## Supplementary Methods

### Computing and Comparing Positional Distributions.

We used the following procedures to compute positional distributions of hexamers and to calculate differences between two profiles using L1 distance. L1 distance, also known as city block distance or Manhattan distance, is simply the sum of the absolute difference between two positional distributions.

1. Create an exon database surrounding all 3′ splice sites (3′ss) and 5′ splice sites (5′ss) annotated in Ensembl. We include up to two hundred nucleotides of intronic sequence flanking either side of the splice site, and up to one hundred nucleotides of exonic sequence flanking either side of the splice site.
2. Each entry in the database must have an ‘AG’ at the 3’ss and a ‘GT’ at the 5’ss.
3. Duplicated entries are removed.
4. For exons with less than 200 nucleotides, the exonic sequence is divided in half and each half assigned to the closest splice site.
5. In rare cases when introns are less than 400 nucleotides, the intronic sequence is divided in half and each half is assigned to the closest splice site.
6. Count the number of occurrences of all 4,096 hexamers at each position in the database.
7. For each hexamer, construct a feature vector of length 600. We use the following indexing of positions:
    a. Positions starting at −200 and ending at -1 correspond to the upstream intronic region (i.e. positions -2 and -1 are the 3’ss with nucleotides ‘AG’).
    b. Positions starting at 0 and ending at 199 correspond to the exonic region.
    c. Positions starting at 200 and ending at 399 correspond to the downstream intronic region (i.e. positions 200 and 201 are the 5’ss with nucleotides ‘GT’).
8. Normalize the counts (i.e., compute a z-score) in each entry of each feature vector.
9. Calculate L1 distance for each pair of positional distributions.

### Preparation of exon databases

An exon/intron database for all analyzed species was made from the Ensembl Genes annotation stored at the UCSC table browser. Each entry in the database consists of at most 600 nucleotides: two 200 nucleotide intronic flanks and two 100 nucleotide exonic flanks on each side of the splice sites. In the case where intronic or exonic length is less than 400 or 200 nucleotides, respectively, the sequence is divided by half and each half is assigned to its nearest splice site. All duplicated entries were screened and removed from the database.

### Constructing orthologous coordinates across species

We used the following procedures to extract orthologous coordinates between all nine fish species and human. For each fish species:

1. Create an intron database from Ensembl annotations.

2. Each entry in the database must start with an ‘GT’ at the 5’ss, end with ‘AG’ at the 3’ss, include ACACAC or CACACA hexamer in the first 50 nucleotides from the 5’ss and include GTGTGT or TGTGTG hexamer in the last 50 nucleotides from the 5’ss.
3. Duplicated entries are removed.
4. For each intron in the database a pair of surrounding it exons is extracted and assigned to this intron.
5. All exon coordinates are converted to human hg19 assembly using UCSC liftOver.
6. Successfully converted exons are grouped by original intron names.
7. Remove entries in which only one exon from the pair was converted or if the pair of converted exons is not located on the same chromosome.
8. For each pair of converted exons check if there is an intron with coordinates corresponding to the end of the upstream exon and start of the downstream exon.

**Structure Prediction**

Zebrafish and human intronic sequences were excised at the 5’ and 3’ splice sites using danRer7 and hg19 annotations, respectively. Introns with a minimum length of 80 nt were included in order to have two 40 nt windows in which to search for hexamers. Introns of maximum length 733 nt were included to account for length limitations of the structure prediction software. All predicted minimum energy structures and corresponding delta G were conducted using RNAfold.

$(AC)_m$-$(GT)_n$ repeat addition simulations were performed on zebrafish introns. AC repeats were added 20 nt downstream of the 5’ss and GT repeats were added 20 nt upstream of the 3’ss. The folded structures were analyzed for hairpins using custom perl scripts. The number of introns in which the AC repeats base-paired to the GT repeats forming a hairpin were counted and plotted. Error bars were determined by sampling with replacement 1000 times and using the 95% confidence intervals.

Human introns were binned by length and folded with RNAfold. The highly structured set of human introns was determined by extracting the most stable 10% of introns in each size bin based on the predicted delta G of folding. Complimentary k-mers in the 40 nt windows immediately downstream of the 5’ss and upstream of the 3’ss were counted.

Introns were shuffled and refolded to test the contribution of repeat pairing towards the overall hairpin structure in both $(AC)_m$-$(GT)_n$ zebrafish introns and $(GGG)_m$-$(CCC)_n$ human introns. Zebrafish with at least one ACACAC-GTGTGT hexamer pair and human introns with at least three GGG-CCC triplet pairs in the forty nucleotide windows downstream of the 5’ss and upstream of the 3’ss were included for this analysis. Introns were shuffled 1000 times maintaining both the nucleotide composition and the concentration of repeats. Other than maintaining the nucleotide composition and the concentration of repeats, the intron sequences were shuffled randomly. The concentration of AC and GT repeats were preserved in zebrafish introns, and the

concentration of GGG and CCC repeats were preserved in human. The average delta G of each shuffled set was used to calculate a *P* value.

**Sequence used for minigene splicing and structure mapping**

For minigene splicing (Figure 4B):

- CONTROL:

GACTCCATCGATCTCTTCCTTGGGAACTATAGCGTGGATGAAACTGATAGTTTGAC
GCCTCTGCACGTTAAGAAGGACTGGAAGTTTCTTTTGgtcagtgctgtcacgatattgcttgatgtt
agagagaatacagaaaaaaaagaggtcattaaaagagaagtttagtccaagctgcctgatacagtcctaacatttataat
tcaacgccacaagctgccgctggtcattccctgcatatttttattcatatgtgtctgttatttgtcctgcagCTTCCTGTTA
TTATGGTGGTGGCTTTCTCCATGTGCATCATCTGTCTCCTAATGGCTG

- AC-TROL:

ACTGCAGACGCTGCTCCTGCACGGAAACATCATCACCACACTGCGCTCGGCTCCT
GCTCACCTGCCGGCACACCTGAGAGTGCTGTCGCTGGCGGAGAACGAGATCCGG
GACCTGACTGAGgtacacacacacacacacacacaaacacacacacacattcaggctcacactgcctcctcct
gtgcggttctgcatgtgagtgtaggcagatgtctgacctg//cctgatacagtcctaacatttataattcaacgccacaagctg
ccgctggtcattccctgcatatttttattcatatgtgtctgttatttgtcctgcagCTTCCTGTTATTATGGTGGTGG
CTTTCTCCATGTGCATCATCTGTCTCCTAATGGCTG

- CON-GT:

GACTCCATCGATCTCTTCCTTGGGAACTATAGCGTGGATGAAACTGATAGTTTGAC
GCCTCTGCACGTTAAGAAGGACTGGAAGTTTCTTTTGgtcagtgctgtcacgatattgcttgatgtt
agagagaatacagaaaaaaaagaggtcattaaaagagaagtttagtccaagctgcctgatacagtcctaa//gcatgtg
agtgtaggcagatgtctgacctgacctcgctgcacagctgaggatgcagatctgtcctgatgtgtgtgtgtgtgtgtgtgtgt
gtgtgcagGTCTGCTATCTGGCTCCGGTGCGGGGTCTGCAGCAGCTCTCCCTCCTCA
GTAACCCGTGTGTGTCGTGTGTGTCCTCCGCTGTGTGTGATTATCGGCCATATGTG
CTCAGCTGGTGCCTCGGCCTGGAGCTGCTAGATGGAGTCGCCGTCACACAGAAG
GAAGG

- AC-GT:

ACTGCAGACGCTGCTCCTGCACGGAAACATCATCACCACACTGCGCTCGGCTCCT
GCTCACCTGCCGGCACACCTGAGAGTGCTGTCGCTGGCGGAGAACGAG
ATCCGGGACCTGACTGAGgtacacacacacacacacacacaaacacacacacacattcaggctcacact
gcctcctcctgtgcggttctgcatgtgagtgtaggcagatgtctgacctgacc
tcgctgcacagctgaggatgcagatctgtcctgatgtgtgtgtgtgtgtgtgtgtgtgtgtgcagGTCTGCTATCTG
GCTCCGGTGCGGGGTCTGCAGCAGCTCTCCCTCCTCAGTAACCCGTGTGTGTCGT
GTGTGTCCTCCGCTGTGTGTGATTATCGGCCATATGTGCTCAGCTGGTGCCTCGG
CCTGGAGCTGCTAGATGGAGTCGCCGTCACACAGAAGGAAGG

For structure mapping (Figure 4C & D):

- AC-GT:

gtacacacacacacacacacacaaacacacacacacattcaggctcacactgcctcctcctgtgcggttctgcatgtgag
tgtaggcagatgtctgacctgacctcgctgcacagctgaggatgcagatctgtcctgatgtgtgtgtgtgtgtgtgtgtgt
gtgcag

- CON-GT:

gtcagtgctgtcacgatattgcttgatgttagagagaatacagaaaaaaaagaggtcattaaaagagaagtttagtccaag
ctgcctgatacagtcctaagcatgtgagtgtaggcagatgtctgacctgacctcgctgcacagctgaggatgcagatctgtc
ctgatgtgtgtgtgtgtgtgtgtgtgtgtgtgtgcag

- CON-pair:

gtcagtgctgtcacgatattgcttgatgttagagagaatacagaaaaaaaagaggtcattaaaagagaagtttagtccaag
ctgcctgatacagtcctaagcatgtgagtgtaggcagatgtctgacctgacctcgctgcacagctgaggatgcagatctgtc
tctctctaacatcaagcaatatcgtgacagcactgag (underlined is the complimentary region.)

**In vitro splicing to test the U1 snRNP and SR protein dependency**

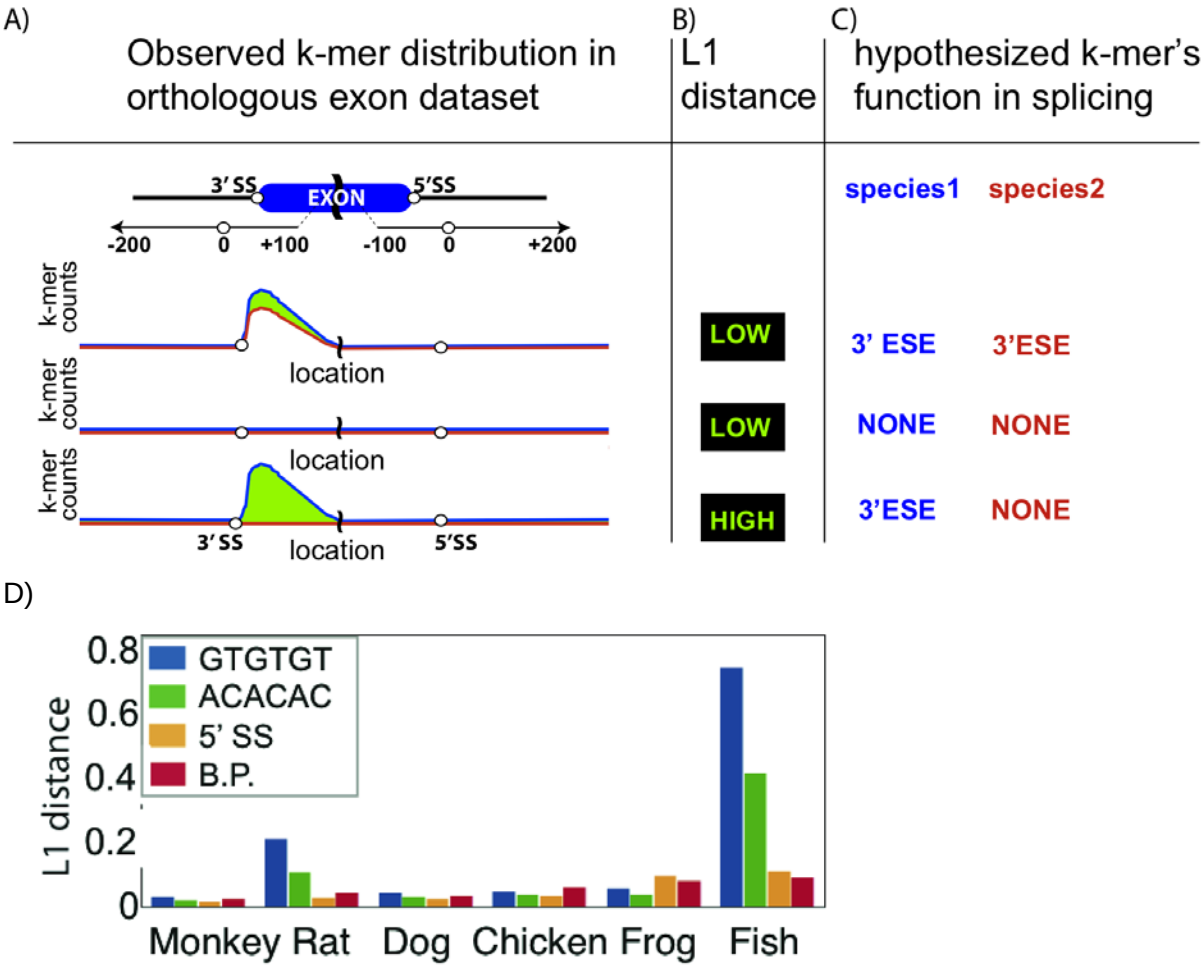
In vitro splicing was performed as described in the main Materials and Methods section. To block U1-dependent splicing, pre-incubate splicing reaction with 9 µM or 47 µM 2'-O-methyl ribonucleotide that targets 1-14 bases of U1 RNA: 5'-mUmGmCmCmAmGmGmUmAmAmGmUmAmU -3', and supplement the reaction with 1 nM substrate. To block SR protein-dependent splicing, pre-incubate reactions with 600 ng or 2 µg antibodies against SR proteins (1H4, Invitrogen).

**Intron size analysis**

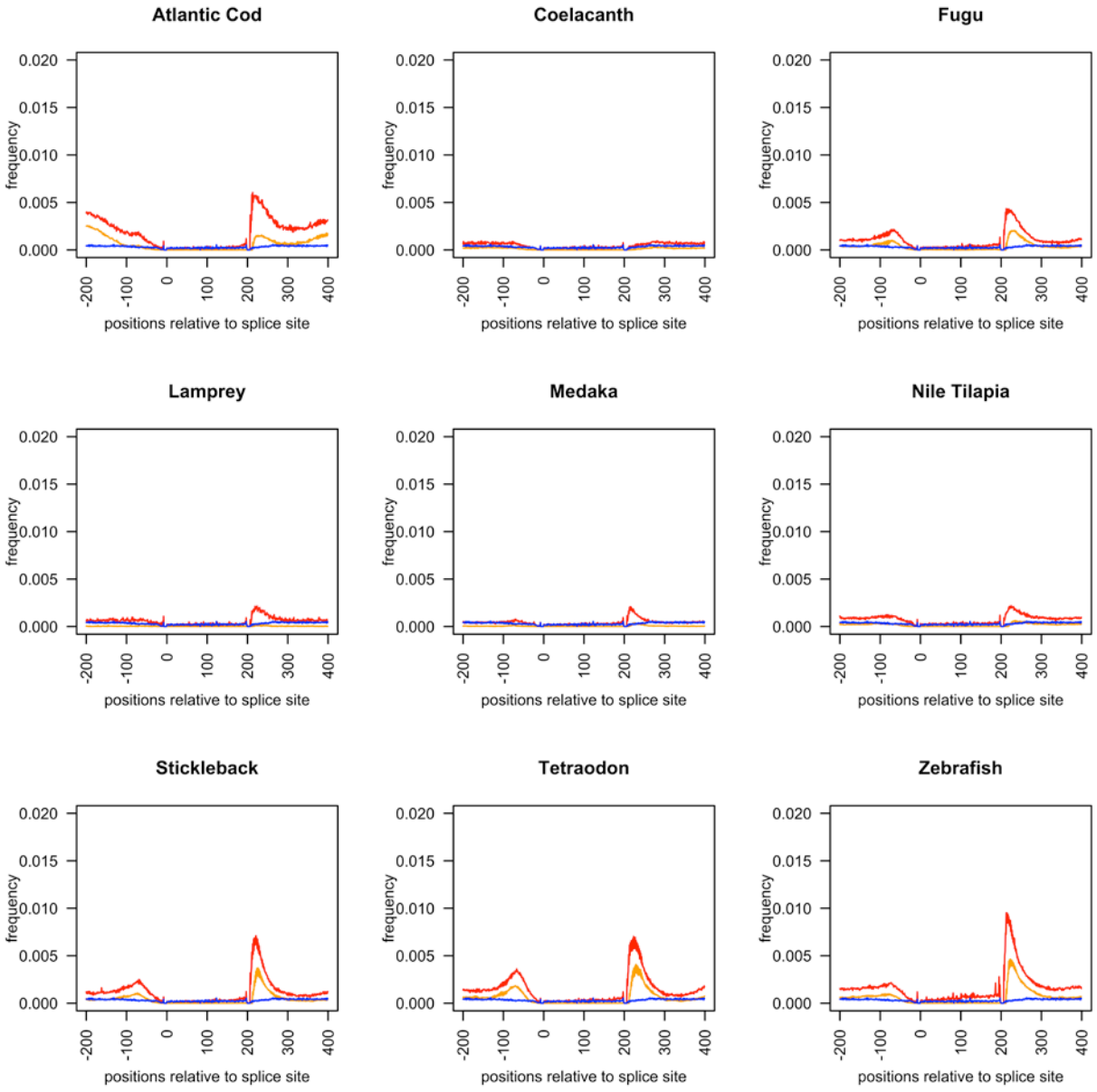
All introns and $(AC)_m$-$(GT)_n$ introns of fish genome were defined as above (see "Constructing orthologous coordinates across species" section). The difference of variance was calculated with one-sided F-test on the ln value of intron length. To test the length difference between $(AC)_m$-$(GT)_n$ and nonACGT introns, non-parametric Wilcoxon test was used (Supplementary Table 1). The correlation of intron length difference with their median intron length (Supplementary Fig. 6B) was calculated by spearman correlation test.

**Polypyrimidine Tract Identification and Scoring**

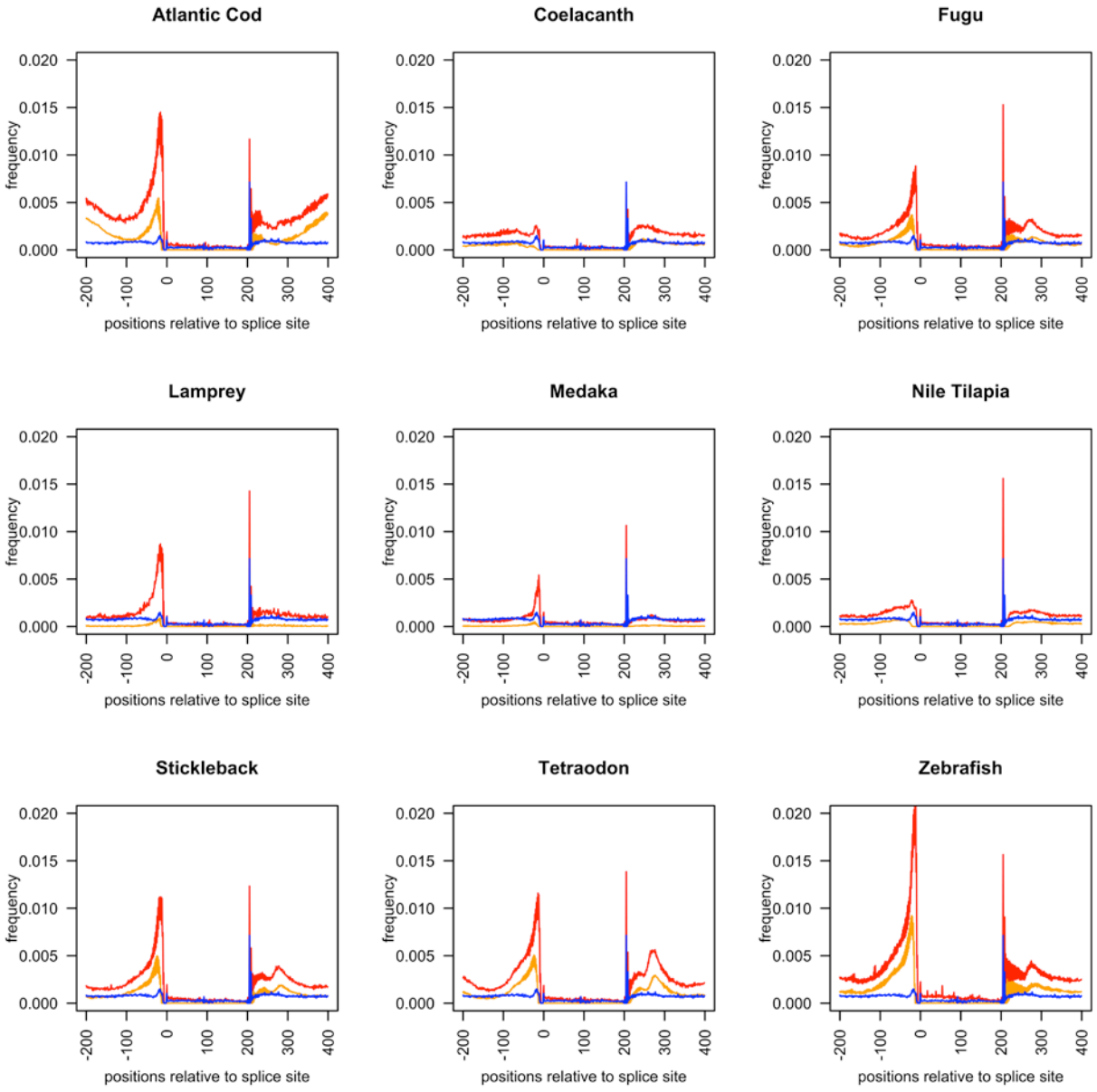
PPT's were identified and scored using the algorithm presented in Clark et al. Briefly, PPT's were identified in windows of 40 nucleotides immediately upstream of the annotated 3' splice sites. For each position in the window, nucleotides were scored (A=-2, G=-2, C = +2, U=+3) and positionally weighted (3 for central position, 2 for +1 and -1 positions, and 1 for +2 and -2 positions). Each position that had a score of at least 2 was considered for existing within a PPT run. Runs of potential PPT positions are required to start and end with a C or U nucleotide. If the run consists of >9 positions, it is flagged as a potential PPT. If the run consists of 5-9 positions and contains at least 5 U's, it is flagged as a potential PPT. Each potential PPT was scored by summing the scores of the nucleotides within the PPT (A=-2,G=-2,C=+2,U=+3), and the highest scoring PPT was selected for the analysis. Introns that did not contain any potential PPTs were assigned a score of 0. To calculate the significance between PPT strength distributions, the non-parametric Wilcoxon test was used.



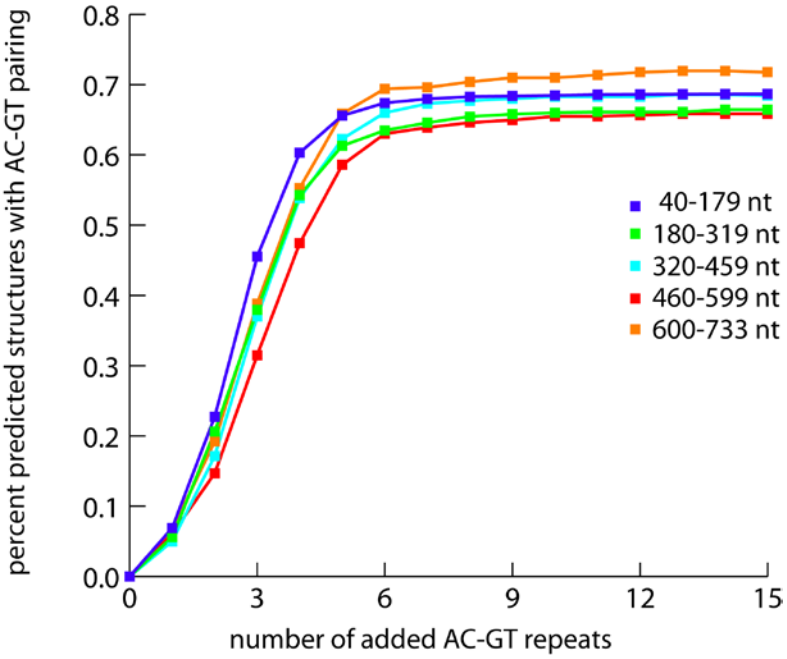

**Supplementary Figure 1. Detecting rapidly evolving cis-elements in pairwise species comparisons.** A) Datasets of all orthologous exon regions was created to study the distribution of k-mers around splice sites (exon diagram shows the size and location of these regions). All 4096 hexamers (k=6) were mapped relative to 3'ss and 5'ss in all exons where an ortholog could be found in the second species. The distribution (i.e. the number of counts plotted on the y-axis at each location (x-axis)) for a hypothetical hexamer was compared for two species (blue and red lines). The three possible outcomes envisioned are illustrated in three rows. The first outcome (top row) describes a functional k-mer with a similar distribution in both species. The area between the two curves (shaded green) is small. B) Therefore, L1 distance is low. C) This distribution is suggestive of a 3'ss ESE as it is enriched in exon positions close to the 3'ss. The second row describes a non-functional k-mer which also has a low L1 distance. The bottom row describes the k-mer whose distribution suggests species-specific function and will subjected to further analysis. D) Summary of L1 distance between human and specified species (x-axis) for AC repeat hexamers, GT repeat hexamers, 5'ss hexamers (GTAAGT) and branchpoint site hexamers (BP, CTAACA).



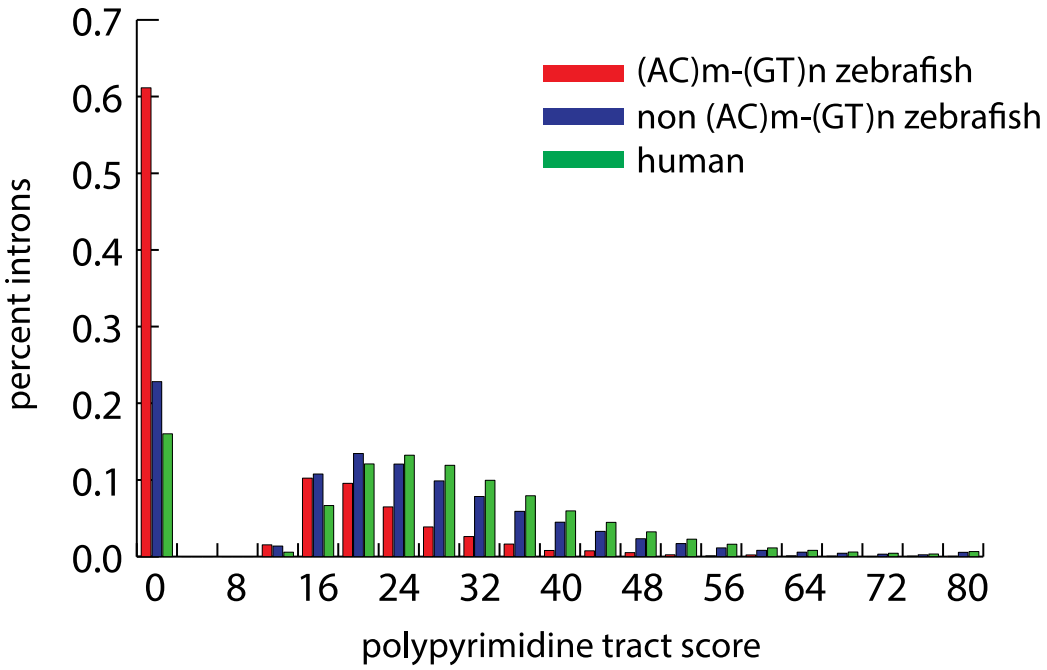

**Supplementary Figure 2. Distribution profiles of AC repeats between available fish genomes, lamprey, and human.** Plot of the frequency (y-axis) of AC repeat hexamers (red line) and 12-mer (orange line) around all 3' and 5'ss and the indicated genome and AC hexamer in human (blue line).



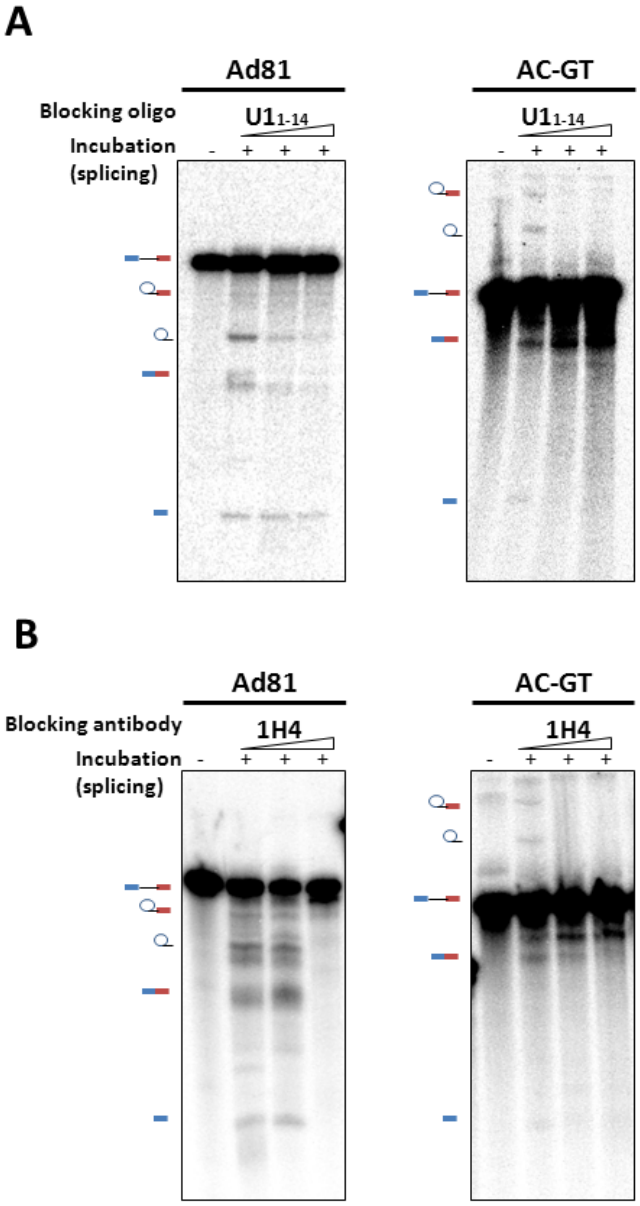

**Supplementary Figure 3. Distribution profiles of GT repeats between available fish genomes, lamprey, and human.** Plot of the frequency (y-axis) of GT repeat hexamers (red line) and 12-mer (orange line) around all 3' and 5'ss and the indicated genome and AC hexamer in human (blue line).



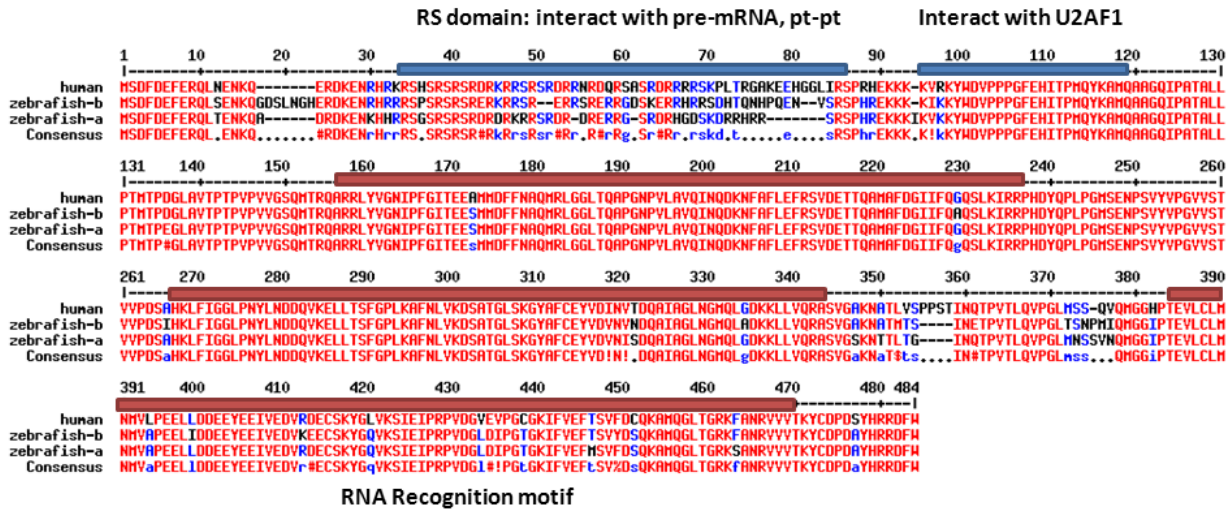

**Supplementary Figure 4. Required number of $(AC)_m$-$(GT)_n$ repeats on hairpin formation is independent of intron length.** $(AC)_m$-$(GT)_n$ repeat simulations were conducted by adding $(AC)_m$-$(GT)_n$ repeats of varying size to the 5'ss/3'ss of zebrafish introns. The percent of introns in which the $(AC)_m$-$(GT)_n$ repeats directed the structure is plotted against the number of added repeats. Binning the introns by length shows no effect of intron size on hairpin formation.



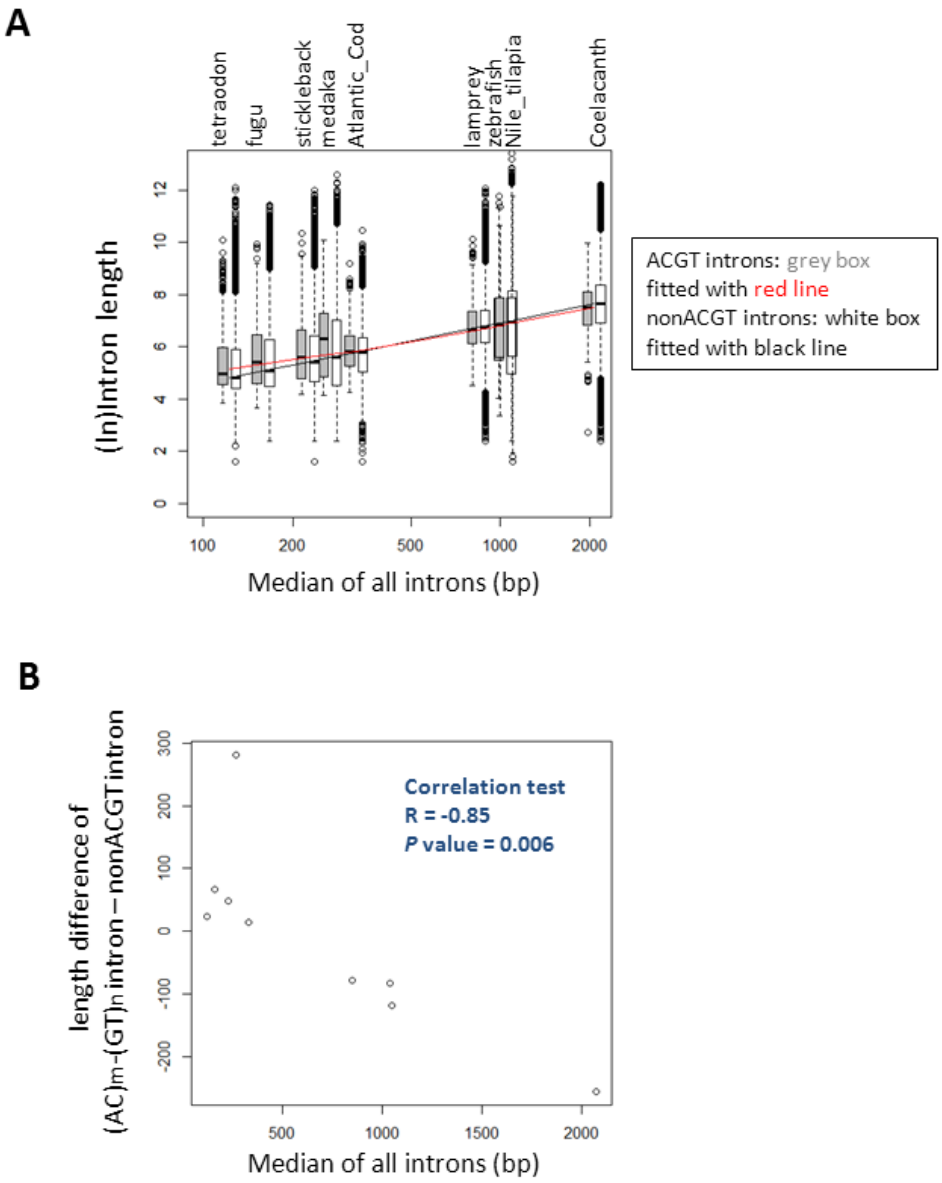

**Supplementary Figure 5. $(AC)_m$-$(GT)_n$ zebrafish introns contain fewer and weaker polypyrimidine tracts than non $(AC)_m$-$(GT)_n$ zebrafish introns.** Normalized histogram of PPT scores for $(AC)_m$-$(GT)_n$ zebrafish introns (red), non $(AC)_m$-$(GT)_n$ zebrafish introns (blue), and human introns (green). Determined by Wilcoxin test, *P* value between $(AC)_m$-$(GT)_n$ introns and non $(AC)_m$-$(GT)_n$ introns in zebrafish is less than $10^{-16}$; *P* value between non $(AC)_m$-$(GT)_n$ zebrafish introns and human introns is less than $10^{-16}$.



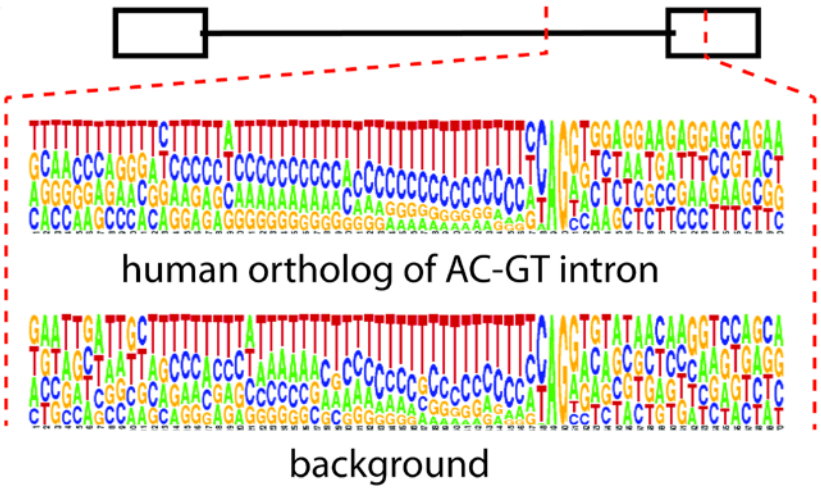

**Supplementary Figure 6. $(AC)_m$-$(GT)_n$ introns and control introns are both spliced in a U1 and SR-protein dependent manner.** In-vitro splicing substrates were prepared as in figure 6. A) The splicing of the Ad81 control was compared to no or increasing dosages of $U1_{1\text{-}14}$ antisense oligo complimentary to the first 14 bases of U1 RNA. The products were resolved on an 8M urea gel and visualized by autoradiography with a phosphorimager. B) The comparison described above was repeated with 1H4 antibody targeting pan-SR proteins.




**Supplementary Figure 7: Comparison of zebrafish and human U2AF65 homologs.** Protein sequence comparison of zebrafish U2AF2a, 2b and human U2AF65. Arg- and Ser-rich (RS) domain and U2AF1-interacting domain are indicated with blue bars and 3 RNA recognition motifs are indicated by red bars.




**Supplementary Figure 8: (AC)m-(GT)n introns evolve toward an optimal range of length.** A) Boxplot of ln(intron length) of (AC)m-(GT)n introns (grey box) and nonACGT introns (white box) of 9 fish genomes, ordered by their median of all intron length (x-axis). The variance of (AC)m-(GT)n introns are smaller than nonACGT introns (judged by the box sizes and F-test), indicating that (AC)m-(GT)n introns have a narrower range of optimal length within species. The smaller slope of the red fitted lines ((AC)m-(GT)n introns) than the black line (nonACGT introns) indicates that (AC)m-(GT)n introns are larger than average in genomes with short introns and shorter than average in genomes with larger introns. B) The length difference between (AC)m-(GT)n and nonACGT introns are plotted against their all intron sizes; one dot represents one fish genome. The significant negative correlation shows that (AC)m-(GT)n introns evolve toward a narrower range of optimal length across species.




**Supplementary Figure 9. Comparison of human (AC)m-(GT)n ortholog introns to control introns.** The aligned sequence composition of human orthologs of (AC)m-(GT)n was compared against the background revealing extended polypyrimidine tracts and purine-rich enhancer like sequences in the downstream exon.

**Supplementary Table 1**

| Genome | Assembly | Number of introns | Median intron size | Number of $(AC)_m$-$(GT)_n$ introns | Percent $(AC)_m$-$(GT)_n$ |
|---|---|---|---|---|---|
| Tetraodon | Mar. 2007 (Genoscope 8.0/tetNig2) | 159742 | 122 | 1524 | 0.95% |
| Fugu | Oct. 2011 (FUGU5/fr3) | 201455 | 168 | 1372 | 0.68% |
| Stickleback | Feb. 2006 (Broad/gasAcu1) | 179305 | 231 | 1955 | 1.09% |
| Medaka | Oct. 2005 (NIG/UT MEDAKA1/oryLat2) | 164026 | 267 | 695 | 0.42% |
| Atlantic cod | May 2010 (Genofisk GadMor_May2010/gadMor1) | 91058 | 329 | 1394 | 1.53% |
| Lamprey | Sep. 2010 (WUGSC 7.0/petMar2) | 60968 | 846 | 613 | 1.01% |
| Nile tilapia | Jan. 2011 (Broad oreNil1.1/oreNil2) | 186848 | 1047 | 822 | 0.44% |
| Zebrafish | Jul. 2010 (Zv9/danRer7) | 238767 | 1072 | 5461 | 2.29% |
| Coelacanth | Aug. 2011 (Broad/latCha1) | 144544 | 2079 | 135 | 0.09% |

**Supplementary Table 1, continued**

| Genome | Number of $(AC)_m$-$(GT)_n$ introns having human orthologs | Percent of $(AC)_m$-$(GT)_n$ introns having human orthologs | Median $(AC)_m$-$(GT)_n$ intron size |
|---|---|---|---|
| Tetraodon | 820 | 53.81% | 145 |
| Fugu | 730 | 53.21% | 226 |
| Stickleback | 1213 | 62.05% | 274 |
| Medaka | 405 | 58.27% | 547 |
| Atlantic cod | 808 | 57.96% | 342 |
| Lamprey | 285 | 46.49% | 769 |
| Nile tilapia | 286 | 34.79% | 928.5 |
| Zebrafish | 2710 | 49.62% | 958 |
| Coelacanth | 60 | 44.44% | 1819 |

**Supplementary Table 1, continued**

| Genome | variance of ln(ACGT) | variance of ln(nonACGT) | P val, test if variance within ACGT is smaller than nonACGT (F-test) | P val, test if ACGT length is bigger than nonACGT | P val, test if ACGT length is smaller than nonACGT |
|---|---|---|---|---|---|
| Tetraodon | 1.113 | 1.261 | 3.88E-04 | 1.29E-11 | 1.00E+00 |
| Fugu | 1.380 | 1.532 | 3.71E-03 | 1.79E-12 | 1.00E+00 |
| Stickleback | 1.437 | 1.439 | 4.89E-01 | 3.46E-09 | 1.00E+00 |
| Medaka | 2.046 | 2.094 | 3.43E-01 | 1.00E-12 | 1.00E+00 |
| Atlantic cod | 0.693 | 0.788 | 4.99E-04 | 7.57E-06 | 1.00E+00 |
| Lamprey | 0.979 | 1.178 | 9.28E-04 | 7.01E-01 | 2.99E-01 |
| Nile tilapia | 2.214 | 2.591 | 1.01E-03 | 9.53E-01 | 4.74E-02 |
| Zebrafish | 2.175 | 2.743 | 3.37E-31 | 3.75E-03 | 9.96E-01 |
| Coelacanth | 1.263 | 1.820 | 2.77E-03 | 9.12E-01 | 8.82E-02 |

**Supplementary Table 3: Co-occurrence frequency of complimentary k-mer motifs in human highly structured introns (HSI)**

| # of paired occurrences[a] | k-mer at 5' region | k-mer at 3' region | fraction HSI[b] | fraction (overall) | enrichment in HSI |
|---|---|---|---|---|---|
| 5 | GG | CC | 0.363941769 | 0.120704467 | 3.01514747 |
| 1 | GGGC | GCCC | 0.335075277 | 0.109781541 | 3.052200509 |
| 1 | CGG | CCG | 0.320766455 | 0.085849288 | 3.736390388 |
| 4 | GC | GC | 0.301480652 | 0.074484536 | 4.04756031 |
| 2 | GGC | GCC | 0.29538385 | 0.08161512 | 3.619229485 |
| 2 | CG | CG | 0.295259425 | 0.070446735 | 4.191243547 |
| 1 | GCG | CGC | 0.27236531 | 0.066003927 | 4.12650158 |
| 1 | CCG | CGG | 0.210277467 | 0.05182867 | 4.057165046 |
| 1 | GGCC | GGCC | 0.207042429 | 0.064003436 | 3.234864256 |
| 6 | GG | CC | 0.187632201 | 0.053129602 | 3.531594304 |
| 3 | GGG | CCC | 0.160009954 | 0.043900344 | 3.644845135 |
| 3 | CG | CG | 0.149931567 | 0.025920471 | 5.784291685 |
| 1 | CGGG | CCCG | 0.143834764 | 0.030841924 | 4.663611854 |
| 1 | CGC | GCG | 0.133134254 | 0.030841924 | 4.316664951 |
| 1 | GCCC | GGGC | 0.131890009 | 0.042169858 | 3.127589613 |
| 1 | GCGG | CCGC | 0.127037452 | 0.024754541 | 5.131884765 |
| 5 | GC | GC | 0.125295508 | 0.023060874 | 5.433250673 |
| 2 | GCC | GGC | 0.12479781 | 0.03365243 | 3.708433833 |
| 1 | GGGGC | GCCCC | 0.103147941 | 0.025319097 | 4.073918669 |

[a] The number of times the k-mer (and also its compliment) is required to occur in the first (and, for compliment, the last) 200 nucleotides of the intron to match the motif

[b] HSI – highly structured intron defined as the bottom ten percentile of introns ranked by free energy of folding.

Table displays all significant enrichments (P value < 0.0005) for k-mer motifs (k<11, min. enrich > 3) that occur in more than 10% of the introns.