Frequent Gain and Loss of Intronic Splicing Regulatory Elements during the Evolution of Vertebrates

Rodger B. Voelker[1], Steffen Erkelenz[2], Vinicio Reynoso[3], Heiner Schaal[2], and J. Andrew Berglund[1,*]

[1]Institute of Molecular Biology, Department of Chemistry, University of Oregon

[2]Institute of Virology, Heinrich-Heine-University, Düsseldorf, Germany

[3]Institute for Cell and Molecular Biology, University of Texas

*Corresponding author: E-mail: aberglund@molbio.uoregon.edu.

Accepted: May 12, 2012

Abstract

Splicing regulatory elements (SREs) are sequences bound by proteins that influence splicing of nearby splice sites. Constitutively spliced introns have evolved to utilize many different splicing factors. The evolutionary processes that influenced which splicing factors are used for splicing of individual introns are generally unclear. We demonstrate that in the lineage that gave rise to mammals, many introns lost U-rich sequences and gained G-rich sequences, both of which resemble known SREs. The apparent conversion of U-rich to G-rich SREs suggests that the associated splicing factors are functionally equivalent. In support of this we demonstrated that U-rich and G-rich SREs are both capable of promoting splicing of an SRE-dependent splicing reporter. Furthermore, we demonstrate, using the heterologous MS2 tethering system (bacterial MS2 coat fusion-protein and its RNA stem-loop binding site), that both the U-rich SRE-binding protein (TIA1) and the G-rich SRE-binding protein (HNRNPF) can promote splicing of the same intron. We also observed that gain of G-rich SREs is significantly associated with G/C-rich genomic isochores, suggesting that gain or loss of SREs was driven by the same processes that ultimately resulted in the formation of mammalian genomic isochores. We propose the following model for the gain and loss of mammalian SREs. Ancestral U-rich SREs located in genomic regions that were experiencing high rates of A/T to G/C conversion would have suffered frequent deleterious mutations. However, this same process resulted in increased formation of functionally equivalent G-rich SREs, and acquisition of new G-rich SREs decreased purifying selection on the U-rich SREs, which were then free to decay.

Key words: introns, evolution, splicing, hnRNP, splicing enhancer, GC-rich.

Introduction

Most vertebrate pre-mRNAs are interrupted by noncoding introns that are removed by the spliceosome during mRNA maturation. Early experiments revealed that the 5′ and 3′ splice sites located at the intron or exon boundaries are compositionally constrained and are directly recognized by the spliceosome (Wahl et al. 2009). A key requirement for proper splicing is that the spliceosome accurately identifies these splice sites before carrying out the splicing reaction. Although the splice site sequences themselves are necessary for splicing, it is now well established that they are not generally sufficient for promoting proper splicing in mammals. Auxiliary signals, so-called splicing regulatory elements (SREs) located near the splice sites, were discovered to support splice site recognition (Matlin et al. 2005; Pozzoli and Sironi 2005; Wang and Burge 2008).

Although SREs have been studied primarily in the context of regulated alternative splicing, they are also required for constitutive splicing. Most constitutively spliced exons contain SREs known as exonic splicing enhancers (ESEs) (Wang and Burge 2008). ESEs are mainly bound by members of the SR protein family, which generally promote splicing and inclusion of the exon (Long and Caceres 2009; Mueller and Hertel 2011). In addition to ESEs, many, if not most, exons are flanked by intronic splicing enhancers (ISEs). Generally, SREs have low sequence complexity (i.e., they are composed primarily of simple repeats). Common examples include U-, C-, G-, UC-, UG-, and CA-repeats (Zhang et al. 2003; Zhang and Chasin 2004). SREs generally function as binding sites for various RNA-binding proteins (Akindahunsi et al. 2005; Martinez-Contreras et al. 2007; Venables et al. 2008). More than a dozen intronic localized SRE-binding proteins have been identified. These proteins can be divided into several

paralogous families. As a group, they represent a diverse set of proteins that share related function, but unlike SR proteins most do not share a common splicing activation domain.

The biological significance for this diversity in splicing factors is not entirely clear. A subset of these proteins may have evolved specialized roles in regulating alternative splicing (e.g., Nova, MBNL, and PTB) (Wang and Burge 2008). However, evidence suggests that some splicing factors also participate in constitutive splicing. For instance, computational studies revealed that intronic regions flanking constitutively spliced exons are generally enriched for sequences resembling SREs (Zhang et al. 2003). More recently, experimental mapping of in vivo binding sites for several splicing factors [e.g., HNRNPF/H (Xiao et al. 2009), TIA1/TIAL (Wang Z, et al. 2010), and HNRNPC (König et al. 2010)] demonstrated that these proteins bind adjacent to many constitutively spliced exons.

Do all constitutively spliced exons utilize the same set of splicing factors? Many proteins have been shown to bind SREs and promote splicing, and different exons appear to utilize different subsets of these proteins. The biological factors that dictate which splicing factor is used are unclear. It is possible that there is "fine-tuning" between splicing and gene expression (e.g., coordinating tissue-specific gene expression with tissue-specific expression of different splicing factors). However, advantages gained by such coordination would have to be strong enough to drive selection for the creation and maintenance of specific SREs in all of the introns or exons within a given transcript. Alternatively, it is possible that, at least with regards to constitutive splicing, some splicing factors are functionally redundant and that the SREs that are downstream of a given exon are simply due to chance selection from among a number of different, but functionally equivalent, splicing factors. According to such a model, random mutations could result in the chance gain of a binding site for an alternate splicing factor resulting in the creation of a functionally redundant binding site. This would relax purifying selection on the original site, which, if it decays, results in the gain of a new SRE. This model is analogous to the "binding site turnover" model that has been proposed to explain the rapid gain or loss of transcription factor binding sites (Doniger and Fay 2007; Zheng et al. 2011).

Although binding site turnover has been explored with regards to transcription, the role that similar processes play in gain or loss of SREs has not been systematically explored. Here we demonstrate that a similar process appears to be operating on SREs downstream of 5′ splice sites. We show that there have been dynamic changes in the SRE composition of orthologous introns during the evolution of vertebrates. In particular, there has been loss of U-rich SRE-like sequences resembling TIA1/TIAL-binding sites in many mammalian introns. Concomitant to this loss was the frequent gain of G-rich sequences resembling HNRNPF/H-binding sites. The majority of these gain or loss events correlate with the split between Synapsida (mammals) and Sauropsida (reptiles and birds). In order to determine if the gain or loss events affected functional SREs, we also examined the subset of introns where splicing was shown to be dependent on HNRNPH (Xiao et al. 2009). As expected, most of the mammalian orthologs of these introns contain HNRNPH-like binding sites; however, we show that these sites are often absent in nonmammalian orthologs. In a similar manner we demonstrate that, between mammals, there has been differential loss of TIA1/TIAL1-binding sites (Wang Z, et al. 2010).

Furthermore, we show that these gain or loss events are highly correlated with introns located in the most G/C-rich regions of mammalian genomes. These regions have higher substitution rates than other regions (Dreszer et al. 2007). This accelerated mutational rate is likely driving this shift in SRE composition. Current evidence indicates that these G/C-rich regions are due to higher rates of fixation of A/T to G/C because of gene conversion (Dreszer et al. 2007; Galtier et al. 2009). Thus, it appears that the identity of the splicing factors that are associated with many constitutively spliced exons is less related to splicing, gene expression, or gene function, than it is to the history of the genomic region within which the intron lies. Finally, these observations suggest that TIA1/TIAL and HNRNPF/H proteins are functionally equivalent, at least with respect to the splicing of many constitutively spliced introns. We provide experimental evidence supporting this hypothesis. Overall, these results indicate that, since the evolutionary split between birds and mammals, binding site turnover of SREs was common and that it appears to have played a major role in shaping the landscape of mammalian SREs.

Materials and Methods

Creation of Database of Orthologous Intronic Sequences

In order to create a database of orthologous intronic sequences, we first extracted the coordinates of all human exons as annotated in the UCSC Known Genes hg18 human gene model (release March 2006) and the NCBI36/hg18 release of the human genome. The human exon coordinates were used to identify coordinates for orthologous exon or intron boundaries based on alignment with human exons as indicated in the UCSC multiple sequence alignment of 44 vertebrate genomes (hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz44way/). These coordinates were used to extract orthologous intronic sequences from the masked genomic sequence files for each organism. Because SREs tend to be located near splice junctions, we extracted intronic sequences corresponding to the first 100 nt. Introns that were shorter than 200 nt were simply cut in half. We also excluded the first 7 nt to avoid biases related to the splice sites themselves. Because many splicing factor binding sites contain mono- and dinucleotide repeats, we

unmasked sequences that were classified as "low complexity" or "simple repeats" as indicated in the Repeat Masker files associated with each sequence. For this analysis, we were primarily interested in introns of the U1/U2 class (which represent >99% of all human introns (Levine and Durbin 2001). We therefore filtered introns to remove those that were less than 30 nt in length and did not begin with the nucleotides "$G^C/_U$" and end with "AG." To remove "GU-AG" U11/U12 introns, we also excluded introns beginning with the sequence "$^G/_A$UAUCCU" (Levine and Durbin 2001). Finally, the remaining sequences were filtered to remove redundant sequences. The final numbers of nonredundant orthologous intronic flanks that were identified are indicated in supplementary figure S1, Supplementary Material online.

Clustering

K-means clustering of the pentamer frequencies was performed using a combination of custom software and the "cluster 3.0" clustering library (de Hoon et al. 2004). The feature vectors consisted of the average frequencies for each pentamer observed in the flanks for each organism. Euclidian distances (de Hoon et al. 2004) were used to calculate the similarities between organisms, and the final number of clusters was set at 16.

Classification of Sequences According to Isochore Type

The isochore type for each intronic flank was assigned by mapping the coordinate of the donor splice site against the consensus isochore coordinates reported by Schmidt and Frishman (2008). Isochore coordinates and types were obtained from the online database "Isobase" (http://www.geneinfo.eu:8080/isobase/).

Plasmids

The HIV-1-derived splicing reporter SV-env/eGFP was generated by substitution of the BamHI/XhoI fragment of SV-env (Kammler et al. 2001) for an amplicon obtained from a PCR with the primer pair #1848/#1849 (supplementary table S1, Supplementary Material online) and pEF eGFP-neo (kindly provided by Dirk Lindemann as a template). For construction of SV GAR^- D1-env/eGFP, a PCR-amplified fragment (#3164/#640, supplementary table S1, Supplementary Material online) was inserted into SV-env/eGFP using restriction sites EcoRI/NdeI. Plasmids SV GAR^- D1 TIA-1(2x)-env/eGFP, SV GAR^- D1 hnRNPF/H(2x)-env/eGFP, SV GAR^- D1 IAS-1(2x)-env/eGFP, and SV GAR^- D1 2MS2-env/eGFP were obtained by insertion of SacI/NdeI-digested PCR products using specific forward primers (#3181/, #3195, #3189, and #3163, supplementary table S1, Supplementary Material online) and the reverse primer #640 (supplementary table S1, Supplementary Material online) into SV-env/eGFP. Cloning of MS2 fusion protein expression vectors was described previously (Singh et al. 2010). For construction of the MS2 single-chain variants SV-scNLS-MS2 ΔFG-TIA-1 and SV-scNLS-MS2 ΔFG-hnRNP F, the BamHI/XhoI of SV-scNLS-MS2 ΔFG-ΔRS (Singh et al. 2010) was replaced by PCR-amplified fragments using primer pairs #2869/#2870 and #2780/#2781 (supplementary table S1, Supplementary Material online), respectively, and pGFP-TIA-1 (kindly provided by Juan Valcarcel) and pFlag-hnRNP F (kindly provided by Massimo Caputi) as templates.

Cell Culture and RT–PCR Assay

HeLa-$T4^+$ cells were cultivated in Dulbecco's high glucose modified Eagle's medium (DMEM Invitrogen) containing 10% fetal calf serum and 100 μg/ml PenStrep (Invitrogen). Transient transfection experiments were performed in six-well plates with 2.5×10^5 cells per well using Trans-IT LT1 (Mirus Corporation) according to the instructions provided by the manufacturer. Total RNA was harvested from the cells 30 h following transfection with the specific constructs and pXGH1 for control. Reverse transcription was performed with 4 μg of total RNA. Traces of plasmid DNA were removed by DNA digestion with 10 U of RNAse-free DNase I (Roche). DNAse I was heat-inactivated at 70°C for 5 min and cDNA synthesis was carried out for 1 h at 50°C and 15 min at 72°C under conditions as follows: 200 U Superscript III RNAse H– Reverse Transcriptase (Invitrogen), 7.5 pmol oligo(dT)12-18 (Invitrogen) as primer, 20 U of RNAsin (Promega), and 10 mM of each deoxynucleoside triphosphate (Qiagen). For semiquantitative analysis of spliced eGFP mRNAs, cDNA was used in a PCR with forward primer #3210 (supplementary table S1, Supplementary Material online) and reverse primer #3211 (supplementary table S1, Supplementary Material online). An additional PCR was performed with primers #1224 (supplementary table S1, Supplementary Material online) and #1225 (supplementary table S1, Supplementary Material online) to specifically detect constitutively spliced GH1 (human growth hormone 1) mRNA expressed from pXGH5. RT–PCR products were separated on 8% nondenaturating polyacrylamide gels and stained with ethidium bromide for exposure with the Lumi-Imager (Roche).

Calculation of the Relative Conservation (C_R)

We used the UCSC Genome Browser primate phyloP scores (Pollard et al. 2010) to represent the conservation at a particular nucleotide (obtained from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phyloP44way/primates). For each kmer, the average phyloP score within the kmer was determined. The relative conservation C_R was then defined to be the difference between the average conservation for all instances of the kmer and the average conservation for all kmers within the set of sequences being analyzed.

Results

In order to explore the evolution of vertebrate SREs, we constructed a database of orthologous introns from 44 vertebrate organisms using the UCSC multiple genome sequence-alignment of 44 vertebrate genomes (Fujita et al. 2011). Orthologous exon or intron boundaries were identified using the human hg18 Known Genes annotation as a reference (see Materials and Methods). Because spliceosome formation is initiated by recognition of the 5′ splice site, and this, in turn, is supported by nearby SREs (Zhang and Chasin 2004), we focused our analysis on donor intronic flanks. Donor flanks were defined to be the intronic positions from 8 to 108 nt (relative to the donor splice junction, fig. 1*A*). The first 7 nt of the introns were ignored because they correspond to the highly constrained sequences that can interact directly with the U1 snRNA. After filtering (see Materials and Methods), we ended up with 203,018 nonredundant human donor intronic flanks.

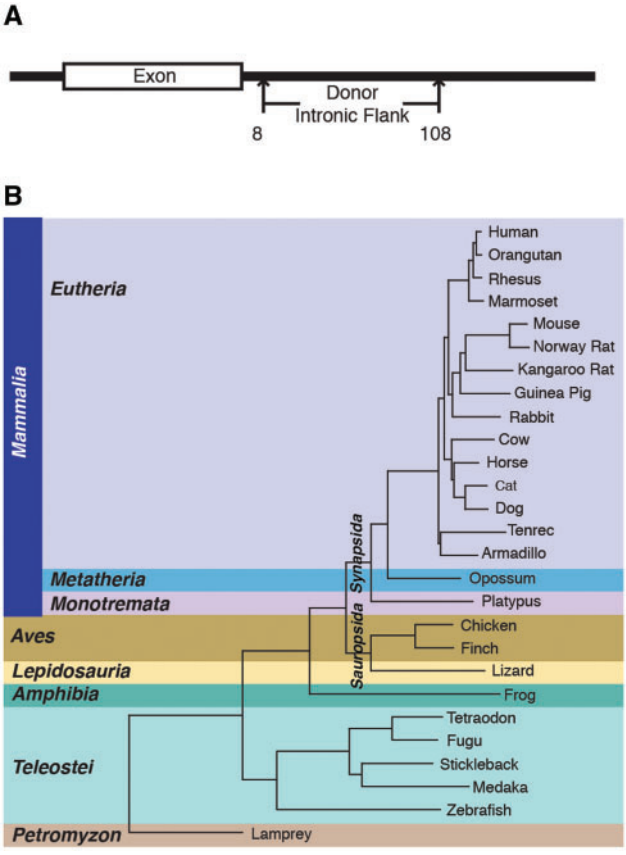

Fig. 1.—Phylogenetic relationship of the 27 vertebrates used in this study. (*A*) Schematic of the intronic region analyzed in this study. The "donor flank" corresponds to nt +8 to +108 relative to the donor splice junction. (*B*) Phylogenetic tree for the organisms used in this study. The tree is based on that presented by Murphy et al. (2001) and the distances are those reported in the UCSC Genome Browser (Kent et al. 2002).

The coordinates for human introns were used to extract the corresponding sequences from other vertebrates. These were then filtered to remove sequences that were unlikely to represent U1/U2 introns (see Materials and Methods). Sequences orthologous to most human introns were identified in the genomes of the other mammals in this study, and orthologs for approximately half of all human introns were identified for nonvertebrates. An exception to this was lamprey where orthologs were identified for only 14% of human introns. Most of the genomes used in this study are not complete, therefore, the differences in the degree of coverage most likely reflects the degree of completeness of sequencing of the various genomes. Nonetheless, in all cases we were able to identify thousands of orthologous intronic flanks for each organism (supplementary fig. S1, Supplementary Material online), which should enable clade-specific trends in intronic composition to be inferred. From the original 43 organisms, 27 organisms were chosen for further analysis (fig. 1*B* and supplementary fig. S1, Supplementary Material online). These 27 were chosen based on completeness of coverage relative to human, error rate, and balancing phylogenetic coverage.

Phylogenetic Variation in SRE Frequencies among Vertebrates

In order to compare the general composition of vertebrate donor intronic flanks, we first calculated the frequencies of all 5-mer oligonucleotides (pentamers) within each set of orthologous sequences. Pentamers were chosen because most RNA-binding proteins have comparably sized binding sites. Pentamer frequencies therefore reflect the frequencies of potential RNA-binding sites. In order to study the variations for specific pentamers and to identify those having clade-specific variations, we used k-means clustering (de Hoon et al. 2004) to cluster pentamers according to their frequencies across the organisms in the study. To establish the optimal number of clusters, we first allowed the number to vary from 2 to 20. A scree-plot of the between-cluster distances revealed that approximately 90% of the total variation is accounted for by using 16 clusters (supplementary fig. S2, Supplementary Material online). We therefore chose 16 clusters for the final analysis. Using this procedure we were able to cluster pentamers having similar frequency profiles across the vertebrate taxa (e.g., see supplementary fig. S3, Supplementary Material online). Heat-map representations of the pentamer frequencies for each organism and the resulting clusters are shown in figure 2. The full list of pentamers composing each cluster is presented in supplementary table S2, Supplementary Material online.

As shown in figure 2, most pentamers have similar frequencies in the intronic flanks of all vertebrates; however, some show remarkable variations between organisms. The most prominent clade-specific variations are seen when comparing

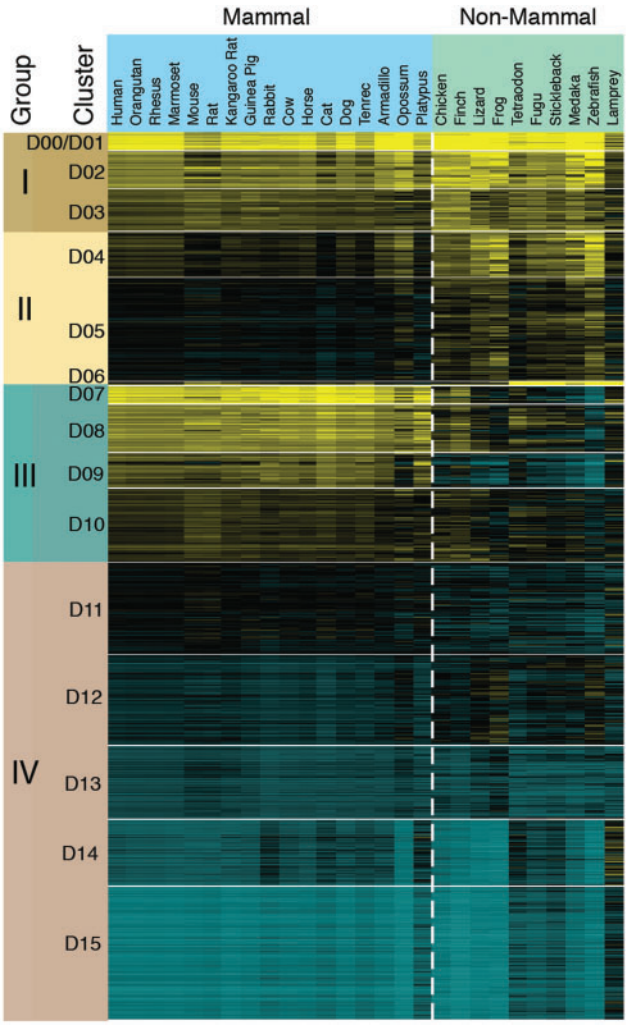

Fig. 2.—Frequencies of pentamers in the donor flanks of 27 vertebrates. Heat-map representation of the frequencies of all pentamers in the donor intronic flanks of the organisms in this study. Prior to clustering, the frequencies for each organism were mean centered using the mean frequency for the organism. Yellow indicates frequencies above the mean and cyan indicates frequencies below the mean. In both cases the intensity of the color corresponds to a greater difference from the mean. Pentamers are arrayed along the *Y*-axis as organized by the clustering algorithm. White horizontal lines separate pentamer clusters. The dashed white line indicates the separation between mammals and nonmammals.

mammals to nonmammals. We therefore grouped the pentamer clusters into the following categories: (I) pentamers that are common in all vertebrate intronic flanks, (II) pentamers that are enriched in nonmammalian flanks but are rare in mammalian flanks, (III) pentamers that are enriched in mammalian flanks but are rare in nonmammalian flanks, and (IV) pentamers that are rare in all vertebrate intronic flanks.

The pentamers that are most relevant for this study are those that occur with high frequency in intronic flanks (groups I–III) because these are more likely to represent common SREs. The G/C content of the pentamers in these groups reveals an overall trend: pentamers that are frequent in all vertebrate flanks (group I) are highly A/U-rich while those that are abundant primarily in mammalian flanks (group III) are highly G/C-rich (supplementary fig. S4, Supplementary Material online). Although the greatest difference between group I and group III pentamers is G/C content, simple G/C content alone does not explain the clustering results. For instance, many of the group III pentamers have an overall G/C content similar to group IV pentamers (e.g., compare D07 and D09 to D14, supplementary fig. S4, Supplementary Material online). Also, clusters D01 and D02 (which are abundant in all vertebrates) have similar A/U content to pentamers in cluster D04, which are more common in nonmammalian flanks (supplementary fig. S4, Supplementary Material online).

Many of the most frequent pentamers are similar to known SREs. For instance, group I contains pentamers with U-repeats (see clusters D00–D03, supplementary table S2, Supplementary Material online). Enrichment of U-repeats in donor-flanks has been previously observed (Aznarez et al. 2008; Gal-Mark et al. 2009), and these sequences have been shown to function as SREs. Several splicing factors are known to bind U-rich sequences. These include HNRNPC (Görlach et al. 1994; König et al. 2010), members of the ELAV/Hu family (Wang H, et al. 2010), and TIA1/TIAL1 (Aznarez et al. 2008; Wang Z, et al. 2010). Although these proteins have been largely studied in the context of alternative splicing, transcriptome wide mapping of TIA1/TIAL1-binding sites in human cells revealed that many binding sites are located in the donor flanks adjacent to constitutively spliced exons, suggesting that these proteins play a general role in splicing (Wang Z, et al. 2010). We observed that 36 out of the top 50 pentamers that are most highly enriched in TIA1/TIAL1-binding sites (determined using iCLIP) (Wang Z, et al. 2010) also clustered in group I. It is therefore reasonable to propose that many of the pentamers in group I represent binding sites for proteins of the TIA1/TIAL1 and/or ELAV/Hu families.

Group II pentamers are more abundant in nonmammalian downstream flanks than in mammalian flanks. In general, the difference in frequencies between mammals and nonmammals for pentamers in group II is less than that seen for group III. Among this group are pentamers containing UG- and AC-repeats (D06). Sequences containing these repeats are known to act as SREs. UG-repeats are bound by CELF proteins and have been shown to influence splicing (Barreau et al. 2006). AC-repeats have also been shown to act as SREs and appear to be targets for HNRNPL (Hui et al. 2005). These pentamers are especially highly enriched in fish introns (fig. 2), which is consistent with a previous study showing that they act as SREs in Fugu (Yeo et al. 2004).

Pentamers in group III show the most significant phylogenetic variation. These are highly abundant in mammalian intronic flanks but are much rarer in nonmammalian intronic flanks (fig. 2). Many of the pentamers in this group contain

G-repeats, which are known to be enriched in mammalian flanks (Nussinov 1988; Zhang et al. 2003), and have been shown by many to function as SREs (Sirand-Pugnet et al. 1995; Cogan et al. 1997; McCullough and Berget 1997; Van Laer et al. 1998; Lew et al. 2004; Masuda et al. 2008; Xiao et al. 2009). Proteins that are members of the HNRNPA/B and HNRNPF/H families bind G-rich sequences and act as splicing regulators (He and Smith 2009; Han et al. 2010). These proteins have especially high affinity for purine-rich sequences containing the triplet "GGG" (Burd and Dreyfuss 1994; Dominguez et al. 2010), and 30 out of the 40 pentamers that contain G-triples are clustered in group III. Many of the pentamers in group III also contain C-repeats that can function as SREs (Murray et al. 2008) and may be bound by poly(rC)-binding proteins such as PCBP1 (Yeap et al. 2002). CU-repeats are also common in this group, and similar sequences have been shown to be targets of the polypyrimidine-tract binding proteins PTBP1/PTBP2, which are well-known regulators of alternative splicing (R Singh et al. 1995; Reid et al. 2009; Llorian et al. 2010).

Group IV (rare in all vertebrate intronic flanks) contains the largest number of pentamers. Not surprisingly, many of these pentamers are similar to splice site sequences (supplementary table S2, Supplementary Material online), which are known to be under-represented in intronic flanks (Zhang et al. 2003). Also common in group IV are pentamers containing CpG di-nucleotides (see clusters D14–D15, supplementary table S2, Supplementary Material online) which are generally underrepresented in vertebrate genomes due to their high methylation induced conversion to TpA (Ehrlich and Wang 1981; Arndt et al. 2005).

An earlier study by Yeo et al. (2004) examined SRE variations between human, mouse, and Fugu introns. We feel that our results support the primary findings of Yeo et al. Both analyses demonstrate that CA- and UG-repeats are more prominent in fish introns than in mammals. Both analyses reveal that G-rich sequences are common in mammalian introns but are rare in fish. Interestingly, our results indicate that U-rich intronic sequences are abundant in all vertebrate introns; however, these were not identified by Yeo et al. This apparent discrepancy could be due to differences in experimental design, and this issue remains to be further explored.

Because more genomic sequences are now available, we can also expand upon these earlier findings by more accurately defining the phylogenetic distribution of G-rich intronic sequences. G-rich sequences occur at relatively low levels in lamprey, fish, frog, and lizard introns (fig. 2). There is a slight increase in avian introns (fig. 2) and then a dramatic increase in all mammalian introns. When considering the vertebrate phylogeny, two possibilities exist. 1) There was an initial expansion of G-rich sequences in the common ancestor of amniotes, but they were then generally lost in reptiles. Meanwhile, there was a second expansion in the lineage that gave rise to mammals. 2) Enrichment of G-rich sequences in birds and mammals is due to two independent events.

The Primary Differences between Mammalian and Nonmammalian Introns Are Related to the Formation of Genomic Isochores

The observations above imply that during the evolution of vertebrates there was differential gain or loss of certain SRE-like sequences in intronic flanks. What were the biological forces that contributed to these variations? Initial analyses of mammalian genomes revealed that mammalian chromosomes generally contain large (>100 kb) regions with very high overall G/C content while other regions have very low overall G/C content. These compositionally distinct regions are often referred to as isochores (Bernardi 2000). Isochores are generally divided into five types based on their overall G/C content: L1, L2, H1, H2, and H3 (listed by increasing G/C content). Although all vertebrate genomes contain some G/C-rich isochores, they represent a significant proportion of mammalian genomes (Costantini et al. 2006, 2009), and the pattern of G-rich enrichment in intronic sequences (fig. 2) closely matches that observed for genomic G/C-rich isochores (Duret and Galtier 2009). The precise mechanisms that created isochores have not been established. Isochores are associated with DNA replication timing (Costantini and Bernardi 2008b; Watanabe et al. 2009) and recombination rates (Fullerton et al. 2001), and there is increasing evidence that their formation is related to G/C-biased gene conversion (Duret and Galtier 2009).

Because introns, exons, and intergenic regions in G/C-rich isochores have a higher G/C content than other regions (Costantini and Bernardi 2008a) it seems reasonable that the formation of intronic G-rich SREs is related to the formation of mammalian isochores. In order to explore this possibility, we first categorized human introns according to the isochore type within which they are found using the consensus isochore assignments of Schmidt and Frishman (Schmidt and Frishman 2008). We then restricted our attention to two sets of pentamers that likely represent binding sites for known splicing factors. The first set (referred to as U-rich, table 1) is composed of the 36 pentamers in group I that were shown to be significantly enriched near TIA1/TIAL1-binding sites (Wang Z, et al. 2010), and therefore may represent binding sites for TIA1/TIAL1 or other U-rich binding factors. The second set, referred to as G-rich (table 1), is composed of the 30 pentamers in group III that contain the trimer GGG, which (as discussed above) are potential binding sites for HNRNPF/H proteins (Schaub et al. 2007).

In order to quantify kmer composition, we developed a statistic referred to as kmer density (D_U or D_G), which is the number of U- or G-rich kmers (from table 1) per 100 nt. Figure 3*A* and 3*B* shows the distributions of D_U and D_G for human introns according to their isochore type. There is a strong

correlation between isochore type and pentamer densities, and most importantly, because D_G increases with increasing overall isochore G/C content (i.e., from L1 to H3), we propose that the majority of the G-rich pentamers present in mammalian introns arose as a consequence of the formation of G/C-rich isochores.

It has been shown that U- and G-rich sequences cluster near 5′ splice sites in the donor flanks of mammalian introns (Nussinov 1988; Zhang et al. 2003), and this enrichment has been interpreted as an indication that these sequences function as SREs. Interestingly, if we examine this clustering according to isochore type it is evident that the degree of enrichment near the 5′ splice site is highly dependent on the type of isochore that the intron is located (fig. 3*C* and 3*D*). In fact, for introns in H2 and H3 isochores there is very little if any enrichment of U-rich pentamers above the background (fig. 3*C*), and conversely there is little if any enrichment of G-rich pentamers above the background for introns in L1 and L2 isochores (fig. 3*D*).

When we examine the D_U and D_G distributions for orthologous sequences from other vertebrates, several interesting evolutionary trends can be observed (figure 4). All organisms show a trend similar to that seen in figure 3 for human donor flanks (i.e., D_U decreases from L1 to H3 and D_G increases from L1 to H3). Because the orthologous flanks were mapped using the human isochore profile, this suggests that some ancient isochore patterns were set up before the divergence of vertebrates and these have been preserved to some degree in other vertebrates. However, the differences in D_U and D_G between isochore types are much greater for mammals than nonmammals (compare the changes in the median values for the different isochores). Thus, although ancient isochore patterns appear to have been set up before the split between Synapsida (mammals) and Sauropsida (reptiles and birds), the divergence in sequence composition between different isochoric regions proceeded to a much greater extent in the mammalian lineage.

Table 1
U-Rich and G-Rich Pentamers Chosen for Closer Analysis

U-rich				
UUUUU	UUUAA	GUUUC	CUUUG	AAUUU
GGUUU	UAUUU	UUUGU	UUUCA	UGUUU
CUUUC	UUUGG	CUUUU	AUUUC	ACUUU
UUUUG	AUUUG	UUUCC	UUUAU	UUUGC
UCUUU	UUUUA	AUUUA	AUUUU	AGUUU
GAUUU	GUUUU	UUUCU	GUUUG	UUUUC
CCUUU	UUUGA	CAUUU	GCUUU	
G-rich				
AGGGG	GGGCA	UGGGU	GUGGG	AUGGG
GGGUC	GGGGA	AGGGA	CUGGG	AGGGU
AGGGC	GGGAC	GGGAG	GGGAU	GGGGU
GGGCU	UUGGG	GAGGG	CCGGG	GGGGG
GGGGC	GGGAA	GGGUU	UGGGA	
GGGCG	GGGCC	CGGGG	GGGUG	
AAGGG	GCGGG	UGGGC	CAGGG	

Note.—The pentamers were chosen as described in the text.

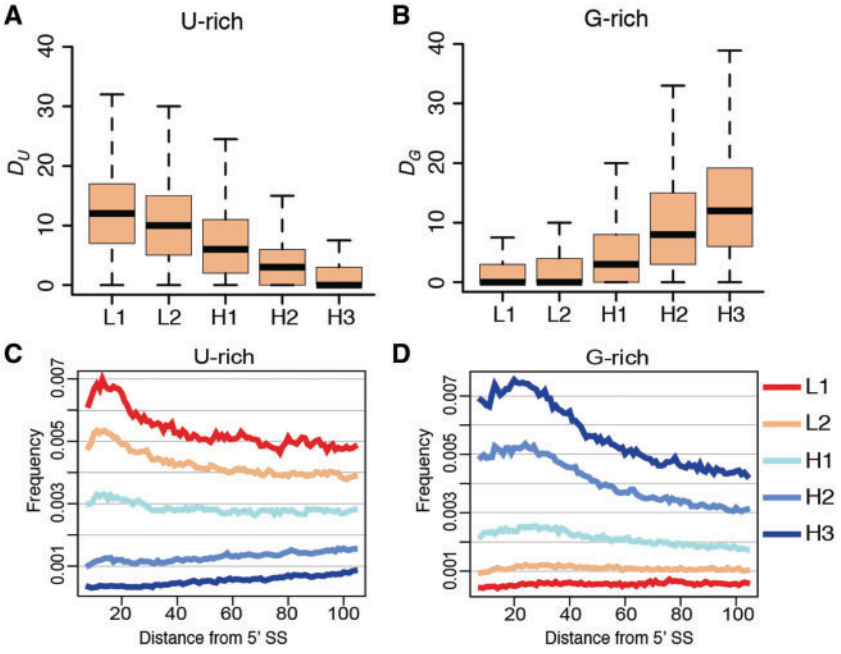

Fig. 3.—Relationship between isochore type and pentamer frequencies in human donor intronic flanks. (*A* and *B*) Boxplots showing the distributions of U- and G-rich pentamer densities (D_U and D_G) for human donor intronic flanks according to the type of isochore that they are located in. (*C* and *D*) Histograms showing the frequencies of U- or G-rich pentamers (from table 1) in human donor flanks relative to the 5′ splice site.

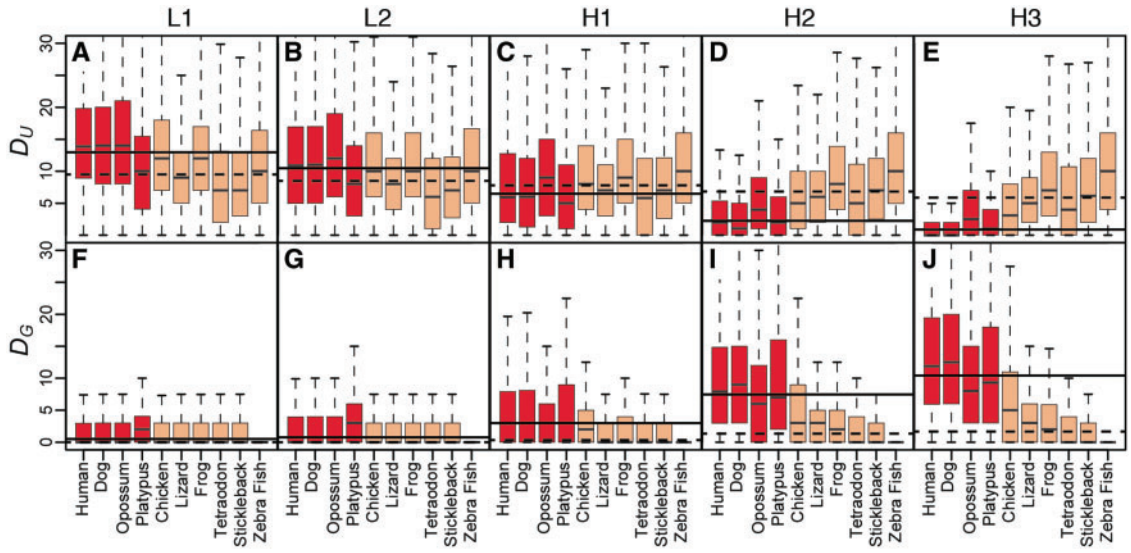

Fig. 4.—Distributions of U- and G-rich pentamer densities for orthologous donor flanks. Each panel shows the distributions of D_U or D_G values for donor flanks that are orthologous to human flanks located in each type of isochore. In all cases the isochore type is that for the human sequences. The median D_U and D_G scores for mammals and nonmammals are indicated as horizontal lines (solid for mammals, dashed for nonmammals).

In all vertebrates G-rich kmers are virtually absent in intronic flanks that are orthologous to those in human L1 and L2 isochores, and in no case do the median D_G values for a nonmammal equal those of mammals (fig. 4 and data not shown). Birds have D_G values that are in between mammals and other vertebrates (see chicken, fig. 4*I* and 4*J* and Finch, data not shown). This is consistent with previous observations that bird genomes also have prominent isochores (Costantini et al. 2007, 2009). However, despite bird genomes having isochores similar to those in mammals, G-rich pentamers are not as highly enriched in bird introns (fig. 2).

Considering the vertebrate phylogeny and that most nonmammals have G-poor introns, we can conclude that the ancestral form of vertebrate intronic flanks must have been generally U-rich and G-poor. The dramatic increase in mammalian introns also indicates that the majority of mammalian G-rich SRE-like sequences were gained in the mammalian lineage after the divergence of Synapsida (mammals) and Sauropsida (reptiles and birds) and that flanks that gained G-rich SRE-like sequences generally lost U-rich SRE-like sequences.

Functional SREs Have Isochore Dependent Distributional Biases

In what way, if any, did the formation of isochores impact ancestral mammalian SREs? Two major possibilities exist. Either, ancestral SREs were preserved through purifying selection or purifying selection was not strong enough to maintain them, and in G/C-rich isochores they were lost entirely or were replaced by new SREs that arose during or after isochore formation. To distinguish between these models, we examined relationships between isochores and two classes of experimentally validated SREs.

First, we examined the set of human donor flanks that are adjacent to exons shown by Xiao et al. to require HNRNPH for normal splicing (Xiao et al. 2009). For this we focused on the 214 exons that showed changes in splicing upon siRNA induced knockdown of HNRNPH. After cross-referencing these against our database we ended up with 203 unique donor intronic flanks. As shown above, G/C-rich isochores are enriched for sequences similar to HNRNPH-binding sites. However, it is possible that these isochore-related G-rich sequences do not function as SREs and are essentially neutral sequences that arose during isochore formation. According to this model we would not expect a correlation between the location of functional HNRNPH SREs and isochore type. In contrast to this model, we observed that intronic flanks with functional HNRNPH SREs are not uniformly distributed throughout the genome, but instead significantly cluster in the most G/C-rich isochores H2 and H3 (fig. 5*A*). It is difficult to imagine an experimental factor that could cause a bias between changes in splicing upon knockdown of HNRNPH and isochore type. Instead, we propose that these SREs evolved from G-rich sequences that were created by the same processes that created genomic isochores.

This hypothesis is further supported in figure 6, which shows heat maps comparing the D_U and D_G values for all pair-wise human and nonhuman orthologs of HNRNPH responsive donor flanks. The majority of the human sequences are very G-rich (high D_G) which is consistent with the binding specificity of HNRNPF/H proteins, and these are highly

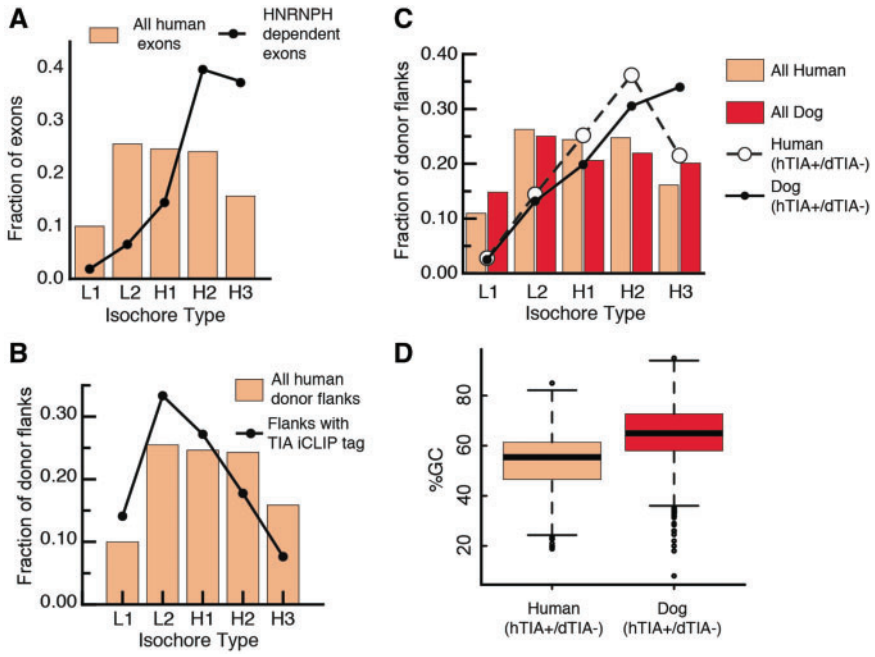

Fig. 5.—Relationships between isochore types and experimentally determined SREs. (*A*) Distribution of all human exons by isochore type and those shown to be dependent on HNRNPH for normal splicing (Xiao et al. 2009). The likelihood that these samples were drawn from the same distribution is $<2.2 \times 10^{-16}$ (according to the χ-test). (*B*) Distribution of all human donor flanks by isochore type and those shown to have TIA1/TIAL1 binding sites (Wang Z, et al. 2010). The likelihood that these samples were drawn from the same distribution is $<2.2 \times 10^{-16}$ (according to the χ^2-test). (*C*) Distribution of all human and dog donor flanks by isochore type, and those that are orthologous to human flanks having TIA1-binding sites but for which the dog flank has no U-rich pentamers (the human orthologous flanks are labeled "Human [hTIA$^+$/dTIA$^-$]"; the dog orthologous flanks are labeled "Dog [hTIA$^+$/dTIA$^-$]"). (*D*) Distributions for the %G/C of the human and dog hTIA+/dTIA- flanks from panel C.

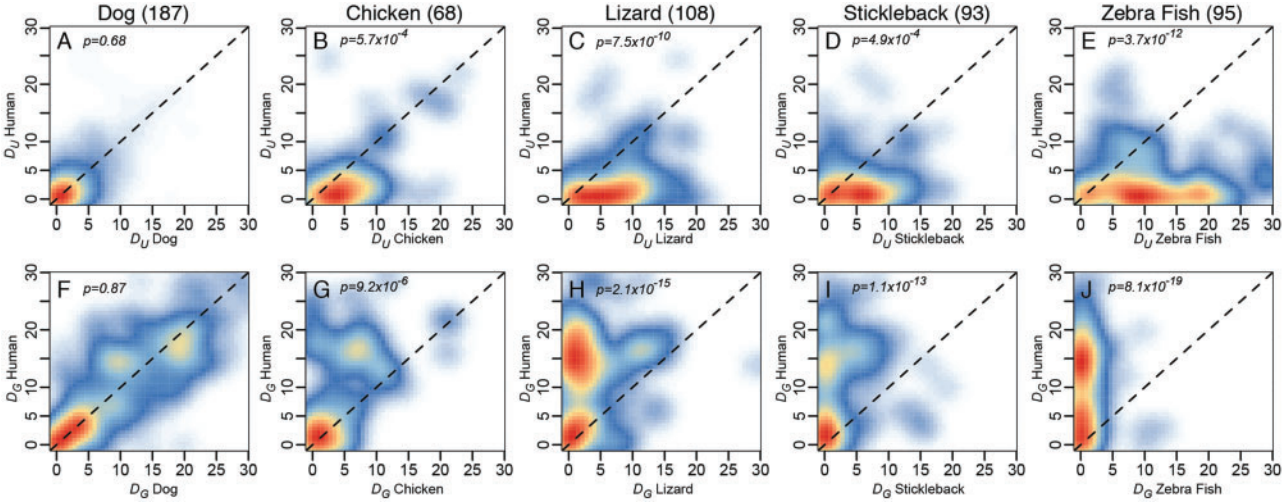

Fig. 6.—Correlations for U- and G-rich pentamer densities between human and nonhuman flanks that are orthologous to human flanks dependent on HNRNPH for normal splicing. The U- and G-rich pentamer densities (D_U and D_G) were calculated for sequences orthologous to the human flanks that were shown by Xio et al. to be dependent on HNRNPH for normal splicing (Xiao et al. 2009). Heat maps were generated by plotting these values versus the values for the corresponding human sequences. The heat maps were generated using the "sm.density" R package and a custom color ramp. The numbers of orthologous flanks (from the 203 human flanks) that were identified for each organism is shown next to the name of the organism. The *P*-values (from a paired *T*-test) are shown in each panel.

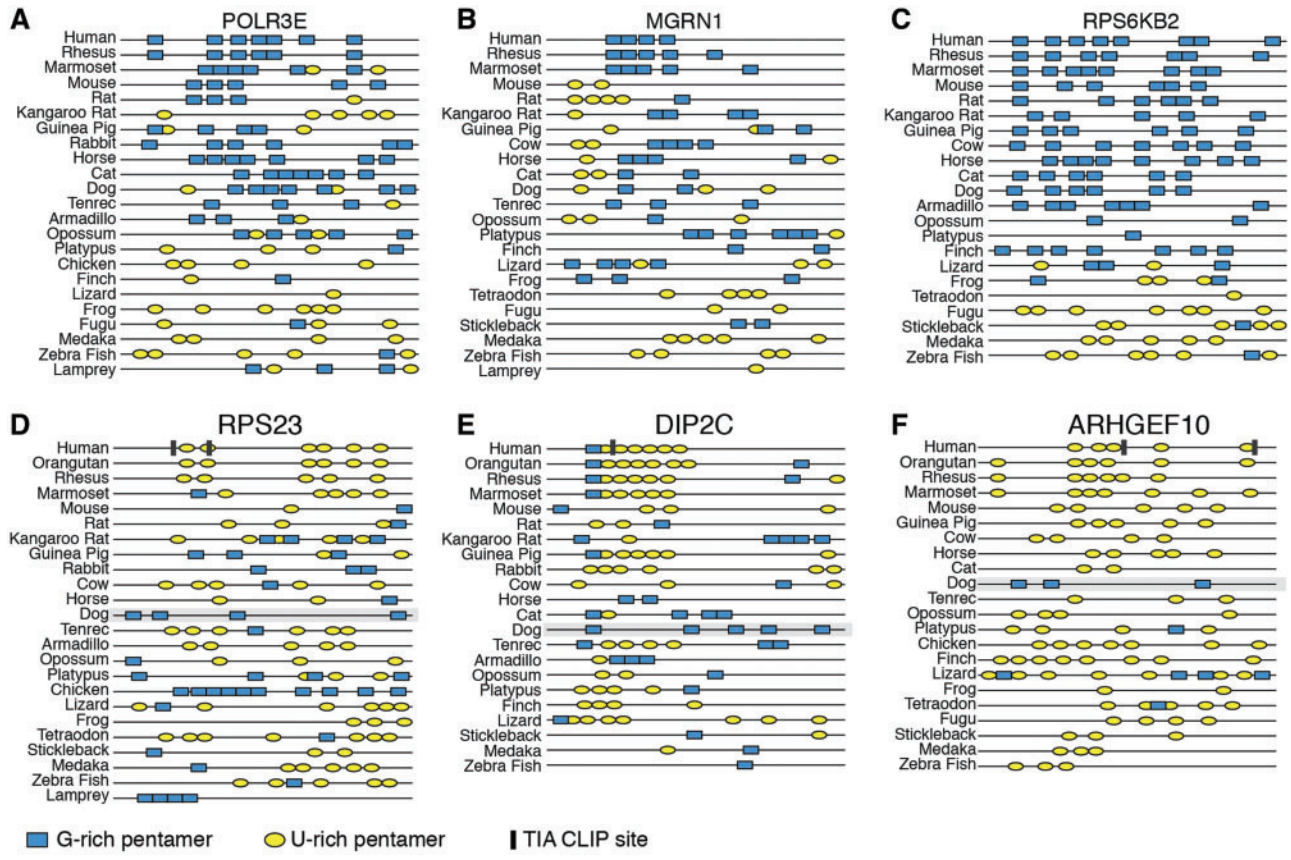

Fig. 7.—Schematic comparison of the positions of U- and G-rich pentamers in several sets of orthologous donor flanks. Shown are schematic depictions of the positions of all U-rich pentamers (as ovals) and G-rich pentamers (as boxes) from table 1 within the first 100 nt for six sets of orthologous donor intronic flanks. The actual sequences and genomic coordinates are presented in supplemental table S3, Supplementary Material online. (*A–C*) Flanks chosen from the set of human donor flanks that were shown to require HNRNPH for normal splicing and that have HNRNPH CLIP tags in the donor intronic flanks (Xiao et al. 2009). (*D–F*) Flanks chosen from the set of 3,557 flanks for which the human flank has a mapped TIA1/TIAL1-binding site (Wang Z, et al. 2010) but the corresponding dog sequence has no U-rich pentamers. The positions of TIA1/TIAL1 iCLIP cross-links are indicated as bars.

correlated with the D_G values observed in other mammals (e.g., see dog fig. 6*F*). In contrast, the nonmammalian flanks generally have low D_G values (fig. 6*G–J*) but have high D_U values (fig. 6*B–E*). The most parsimonious explanation for these data is that the majority of human HNRNPH SREs are not ancestral to vertebrates but evolved from G-rich sequences created during the formation of mammalian genomic isochores.

Figure 7*A–C* show schematic representations of the locations of all G- and U-rich pentamers (from table 1) within orthologs of a subset of the donor flanks that showed both a change in splicing upon knockdown of HNRNPH and also had multiple HNRNPH CLIP tags (Xiao et al. 2009). Within these examples there are many instances of gain/loss of individual SRE-like sequences, and within each there are nonmammalian orthologs that completely lack G-rich pentamers. If we assume that the orthologous HNRNPF/H proteins have similar binding specificity, then in contrast to human, splicing of the orthologous introns must not be dependent on HNRNPF/H proteins. In place of G-rich pentamers, the nonmammalian orthologs generally have U-rich sequences, which could be functioning as U-rich SREs.

The data presented above indicate that the mutagenic process associated with formation of G/C-rich isochores provided a forward process for the creation of new mammalian-specific G-rich SREs. What was the fate of ancestral SREs that were present before isochore formation? Were they preserved by purifying selection? As discussed above, ancestral introns were likely to have been U-rich, and it is reasonable to assume that some contained ancestral U-rich SREs. If these were maintained by purifying selection then U-rich SREs would be uniformly distributed throughout the genome. To address this we examined the distribution of donor flanks containing TIA1/TIAL1 iCLIP tags (Wang Z, et al. 2010). From the dataset of Wang *et al*. who used iCLIP to map TIA1/TIAL1 sites in human RNAs (Wang Z, et al. 2010), we identified all human donor flanks having a TIA1/TIAL1-binding site within 100 nt of the 5′ splice site. This yielded a set of 59,777 unique donor intronic

flanks. Figure 5B shows that these also have an isochore-dependent distributional bias, but are biased toward the most A/T-rich isochores L1 and L2. If before isochore formation, ancestral introns in H2 and H3 isochores had TIA-binding sites at levels comparable to those in other regions then the relative decrease in these regions reveals a biased loss of TIA sites in these regions. The loss of such sites indicates that ancestral TIA sites were not maintained by purifying selection. We must therefore conclude that either the majority of TIA CLIP sites are not functional (and not experiencing purifying selection) or that a significant proportion of functional TIA SREs were lost in introns lying in regions that became H2 and H3 isochores.

Although isochore patterns are generally conserved between mammals, there are clade- and species-specific differences in the degree of G/C enrichment for orthologous regions (Romiguier et al. 2010). In the same way that isochore formation contributed to differences between mammalian and nonmammalian introns, differences between mammalian regional isochore profiles could have contributed to clade- and species-specific differences between mammals. Examples of this can be seen in figure 7*A* and 7*B* where the kangaroo rat POLR3E and mouse MGRN1 sequences lack G-rich pentamers but instead have U-rich pentamers.

To look for additional evidence that mammalian-specific differences in SREs are related to species-specific differences in isochore formation, we examined dog sequences that are orthologous to human flanks having TIA CLIP tags to identify dog flanks that have no U-rich pentamers and are therefore unlikely to possess a TIA-binding site. From the set of donor sequences with TIA1/TIAL1-binding sites, 3,557 of the dog orthologs completely lack U-rich pentamers. These exceptional dog sequences have a higher G/C-content than the corresponding human sequences (fig. 5*D*). Schematic representations of the locations of U- and G-rich pentamers and mapped TIA1/TIAL1 sites for three of these are shown in figure 7*D*–*F*. In each of these examples most of the organisms have U-rich SREs; however, these are absent in dog, which instead have G-rich sequences. In both human and dog, these introns generally reside in H2 and H3 isochores (fig. 5*C*), suggesting that the differences between the human and dog sequences are due to different levels of A/T to G/C conversion during or after isochore formation. Furthermore, if we assume that at least some of the human TIA-binding sites are functional SREs, then the dog introns that are lacking these sites must not require this protein for normal splicing.

Functional Equivalency of U-Rich and G-Rich SREs

The trend that we observe is that mammalian introns in H2 and H3 isochores generally lost U-rich sequences and gained G-rich sequences. This suggests that the loss of a U-rich SRE could be mitigated by the gain of a G-rich SRE, and, at least for some introns, TIA1/TIAL1 and HNRNPF/H are functionally equivalent. In order to directly verify this hypothesis, we experimentally tested the abilities of high-affinity protein-binding sites for the splicing regulatory proteins TIA1/TIAL1 or HNRNPF/H to rescue splicing when placed downstream of an SRE-dependent 5′ splice site. For this we used an HIV-1-based splicing reporter (fig. 8*A*) whose 5′ splice site has been previously shown to be dependent on a purine-rich ESE, called GAR (Kammler et al. 2001; Caputi et al. 2004). Inactivation of this ESE leads to loss of 5′ splice site recognition and therefore we termed this sequence as "neutral." To test whether TIA1/TIAL1- or HNRNPF/H-binding sites downstream of this 5′ splice site can restore its recognition, we inserted two copies of each binding site into the downstream intronic flank (fig. 8*B*) and analyzed their impact on 5′ splice site recognition. As previously shown, inactivation of the GAR ESE led to loss of 5′ splice site recognition (fig. 8*D*, lane 1 and 8*E*, neutral). On the other hand, insertion of two copies of any binding site for TIA1/TIAL1 (TIA1 or IAS-1) or HNRNPF/H (HNRNPF/H) restored 5′ splice site recognition (fig. 8*D*, compare lanes 1 with lanes 2 and 3 and fig. 8*E*) demonstrating that D1 can be activated by an SRE equally well from the exonic or intronic flank. Because the exemplary chosen binding site for HNRNPF/H, however, seemed to be somewhat less supportive for splice site activation (fig. 8*E*, cf. TIA1 and IAS-1 with HNRNPF/H), we decided to compare TIA1 and HNRNPF for their 5′ splice site supportive activity independently from their tested binding sites by recruiting these proteins via the MS2-tethering system to the intronic position. Thus, we inserted two MS2 RNA hairpins into the downstream flank (fig. 8*C*) and cotransfected HeLa cells with this splicing reporter (SV GAR^{-} D1 2MS2-env/eGFP) and a plasmid coding for an MS2 coat fusion protein to monitor splicing-dependent eGFP fusion protein expression. Indeed, recruiting TIA1 and HNRNPF in a comparable manner to the RNA downstream of the 5′ splice site revealed that they were functionally equivalent and thus interchangeable (fig. 8*F*, cf. MS2-HNRNPF with MS2-TIA1). We therefore feel it is reasonable to propose that this functional equivalency extends to constitutively spliced introns in general and that gain of G-rich HNRNPF/H-binding sites could alleviate the otherwise deleterious effects of the loss of U-rich SREs.

Model for the Gain and Loss of Mammalian SREs

The evolutionary changes in vertebrate SREs offer interesting insights into the mechanisms by which new *cis* sequence elements can be acquired. Gain of new *cis* elements requires a forward mutational process to create sites capable of binding *trans*-acting factors. Upon exposure to selective pressures, a subset of these may evolve to become functional regulatory elements. The correspondence between the frequencies of the G-rich pentamers in table 1 and functional HNRNPH sites (i.e., both are enriched in H2 and H3 isochores) demonstrates that D_G is a good proxy for the frequency of sites capable of binding HNRNPH. When we examine the frequencies of G-rich pentamers in other nongenic regions

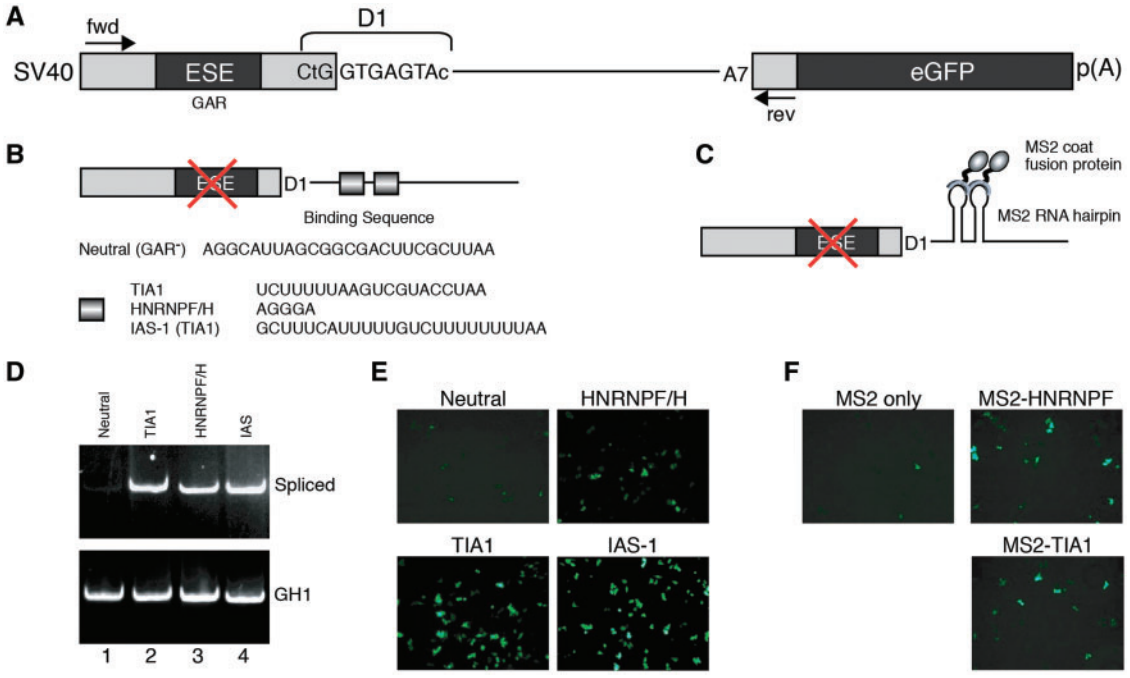

Fig. 8.—TIA-1 and hnRNP F/H can functionally substitute each other to activate splicing of an upstream localized 5′ splice site. (*A*) Schematic of the one-intron HIV-1 based SV D1-env/eGFP splicing reporter. Activation of the test 5′ splice site D1 is dependent on an SRE. eGFP is expressed from the spliced mRNA. (*B*) Inactivation of the exonic splicing enhancer GAR (GAR^{-}; neutral, N) leads to loss of U1 snRNP binding to the 5′ss D1 (Kammler et al. 2001; Caputi et al. 2004). High-affinity-protein binding sites for the splicing regulatory proteins TIA-1 (Dember et al. 1996; Del Gatto-Konczak et al. 2000) or hnRNP F/H (Schaub et al. 2007) were inserted immediately downstream of D1. The sequences are given below. (*C*) Fluorescence microscopy of transfected HeLa cells to monitor splicing-dependent eGFP expression for each of the constructs. (*D*) RT–PCR analysis of RNA isolated from HeLa cells, which were cotransfected with the splicing constructs and pXGH5 (human growth hormone 1, GH1) to normalize the transfection efficiency. Positions of the primers used are given in (*A*). (*E*) High-affinity RNA-binding sites were replaced by two MS2 RNA hairpins allowing site-directed tethering of MS2 coat fusion proteins downstream of 5′ss D1. (*F*) Fluorescence microscopy of HeLa cells cotransfected with the SV GAR^{-} D1 2MS2-env/eGFP splicing reporter, SVcrev and the respective plasmid for an MS2 coat fusion protein to monitor splicing-dependent eGFP expression. SV40: Simian Virus 40 early promoter, p(A): polyadenylation signal.

(e.g., deeper intronic and intergenic sequences), we see that these same pentamers are enriched throughout G/C-rich isochores, not just in donor flanks (fig. 9*A* and *B*). This demonstrates that the processes that created isochores, though unrelated to splicing, also resulted in the creation of large numbers of sequences similar to binding sites for HNRNPF/H proteins. Although G-rich pentamers occur at high frequencies throughout G/C-rich isochores, they are most abundant in donor intronic flanks (fig. 9*A* and *B*). This increased enrichment suggests that some of the G-rich sequences that arose near 5′ splice sites are experiencing a greater degree of selection while those in more distal intronic regions or in intergenic regions are experiencing increased rates of decay.

If this is true we would expect that G-rich pentamers located near 5′ splice sites to be more highly conserved than those located deeper within the intron. In order to measure this, we developed a score referred to as relative conservation (C_R) (see Materials and Methods). The C_R score is based on the phyloP score (Pollard et al. 2010), which is a phylogenetically weighted measure of conservation at each nucleotide using genomic multiple sequence alignments. The C_R score is calculated for a kmer by comparing the distributions of the average phyloP score for all instances of the kmer versus the average phyloP scores for all kmers in the set of sequences. C_R is therefore a measure of how the mean conservation of a kmer differs from the background conservation within the sequences being analyzed. For this analysis, we used the primate phyloP scores and calculated C_R for all kmers in the first 50 nt of donor flanks and compared these against the values for the last 50 nt.

As shown in figure 9*C* and *D*, U-rich pentamers on average are more highly conserved than the background in all isochores; however, the conservation decreases in the most G/C-rich isochores. With the exception of H3 isochores, U-rich pentamers that are proximal to 5′ splice sites have higher C_R scores than those that are more distal, suggesting that these are experiencing stronger selection. Meanwhile in L1 and L2 isochores, G-rich pentamers have a lower than average phyloP score (fig. 9*D*). However, both proximal and distal G-rich pentamers are conserved above background in H1–H3 isochores,

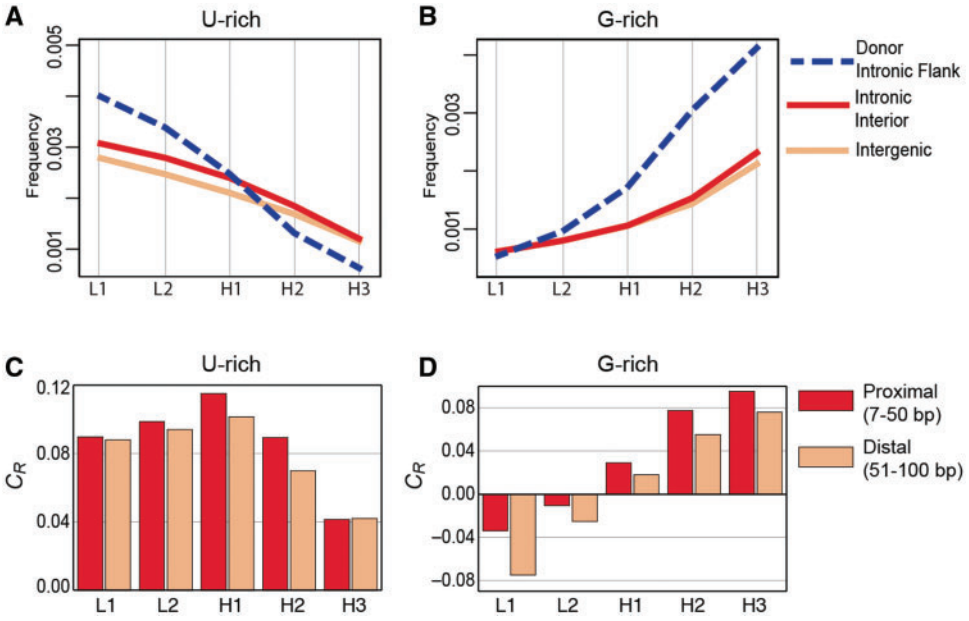

Fig. 9.—U- and G-rich pentamers are more highly enriched in donor intronic flanks than in other noncoding regions and are more highly conserved adjacent to 5′ splice sites. (*A* and *B*) Shown are the average frequencies of the U- and G-rich pentamers from table 1 in donor intronic flanks, deeper intronic sequences (intronic sequences minus 100 nt from both the donor and acceptor flanks), and intergenic regions (genomic regions in between annotated genes excluding the 500 nt upstream of start codons). The sequences were separated according to isochore type (Schmidt and Frishman 2008). (*C*–*D*) Barplots showing the average relative conservation scores C_R (see Materials and Methods) for the U- and G-rich pentamers from table 1. Scores were calculated for the first 7–50 nt (proximal) and 51–100 nt (distal) of donor intronic flanks within each type of isochore.

and consistent with our hypothesis those that are proximal to 5′ splice sites are more highly conserved than those that are more distal (fig. 9*D*). These observations are consistent with the hypothesis that some of the G-rich pentamers that were gained as a result of isochorization have evolved to become functional *cis* elements and are experiencing purifying selection.

Given the data presented we propose the following model for the gain or loss of SREs during the evolution of mammals. The ancestral forms of most mammalian introns probably resembled those in nonmammalian vertebrates; that is, they typically had U-rich SREs and generally lacked G-rich SREs. Due to processes that are still unclear, some chromosomal regions began to experience high levels of A/T to G/C substitutions. Introns that were located in these regions experienced these same directional pressures, which, by chance, lead to the creation of G-rich sequences capable of binding HNRNPF/H proteins. Because, as we have shown, HNRNPF/H can, at least in some cases, functionally replace splicing factors such as TIA1 that bind U-rich sequences, the gain of a new HNRNPF/H site could make the ancestral U-rich SRE redundant. This would relax purifying selection on the original U-rich SRE, and, because these regions experienced a general conversion of A/T to G/C, U-rich SREs would be more likely to suffer deleterious mutations that would ultimately lead to their decay and loss. This model is similar to the "binding site turnover" model that has been proposed to explain gain or loss of transcription factor binding sites (Zheng et al. 2011).

Discussion

The divergence of mammals appears to have been accompanied by divergence in the architecture of thousands of introns. Interestingly, this conversion appears to have been driven by the same mutational pressures that ultimately lead to the formation of mammalian genomic isochores. Although most isochore-related mutations are probably neutral, isochore formation contributed to a significant proportion of mammalian specific mutations in coding sequences (Galtier et al. 2009; Capra and Pollard 2011). In addition, this process appears to have driven the loss of U-rich intronic SREs coupled with the gain of new G-rich SREs. An important component of this conversion is that U-rich and G-rich SREs are functionally interchangeable, and our in vivo splicing data (fig. 8) support this possibility. If these two classes of SREs are generally interchangeable, then their functional redundancy could have played an important role in buffering the otherwise deleterious effects that would have been associated with the large-scale A/T to G/C mutations that accompanied isochore formation. Although we focused on TIA1/TIAL1 and HNRNPF/H, other splicing factors bind similarly U-rich and G-rich SREs. For instance HNRNPC and ELAV proteins bind U-rich

sequences, and HNRNPA proteins bind G-rich sequences. The degree to which these and other splicing factors can functionally replace each other remains to be further explored.

Our in vivo splicing results (fig. 8) demonstrate that TIA1 and HNRNPF can equally support splicing of the same intron. It is important to acknowledge, however, that our assay uses a single intron reporter, whereas the majority of mammalian transcripts have multiple introns, and according to the exon-definition model (Robberson et al. 1990; Hertel 2008), splicing of multiexon transcripts involves interactions from signals in both the upstream and downstream intronic flanks. It is possible that gain or loss of SREs in this context is more complicated than simply gain or loss of signals in the donor flank. This possibility is supported by previous observations. For instance, exons with donor-side G-rich motifs resembling HNRNPF/H-binding sites tend to have additional HNRNPF/H or HNRNPA1/A2 sites in upstream acceptor flanks (Ke and Chasin 2010). In addition, it has been demonstrated that splicing of exons is enhanced when HNRNPF/H sites are present in both the acceptor and donor flanks surrounding the exon (Martinez-Contreras et al. 2006; Ke and Chasin 2010). Such observations imply that interactions across exons are important for HNRNPF/H function. Therefore, the successful gain of donor-side HNRNPF/H SREs might require the gain of additional sites in the upstream acceptor flanks. Because isochores are large and encompass entire genes, upstream regions would also experience similar levels of A/T to G/C mutations, which would increase the chance formation of G/C-rich motifs similar to those observed by Ke *et al*. (Ke and Chasin 2010). These functionally important pairings would only strengthen the role of newly acquired HNRNPF/H sites and could accelerate the loss of ancestral U-rich SREs.

There is clearly a strong correlation between the gain of HNRNPF/H sites in mammalian introns and the formation of G/C-rich isochores (fig. 5*A*). What is less clear is why this phenomenon was so prevalent in mammalian introns but not in other vertebrates with genomic isochores. Regional variations in G/C-content are a common feature of all vertebrate genomes. However, the degree of variation and the proportions of the genome that are contained in the most A/T- and G/C-rich regions differ widely among vertebrates (Costantini et al. 2009). Of other vertebrates only birds have isochore profiles similar to mammals (Costantini et al. 2007, 2009). Currently, we can only speculate about why we do not observe similar levels of gain of G-rich ISREs in bird introns. But, it is interesting to note that in mammals the HNRNPF/H family comprises five paralogous proteins, whereas in birds there are only three (Barbosa-Morais et al. 2006). It is possible that one or more of these proteins are expressed at higher levels in mammals or that one or more of the mammalian-specific paralogs acquired specialized functions.

Another interesting question is what roles did selective processes contribute to the gain or loss of intronic SREs? Because introns that gained G-rich SREs are generally located in G/C-rich isochores, and the formation of isochores is the result of mutational processes operating at the genomic level (i.e., is unrelated to splicing), the gain of G/C-rich SREs appears to have less to do with gene function or splicing, and instead appears to have been dictated primarily by genomic location. This implies that, in many cases, gain of a G-rich SRE and loss of a U-rich SRE is effectively neutral. This interpretation is supported by the observed variation in SREs even between mammals (for instance, see figs. 5*D* and 7). Nevertheless, in certain cases the gain of a G-rich SRE may have provided a fitness advantage. In support of this is the observation that genes containing exons that require HNRNPH for normal splicing are enriched for certain gene-ontology terms (Xiao et al. 2009). And, genes that have cassette exons that are flanked by HNRNPH-binding sites are enriched for genes with tissue-specific expression patterns (Xiao et al. 2009). Thus, isochore-related gain or loss of SREs could have played an important role in setting up species-specific differences in splicing.

Supplementary Material

Supplementary figures S1–S4 and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

Thanks to Jamie Purcell, Stacey Wagner, Pascale Voelker, and Dave Anderson for helpful comments with the manuscript. Thanks to Dan Graham for helpful suggestions regarding database design and computer programming. This work was supported by grants from the Stiftung für AIDS-Forschung, Düsseldorf (H.S.), the Jürgen-Manchot-Stiftung (H.S.), NSF grant 0616264-MCB, and NIH grant AR053903.

Literature Cited

Akindahunsi AA, Bandiera A, Manzini G. 2005. Vertebrate 2xRBD hnRNP proteins: a comparative analysis of genome, mRNA and protein sequences. Comput Biol Chem. 29:13–23.

Arndt PF, Hwa T, Petrov DA. 2005. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. J Mol Evol. 60:748–763.

Aznarez I, et al. 2008. A systematic analysis of intronic sequences downstream of 5′ splice sites reveals a widespread role for U-rich motifs and TIA1/TIAL1 proteins in alternative splicing regulation. Genome Res. 18:1247–1258.

Barbosa-Morais NL, Carmo-Fonseca M, Aparício S. 2006. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. Genome Res. 16:66–77.

Barreau C, Paillard L, Méreau A, Osborne HB. 2006. Mammalian CELF/Bruno-like RNA-binding proteins: molecular characteristics and biological functions. Biochimie. 88:515–525.

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241:3–17.

Burd CG, 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.

Capra JA, Pollard KS. 2011. Substitution patterns are GC-biased in divergent sequences across the metazoans. Genome Biol Evol. 3:516–527.
Caputi M, Freund M, Kammler S, Asang C, Schaal H. 2004. A bidirectional SF2/ASF- and SRp40-dependent splicing enhancer regulates human immunodeficiency virus type 1 rev, env, vpu, and nef gene expression. J Virol. 78:6517–6526.
Cogan JD, et al. 1997. A novel mechanism of aberrant pre-mRNA splicing in humans. Hum Mol Genet. 6:909–912.
Costantini M, Bernardi G. 2008a. Correlations between coding and contiguous non-coding sequences in isochore families from vertebrate genomes. Gene 410:241–248.
Costantini M, Bernardi G. 2008b. Replication timing, chromosomal bands, and isochores. Proc Natl Acad Sci USA. 105:3433–3437.
Costantini M, Cammarano R, Bernardi G. 2009. The evolution of isochore patterns in vertebrate genomes. BMC Genomics 10:146.
Costantini M, Clay O, Auletta F, Bernardi G. 2006. An isochore map of human chromosomes. Genome Res. 16:536–541.
Costantini M, Di Filippo M, Auletta F, Bernardi G. 2007. Isochore pattern and gene distribution in the chicken genome. Gene 400: 9–15.
de Hoon MJL, Imoto S, Nolan J, Miyano S. 2004. Open source clustering software. Bioinformatics 20:1453–1454.
Del Gatto-Konczak F, et al. 2000. The RNA-binding protein TIA-1 is a novel mammalian splicing regulator acting through intron sequences adjacent to a 5′ splice site. Mol Cell Biol. 20:6287–6299.
Dember LM, Kim ND, Liu KQ, Anderson P. 1996. Individual RNA recognition motifs of TIA-1 and TIAR have different RNA binding specificities. J Biol Chem. 271:2783–2788.
Dominguez C, Fisette J-F, Chabot B, Allain FH-T. 2010. Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. Nat Struct Mol Biol. 17:853–861.
Doniger SW, Fay JC. 2007. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol. 3:e99.
Dreszer TR, Wall GD, Haussler D, Pollard KS. 2007. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. Genome Res. 17:1420–1430.
Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 10:285–311.
Ehrlich M, Wang RY. 1981. 5-Methylcytosine in eukaryotic DNA. Science 212:1350–1357.
Fujita PA, et al. 2011. The UCSC Genome Browser database: update 2011. Nucleic Acids Res. 39:D876–D882.
Fullerton SM, Bernardo Carvalho A, Clark AG. 2001. Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol. 18:1139–1142.
Gal-Mark N, Schwartz S, Ram O, Eyras E, Ast G. 2009. The pivotal roles of TIA proteins in 5′ splice-site selection of alu exons and across evolution. PLoS Genet. 5:e1000717.
Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet. 25:1–5.
Görlach M, Burd CG, Dreyfuss G. 1994. The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins. J Biol Chem. 269:23074–23078.
Han SP, Tang YH, Smith R. 2010. Functional diversity of the hnRNPs: past, present and perspectives. Biochem J. 430:379–392.
He Y, Smith R. 2009. Nuclear functions of heterogeneous nuclear ribonucleoproteins A/B. Cell Mol Life Sci. 66:1239–1256.
Hertel KJ. 2008. Combinatorial control of exon recognition. J Biol Chem. 283:1211–1215.
Hui J, et al. 2005. Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J. 24: 1988–1998.
Kammler S, et al. 2001. The sequence complementarity between HIV-1 5′ splice site SD4 and U1 snRNA determines the steady-state level of an unstable env pre-mRNA. RNA 7:421–434.
Ke S, Chasin LA. 2010. Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. Genome Biol. 11:R84.
Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.
König J, et al. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol. 17:909–915.
Levine A, Durbin R. 2001. A computational scan for U12-dependent introns in the human genome sequence. Nucleic Acids Res. 29: 4006–4013.
Lew JM, et al. 2004. CDKN1C mutation in Wiedemann-Beckwith syndrome patients reduces RNA splicing efficiency and identifies a splicing enhancer. Am J Med Genet A. 127A:268–276.
Llorian M, et al. 2010. Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. Nat Struct Mol Biol. 17:1114–1123.
Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. Biochem J. 417:15–27.
Martinez-Contreras R, et al. 2006. Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. PLoS Biol. 4:e21.
Martinez-Contreras R, et al. 2007. hnRNP proteins and splicing control. Adv Exp Med Biol. 623:123–147.
Masuda A, et al. 2008. hnRNP H enhances skipping of a nonfunctional exon P3A in CHRNA1 and a mutation disrupting its binding causes congenital myasthenic syndrome. Hum Mol Genet. 17: 4022–4035.
Matlin AJ, Clark F, Smith CWJ. 2005. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 6:386–398.
McCullough AJ, Berget SM. 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.
Mueller WF, Hertel KJ. 2011. The role of SR and SR-related proteins in pre-mRNA splicing. In: Z Lorkovi, editor. RNA binding proteins. New York: Landes Bioscience and Springer Science, p. 1–21.
Murphy WJ, et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351.
Murray JI, et al. 2008. Identification of motifs that function in the splicing of non-canonical introns. Genome Biol. 9:R97.
Nussinov R. 1988. Conserved quartets near 5′ intron junctions in primate nuclear pre-mRNA. J Theor Biol. 133:73–84.
Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20:110–121.
Pozzoli U, Sironi M. 2005. Silencers regulate both constitutive and alternative splicing events in mammals. Cell Mol Life Sci. 62: 1579–1604.
Reid DC, et al. 2009. Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence. RNA 15:2385–2397.
Robberson BL, Cote GJ, Berget SM. 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol. 10: 84–94.
Romiguier J, Ranwez V, Douzery EJP, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. 20:1001–1009.
Schaub MC, Lopez SR, Caputi M. 2007. Members of the heterogeneous nuclear ribonucleoprotein H family activate splicing of an HIV-1 splicing substrate by promoting formation of ATP-dependent spliceosomal complexes. J Biol Chem. 282:13617–13626.
Schmidt T, Frishman D. 2008. Assignment of isochores for all completely sequenced vertebrate genomes using a consensus. Genome Biol. 9: R104.

Singh KK, et al. 2010. Human SAP18 mediates assembly of a splicing regulatory multiprotein complex via its ubiquitin-like fold. RNA 16: 2442–2454.
Singh R, Valcárcel J, Green MR. 1995. Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science 268:1173–1176.
Sirand-Pugnet P, Durosay P, Brody E, Marie J. 1995. An intronic (A/U)GGG repeat enhances the splicing of an alternative intron of the chicken beta-tropomyosin pre-mRNA. Nucleic Acids Res. 23:3501–3507.
Van Laer L, et al. 1998. Nonsyndromic hearing impairment is associated with a mutation in DFNA5. Nat Genet. 20:194–197.
Venables JP, et al. 2008. Multiple and specific mRNA processing targets for the major human hnRNP proteins. Mol Cell Biol. 28:6033–6043.
Wahl MC, Will CL, Lührmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. Cell 136:701–718.
Wang H, Molfenter J, Zhu H, Lou H. 2010. Promotion of exon 6 inclusion in HuD pre-mRNA by Hu protein family members. Nucleic Acids Res. 38:3760–3770.
Wang Z, et al. 2010. iCLIP predicts the dual splicing effects of TIA-RNA interactions. PLoS Biol. 8:e1000530.
Wang Z, Burge CB. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA 14:802–813.
Watanabe Y, Abe T, Ikemura T, Maekawa M. 2009. Relationships between replication timing and GC content of cancer-related genes on human chromosomes 11q and 21q. Gene 433:26–31.
Xiao X, et al. 2009. Splice site strength–dependent activity and genetic buffering by poly-G runs. Nat Struct Mol Biol. 16:1094–1100.
Yeap BB, et al. 2002. Novel binding of HuR and poly(C)-binding protein to a conserved UC-rich motif within the 3′-untranslated region of the androgen receptor messenger RNA. J Biol Chem. 277:27183–27192.
Yeo G, Hoon S, Venkatesh B, Burge CB. 2004. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc Natl Acad Sci USA. 101:15700–15705.
Zhang XH-F, Chasin LA. 2004. Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev. 18:1241–1250.
Zhang XH-F, Heller KA, Hefter I, Leslie CS, Chasin LA. 2003. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. Genome Res. 13: 2637–2650.
Zheng W, Gianoulis TA, Karczewski KJ, Zhao H, Snyder M. 2011. Regulatory variation within and between species. Annu Rev Genomics Hum Genet. 12:327–346.

Associate editor: Bill Martin