007933 | 0.004724 | 0.013292 | 1.66665 | 0.095584 |
| CGGG | 4 | 18 | 28018 | 500 | 1799620 | 0.036 | 0.015569 | 0.009896 | 0.024414 | 3.68909 | 0.000225 |
| CGGT | 4 | 10 | 21095 | 500 | 1799620 | 0.02 | 0.011722 | 0.006958 | 0.019683 | 1.71939 | 0.085544 |
| GACC | 4 | 12 | 21062 | 500 | 1799620 | 0.024 | 0.011704 | 0.006944 | 0.01966 | 2.55586 | 0.010593 |
| GAGG | 4 | 26 | 28015 | 750 | 2699430 | 0.034667 | 0.010378 | 0.006591 | 0.016305 | 6.56052 | 5.36E-11 |
| GAGT | 4 | 16 | 21335 | 500 | 1799620 | 0.032 | 0.011855 | 0.007058 | 0.019849 | 4.16024 | 3.18E-05 |
| GATG | 4 | 13 | 28209 | 750 | 2699430 | 0.017333 | 0.01045 | 0.006647 | 0.016392 | 1.85334 | 0.063834 |
| GCGG | 4 | 18 | 28223 | 750 | 2699430 | 0.024 | 0.010455 | 0.006651 | 0.016398 | 3.64572 | 0.000267 |
| GGAG | 4 | 19 | 28339 | 750 | 2699430 | 0.025333 | 0.010498 | 0.006685 | 0.016451 | 3.98487 | 6.75E-05 |
| GGCG | 4 | 15 | 28442 | 750 | 2699430 | 0.02 | 0.010536 | 0.006715 | 0.016497 | 2.53766 | 0.01116 |
| GGCT | 4 | 11 | 21106 | 500 | 1799620 | 0.022 | 0.011728 | 0.006962 | 0.019691 | 2.13292 | 0.032931 |
| GGGC | 4 | 43 | 49067 | 500 | 1799620 | 0.086 | 0.027265 | 0.019367 | 0.038259 | 8.06106 | 7.56E-16 |
| GGGG | 4 | 42 | 28105 | 2000 | 7198480 | 0.021 | 0.003904 | 0.002479 | 0.006143 | 12.2506 | 1.67E-34 |
| GGGT | 4 | 11 | 21111 | 500 | 1799620 | 0.022 | 0.011731 | 0.006965 | 0.019694 | 2.1321 | 0.032999 |
| GGTG | 4 | 21 | 27984 | 750 | 2699430 | 0.028 | 0.010367 | 0.006583 | 0.016291 | 4.76593 | 1.88E-06 |
| GGTT | 4 | 10 | 21126 | 500 | 1799620 | 0.02 | 0.011739 | 0.006971 | 0.019705 | 1.71456 | 0.086425 |
| GTAG | 4 | 24 | 55917 | 750 | 2699430 | 0.032 | 0.020714 | 0.015022 | 0.028501 | 2.16957 | 0.03004 |
| GTGA | 4 | 20 | 49127 | 500 | 1799620 | 0.04 | 0.027299 | 0.019395 | 0.038298 | 1.74258 | 0.081408 |
| GTGG | 4 | 44 | 56695 | 750 | 2699430 | 0.058667 | 0.021003 | 0.015265 | 0.028833 | 7.19059 | 6.45E-13 |
| GTGT | 4 | 26 | 49154 | 1000 | 3599240 | 0.026 | 0.013657 | 0.009687 | 0.019221 | 3.36222 | 0.000773 |
| TAGA | 4 | 14 | 27883 | 500 | 1799620 | 0.028 | 0.015494 | 0.009837 | 0.024323 | 2.26367 | 0.023595 |
| TAGG | 4 | 16 | 35047 | 500 | 1799620 | 0.032 | 0.019475 | 0.012986 | 0.02911 | 2.02634 | 0.04273 |
| TAGT | 4 | 13 | 27981 | 750 | 2699430 | 0.017333 | 0.010366 | 0.006582 | 0.016289 | 1.88362 | 0.059616 |
| TATG | 4 | 15 | 35246 | 500 | 1799620 | 0.03 | 0.019585 | 0.013075 | 0.029241 | 1.68025 | 0.09291 |
| TGAG | 4 | 15 | 35211 | 500 | 1799620 | 0.03 | 0.019566 | 0.013059 | 0.029218 | 1.6842 | 0.092142 |
| TGGA | 4 | 15 | 28342 | 500 | 1799620 | 0.03 | 0.015749 | 0.010036 | 0.024633 | 2.55883 | 0.010502 |
| TGGC | 4 | 40 | 56418 | 500 | 1799620 | 0.08 | 0.03135 | 0.022796 | 0.042973 | 6.24044 | 4.36E-10 |
| TGGG | 4 | 24 | 35134 | 500 | 1799620 | 0.048 | 0.019523 | 0.013025 | 0.029167 | 4.60088 | 4.21E-06 |
| TGTG | 4 | 17 | 34998 | 1000 | 3599240 | 0.017 | 0.009724 | 0.006474 | 0.014582 | 2.34428 | 0.019064 |
| AAAGA | 5 | 4 | 4874 | 500 | 1799620 | 0.008 | 0.002708 | 0.000946 | 0.007727 | 2.2758 | 0.022858 |
| AACAC | 5 | 4 | 4855 | 375 | 1349720 | 0.010667 | 0.003597 | 0.001255 | 0.010269 | 2.28582 | 0.022265 |
| AAGAC | 5 | 5 | 4829 | 375 | 1349720 | 0.013333 | 0.003578 | 0.001245 | 0.01024 | 3.16238 | 0.001565 |
| AAGGC | 5 | 8 | 11863 | 375 | 1349720 | 0.021333 | 0.008789 | 0.004406 | 0.017456 | 2.60165 | 0.009278 |
| AAGGG | 5 | 7 | 6614 | 375 | 1349720 | 0.018667 | 0.0049 | 0.001968 | 0.012151 | 3.81559 | 0.000136 |
| AAGTG | 5 | 7 | 6506 | 375 | 1349720 | 0.018667 | 0.00482 | 0.001922 | 0.012037 | 3.8693 | 0.000109 |
| AATTG | 5 | 6 | 6471 | 375 | 1349720 | 0.016 | 0.004794 | 0.001907 | 0.012 | 3.14001 | 0.001689 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ACCAA | 5 | 4 | 4776 | 500 | 1799620 | 0.008 | 0.002654 | 0.000918 | 0.007646 | 2.32261 | 0.0202 |
| ACCGT | 5 | 4 | 4779 | 375 | 1349720 | 0.010667 | 0.003541 | 0.001225 | 0.010186 | 2.32219 | 0.020223 |
| ACCTC | 5 | 5 | 4900 | 375 | 1349720 | 0.013333 | 0.00363 | 0.001272 | 0.010318 | 3.12256 | 0.001793 |
| AGAAT | 5 | 5 | 4745 | 375 | 1349720 | 0.013333 | 0.003516 | 0.001213 | 0.010149 | 3.21047 | 0.001325 |
| AGACC | 5 | 4 | 4807 | 375 | 1349720 | 0.010667 | 0.003561 | 0.001236 | 0.010216 | 2.30871 | 0.02096 |
| AGAGG | 5 | 7 | 6553 | 375 | 1349720 | 0.018667 | 0.004855 | 0.001942 | 0.012087 | 3.84579 | 0.00012 |
| AGAGT | 5 | 4 | 4903 | 375 | 1349720 | 0.010667 | 0.003633 | 0.001273 | 0.010321 | 2.26321 | 0.023623 |
| AGGGC | 5 | 7 | 11795 | 375 | 1349720 | 0.018667 | 0.008739 | 0.004372 | 0.01739 | 2.06499 | 0.038924 |
| AGGGG | 5 | 8 | 6582 | 375 | 1349720 | 0.021333 | 0.004877 | 0.001954 | 0.012117 | 4.57194 | 4.83E-06 |
| AGGTG | 5 | 5 | 6563 | 375 | 1349720 | 0.013333 | 0.004863 | 0.001946 | 0.012097 | 2.35725 | 0.018411 |
| AGTGA | 5 | 6 | 4826 | 500 | 1799620 | 0.012 | 0.002682 | 0.000932 | 0.007687 | 4.02655 | 5.66E-05 |
| AGTGG | 5 | 10 | 6699 | 375 | 1349720 | 0.026667 | 0.004963 | 0.002004 | 0.01224 | 5.97609 | 2.29E-09 |
| ATATG | 5 | 6 | 6591 | 375 | 1349720 | 0.016 | 0.004883 | 0.001958 | 0.012127 | 3.08677 | 0.002023 |
| ATGGC | 5 | 14 | 11824 | 375 | 1349720 | 0.037333 | 0.00876 | 0.004387 | 0.017418 | 5.93424 | 2.95E-09 |
| ATTAT | 5 | 4 | 4997 | 625 | 2249520 | 0.0064 | 0.002221 | 0.000785 | 0.006267 | 2.21807 | 0.02655 |
| CAAGG | 5 | 10 | 6657 | 375 | 1349720 | 0.026667 | 0.004932 | 0.001986 | 0.012196 | 6.00339 | 1.93E-09 |
| CACCT | 5 | 4 | 4692 | 375 | 1349720 | 0.010667 | 0.003476 | 0.001192 | 0.01009 | 2.36473 | 0.018043 |
| CAGTG | 5 | 6 | 6656 | 375 | 1349720 | 0.016 | 0.004931 | 0.001985 | 0.012195 | 3.05845 | 0.002225 |
| CCAAG | 5 | 7 | 6510 | 375 | 1349720 | 0.018667 | 0.004823 | 0.001924 | 0.012041 | 3.86729 | 0.00011 |
| CCGAG | 5 | 5 | 6559 | 375 | 1349720 | 0.013333 | 0.00486 | 0.001944 | 0.012093 | 2.35878 | 0.018335 |
| CCGAT | 5 | 4 | 4808 | 375 | 1349720 | 0.010667 | 0.003562 | 0.001237 | 0.010217 | 2.30823 | 0.020987 |
| CCTCC | 5 | 4 | 4871 | 625 | 2249520 | 0.0064 | 0.002165 | 0.000756 | 0.006184 | 2.2766 | 0.02281 |
| CGAGT | 5 | 4 | 4886 | 375 | 1349720 | 0.010667 | 0.00362 | 0.001266 | 0.010303 | 2.27118 | 0.023136 |
| CGATG | 5 | 8 | 6496 | 375 | 1349720 | 0.021333 | 0.004813 | 0.001918 | 0.012027 | 4.61974 | 3.84E-06 |
| CGGGA | 5 | 4 | 4850 | 375 | 1349720 | 0.010667 | 0.003593 | 0.001253 | 0.010263 | 2.28819 | 0.022127 |
| CGGGG | 5 | 5 | 6625 | 375 | 1349720 | 0.013333 | 0.004908 | 0.001972 | 0.012162 | 2.33353 | 0.01962 |
| CGGGT | 5 | 4 | 4712 | 375 | 1349720 | 0.010667 | 0.003491 | 0.0012 | 0.010112 | 2.35486 | 0.01853 |
| CGGTG | 5 | 5 | 6503 | 375 | 1349720 | 0.013333 | 0.004818 | 0.001921 | 0.012034 | 2.38046 | 0.017291 |
| CGGTT | 5 | 4 | 4834 | 375 | 1349720 | 0.010667 | 0.003582 | 0.001246 | 0.010246 | 2.2958 | 0.021688 |
| CTAGG | 5 | 5 | 6483 | 375 | 1349720 | 0.013333 | 0.004803 | 0.001912 | 0.012013 | 2.38826 | 0.016928 |
| CTCGG | 5 | 5 | 6722 | 375 | 1349720 | 0.013333 | 0.00498 | 0.002014 | 0.012264 | 2.29696 | 0.021621 |
| CTGGG | 5 | 6 | 6289 | 375 | 1349720 | 0.016 | 0.00466 | 0.001831 | 0.011808 | 3.22319 | 0.001268 |
| GACCA | 5 | 4 | 4763 | 375 | 1349720 | 0.010667 | 0.003529 | 0.001219 | 0.010168 | 2.32994 | 0.019809 |
| GAGGA | 5 | 5 | 4770 | 625 | 2249520 | 0.008 | 0.00212 | 0.000733 | 0.006118 | 3.19377 | 0.001404 |
| GAGGC | 5 | 10 | 11892 | 375 | 1349720 | 0.026667 | 0.008811 | 0.00442 | 0.017485 | 3.69855 | 0.000217 |
| GAGGG | 5 | 8 | 6569 | 500 | 1799620 | 0.016 | 0.00365 | 0.001461 | 0.009089 | 4.57631 | 4.73E-06 |
| GAGTA | 5 | 4 | 4835 | 375 | 1349720 | 0.010667 | 0.003582 | 0.001247 | 0.010247 | 2.29532 | 0.021715 |
| GAGTG | 5 | 7 | 6725 | 500 | 1799620 | 0.014 | 0.003737 | 0.001511 | 0.009212 | 3.7592 | 0.00017 |
| GATAT | 5 | 4 | 4827 | 375 | 1349720 | 0.010667 | 0.003576 | 0.001244 | 0.010238 | 2.29914 | 0.021497 |
| GATGG | 5 | 10 | 6628 | 500 | 1799620 | 0.02 | 0.003683 | 0.00148 | 0.009136 | 6.01865 | 1.76E-09 |
| GCACC | 5 | 4 | 4837 | 375 | 1349720 | 0.010667 | 0.003584 | 0.001248 | 0.010249 | 2.29437 | 0.021769 |
| GCGGG | 5 | 10 | 6456 | 500 | 1799620 | 0.02 | 0.003587 | 0.001425 | 0.009 | 6.1336 | 8.59E-10 |
| GCGGT | 5 | 4 | 4843 | 375 | 1349720 | 0.010667 | 0.003588 | 0.00125 | 0.010256 | 2.29151 | 0.021934 |
| GCTGG | 5 | 8 | 6494 | 500 | 1799620 | 0.016 | 0.003609 | 0.001437 | 0.00903 | 4.61807 | 3.87E-06 |
| GGACC | 5 | 5 | 4794 | 375 | 1349720 | 0.013333 | 0.003552 | 0.001231 | 0.010202 | 3.18228 | 0.001461 |
| GGAGG | 5 | 7 | 6609 | 625 | 2249520 | 0.0112 | 0.002938 | 0.001179 | 0.007303 | 3.8143 | 0.000137 |
| GGAGT | 5 | 4 | 4919 | 375 | 1349720 | 0.010667 | 0.003644 | 0.001279 | 0.010339 | 2.25573 | 0.024087 |
| GGATA | 5 | 4 | 4790 | 375 | 1349720 | 0.010667 | 0.003549 | 0.00123 | 0.010198 | 2.31688 | 0.02051 |
| GGCGG | 5 | 11 | 6706 | 625 | 2249520 | 0.0176 | 0.002981 | 0.001204 | 0.007364 | 6.69827 | 2.11E-11 |
| GGCTA | 5 | 5 | 4800 | 375 | 1349720 | 0.013333 | 0.003556 | 0.001233 | 0.010209 | 3.17886 | 0.001479 |
| GGGAG | 5 | 6 | 6735 | 500 | 1799620 | 0.012 | 0.003742 | 0.001514 | 0.00922 | 3.02259 | 0.002506 |
| GGGCA | 5 | 4 | 4761 | 375 | 1349720 | 0.010667 | 0.003527 | 0.001219 | 0.010166 | 2.33091 | 0.019758 |
| GGGCG | 5 | 5 | 6655 | 500 | 1799620 | 0.01 | 0.003698 | 0.001488 | 0.009157 | 2.32071 | 0.020302 |
| GGGGC | 5 | 21 | 11858 | 375 | 1349720 | 0.056 | 0.008786 | 0.004404 | 0.017451 | 9.78905 | 1.25E-22 |
| GGGGG | 5 | 17 | 6589 | 1875 | 6748580 | 0.009067 | 0.000976 | 0.000391 | 0.002435 | 11.2025 | 3.96E-29 |
| GGGTG | 5 | 7 | 6537 | 500 | 1799620 | 0.014 | 0.003632 | 0.001451 | 0.009064 | 3.85142 | 0.000117 |
| GGTGG | 5 | 14 | 6731 | 625 | 2249520 | 0.0224 | 0.002992 | 0.00121 | 0.00738 | 8.87407 | 7.05E-19 |
| GGTTG | 5 | 6 | 6641 | 500 | 1799620 | 0.012 | 0.00369 | 0.001484 | 0.009146 | 3.06306 | 0.002191 |
| GTAAA | 5 | 7 | 11813 | 375 | 1349720 | 0.018667 | 0.008752 | 0.004381 | 0.017407 | 2.06066 | 0.039336 |
| GTAGA | 5 | 10 | 11703 | 375 | 1349720 | 0.026667 | 0.008671 | 0.004327 | 0.0173 | 3.75724 | 0.000172 |
| GTGAG | 5 | 9 | 13633 | 500 | 1799620 | 0.018 | 0.007575 | 0.003971 | 0.014404 | 2.68747 | 0.0072 |
| GTGGC | 5 | 20 | 19065 | 375 | 1349720 | 0.053333 | 0.014125 | 0.008168 | 0.024319 | 6.4307 | 1.27E-10 |
| GTGGG | 5 | 12 | 13696 | 500 | 1799620 | 0.024 | 0.007611 | 0.003995 | 0.01445 | 4.21516 | 2.50E-05 |
| GTGTA | 5 | 8 | 11859 | 375 | 1349720 | 0.021333 | 0.008786 | 0.004404 | 0.017452 | 2.6027 | 0.009249 |
| GTGTT | 5 | 7 | 11954 | 375 | 1349720 | 0.018667 | 0.008857 | 0.004451 | 0.017546 | 2.027 | 0.042663 |
| GTTGG | 5 | 8 | 13697 | 500 | 1799620 | 0.016 | 0.007611 | 0.003996 | 0.014451 | 2.15776 | 0.030947 |
| TAAGT | 5 | 5 | 6561 | 500 | 1799620 | 0.01 | 0.003646 | 0.001459 | 0.009083 | 2.35658 | 0.018444 |
| TAATT | 5 | 5 | 6619 | 500 | 1799620 | 0.01 | 0.003678 | 0.001477 | 0.009129 | 2.33437 | 0.019576 |
| TAGAA | 5 | 6 | 6486 | 375 | 1349720 | 0.016 | 0.004805 | 0.001913 | 0.012016 | 3.13329 | 0.001729 |
| TAGAG | 5 | 7 | 8315 | 375 | 1349720 | 0.018667 | 0.006161 | 0.002716 | 0.013912 | 3.09377 | 0.001976 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TAGGC | 5 | 8 | 13605 | 375 | 1349720 | 0.021333 | 0.01008 | 0.005283 | 0.01915 | 2.18095 | 0.029187 |
| TAGTA | 5 | 5 | 6516 | 625 | 2249520 | 0.008 | 0.002897 | 0.001155 | 0.007244 | 2.37311 | 0.017639 |
| TATGA | 5 | 5 | 6628 | 375 | 1349720 | 0.013333 | 0.004911 | 0.001974 | 0.012166 | 2.33239 | 0.01968 |
| TATGG | 5 | 9 | 8411 | 375 | 1349720 | 0.024 | 0.006232 | 0.00276 | 0.01401 | 4.37004 | 1.24E-05 |
| TCAGT | 5 | 5 | 6748 | 500 | 1799620 | 0.01 | 0.00375 | 0.001518 | 0.009231 | 2.28584 | 0.022264 |
| TCCGA | 5 | 6 | 6654 | 375 | 1349720 | 0.016 | 0.00493 | 0.001985 | 0.012193 | 3.05931 | 0.002218 |
| TGAGG | 5 | 9 | 8379 | 375 | 1349720 | 0.024 | 0.006208 | 0.002745 | 0.013977 | 4.38416 | 1.16E-05 |
| TGGAC | 5 | 5 | 6631 | 375 | 1349720 | 0.013333 | 0.004913 | 0.001975 | 0.012169 | 2.33125 | 0.01974 |
| TGGCT | 5 | 5 | 6581 | 500 | 1799620 | 0.01 | 0.003657 | 0.001465 | 0.009099 | 2.34889 | 0.018829 |
| TGGGC | 5 | 10 | 13583 | 375 | 1349720 | 0.026667 | 0.010064 | 0.005271 | 0.019129 | 3.22006 | 0.001282 |
| TGGGG | 5 | 12 | 8309 | 375 | 1349720 | 0.032 | 0.006156 | 0.002713 | 0.013906 | 6.39366 | 1.62E-10 |
| TGTGG | 5 | 6 | 8242 | 375 | 1349720 | 0.016 | 0.006106 | 0.002683 | 0.013837 | 2.45835 | 0.013958 |
| TGTGT | 5 | 6 | 6606 | 1000 | 3599240 | 0.006 | 0.001835 | 0.000736 | 0.004568 | 3.07548 | 0.002102 |
| TTAGT | 5 | 5 | 6596 | 500 | 1799620 | 0.01 | 0.003665 | 0.00147 | 0.009111 | 2.34315 | 0.019122 |
| TTATG | 5 | 6 | 8434 | 375 | 1349720 | 0.016 | 0.006249 | 0.00277 | 0.014033 | 2.39546 | 0.0166 |
| TTGTG | 5 | 7 | 8242 | 375 | 1349720 | 0.018667 | 0.006106 | 0.002683 | 0.013837 | 3.12078 | 0.001804 |
| AACACC | 6 | 3 | 1051 | 375 | 1349720 | 0.008 | 0.000779 | 0.000104 | 0.005807 | 5.00613 | 5.55E-07 |
| AAGACC | 6 | 4 | 1099 | 375 | 1349720 | 0.010667 | 0.000814 | 0.000112 | 0.005869 | 6.67681 | 2.44E-11 |
| AAGGGC | 6 | 3 | 2802 | 375 | 1349720 | 0.008 | 0.002076 | 0.000539 | 0.007955 | 2.51905 | 0.011767 |
| AAGGGG | 6 | 3 | 1567 | 375 | 1349720 | 0.008 | 0.001161 | 0.000208 | 0.006465 | 3.88538 | 0.000102 |
| AAGTGG | 6 | 3 | 1570 | 375 | 1349720 | 0.008 | 0.001163 | 0.000208 | 0.006469 | 3.88041 | 0.000104 |
| AATCGG | 6 | 3 | 1493 | 375 | 1349720 | 0.008 | 0.001106 | 0.000191 | 0.006372 | 4.01211 | 6.02E-05 |
| AATTGT | 6 | 3 | 1098 | 375 | 1349720 | 0.008 | 0.000814 | 0.000112 | 0.005868 | 4.87458 | 1.09E-06 |
| ACACCT | 6 | 3 | 1070 | 375 | 1349720 | 0.008 | 0.000793 | 0.000107 | 0.005832 | 4.95198 | 7.35E-07 |
| ACACGT | 6 | 3 | 1118 | 375 | 1349720 | 0.008 | 0.000828 | 0.000116 | 0.005894 | 4.82098 | 1.43E-06 |
| ACCAAG | 6 | 3 | 1566 | 375 | 1349720 | 0.008 | 0.00116 | 0.000207 | 0.006464 | 3.88703 | 0.000101 |
| ACCGTT | 6 | 3 | 1118 | 375 | 1349720 | 0.008 | 0.000828 | 0.000116 | 0.005894 | 4.82098 | 1.43E-06 |
| ACCTCC | 6 | 3 | 1103 | 375 | 1349720 | 0.008 | 0.000817 | 0.000113 | 0.005875 | 4.86105 | 1.17E-06 |
| AGAGGA | 6 | 3 | 1082 | 375 | 1349720 | 0.008 | 0.000802 | 0.000109 | 0.005847 | 4.91846 | 8.72E-07 |
| AGAGTG | 6 | 3 | 1607 | 375 | 1349720 | 0.008 | 0.001191 | 0.000217 | 0.006515 | 3.82024 | 0.000133 |
| AGGGAG | 6 | 3 | 1551 | 500 | 1799620 | 0.006 | 0.000862 | 0.000153 | 0.004839 | 3.91151 | 9.17E-05 |
| AGGGGC | 6 | 5 | 2965 | 375 | 1349720 | 0.013333 | 0.002197 | 0.00059 | 0.008145 | 4.60244 | 4.18E-06 |
| AGGTGG | 6 | 3 | 1554 | 375 | 1349720 | 0.008 | 0.001151 | 0.000205 | 0.006449 | 3.90704 | 9.34E-05 |
| AGTAGC | 6 | 3 | 2827 | 375 | 1349720 | 0.008 | 0.002095 | 0.000547 | 0.007984 | 2.50008 | 0.012416 |
| AGTGAG | 6 | 4 | 1587 | 500 | 1799620 | 0.008 | 0.000882 | 0.000159 | 0.004873 | 5.35549 | 8.53E-08 |
| AGTGGC | 6 | 4 | 2920 | 375 | 1349720 | 0.010667 | 0.002163 | 0.000576 | 0.008093 | 3.54163 | 0.000398 |
| AGTGGG | 6 | 4 | 1585 | 375 | 1349720 | 0.010667 | 0.001174 | 0.000212 | 0.006487 | 5.36048 | 8.30E-08 |
| AGTGTA | 6 | 3 | 1106 | 375 | 1349720 | 0.008 | 0.000819 | 0.000114 | 0.005879 | 4.85298 | 1.22E-06 |
| ATATGG | 6 | 6 | 1546 | 375 | 1349720 | 0.016 | 0.001145 | 0.000203 | 0.006439 | 8.48791 | 2.10E-17 |
| ATCGGC | 6 | 3 | 2780 | 375 | 1349720 | 0.008 | 0.00206 | 0.000533 | 0.007929 | 2.53593 | 0.011215 |
| ATTGTG | 6 | 3 | 1470 | 375 | 1349720 | 0.008 | 0.001089 | 0.000186 | 0.006343 | 4.05327 | 5.05E-05 |
| CAAGGC | 6 | 5 | 2932 | 375 | 1349720 | 0.013333 | 0.002172 | 0.00058 | 0.008107 | 4.63833 | 3.51E-06 |
| CAAGGG | 6 | 3 | 1539 | 375 | 1349720 | 0.008 | 0.00114 | 0.000201 | 0.00643 | 3.93234 | 8.41E-05 |
| CACCTC | 6 | 3 | 1030 | 375 | 1349720 | 0.008 | 0.000763 | 0.0001 | 0.00578 | 5.06762 | 4.03E-07 |
| CCAAGG | 6 | 7 | 1610 | 375 | 1349720 | 0.018667 | 0.001193 | 0.000217 | 0.006519 | 9.78205 | 1.34E-22 |
| CCAGGC | 6 | 3 | 2816 | 375 | 1349720 | 0.008 | 0.002086 | 0.000544 | 0.007971 | 2.5084 | 0.012128 |
| CCGATG | 6 | 3 | 1597 | 375 | 1349720 | 0.008 | 0.001183 | 0.000214 | 0.006502 | 3.83631 | 0.000125 |
| CCTCGG | 6 | 3 | 1594 | 375 | 1349720 | 0.008 | 0.001181 | 0.000214 | 0.006499 | 3.84116 | 0.000122 |
| CGATGG | 6 | 7 | 1576 | 375 | 1349720 | 0.018667 | 0.001168 | 0.00021 | 0.006476 | 9.90067 | 4.14E-23 |
| CGCTGG | 6 | 3 | 1512 | 375 | 1349720 | 0.008 | 0.00112 | 0.000195 | 0.006396 | 3.97876 | 6.93E-05 |
| CGGGAT | 6 | 3 | 1101 | 375 | 1349720 | 0.008 | 0.000816 | 0.000113 | 0.005872 | 4.86645 | 1.14E-06 |
| CGGGGC | 6 | 4 | 2873 | 375 | 1349720 | 0.010667 | 0.002129 | 0.000561 | 0.008038 | 3.585 | 0.000337 |
| CGGGTG | 6 | 4 | 1488 | 375 | 1349720 | 0.010667 | 0.001102 | 0.00019 | 0.006366 | 5.57367 | 2.49E-08 |
| CGGTGG | 6 | 4 | 1557 | 375 | 1349720 | 0.010667 | 0.001154 | 0.000205 | 0.006452 | 5.42011 | 5.96E-08 |
| CGGTTG | 6 | 3 | 1558 | 375 | 1349720 | 0.008 | 0.001154 | 0.000206 | 0.006454 | 3.90035 | 9.61E-05 |
| CTAGGC | 6 | 3 | 2821 | 375 | 1349720 | 0.008 | 0.00209 | 0.000545 | 0.007977 | 2.50462 | 0.012258 |
| CTCGGG | 6 | 3 | 1591 | 375 | 1349720 | 0.008 | 0.001179 | 0.000213 | 0.006495 | 3.84602 | 0.00012 |
| CTGGGC | 6 | 4 | 2641 | 375 | 1349720 | 0.010667 | 0.001957 | 0.000491 | 0.007766 | 3.81387 | 0.000137 |
| GACATG | 6 | 3 | 1553 | 375 | 1349720 | 0.008 | 0.001151 | 0.000204 | 0.006447 | 3.90871 | 9.28E-05 |
| GACCAA | 6 | 3 | 1126 | 375 | 1349720 | 0.008 | 0.000834 | 0.000117 | 0.005904 | 4.79991 | 1.59E-06 |
| GACCTG | 6 | 3 | 1547 | 375 | 1349720 | 0.008 | 0.001146 | 0.000203 | 0.00644 | 3.91881 | 8.90E-05 |
| GAGTGG | 6 | 4 | 1582 | 375 | 1349720 | 0.010667 | 0.001172 | 0.000211 | 0.006484 | 5.3668 | 8.01E-08 |
| GATATG | 6 | 4 | 1550 | 375 | 1349720 | 0.010667 | 0.001148 | 0.000204 | 0.006444 | 5.43525 | 5.47E-08 |
| GATGGC | 6 | 8 | 2823 | 375 | 1349720 | 0.021333 | 0.002092 | 0.000546 | 0.00798 | 8.14456 | 3.81E-16 |
| GCAAGG | 6 | 3 | 1535 | 375 | 1349720 | 0.008 | 0.001137 | 0.0002 | 0.006425 | 3.93915 | 8.18E-05 |
| GCTGGA | 6 | 3 | 1087 | 375 | 1349720 | 0.008 | 0.000805 | 0.00011 | 0.005854 | 4.90465 | 9.36E-07 |
| GCTGGG | 6 | 4 | 1480 | 375 | 1349720 | 0.010667 | 0.001097 | 0.000188 | 0.006356 | 5.59212 | 2.24E-08 |
| GGAGGC | 6 | 4 | 2891 | 375 | 1349720 | 0.010667 | 0.002142 | 0.000567 | 0.008059 | 3.56828 | 0.000359 |
| GGAGGG | 6 | 3 | 1564 | 500 | 1799620 | 0.006 | 0.000869 | 0.000155 | 0.004851 | 3.88979 | 0.0001 |
| GGATAT | 6 | 3 | 1087 | 375 | 1349720 | 0.008 | 0.000805 | 0.00011 | 0.005854 | 4.90465 | 9.36E-07 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GGCGGG | 6 | 6 | 1527 | 500 | 1799620 | 0.012 | 0.000849 | 0.000149 | 0.004816 | 8.54717 | 1.26E-17 |
| GGCGGT | 6 | 3 | 1116 | 375 | 1349720 | 0.008 | 0.000827 | 0.000116 | 0.005892 | 4.82628 | 1.39E-06 |
| GGCTAG | 6 | 3 | 1543 | 375 | 1349720 | 0.008 | 0.001143 | 0.000202 | 0.006435 | 3.92556 | 8.65E-05 |
| GGCTGG | 6 | 3 | 1475 | 500 | 1799620 | 0.006 | 0.00082 | 0.00014 | 0.004767 | 4.04369 | 5.26E-05 |
| GGGAGG | 6 | 3 | 1494 | 500 | 1799620 | 0.006 | 0.00083 | 0.000144 | 0.004785 | 4.00978 | 6.08E-05 |
| GGGATA | 6 | 3 | 1099 | 375 | 1349720 | 0.008 | 0.000814 | 0.000112 | 0.005869 | 4.87187 | 1.11E-06 |
| GGGCGG | 6 | 3 | 1606 | 500 | 1799620 | 0.006 | 0.000892 | 0.000162 | 0.004891 | 3.82127 | 0.000133 |
| GGGGCG | 6 | 3 | 1516 | 375 | 1349720 | 0.008 | 0.001123 | 0.000196 | 0.006401 | 3.97181 | 7.13E-05 |
| GGGGGC | 6 | 6 | 2780 | 375 | 1349720 | 0.016 | 0.00206 | 0.000533 | 0.007929 | 5.94795 | 2.72E-09 |
| GGGGGG | 6 | 10 | 1546 | 1750 | 6298670 | 0.005714 | 0.000245 | 4.35E-05 | 0.001385 | 14.5575 | 5.23E-48 |
| GGGTGG | 6 | 5 | 1591 | 500 | 1799620 | 0.01 | 0.000884 | 0.00016 | 0.004877 | 6.84782 | 7.50E-12 |
| GGTGGC | 6 | 10 | 2938 | 375 | 1349720 | 0.026667 | 0.002177 | 0.000581 | 0.008114 | 10.1586 | 3.03E-24 |
| GGTTGG | 6 | 3 | 1544 | 500 | 1799620 | 0.006 | 0.000858 | 0.000152 | 0.004832 | 3.92331 | 8.73E-05 |
| GTAAAG | 6 | 5 | 3287 | 375 | 1349720 | 0.013333 | 0.002435 | 0.000693 | 0.008517 | 4.27842 | 1.88E-05 |
| GTAATT | 6 | 5 | 2886 | 375 | 1349720 | 0.013333 | 0.002138 | 0.000565 | 0.008053 | 4.68928 | 2.74E-06 |
| GTACAG | 6 | 4 | 3303 | 375 | 1349720 | 0.010667 | 0.002447 | 0.000699 | 0.008535 | 3.21956 | 0.001284 |
| GTAGAA | 6 | 5 | 2839 | 375 | 1349720 | 0.013333 | 0.002103 | 0.000551 | 0.007998 | 4.74249 | 2.11E-06 |
| GTAGAG | 6 | 5 | 3255 | 375 | 1349720 | 0.013333 | 0.002412 | 0.000683 | 0.00848 | 4.30867 | 1.64E-05 |
| GTATAC | 6 | 3 | 2834 | 375 | 1349720 | 0.008 | 0.0021 | 0.000549 | 0.007992 | 2.49481 | 0.012602 |
| GTGAGG | 6 | 6 | 3343 | 375 | 1349720 | 0.016 | 0.002477 | 0.000712 | 0.008581 | 5.26377 | 1.41E-07 |
| GTGGCT | 6 | 5 | 2866 | 375 | 1349720 | 0.013333 | 0.002123 | 0.000559 | 0.00803 | 4.71178 | 2.46E-06 |
| GTGGGC | 6 | 4 | 4664 | 375 | 1349720 | 0.010667 | 0.003456 | 0.001182 | 0.01006 | 2.37863 | 0.017377 |
| GTGGGG | 6 | 8 | 3311 | 375 | 1349720 | 0.021333 | 0.002453 | 0.000701 | 0.008544 | 7.38201 | 1.56E-13 |
| GTGTAG | 6 | 4 | 3242 | 375 | 1349720 | 0.010667 | 0.002402 | 0.000679 | 0.008465 | 3.26747 | 0.001085 |
| GTTATG | 6 | 4 | 3288 | 375 | 1349720 | 0.010667 | 0.002436 | 0.000694 | 0.008518 | 3.23123 | 0.001233 |
| GTTGGA | 6 | 3 | 2833 | 375 | 1349720 | 0.008 | 0.002099 | 0.000549 | 0.007991 | 2.49556 | 0.012576 |
| TAAGTG | 6 | 4 | 1947 | 375 | 1349720 | 0.010667 | 0.001443 | 0.000299 | 0.006934 | 4.70162 | 2.58E-06 |
| TAATTG | 6 | 4 | 1923 | 375 | 1349720 | 0.010667 | 0.001425 | 0.000293 | 0.006905 | 4.73989 | 2.14E-06 |
| TAGAAT | 6 | 4 | 1532 | 375 | 1349720 | 0.010667 | 0.001135 | 0.0002 | 0.006421 | 5.47463 | 4.38E-08 |
| TAGAGG | 6 | 3 | 1937 | 375 | 1349720 | 0.008 | 0.001435 | 0.000296 | 0.006922 | 3.35564 | 0.000792 |
| TAGAGT | 6 | 4 | 1550 | 375 | 1349720 | 0.010667 | 0.001148 | 0.000204 | 0.006444 | 5.43525 | 5.47E-08 |
| TAGGGA | 6 | 3 | 1529 | 375 | 1349720 | 0.008 | 0.001133 | 0.000199 | 0.006417 | 3.9494 | 7.83E-05 |
| TAGTAG | 6 | 3 | 1945 | 625 | 2249520 | 0.0048 | 0.000865 | 0.000179 | 0.004167 | 3.34476 | 0.000824 |
| TATGAC | 6 | 3 | 1566 | 375 | 1349720 | 0.008 | 0.00116 | 0.000207 | 0.006464 | 3.88703 | 0.000101 |
| TATGGC | 6 | 5 | 3316 | 375 | 1349720 | 0.013333 | 0.002457 | 0.000703 | 0.00855 | 4.25135 | 2.12E-05 |
| TCAGTG | 6 | 4 | 2000 | 375 | 1349720 | 0.010667 | 0.001482 | 0.000312 | 0.006999 | 4.61938 | 3.85E-06 |
| TCCGAG | 6 | 3 | 1961 | 375 | 1349720 | 0.008 | 0.001453 | 0.000302 | 0.006951 | 3.32607 | 0.000881 |
| TCCGAT | 6 | 3 | 1595 | 375 | 1349720 | 0.008 | 0.001182 | 0.000214 | 0.0065 | 3.83954 | 0.000123 |
| TCGGCG | 6 | 3 | 1999 | 375 | 1349720 | 0.008 | 0.001481 | 0.000312 | 0.006998 | 3.28023 | 0.001037 |
| TCGGGG | 6 | 3 | 1959 | 375 | 1349720 | 0.008 | 0.001451 | 0.000302 | 0.006949 | 3.32851 | 0.000873 |
| TGACAT | 6 | 3 | 1546 | 375 | 1349720 | 0.008 | 0.001145 | 0.000203 | 0.006439 | 3.92049 | 8.84E-05 |
| TGAGGC | 6 | 4 | 3245 | 375 | 1349720 | 0.010667 | 0.002404 | 0.00068 | 0.008468 | 3.26508 | 0.001094 |
| TGGACC | 6 | 4 | 1539 | 375 | 1349720 | 0.010667 | 0.00114 | 0.000201 | 0.00643 | 5.45924 | 4.78E-08 |
| TGGCGG | 6 | 4 | 1988 | 375 | 1349720 | 0.010667 | 0.001473 | 0.000309 | 0.006984 | 4.63774 | 3.52E-06 |
| TGGCTG | 6 | 3 | 1996 | 500 | 1799620 | 0.006 | 0.001109 | 0.000233 | 0.005252 | 3.2832 | 0.001026 |
| TGGGGC | 6 | 6 | 3240 | 375 | 1349720 | 0.016 | 0.002401 | 0.000678 | 0.008463 | 5.3766 | 7.59E-08 |
| TGGGGG | 6 | 4 | 2007 | 375 | 1349720 | 0.010667 | 0.001487 | 0.000314 | 0.007007 | 4.60874 | 4.05E-06 |
| TGTCAG | 6 | 3 | 1914 | 375 | 1349720 | 0.008 | 0.001418 | 0.00029 | 0.006894 | 3.38444 | 0.000713 |
| TGTGGG | 6 | 3 | 1953 | 375 | 1349720 | 0.008 | 0.001447 | 0.0003 | 0.006942 | 3.33587 | 0.00085 |
| TGTGTT | 6 | 4 | 1581 | 375 | 1349720 | 0.010667 | 0.001171 | 0.000211 | 0.006482 | 5.36891 | 7.92E-08 |
| TGTTCG | 6 | 3 | 2005 | 375 | 1349720 | 0.008 | 0.001486 | 0.000314 | 0.007005 | 3.2731 | 0.001064 |
| TTATGA | 6 | 4 | 1553 | 375 | 1349720 | 0.010667 | 0.001151 | 0.000204 | 0.006447 | 5.42875 | 5.68E-08 |
| TTGGAC | 6 | 3 | 1535 | 375 | 1349720 | 0.008 | 0.001137 | 0.0002 | 0.006425 | 3.93915 | 8.18E-05 |
| TTGTGT | 6 | 3 | 1550 | 375 | 1349720 | 0.008 | 0.001148 | 0.000204 | 0.006444 | 3.91375 | 9.09E-05 |
| TTTGCG | 6 | 3 | 2005 | 375 | 1349720 | 0.008 | 0.001486 | 0.000314 | 0.007005 | 3.2731 | 0.001064 |

**Table S7. GCCS clusters derived from the ISRE enriched n-mers**

| Field | Description |
|---|---|
| n-mer | The n-mer (4-6mers) |
| clustID | ClusterID |
| GCS | Greatest Common Substring |
| Len | n-mer length |
| aligned | Aligned n-mers |
| wWeight | Edge weight |
| count | Count of n-mer in ISS dataset |
| Zscore | Z-score for n-mer |
| round | Clustering round in which produced cluster |
| vDegree | Vertex degree (number of other vertices attached) |
| TA | Association score |

| n-mer | clustID | GCS | len | aligned | wWeight | count | Zscore | round | vDegree | TA |
|---|---|---|---|---|---|---|---|---|---|---|
| AGGTG | 1 | GGTG | 5 | 'AGGTG-- | 2.35725 | 5 | 2.35725 | 1 | 1 | 0 |
| AGGTGG | 1 | GGTG | 6 | 'AGGTGG- | 3.90704 | 3 | 3.90704 | 1 | 5 | 0.6 |
| CGGTG | 1 | GGTG | 5 | 'CGGTG-- | 2.38046 | 5 | 2.38046 | 1 | 1 | 0 |
| CGGTGG | 1 | GGTG | 6 | 'CGGTGG- | 5.42011 | 4 | 5.42011 | 1 | 5 | 0.6 |
| GGGTG | 1 | GGTG | 5 | 'GGGTG-- | 3.85142 | 7 | 3.85142 | 1 | 1 | 0 |
| GGGTGG | 1 | GGTG | 6 | 'GGGTGG- | 6.84782 | 5 | 6.84782 | 1 | 5 | 0.6 |
| GGTGG | 1 | GGTG | 5 | '-GGTGG- | 8.87407 | 14 | 8.87407 | 1 | 4 | 1 |
| GGTGGC | 1 | GGTG | 6 | '-GGTGGC | 10.1586 | 10 | 10.1586 | 1 | 4 | 1 |
| AAGGC | 2 | AAGG | 5 | '--AAGGC- | 2.60165 | 8 | 2.60165 | 1 | 1 | 0 |
| AAGGG | 2 | AAGG | 5 | '--AAGGG- | 3.81559 | 7 | 3.81559 | 1 | 2 | 1 |
| AAGGGC | 2 | AAGG | 6 | '--AAGGGC | 2.51905 | 3 | 2.51905 | 1 | 2 | 1 |
| CAAGG | 2 | AAGG | 5 | '-CAAGG-- | 6.00339 | 10 | 6.00339 | 1 | 4 | 1 |
| CAAGGC | 2 | AAGG | 6 | '-CAAGGC- | 4.63833 | 5 | 4.63833 | 1 | 5 | 0.6 |
| CAAGGG | 2 | AAGG | 6 | '-CAAGGG- | 3.93234 | 3 | 3.93234 | 1 | 6 | 0.47 |
| CCAAGG | 2 | AAGG | 6 | 'CCAAGG-- | 9.78205 | 7 | 9.78205 | 1 | 4 | 1 |
| GCAAGG | 2 | AAGG | 6 | 'GCAAGG-- | 3.93915 | 3 | 3.93915 | 1 | 4 | 1 |
| CGGCGG | 3 | GGCGG | 6 | 'CGGCGG- | 3.80021 | 3 | 3.80021 | 1 | 5 | 1 |
| GGCGG | 3 | GGCGG | 5 | '-GGCGG- | 6.69827 | 11 | 6.69827 | 1 | 5 | 1 |
| GGCGGG | 3 | GGCGG | 6 | '-GGCGGG | 8.54717 | 6 | 8.54717 | 1 | 5 | 1 |
| GGCGGT | 3 | GGCGG | 6 | '-GGCGGT | 4.82628 | 3 | 4.82628 | 1 | 5 | 1 |
| GGGCGG | 3 | GGCGG | 6 | 'GGGCGG- | 3.82127 | 3 | 3.82127 | 1 | 5 | 1 |
| TGGCGG | 3 | GGCGG | 6 | 'TGGCGG- | 4.63774 | 4 | 4.63774 | 1 | 5 | 1 |
| AGGGGC | 4 | GGGGC | 6 | 'AGGGGC- | 4.60244 | 5 | 4.60244 | 1 | 4 | 1 |
| CGGGGC | 4 | GGGGC | 6 | 'CGGGGC- | 3.585 | 4 | 3.585 | 1 | 4 | 1 |
| GGGGC | 4 | GGGGC | 5 | '-GGGGC- | 9.78905 | 21 | 9.78905 | 1 | 4 | 1 |
| GGGGCG | 4 | GGGGC | 6 | '-GGGGCG | 3.97181 | 3 | 3.97181 | 1 | 4 | 1 |
| TGGGGC | 4 | GGGGC | 6 | 'TGGGGC- | 5.3766 | 6 | 5.3766 | 1 | 4 | 1 |
| CGCTGG | 5 | GCTGG | 6 | 'CGCTGG- | 3.97876 | 3 | 3.97876 | 1 | 4 | 1 |
| GCTGG | 5 | GCTGG | 5 | '-GCTGG- | 4.61807 | 8 | 4.61807 | 1 | 4 | 1 |
| GCTGGA | 5 | GCTGG | 6 | '-GCTGGA | 4.90465 | 3 | 4.90465 | 1 | 4 | 1 |
| GCTGGG | 5 | GCTGG | 6 | '-GCTGGG | 5.59212 | 4 | 5.59212 | 1 | 4 | 1 |
| GGCTGG | 5 | GCTGG | 6 | 'GGCTGG- | 4.04369 | 3 | 4.04369 | 1 | 4 | 1 |
| AGTGGG | 6 | GTGGG | 6 | 'AGTGGG- | 5.36048 | 4 | 5.36048 | 1 | 4 | 1 |
| GTGGG | 6 | GTGGG | 5 | '-GTGGG- | 4.21516 | 12 | 4.21516 | 1 | 4 | 1 |
| GTGGGC | 6 | GTGGG | 6 | '-GTGGGC | 2.37863 | 4 | 2.37863 | 1 | 4 | 1 |
| GTGGGG | 6 | GTGGG | 6 | '-GTGGGG | 7.38201 | 8 | 7.38201 | 1 | 4 | 1 |
| TGTGGG | 6 | GTGGG | 6 | 'TGTGGG- | 3.33587 | 3 | 3.33587 | 1 | 4 | 1 |
| GTAGAG | 7 | TAGAG | 6 | 'GTAGAG- | 4.30867 | 5 | 4.30867 | 1 | 3 | 1 |
| TAGAG | 7 | TAGAG | 5 | '-TAGAG- | 3.09377 | 7 | 3.09377 | 1 | 3 | 1 |
| TAGAGG | 7 | TAGAG | 6 | '-TAGAGG | 3.35564 | 3 | 3.35564 | 1 | 3 | 1 |
| TAGAGT | 7 | TAGAG | 6 | '-TAGAGT | 5.43525 | 4 | 5.43525 | 1 | 3 | 1 |
| ATTGTG | 8 | TGTG | 6 | 'ATTGTG-- | 4.05327 | 3 | 4.05327 | 1 | 2 | 1 |
| TGTGT | 8 | TGTG | 5 | '--TGTGT- | 3.07548 | 6 | 3.07548 | 1 | 2 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| TGTGTT | 8 | TGTG | 6 | '--TGTGTT | 5.36891 | 4 | 5.36891 | 1 | 2 | 1 |
| TTGTG | 8 | TGTG | 5 | '-TTGTG-- | 3.12078 | 7 | 3.12078 | 1 | 2 | 1 |
| TTGTGT | 8 | TGTG | 6 | '-TTGTGT- | 3.91375 | 3 | 3.91375 | 1 | 4 | 0.33 |
| GGGGG | 9 | GGGGG | 5 | '-GGGGG- | 11.2025 | 17 | 11.2025 | 1 | 3 | 1 |
| GGGGGC | 9 | GGGGG | 6 | '-GGGGGC | 5.94795 | 6 | 5.94795 | 1 | 3 | 1 |
| GGGGGG | 9 | GGGGG | 6 | '-GGGGGG | 14.5575 | 10 | 14.5575 | 1 | 3 | 1 |
| TGGGGG | 9 | GGGGG | 6 | 'TGGGGG- | 4.60874 | 4 | 4.60874 | 1 | 3 | 1 |
| AAGTGG | 10 | AGTG | 6 | '-AAGTGG | 3.88041 | 3 | 3.88041 | 2 | 2 | 1 |
| AGAGTG | 10 | AGTG | 6 | 'AGAGTG- | 3.82024 | 3 | 3.82024 | 2 | 2 | 1 |
| AGTGG | 10 | AGTG | 5 | '--AGTGG | 5.97609 | 10 | 5.97609 | 2 | 2 | 1 |
| GAGTG | 10 | AGTG | 5 | '-GAGTG- | 3.7592 | 7 | 3.7592 | 2 | 2 | 1 |
| GAGTGG | 10 | AGTG | 6 | '-GAGTGG | 5.3668 | 4 | 5.3668 | 2 | 4 | 0.33 |
| AGAGG | 11 | AGG | 5 | '-AGAGG--- | 3.84579 | 7 | 3.84579 | 4 | 13 | 1 |
| AGAGGA | 11 | AGG | 6 | '-AGAGGA-- | 4.91846 | 3 | 4.91846 | 4 | 13 | 1 |
| AGGC | 11 | AGG | 4 | '---AGGC-- | 5.02257 | 32 | 5.02257 | 4 | 6 | 1 |
| AGGG | 11 | AGG | 4 | '---AGGG-- | 4.73884 | 21 | 4.73884 | 4 | 4 | 1 |
| AGGGAG | 11 | AGG | 6 | '---AGGGAG | 3.91151 | 3 | 3.91151 | 4 | 8 | 0.86 |
| CCAGGC | 11 | AGG | 6 | '-CCAGGC-- | 2.5084 | 3 | 2.5084 | 4 | 6 | 1 |
| CTAGGC | 11 | AGG | 6 | '-CTAGGC-- | 2.50462 | 3 | 2.50462 | 4 | 6 | 1 |
| GAGG | 11 | AGG | 4 | '--GAGG--- | 6.56052 | 26 | 6.56052 | 4 | 13 | 1 |
| GAGGA | 11 | AGG | 5 | '--GAGGA-- | 3.19377 | 5 | 3.19377 | 4 | 13 | 1 |
| GAGGAG | 11 | AGG | 6 | '--GAGGAG- | 5.56098 | 4 | 5.56098 | 4 | 14 | 0.92 |
| GAGGC | 11 | AGG | 5 | '--GAGGC-- | 3.69855 | 10 | 3.69855 | 4 | 17 | 0.68 |
| GAGGG | 11 | AGG | 5 | '--GAGGG-- | 4.57631 | 8 | 4.57631 | 4 | 15 | 0.83 |
| GAGGGG | 11 | AGG | 6 | '--GAGGGG- | 5.4483 | 4 | 5.4483 | 4 | 15 | 0.83 |
| GGAGG | 11 | AGG | 5 | '-GGAGG--- | 3.8143 | 7 | 3.8143 | 4 | 14 | 0.92 |
| GGAGGC | 11 | AGG | 6 | '-GGAGGC-- | 3.56828 | 4 | 3.56828 | 4 | 18 | 0.64 |
| GGAGGG | 11 | AGG | 6 | '-GGAGGG-- | 3.88979 | 3 | 3.88979 | 4 | 15 | 0.83 |
| GGGAGG | 11 | AGG | 6 | 'GGGAGG--- | 4.00978 | 3 | 4.00978 | 4 | 14 | 0.92 |
| TAGGC | 11 | AGG | 5 | '--TAGGC-- | 2.18095 | 8 | 2.18095 | 4 | 6 | 1 |
| TGAGG | 11 | AGG | 5 | '-TGAGG--- | 4.38416 | 9 | 4.38416 | 4 | 13 | 1 |
| TGAGGC | 11 | AGG | 6 | '-TGAGGC-- | 3.26508 | 4 | 3.26508 | 4 | 17 | 0.68 |
| ATATG | 12 | TATG | 5 | '-ATATG-- | 3.08677 | 6 | 3.08677 | 4 | 10 | 1 |
| ATATGG | 12 | TATG | 6 | '-ATATGG- | 8.48791 | 6 | 8.48791 | 4 | 10 | 1 |
| GATATG | 12 | TATG | 6 | 'GATATG-- | 5.43525 | 4 | 5.43525 | 4 | 10 | 1 |
| GTTATG | 12 | TATG | 6 | 'GTTATG-- | 3.23123 | 4 | 3.23123 | 4 | 10 | 1 |
| TATG | 12 | TATG | 4 | '--TATG-- | 1.68025 | 15 | 1.68025 | 4 | 10 | 1 |
| TATGA | 12 | TATG | 5 | '--TATGA- | 2.33239 | 5 | 2.33239 | 4 | 10 | 1 |
| TATGAC | 12 | TATG | 6 | '--TATGAC | 3.88703 | 3 | 3.88703 | 4 | 10 | 1 |
| TATGG | 12 | TATG | 5 | '--TATGG- | 4.37004 | 9 | 4.37004 | 4 | 10 | 1 |
| TATGGC | 12 | TATG | 6 | '--TATGGC | 4.25135 | 5 | 4.25135 | 4 | 10 | 1 |
| TTATG | 12 | TATG | 5 | '-TTATG-- | 2.39546 | 6 | 2.39546 | 4 | 10 | 1 |
| TTATGA | 12 | TATG | 6 | '-TTATGA- | 5.42875 | 4 | 5.42875 | 4 | 10 | 1 |
| AAGACC | 13 | ACC | 6 | 'AAGACC--- | 6.67681 | 4 | 6.67681 | 4 | 6 | 1 |
| ACCAA | 13 | ACC | 5 | '---ACCAA- | 2.32261 | 4 | 2.32261 | 4 | 3 | 1 |
| ACCAAG | 13 | ACC | 6 | '---ACCAAG | 3.88703 | 3 | 3.88703 | 4 | 3 | 1 |
| AGACC | 13 | ACC | 5 | '-AGACC--- | 2.30871 | 4 | 2.30871 | 4 | 6 | 1 |
| GACC | 13 | ACC | 4 | '--GACC--- | 2.55586 | 12 | 2.55586 | 4 | 6 | 1 |
| GACCA | 13 | ACC | 5 | '--GACCA-- | 2.32994 | 4 | 2.32994 | 4 | 8 | 0.64 |
| GACCAA | 13 | ACC | 6 | '--GACCAA- | 4.79991 | 3 | 4.79991 | 4 | 8 | 0.64 |
| GACCTG | 13 | ACC | 6 | '--GACCTG- | 3.91881 | 3 | 3.91881 | 4 | 6 | 1 |
| GGACC | 13 | ACC | 5 | '-GGACC--- | 3.18228 | 5 | 3.18228 | 4 | 6 | 1 |
| AATCGG | 14 | CGG | 6 | 'AATCGG-- | 4.01211 | 3 | 4.01211 | 4 | 5 | 1 |
| ATCGGC | 14 | CGG | 6 | '-ATCGGC- | 2.53593 | 3 | 2.53593 | 4 | 5 | 1 |
| CGGGG | 14 | CGG | 5 | '---CGGGG | 2.33353 | 5 | 2.33353 | 4 | 2 | 1 |
| CTCGG | 14 | CGG | 5 | '-CTCGG-- | 2.29696 | 5 | 2.29696 | 4 | 5 | 1 |
| CTCGGG | 14 | CGG | 6 | '-CTCGGG- | 3.84602 | 3 | 3.84602 | 4 | 6 | 0.73 |
| TCGGCG | 14 | CGG | 6 | '--TCGGCG | 3.28023 | 3 | 3.28023 | 4 | 5 | 1 |
| TCGGGG | 14 | CGG | 6 | '--TCGGGG | 3.32851 | 3 | 3.32851 | 4 | 6 | 0.73 |
| AGAA | 15 | AGAA | 4 | '--AGAA- | 2.16143 | 11 | 2.16143 | 4 | 4 | 1 |
| AGAAT | 15 | AGAA | 5 | '--AGAAT | 3.21047 | 5 | 3.21047 | 4 | 4 | 1 |
| GTAGAA | 15 | AGAA | 6 | 'GTAGAA- | 4.74249 | 5 | 4.74249 | 4 | 4 | 1 |
| TAGAA | 15 | AGAA | 5 | '-TAGAA- | 3.13329 | 6 | 3.13329 | 4 | 4 | 1 |
| TAGAAT | 15 | AGAA | 6 | '-TAGAAT | 5.47463 | 4 | 5.47463 | 4 | 4 | 1 |
| AATTG | 16 | AATT | 5 | '--AATTG- | 3.14001 | 6 | 3.14001 | 4 | 4 | 1 |
| AATTGT | 16 | AATT | 6 | '--AATTGT | 4.87458 | 3 | 4.87458 | 4 | 4 | 1 |
| GTAATT | 16 | AATT | 6 | 'GTAATT-- | 4.68928 | 5 | 4.68928 | 4 | 4 | 1 |
| TAATT | 16 | AATT | 5 | '-TAATT-- | 2.33437 | 5 | 2.33437 | 4 | 4 | 1 |
| TAATTG | 16 | AATT | 6 | '-TAATTG- | 4.73989 | 4 | 4.73989 | 4 | 4 | 1 |
| ACCTC | 17 | CCTC | 5 | '-ACCTC-- | 3.12256 | 5 | 3.12256 | 4 | 5 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ACCTCC | 17 | CCTC | 6 | '-ACCTCC- | 4.86105 | 3 | 4.86105 | 4 | 5 | 1 |
| CACCTC | 17 | CCTC | 6 | 'CACCTC-- | 5.06762 | 3 | 5.06762 | 4 | 5 | 1 |
| CCTC | 17 | CCTC | 4 | '--CCTC-- | 1.66665 | 10 | 1.66665 | 4 | 5 | 1 |
| CCTCC | 17 | CCTC | 5 | '--CCTCC- | 2.2766 | 4 | 2.2766 | 4 | 5 | 1 |
| CCTCGG | 17 | CCTC | 6 | '--CCTCGG | 3.84116 | 3 | 3.84116 | 4 | 5 | 1 |
| AGGGC | 18 | GGGC | 5 | '-AGGGC- | 2.06499 | 7 | 2.06499 | 4 | 5 | 1 |
| CTGGGC | 18 | GGGC | 6 | 'CTGGGC- | 3.81387 | 4 | 3.81387 | 4 | 5 | 1 |
| GGGC | 18 | GGGC | 4 | '--GGGC- | 8.06106 | 43 | 8.06106 | 4 | 5 | 1 |
| GGGCA | 18 | GGGC | 5 | '--GGGCA | 2.33091 | 4 | 2.33091 | 4 | 5 | 1 |
| GGGCG | 18 | GGGC | 5 | '--GGGCG | 2.32071 | 5 | 2.32071 | 4 | 5 | 1 |
| TGGGC | 18 | GGGC | 5 | '-TGGGC- | 3.22006 | 10 | 3.22006 | 4 | 5 | 1 |
| AGTA | 19 | AGTA | 4 | '-AGTA-- | 2.15759 | 11 | 2.15759 | 4 | 4 | 1 |
| AGTAGC | 19 | AGTA | 6 | '-AGTAGC | 2.50008 | 3 | 2.50008 | 4 | 4 | 1 |
| GAGTA | 19 | AGTA | 5 | 'GAGTA-- | 2.29532 | 4 | 2.29532 | 4 | 4 | 1 |
| TAGTA | 19 | AGTA | 5 | 'TAGTA-- | 2.37311 | 5 | 2.37311 | 4 | 4 | 1 |
| TAGTAG | 19 | AGTA | 6 | 'TAGTAG- | 3.34476 | 3 | 3.34476 | 4 | 4 | 1 |
| AGTG | 20 | GTG | 4 | '-AGTG-- | 5.8182 | 24 | 5.8182 | 4 | 3 | 1 |
| AGTGA | 20 | GTG | 5 | '-AGTGA- | 4.02655 | 6 | 4.02655 | 4 | 5 | 0.6 |
| AGTGAG | 20 | GTG | 6 | '-AGTGAG | 5.35549 | 4 | 5.35549 | 4 | 5 | 0.6 |
| CAGTG | 20 | GTG | 5 | 'CAGTG-- | 3.05845 | 6 | 3.05845 | 4 | 3 | 1 |
| GTGA | 20 | GTG | 4 | '--GTGA- | 1.74258 | 20 | 1.74258 | 4 | 3 | 1 |
| GTGAG | 20 | GTG | 5 | '--GTGAG | 2.68747 | 9 | 2.68747 | 4 | 3 | 1 |
| CGGTT | 21 | GGTT | 5 | 'CGGTT-- | 2.2958 | 4 | 2.2958 | 4 | 4 | 1 |
| CGGTTG | 21 | GGTT | 6 | 'CGGTTG- | 3.90035 | 3 | 3.90035 | 4 | 4 | 1 |
| GGTT | 21 | GGTT | 4 | '-GGTT-- | 1.71456 | 10 | 1.71456 | 4 | 4 | 1 |
| GGTTG | 21 | GGTT | 5 | '-GGTTG- | 3.06306 | 6 | 3.06306 | 4 | 4 | 1 |
| GGTTGG | 21 | GGTT | 6 | '-GGTTGG | 3.92331 | 3 | 3.92331 | 4 | 4 | 1 |
| AGTGTA | 22 | GTGT | 6 | 'AGTGTA- | 4.85298 | 3 | 4.85298 | 4 | 4 | 1 |
| GTGT | 22 | GTGT | 4 | '-GTGT-- | 3.36222 | 26 | 3.36222 | 4 | 4 | 1 |
| GTGTA | 22 | GTGT | 5 | '-GTGTA- | 2.6027 | 8 | 2.6027 | 4 | 4 | 1 |
| GTGTAG | 22 | GTGT | 6 | '-GTGTAG | 3.26747 | 4 | 3.26747 | 4 | 4 | 1 |
| GTGTT | 22 | GTGT | 5 | '-GTGTT- | 2.027 | 7 | 2.027 | 4 | 4 | 1 |
| CGGG | 23 | CGGG | 4 | 'CGGG-- | 3.68909 | 18 | 3.68909 | 4 | 3 | 1 |
| CGGGA | 23 | CGGG | 5 | 'CGGGA- | 2.28819 | 4 | 2.28819 | 4 | 3 | 1 |
| CGGGT | 23 | CGGG | 5 | 'CGGGT- | 2.35486 | 4 | 2.35486 | 4 | 3 | 1 |
| CGGGTG | 23 | CGGG | 6 | 'CGGGTG | 5.57367 | 4 | 5.57367 | 4 | 3 | 1 |
| GCGG | 24 | GCGG | 4 | 'GCGG-- | 3.64572 | 18 | 3.64572 | 4 | 4 | 1 |
| GCGGG | 24 | GCGG | 5 | 'GCGGG- | 6.1336 | 10 | 6.1336 | 4 | 4 | 1 |
| GCGGGC | 24 | GCGG | 6 | 'GCGGGC | 3.67273 | 4 | 3.67273 | 4 | 4 | 1 |
| GCGGGT | 24 | GCGG | 6 | 'GCGGGT | 5.00325 | 3 | 5.00325 | 4 | 4 | 1 |
| GCGGT | 24 | GCGG | 5 | 'GCGGT- | 2.29151 | 4 | 2.29151 | 4 | 4 | 1 |
| AACAC | 25 | ACAC | 5 | 'AACAC-- | 2.28582 | 4 | 2.28582 | 4 | 3 | 1 |
| AACACC | 25 | ACAC | 6 | 'AACACC- | 5.00613 | 3 | 5.00613 | 4 | 3 | 1 |
| ACACCT | 25 | ACAC | 6 | '-ACACCT | 4.95198 | 3 | 4.95198 | 4 | 3 | 1 |
| ACACGT | 25 | ACAC | 6 | '-ACACGT | 4.82098 | 3 | 4.82098 | 4 | 3 | 1 |
| AAGT | 26 | AAGT | 4 | '-AAGT- | 1.74801 | 10 | 1.74801 | 4 | 3 | 1 |
| AAGTG | 26 | AAGT | 5 | '-AAGTG | 3.8693 | 7 | 3.8693 | 4 | 3 | 1 |
| TAAGT | 26 | AAGT | 5 | 'TAAGT- | 2.35658 | 5 | 2.35658 | 4 | 3 | 1 |
| TAAGTG | 26 | AAGT | 6 | 'TAAGTG | 4.70162 | 4 | 4.70162 | 4 | 3 | 1 |
| AGTGGC | 27 | GGC | 6 | 'AGTGGC--- | 3.54163 | 4 | 3.54163 | 5 | 7 | 1 |
| ATGGC | 27 | GGC | 5 | '-ATGGC--- | 5.93424 | 14 | 5.93424 | 5 | 7 | 1 |
| GATGGC | 27 | GGC | 6 | 'GATGGC--- | 8.14456 | 8 | 8.14456 | 5 | 7 | 1 |
| GGCT | 27 | GGC | 4 | '---GGCT-- | 2.13292 | 11 | 2.13292 | 5 | 5 | 1 |
| GGCTA | 27 | GGC | 5 | '---GGCTA- | 3.17886 | 5 | 3.17886 | 5 | 5 | 1 |
| GGCTAG | 27 | GGC | 6 | '---GGCTAG | 3.92556 | 3 | 3.92556 | 5 | 5 | 1 |
| GTGGC | 27 | GGC | 5 | '-GTGGC--- | 6.4307 | 20 | 6.4307 | 5 | 7 | 1 |
| GTGGCT | 27 | GGC | 6 | '-GTGGCT-- | 4.71178 | 5 | 4.71178 | 5 | 10 | 0.67 |
| TGGC | 27 | GGC | 4 | '--TGGC--- | 6.24044 | 40 | 6.24044 | 5 | 7 | 1 |
| TGGCT | 27 | GGC | 5 | '--TGGCT-- | 2.34889 | 5 | 2.34889 | 5 | 10 | 0.67 |
| TGGCTG | 27 | GGC | 6 | '--TGGCTG- | 3.2832 | 3 | 3.2832 | 5 | 10 | 0.67 |
| CCGA | 28 | CGA | 4 | '-CCGA--- | 2.13739 | 11 | 2.13739 | 5 | 6 | 1 |
| CCGAG | 28 | CGA | 5 | '-CCGAG-- | 2.35878 | 5 | 2.35878 | 5 | 6 | 1 |
| CCGAT | 28 | CGA | 5 | '-CCGAT-- | 2.30823 | 4 | 2.30823 | 5 | 8 | 0.71 |
| CCGATG | 28 | CGA | 6 | '-CCGATG- | 3.83631 | 3 | 3.83631 | 5 | 8 | 0.71 |
| CGATG | 28 | CGA | 5 | '--CGATG- | 4.61974 | 8 | 4.61974 | 5 | 4 | 1 |
| CGATGG | 28 | CGA | 6 | '--CGATGG | 9.90067 | 7 | 9.90067 | 5 | 4 | 1 |
| TCCGA | 28 | CGA | 5 | 'TCCGA--- | 3.05931 | 6 | 3.05931 | 5 | 6 | 1 |
| TCCGAG | 28 | CGA | 6 | 'TCCGAG-- | 3.32607 | 3 | 3.32607 | 5 | 6 | 1 |
| TCCGAT | 28 | CGA | 6 | 'TCCGAT-- | 3.83954 | 3 | 3.83954 | 5 | 8 | 0.71 |
| AAGGGG | 29 | GGG | 6 | 'AAGGGG | 3.88538 | 3 | 3.88538 | 5 | 3 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| AGGGG | 29 | GGG | 5 | '-AGGGG | 4.57194 | 8 | 4.57194 | 5 | 3 | 1 |
| CTGGG | 29 | GGG | 5 | 'CTGGG- | 3.22319 | 6 | 3.22319 | 5 | 2 | 1 |
| GGGG | 29 | GGG | 4 | '--GGGG | 12.2506 | 42 | 12.2506 | 5 | 3 | 1 |
| TGGG | 29 | GGG | 4 | '-TGGG- | 4.60088 | 24 | 4.60088 | 5 | 2 | 1 |
| TGGGG | 29 | GGG | 5 | '-TGGGG | 6.39366 | 12 | 6.39366 | 5 | 5 | 0.4 |
| GTTGGA | 30 | TGGA | 6 | 'GTTGGA-- | 2.49556 | 3 | 2.49556 | 5 | 4 | 1 |
| TGGA | 30 | TGGA | 4 | '--TGGA-- | 2.55883 | 15 | 2.55883 | 5 | 4 | 1 |
| TGGAC | 30 | TGGA | 5 | '--TGGAC- | 2.33125 | 5 | 2.33125 | 5 | 4 | 1 |
| TGGACC | 30 | TGGA | 6 | '--TGGACC | 5.45924 | 4 | 5.45924 | 5 | 4 | 1 |
| TTGGAC | 30 | TGGA | 6 | '-TTGGAC- | 3.93915 | 3 | 3.93915 | 5 | 4 | 1 |

**Table S8. Summary of the enriched ISRE n-mers and GCCS clustering performance**

| | ISRE sequences | Random Sample |
|---|---|---|
| Total n-mers | 5376 | 5376 |
| Probability >$Cl_{high}$ (2) | 241 | 91 |
| Clustered | 193 | 64 |
| % clustered | 80.1% | 70.33% |
| Number of Clusters | 30 | 11 |

**Table S9. Detailed comparison of GCCS clusters consensus motifs to known trans-acting factor binding sites**

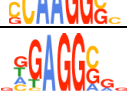
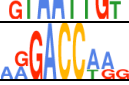
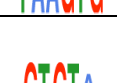
| Class | Pictogram | Similar To |
|---|---|---|
| G-rich |  | hnRNP F/H consensus binding site (GGGGG) (6), which functions as either a splicing enhancer or silencer (7). Contains a G-triplet, a known ISE sequence (8) that is abundant in mammalian introns (9). |
| G-rich |  | High affinity hnRNPA1 binding site (TAGGG) identified by SELEX (10). Contains a G-triplet, a known ISE sequence (8) that is abundant in mammalian introns (9). |
| Other |  | hnRNP A1 binding site (TAGAGT) (11) |
| Other |  | High affinity hnRNP L binding site (CA-rich) identified by SELEX and an ISE element comprised of variable-length CA repeats (12). A/C-rich ESSs (13). |
| Other |  | CTCC and CCTCCC repeats identified by computational analysis of introns flanking skipped exons (14). CT-rich intronic sequences that act as PTB binding sites (15,16). |
| Other |  | SRp40 binding site (ACAAG) (17). |
| Other |  | SC35 binding site (AGGAGAT) (18). A purine-rich element (AGGG) identified in introns flanking skipped exons (14). |
| Other |  | Sam68 binding site (TAAA) (19,20). |
| Other |  | 9G8 high-affinity binding site (GAC) identified by SELEX (18). |
| Other |  | SF2/ASF high-affinity binding site (GAAGAA) identified by SELEX (21). Tra2β high-affinity binding site $(GAA)_n$ identified by SELEX (22). |
| Other |  | Srp30c consensus sequence (CTGGATT) (23). |
| GT-rich |  | hnRNP G binding motif (AAGT) (24). |
| GT-rich |  | CELF/Bruno-like family of proteins that bind GT repeats with high affinity (25). CUG-BP1 binding sites consisting of TGT-repeats (25). hnRNP M binding sites consisting of poly(G) and poly(T) homopolymers (26). |
| GT-rich |  | CELF/Bruno-like family of proteins that bind GT repeats with high affinity (25). |
| GT-rich |  | CELF/Bruno-like family of proteins that bind GT repeats with high affinity (25). CUG-BP1 binding sites consisting of TGT-repeats (25). |
| GT-rich |  | CELF/Bruno-like family of proteins that bind GT repeats with high affinity (25). |

**Table S10. ISRE pentamers that do not resemble known splicing regulatory elements**

| Field | Description |
|---|---|
| n-mer | The n-mer (5mers) |
| GCS | Greatest Common Substring |
| clustID | ClusterID |

| Both Intronic and Exonic | | | Intronic elements | | | Exonic elements | | |
|---|---|---|---|---|---|---|---|---|
| **n-mer** | **GCS** | **clustID** | **n-mer** | **GCS** | **clustID** | **n-mer** | **GCS** | **clustID** |
| CGATG | CGA | 28 | GTGGC | GGC | 27 | GGGGC | GGGGC | 4 |
| *TAGAG | TAGAG | 7 | CAAGG | AAGG | 2 | GCGGG | GCGG | 24 |
| AAGGC | AAGG | 2 | AGTGA | GTG | 20 | TGGGC | GGGC | 18 |
| CCGAT | CGA | 28 | AAGTG | AAGT | 26 | GGCTA | GGC | 27 |
| CGGTT | GGTT | 21 | AGAAT | AGAA | 15 | AATTG | AATT | 16 |
| GCGGT | GCGG | 24 | GAGGA | AGG | 11 | CGGGT | CGGG | 23 |
| | | | GGACC | ACC | 13 | GGGCG | GGGC | 18 |
| | | | ACCTC | CCTC | 17 | AGGGC | GGGC | 18 |
| | | | ATATG | TATG | 12 | | | |
| | | | TCCGA | CGA | 28 | | | |
| | | | CAGTG | GTG | 20 | | | |
| | | | GTGTA | GTGT | 22 | | | |
| | | | CGGTG | GGTG | 1 | | | |
| | | | TAGTA | AGTA | 19 | | | |
| | | | TATGA | TATG | 12 | | | |
| | | | TGGAC | TGGA | 30 | | | |
| | | | GACCA | ACC | 13 | | | |
| | | | ACCAA | ACC | 13 | | | |
| | | | AGACC | ACC | 13 | | | |
| | | | CTCGG | CGG | 14 | | | |
| | | | GAGTA | AGTA | 19 | | | |
| | | | AACAC | ACAC | 25 | | | |
| | | | CCTCC | CCTC | 17 | | | |

n-mers which do not overlap with known intronic and exonic regulatory elements were placed under the 'both intronic and exonic' heading. Enriched ISRE pentamers that do not overlap with known intronic regulatory elements were placed under the intronic element heading and similarly for elements which do not resemble exonic regulatory elements. *The TAGAG was found to overlap with an element identified upstream of constitutively spliced exons (5).

## Table S11. Overlap of enriched hexamers with extended recovered ISRE sequences

| Extended ISS sequence | Enriched Hexamers | | | |
|---|---|---|---|---|
| GTTCGAATCTCTCCAGTGC | | | | |
| GTCCTACGCTCATTATTGC | | | | |
| GTTCTTCTCTTCTCTTCGC | | | | |
| GTTGTTCGCACCGCTGGGC | CGCTGG | CTGGGC | GCTGGG | TGTTCG |
| GTTGTTCGCACCACTGAGC | TGTTCG | | | |
| GTAGTCACCTATTATAGGC | | | | |
| GTGTTAACCAACGATGGGC | CGATGG | | | |
| GTGGTATCGAAAGTTGTGC | | | | |
| GTTACATCCAGAAGTCGGC | | | | |
| GTTACATCCCTCGGTTGGC | CCTCGG | CGGTTG | GGTTGG | |
| GTTGGACCAGGCGTACGGC | CCAGGC | GTTGGA | TGGACC | TTGGAC |
| GTTGGACACGTCAGTCAGC | ACACGT | GTTGGA | TTGGAC | |
| GTCACACGTGAGAGAGAGC | ACACGT | | | |
| GTGAAGGGCGACAGATAGC | AAGGGC | | | |
| GTAGAACGCTGGATTAAGC | CGCTGG | GCTGGA | GTAGAA | |
| GTTTACTTTAAGGATAAGC | | | | |
| GTATACGGAAAGGCCTTGC | GTATAC | | | |
| GTGTGCTTATATGGGTTGC | ATATGG | | | |
| GTTTAGTCCCATTCCGAGC | TCCGAG | | | |
| GTCCACTTCGGTTGCCTGC | CGGTTG | | | |
| GTACGTCCGTCGTGGATGC | | | | |
| GTACCTCGAGGTCTGAAGC | | | | |
| GTACCTCAGGCTCTGAAGC | | | | |
| GTAAGGCTAGTTTAGTAGC | AGTAGC | GGCTAG | | |
| GTAAGGCTAGATTAGTAGC | AGTAGC | GGCTAG | | |
| GTAGAGGAGTCGTGTCAGC | AGAGGA | GTAGAG | TAGAGG | TGTCAG |
| GTAGTGGAATCGTATCAGC | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| GTGGTCGAGTCGCAAGGGC | AAGGGC | CAAGGG | GCAAGG | | | |
| GTATTCCAGCTGGAGCTGC | GCTGGA | | | | | |
| GTAGTATATGGTGAGGAGC | ATATGG | GTGAGG | | | | |
| GTGCCGAGTAAAGTGTAGC | AGTGTA | GTAAAG | GTGTAG | | | |
| GTTCTGACTCAATAGTAGC | AGTAGC | | | | | |
| GTCTTGAGTACCCCCGAGC | | | | | | |
| GTCATGCACCGACCAAGGC | ACCAAG | CAAGGC | CCAAGG | GACCAA | | |
| GTAATTGTGTTTGTGATGC | AATTGT | ATTGTG | GTAATT | TAATTG | TGTGTT | TTGTGT |
| GTGACTGTGTTAGGCGGGC | GGCGGG | TGTGTT | | | | |
| GTAATTGGGTTTGGGGGGC | GGGGGC | GGGGGG | GTAATT | TAATTG | TGGGGG | |
| GTAATTGTGTTCGGTGGGC | AATTGT | ATTGTG | CGGTGG | GTAATT | GTGGGC | TAATTG |
| | TGTGTT | TGTTCG | TTGTGT | | | |
| GTAATTGTGTTTGGCGGGC | AATTGT | ATTGTG | GGCGGG | GTAATT | TAATTG | TGGCGG |
| | TGTGTT | TTGTGT | | | | |
| GTTATGACATGTGGGGAGC | GACATG | GTGGGG | GTTATG | TATGAC | TGACAT | TGTGGG |
| | TTATGA | | | | | |
| GTTATGACGTGTGGGGGGC | GGGGGC | GGGGGG | GTGGGG | GTTATG | TATGAC | TGGGGG |
| | TGTGGG | TTATGA | | | | |
| GTTATGACATGTGGGGGGC | GACATG | GGGGGC | GGGGGG | GTGGGG | GTTATG | TATGAC |
| | TGACAT | TGGGGG | TGTGGG | TTATGA | | |
| GTCAATTGAGTTGGTGTGC | | | | | | |
| GTCGATGGGGCAGGGGAGC | CGATGG | TGGGGC | | | | |
| GTCAGTGAACTTTGCGAGC | TCAGTG | TTTGCG | | | | |
| GTCCTTGGTCCTGACATGC | GACATG | TGACAT | | | | |
| GTCCGAGTGCGACGGTGGC | CGGTGG | GGTGGC | TCCGAG | | | |
| GTGAGTGGCCTAGGGAGGC | AGGGAG | AGTGGC | GAGTGG | GGAGGC | GGGAGG | TAGGGA |
| | TAGTAG | | | | | |
| GTGGCTGGGCTAGGATGGC | CTGGGC | GATGGC | GCTGGG | GGCTAG | GGCTGG | GTGGCT |
| | TGGCTG | | | | | |
| GTGATATGGCGAGGGTGGC | ATATGG | GATATG | GGGTGG | GGTGGC | TATGGC | |
| GTAAGTGGGCACGGTTGGC | AAGTGG | AGTGGG | CGGTTG | GGTTGG | GTGGGC | TAAGTG |
| GTAGGTAGCCACCGTTGGC | ACCGTT | | | | | |
| GTGGGGGGGTCACTTAGGC | GGGGGG | GTGGGG | TGGGGG | | | |
| GTTGGTTGGACCCGTAGGC | GGTTGG | GTTGGA | TGGACC | TTGGAC | | |
| GTCCCTATGGTTCCTCGGC | CCTCGG | | | | | |
| GTCAGAGGAGTCTCTAGGC | AGAGGA | CTAGGC | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| GTTTATGGAGTTCCTAGGC | CTAGGC | | | | | |
| GTAAATAGAGGCCCCAGGC | CCAGGC | TAGAGG | | | | |
| GTCTAGTAACCAGCCAGGC | CCAGGC | | | | | |
| GTCTAAGCACCACTGAGGC | TGAGGC | | | | | |
| GTTGTTTTGCGTCCAAGGC | CAAGGC | CCAAGG | TTTGCG | | | |
| GTCATGTCAGGACCAAGGC | ACCAAG | CAAGGC | CCAAGG | GACCAA | TGTCAG | |
| GTCATGGACCGACCAAGGC | ACCAAG | CAAGGC | CCAAGG | GACCAA | TGGACC | |
| GTTATGCCTCCCCGATAGC | GTTATG | | | | | |
| GTCGAAGAACCCCAAGGGC | AAGGGC | CAAGGG | CCAAGG | | | |
| GTCGGAGAAACCGGAGGGC | GGAGGG | | | | | |
| GTCCGAGGAACCATAGGGC | TCCGAG | | | | | |
| GTCTATCTCCTTCTATGGC | TATGGC | | | | | |
| GTTTAACACCTCCCAAGGC | AACACC | ACACCT | ACCTCC | CAAGGC | CACCTC | CCAAGG |
| GTCAAAGACCTGCGATGGC | AAGACC | CGATGG | GACCTG | GATGGC | | |
| GTCAAACACGTCCGATGGC | ACACGT | CCGATG | CGATGG | GATGGC | TCCGAT | |
| GTCTAACACCTCCGATGGC | AACACC | ACACCT | ACCTCC | CACCTC | CCGATG | CGATGG |
| | GATGGC | TCCGAT | | | | |
| GTCAAACACCTCCGATGGC | AACACC | ACACCT | ACCTCC | CACCTC | CCGATG | CGATGG |
| | GATGGC | TCCGAT | | | | |
| GTGTGGCTATGAATTTGGC | GTGGCT | | | | | |
| GTGTGGCTAAGAATTGGGC | GTGGCT | | | | | |
| GTGGCTGGAAGACCTGCGC | AAGACC | GACCTG | GCTGGA | GGCTGG | GTGGCT | TGGCTG |
| GTGTAAAGGGTGTCAGTGC | GTAAAG | TCAGTG | TGTCAG | | | |
| GTATTAATAATACTGGGGC | TGGGGC | | | | | |
| GTGTTAATAGCGCGGGAGC | | | | | | |
| GTTTGTAAGGTGCTGGGGC | GCTGGG | TGGGGC | | | | |
| GTTGTGGTCGCGACCTGGC | GACCTG | | | | | |
| GTGGCGGTCGAGTACAGGC | GGCGGT | GTACAG | TGGCGG | | | |
| GTGTTGTGAAAGAGGAGGC | AGAGGA | GGAGGC | | | | |
| GTGGTGGCAGACACGATGC | GGTGGC | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| GTGCGGTTTGCGGGCGGGC | GGCGGG | GGGCGG | TTTGCG | | | |
| GTGGGGCGCGCGGGGGGGC | GGGGCG | GGGGGC | GGGGGG | GTGGGG | TGGGGC | |
| GTGAGGGCAGTCCGTGGGC | GTGAGG | GTGGGC | | | | |
| GTGACGGGTGCCTCGGGGC | CCTCGG | CGGGGC | CGGGTG | CTCGGG | TCGGGG | |
| GTTAGGTGTGTCTCGGGGC | CGGGGC | CTCGGG | TCGGGG | | | |
| GTGACGTGTGTCTCGGGGC | CGGGGC | CTCGGG | TCGGGG | | | |
| GTGACGGAGCCGTCTGGGC | CTGGGC | | | | | |
| GTGCATGGCCCCGCTGGGC | CGCTGG | CTGGGC | GCTGGG | | | |
| GTGCAAGGTCCCTCTAGGC | CTAGGC | GCAAGG | | | | |
| GTGCACTAGAATCTGAGGC | TAGAAT | TGAGGC | | | | |
| GTGCAGTACGGGCTTAGGC | | | | | | |
| GTCGAGCGGCTTTAGAGGC | TAGAGG | | | | | |
| GTAGAGTGGGGCGGGTGGC | AGAGTG | AGTGGG | CGGGTG | GAGTGG | GGCGGG | GGGCGG |
| | GGGGCG | GGGTGG | GGTGGC | GTAGAG | GTGGGG | TAGAGT |
| | TGGGGC | | | | | |
| GTATAGTGGCGGTGGAGGC | AGTGGC | CGGTGG | GGAGGC | GGCGGT | TGGCGG | |
| GTAGAGTGGCGGTGGAGGC | AGAGTG | AGTGGC | CGGTGG | GAGTGG | GGAGGC | GGCGGT |
| | TAGAGT | TGGCGG | GTAGAG | | | |
| GTAGAATGGACCGTGAGGC | GTAGAA | GTGAGG | TAGAAT | TGAGGC | TGGACC | |
| GTGGAGTGGCTGGTTCGGC | AGTGGC | GAGTGG | GAGTGG | GGCTGG | GTGGCT | TGGCTG |
| GTGTACAGCGGAGAGGGGC | AGGGGC | GTACAG | | | | |
| GTGTACGGTGCAGAGGGGC | AGGGGC | | | | | |
| GTGTAGTGTAGGGAGGGGC | AGGGAG | AGGGGC | AGTGTA | GGAGGG | GGGAGG | GTGTAG |
| | TAGTAG | TAGGGA | | | | |
| GTGAAGTGTAGGGAGGGGC | AGGGAG | AGGGGC | AGTGTA | GGAGGG | GGGAGG | GTGTAG |
| | TAGTAG | TAGGGA | | | | |
| GTATACCGTTCAGTGGGGC | ACCGTT | AGTGGG | GTATAC | GTGGGG | TCAGTG | TGGGGC |
| GTATACCGTTCAGTGAGGC | ACCGTT | AGTGAG | GTATAC | GTGAGG | TCAGTG | TGAGGC |
| GTAAAGGGGCAAGGTGGGC | AAGGGG | AGGGGC | AGGTGG | GCAAGG | GTAAAG | GTGGGC |
| GTAGAGTGCGAAGCGGGGC | AGAGTG | CGGGGC | GTAGAG | TAGAGT | | |
| GTACAGTGCTAAGTAGGGC | GTACAG | | | | | |
| GTGTAAATCGGCGGGTGGC | AATCGG | ATCGGC | CGGGTG | GGCGGG | GGGTGG | GGTGGC |
| | TCGGCG | GGTGGC | | | | |
| GTGGAAATCGGCGGATGGC | AATCGG | ATCGGC | GATGGC | TCGGCG | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| GTGGCAATCGGCGGGTGGC | AATCGG TCGGCG | ATCGGC | CGGGTG | GGCGGG | GGGTGG | GGTGGC |
| GTAAAGAACGGGATATGGC | ATATGG TATGGC | CGGGAT | GATATG | GGATAT | GGGATA | GTAAAG |
| GTCAAGACCGGGATATGGC | AAGACC TATGGC | ATATGG | CGGGAT | GATATG | GGATAT | GGGATA |
| GTAAAGACCGGGATATGGC | AAGACC TATGGC | ATATGG GTAAAG | CGGGAT | GATATG | GGATAT | GGGATA |
| GTAATTATTAGTCGATGGC | CGATGG | GATGGC | GTAATT | | | |
| GTGCTTAGTGAGTGATGGC | AGTGAG | GATGGC | | | | |
| GTACAGGCCAAGGGGGGGC | AAGGGG | CAAGGG | CCAAGG | GGGGGC | GGGGGG | GTACAG |
| GTAGAAGACAAGTGGTGGC | AAGTGG | GGTGGC | GTAGAA | | | |
| GTGGTTGAAGGGGGGCGGC | AAGGGG | GGGCGG | GGGGCG | GGGGGC | GGGGGG | |
| GTACATTATGAGGGTCGGC | TTATGA | | | | | |
| GTAGAGTAAGTGAGGTGGC | AGGTGG TAGAGT | AGTGAG | GGTGGC | GTAGAG | GTGAGG | TAAGTG |
| GTAGAATAAGTGAGGTGGC | AGGTGG TAGAAT | AGTGAG | GGTGGC | GTAGAA | GTGAGG | TAAGTG |
| GTAGAATAAGTGGGGTGGC | AAGTGG TAGAAT | AGTGGG TAAGTG | GGGTGG | GGTGGC | GTAGAA | GTGGGG |

## References


1. Blencowe, B.J. (2006) Alternative splicing: new insights from global analyses. *Cell*, **126**, 37-47.
2. Voelker, R.B. and Berglund, J.A. (2007) A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. *Genome Res*, **17**, 1023-1033.
3. Fairbrother, W.G. and Chasin, L.A. (2000) Human genomic sequences that inhibit splicing. *Mol Cell Biol*, **20**, 6816-6825.
4. Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M. and Burge, C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831-845.
5. Zhang, X.H., Heller, K.A., Hefter, I., Leslie, C.S. and Chasin, L.A. (2003) Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res*, **13**, 2637-2650.
6. Markovtsov, V., Nikolic, J.M., Goldman, J.A., Turck, C.W., Chou, M.Y. and Black, D.L. (2000) Cooperative assembly of an hnRNP complex induced by a tissue-specific homolog of polypyrimidine tract binding protein. *Mol Cell Biol*, **20**, 7463-7479.
7. Chou, M.Y., Rooke, N., Turck, C.W. and Black, D.L. (1999) hnRNP H is a component of a splicing enhancer complex that activates a c-src alternative exon in neuronal cells. *Mol Cell Biol*, **19**, 69-77.
8. McCullough, A.J. and Berget, S.M. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol Cell Biol*, **17**, 4562-4571.
9. Yeo, G., Hoon, S., Venkatesh, B. and Burge, C.B. (2004) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A*, **101**, 15700-15705.
10. Burd, C.G. and Dreyfuss, G. (1994) RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. *Embo J*, **13**, 1197-1204.
11. Hutchison, S., LeBel, C., Blanchette, M. and Chabot, B. (2002) Distinct sets of adjacent heterogeneous nuclear ribonucleoprotein (hnRNP) A1/A2 binding sites control 5' splice site selection in the hnRNP A1 mRNA precursor. *J Biol Chem*, **277**, 29745-29752.
12. Hui, J., Stangl, K., Lane, W.S. and Bindereif, A. (2003) HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nat Struct Biol*, **10**, 33-37.
13. Coulter, L.R., Landree, M.A. and Cooper, T.A. (1997) Identification of a new class of exonic splicing enhancers by in vivo selection. *Mol Cell Biol*, **17**, 2143-2150.
14. Miriami, E., Margalit, H. and Sperling, R. (2003) Conserved sequence elements associated with exon skipping. *Nucleic Acids Res*, **31**, 1974-1983.
15. Chan, R.C. and Black, D.L. (1995) Conserved intron elements repress splicing of a neuron-specific c-src exon in vitro. *Mol Cell Biol*, **15**, 6377-6385.
16. Chou, M.Y., Underwood, J.G., Nikolic, J., Luu, M.H. and Black, D.L. (2000) Multisite RNA binding and release of polypyrimidine tract binding protein during the regulation of c-src neural-specific splicing. *Mol Cell*, **5**, 949-957.
17. Liu, H.X., Zhang, M. and Krainer, A.R. (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev*, **12**, 1998-2012.

18. Cavaloc, Y., Bourgeois, C.F., Kister, L. and Stevenin, J. (1999) The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *Rna*, **5**, 468-483.
19. Itoh, H., Washio, T. and Tomita, M. (2004) Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. *Rna*, **10**, 1005-1018.
20. Paronetto, M.P., Achsel, T., Massiello, A., Chalfant, C.E. and Sette, C. (2007) The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x. *J Cell Biol*, **176**, 929-939.
21. Tacke, R. and Manley, J.L. (1995) The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *Embo J*, **14**, 3540-3551.
22. Tacke, R., Tohyama, M., Ogawa, S. and Manley, J.L. (1998) Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing. *Cell*, **93**, 139-148.
23. Simard, M.J. and Chabot, B. (2002) SRp30c is a repressor of 3' splice site utilization. *Mol Cell Biol*, **22**, 4001-4010.
24. Nasim, M.T., Chernova, T.K., Chowdhury, H.M., Yue, B.G. and Eperon, I.C. (2003) HnRNP G and Tra2beta: opposite effects on splicing matched by antagonism in RNA binding. *Hum Mol Genet*, **12**, 1337-1348.
25. Marquis, J., Paillard, L., Audic, Y., Cosson, B., Danos, O., Le Bec, C. and Osborne, H.B. (2006) CUG-BP1/CELF1 requires UGU-rich sequences for high-affinity binding. *Biochem J*, **400**, 291-301.
26. Hovhannisyan, R.H. and Carstens, R.P. (2007) Heterogeneous ribonucleoprotein m is a splicing regulatory protein that can enhance or silence splicing of alternatively spliced exons. *J Biol Chem*, **282**, 36265-36274.