Cell Reports
Article

Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition

Maayan Amit,[1,4] Maya Donyo,[1,4] Dror Hollander,[1,4] Amir Goren,[1,4] Eddo Kim,[1] Sahar Gelfman,[1] Galit Lev-Maor,[1] David Burstein,[2] Schraga Schwartz,[3] Benny Postolsky,[1] Tal Pupko,[2] and Gil Ast[1,*]
[1]Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel-Aviv University, Ramat Aviv 69978, Israel
[2]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel
[3]Broad Institute, 7 Cambridge Center, Cambridge MA 02142, USA
[4]These authors contributed equally to this work
*Correspondence: gilast@post.tau.ac.il
DOI 10.1016/j.celrep.2012.03.013

SUMMARY

During evolution segments of homeothermic genomes underwent a GC content increase. Our analyses reveal that two exon-intron architectures have evolved from an ancestral state of low GC content exons flanked by short introns with a lower GC content. One group underwent a GC content elevation that abolished the differential exon-intron GC content, with introns remaining short. The other group retained the overall low GC content as well as the differential exon-intron GC content, and is associated with longer introns. We show that differential exon-intron GC content regulates exon inclusion level in this group, in which disease-associated mutations often lead to exon skipping. This group's exons also display higher nucleosome occupancy compared to flanking introns and exons of the other group, thus "marking" them for spliceosomal recognition. Collectively, our results reveal that differential exon-intron GC content is a previously unidentified determinant of exon selection and argue that the two GC content architectures reflect the two mechanisms by which splicing signals are recognized: exon definition and intron definition.

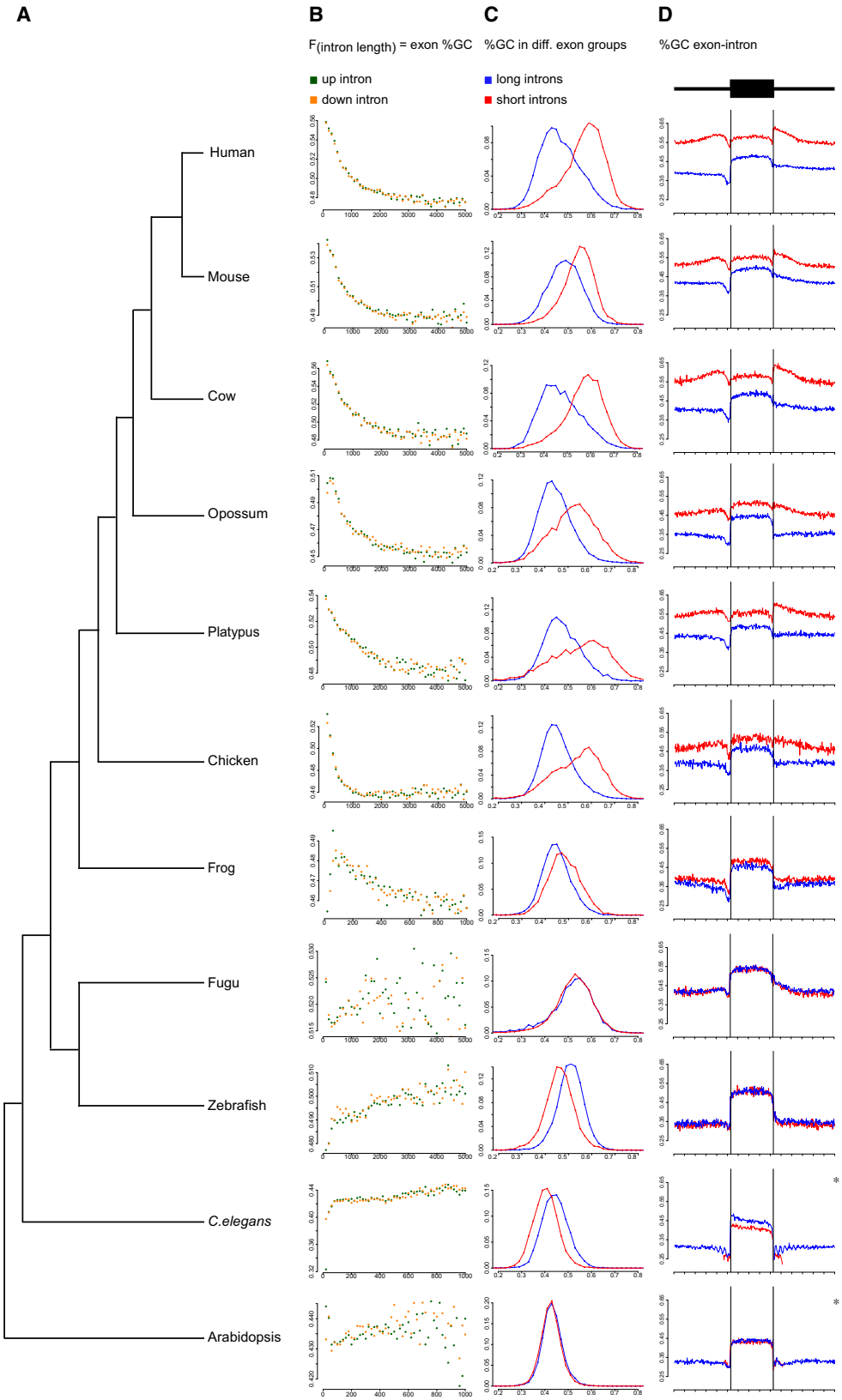
INTRODUCTION

The human genome consists of regions that differ in their GC content. Regions greater than 3 kb in length with a high degree of GC content uniformity are termed isochores. In general, genomes of higher eukaryotes contain regions of high GC content, absent from genomes of lower organisms (Bernardi, 1993). During the evolution of homeotherms, the gene-rich and moderately GC-rich isochores of the poikilothermic ancestors underwent a major GC increase (Bernardi et al., 1985). This increase was maintained during the evolution of mammalian and avian orders (Bernardi, 2000). The genomic organization of GC-rich regions differs significantly from that of GC-poor regions. For example, it has been found that chromosomal regions of high GC content exhibit higher gene densities (Bernardi, 2000; Lander et al., 2001) and, hence, higher CpG island densities (Cross et al., 2000). In addition, the density of short interspersed repetitive DNA elements is higher in GC-rich genomic regions (Bernardi, 2000). Several studies revealed that gene structure also differs significantly between genes located within regions of high GC content versus genes located within low GC content regions. Human housekeeping genes, for example, contain a relatively high GC content and were found to include short introns (Bernardi, 2001; Eisenberg and Levanon, 2003; Lercher et al., 2003).

Splicing is an ancient molecular mechanism, existing even in unicellular eukaryotes such as yeasts. Splicing regulatory proteins are also highly conserved from human and mouse to insects and plants (Barbosa-Morais et al., 2006). Moreover, several experimental SELEX (Systematic Evolution of Ligands by Exponential Enrichment) (Cartegni et al., 2003) and HITS-CLIP (HIgh-Throughput Sequencing of RNA isolated by Cross-Linking ImmunoPrecipitation) (Licatalosi et al., 2008; Zhang and Darnell, 2011) studies revealed that these splicing regulatory proteins identify specific sequence motifs, which obviously contain specific GC contents. Since density of sequence motifs contributing to splice-site recognition is high throughout the transcriptome, the splicing mechanism must have adapted as genes of higher eukaryotes became more GC rich.

Current knowledge suggests that virtually all human multiple-exon genes are alternatively spliced (Pan et al., 2008; Wang et al., 2008). Understanding why some exons are constitutively recognized by the splicing machinery while others are not could shed light on the basal factors that govern splice-site recognition. Since GC content affects characteristics involved in splice-site recognition, it is plausible that GC content may also directly affect splice-site recognition. Two models of the recognition of the splicing unit were proposed: intron definition and exon definition. In the first, the splicing machinery recognizes the intron as the splicing unit and places the basal splicing machinery across introns. As the splicing machinery is limited in its ability to recognize introns of a certain length, these introns are probably under evolutionary selection to remain short.

Cell PRESS
Open ACCESS



Indeed, in lower eukaryotes, where introns are generally small and can be directly recognized by the splicing machinery, intron definition is probably the dominant mode of splicing. On the other hand, in higher eukaryotes, such as vertebrates, where most introns were lengthened, the splicing machinery had to adapt its mode of recognition to identify the short exons among the long introns—the exon definition model (Ast, 2004; Berget, 1995; Gelfman et al., 2012; Hertel, 2008; Keren et al., 2010; Niu, 2008; Ram and Ast, 2007).

In this study, we examined effects of GC content on splice-site recognition. We found that exons in low GC content regions are flanked by introns with markedly lower GC content, whereas exons located in high GC content regions have roughly the same GC content as do their flanking introns. Intron lengthening during mammalian evolution occurred mainly within lower GC content gene environments and disease-associated mutations leading to exon skipping are predominantly located within these genes. Presumably, this is the result of different ways by which exons and introns are recognized, namely, exon versus intron definition. Exons in low GC content regions are flanked by long introns and are thus arguably spliced via exon definition. We hypothesize that the difference in GC content between the exons and their flanking introns constitutes a "flag" meant to assist the recognition of the short exonic sequence by the splicing machinery in the vast intronic landscape. However, such a flag is unnecessary in the case of exons flanked by short introns, which are presumably spliced via intron definition. We validated our hypothesis both by bioinformatic analyses and by experimental examinations.

RESULTS

GC Content and Intron Length

Genomes of higher eukaryotes contain isochores, namely regions of uniform GC content that differs from adjacent regions. To illustrate how the elevation in GC content has affected mammalian and avian genomes, we studied gene structure in eleven organisms. The genomes we analyzed are of mammals (human, cow, mouse, opossum, and platypus), bird (chicken), other vertebrates (frog, fugu, and zebrafish), invertebrates (*Caenorhabditis elegans*), and plants (*Arabidopsis thaliana*) (see Figure 1A for the evolutionary relationships among the species). First, we analyzed the relationship between GC content and intron length in different species. The results revealed that GC content of mammalian, avian, and frog exons negatively correlated with the length of their flanking introns, whereas in other species the opposite trend was observed (Figure 1B). Moreover, mammalian and avian genomes similarly exhibit a separation to two groups of exons: GC-rich exons flanked by short introns and GC-poor exons flanked by long introns (Figure 1C; red and blue lines, respectively). The genome of the vertebrate zebrafish reveals the opposite trend in this respect, and is much more similar to the trend revealed in the *C. elegans* genome (Figure 1C). These results are consistent with previous studies concerning GC content elevation in mammalian and avian genomes (Bernardi, 2000; Bernardi et al., 1985). We assume that the separation into these two exon groups occurred in concert with the GC content elevation.

GC Content and Gene Structure

Previously, exon-intron architecture has been shown to influence splice-site recognition (Berget, 1995; Fox-Walsh et al., 2005; Gelfman et al., 2012; Schwartz et al., 2008; Sterner et al., 1996; Talerico and Berget, 1994). It has also been suggested that intron length significantly affects the efficiency of the pre-mRNA splicing reaction and splice-site recognition (Hertel, 2008). Our results (Figures 1B and 1C) imply that the GC content elevation was accompanied by a change in gene structure, to which the splicing machinery must have adapted in order to maintain proper splicing. We therefore examined the exon-intron architecture of the human genome with respect of its GC content "structure." Remarkably, the results revealed that introns flanking exons of low GC content are not only long, but also contain markedly lower GC content than do the exons. However, introns in GC-rich regions exhibit similar levels of GC content compared to the exons they flank (see Figure 1D–"Human" panel). This finding was also observed for the orthologous exons and introns of other mammalian and avian genomes including those of the cow, mouse, opossum, platypus, and chicken (Figure 1D). This differentiation was not observed in frog, fugu, zebrafish, *C. elegans*, or *A. thaliana*.

Previous studies showing that introns have undergone lengthening throughout the course of evolution (Collins and Penny, 2005; Deutsch and Long, 1999; Gelfman et al., 2012; Lynch and Conery, 2003; Roy and Gilbert, 2006; Yandell et al., 2006), together with an examination of our results along the evolutionary tree, suggest that the ancestral form most likely contained genes with short introns that had significantly lower GC levels than the exons they flanked. Indeed when using maximum parsimony to reconstruct the ancestral gene structure GC state, this assumption was confirmed (Extended Experimental Procedures, Figure S1 available online). These results suggest that differences in GC content between exons and introns "mark" the exon for the splicing machinery. This might explain how exons flanked by long intronic sequences are recognized by the splicing machinery: since the difference in GC content discriminates intronic sequences from exons, it could "flag" the exons and compensate for longer intronic sequences that

Figure 1. The Relationship between Intron Length and GC Content

(A) Evolutionary tree of eleven organisms.

(B) The GC content of exons (y axis) is plotted against the length of their upstream (green) and downstream (orange) introns.

(C) Density function (y axis) of the number of exons having a specific GC content (x axis) for exons flanked by long (blue) and short (red) introns.

(D) A plot of the average GC content in each position within exons extending 150 nt into their upstream (left) and downstream (right) introns (exons were scaled to fit in a 100 bp window; splice-site signals were omitted; exon/intron boundaries are marked by black vertical lines). The plot depicts the averages for exons flanked by long (blue) and short (red) introns. In cases marked by an asterisk we used data from a whole exome instead of orthologs (due to lack of data). See Figure S1.

would otherwise mask the exons. On the other hand, in high GC content regions such a difference in GC content does not exist, and therefore the introns were presumably under a stronger selective pressure to remain short so that splicing unit recognition is not obstructed. The short introns in high GC content regions are presumably recognized through the intron definition mechanism, and a marking of exons by sequence composition is not required.

Evolutionary Constraints on Splice-Site Signals of Exons Flanked by Long Introns

To further examine the differences between the two exon groups we next examined constraints acting on the splice-site signals of exons flanked by long introns ("low GC" group) compared to exons flanked by short introns ("high GC" group; see Extended Results). Long introns lead to a stronger dependency on core splicing signals such as the 3′ and 5′ splice-site signals and the branch site sequence (Figure S2).

Connection between GC Content and Splicing Unit Recognition

Since GC content correlates with characteristics involved in the recognition of the splicing unit, such as intron length and splice-site signal strength, we suspected that GC content may also directly affect the recognition of the splicing unit. Splicing mutations constitute at least 15% of disease-causing mutations (Blencowe, 2000; Cooper and Mattox, 1997; Cooper et al., 2009; Fairbrother et al., 2004; Hastings and Krainer, 2001; Krawczak et al., 1992). Therefore, understanding the molecular mechanisms that differentiate between intron retention and exon skipping is of great importance for biomedical purposes. Our data support the assumption that the shift in GC content led to the appearance of two types of exon-intron structures and that the splicing machinery handles each group differently. We further hypothesized that splicing-disrupting mutations adjacent to introns that are most probably selected via the intron-definition mechanism might result in a retained intron, whereas similar mutations near exons that are selected via the exon-definition mechanism lead to exon skipping (Ast, 2004; Ram and Ast, 2007). If so, does GC content correlate with the outcome of mutations and can we predict the fashion in which a mutation affects the splicing pattern of an exon by analyzing its structure?

To this end, we compiled a data set of 199 intronic disease-associated mutations that lead to the skipping of the adjacent exon (Experimental Procedures). Skipping of these exons causes a disease, indicating that these exons are of functional importance. Remarkably, we were unable to construct a similar data set of disease-associated mutations that lead to intron retention events. We assume that this is due to the fact that intron retention is the least prevalent type of alternative splicing (Kim et al., 2007; Wang et al., 2008) and that mutations leading to intron retention are likely to cause the mature mRNA to be treated as unspliced mRNA by the exosome and therefore be degraded. Instead, we compiled a data set that includes 584 highly significant tissue-specific intron retention events (Experimental Procedures). Next, we determined the GC contents of the skipped exons and their flanking introns and the GC contents of the retained introns and their corresponding upstream and downstream exons. The overall GC content level of the retained introns and their flanking exons is considerably higher than that of the overall GC content level of the skipped exons and their flanking introns (Figure 2). The GC content of the disease-associated skipped exons was higher than that of their upstream and downstream introns, with exon GC content roughly 7% higher. This difference is similar to the trend observed in exons exhibiting low GC content and flanked by long intronic sequences (Figure 2, compare panels A and C, respectively). On the other hand, GC content was only slightly lower in retained introns relative to their upstream and downstream exons (by 4.2% and 2.1%, respectively). This resembles the phenomenon observed in exons exhibiting high GC content and flanked by short introns (Figure 2, compare B and D, respectively). This analysis demonstrates that the GC content profiles of exon skipping events in diseases and intron retention events highly resemble those of the two distinctive GC content groups we identified. It therefore indicates toward the existence of a relation between GC content architecture and the splicing product. Moreover, our results imply that some of the differences in gene structure are related to different ways by which the splicing machinery identifies introns and exons, namely exon definition versus intron definition.

Exon-Intron Differential GC Content Specifies Splicing Pattern Despite Intron Length Changes

To further test the validity of our hypothesis that GC content contributes to recognition by the splicing machinery, we cloned seven representative human exons along with their flanking introns and exons (see Experimental Procedures and Extended Experimental Procedures) into a minigene reporter system. The minigenes were composed of three exons separated by two introns, and were cloned into a pEGFP-C3 vector to test their splicing pattern in vivo. The minigenes evaluated represent the two distinct GC content groups as well as the ancestral form. The first group is characterized by (1) long introns, (2) overall low GC content, and (3) differential GC content between exons and introns (*CA3*, *THSD7B, MDH1*, and *DDX60*). The second group represents the ancestral gene structure and is characterized by (1) short introns, (2) overall low GC content, and (3) differential GC content between the middle exon and its flanking introns (*MYH1* and *TTN*). The third group is characterized by (1) short introns and (2) an overall high level of GC content with (3) no differential GC content between exons and introns (*PLXNB1*). (See additional information in Table S1.)

Our findings suggest that exon-intron differential GC content contributes to exon selection. In order to test this hypothesis, we first set out to examine the effect of intron length on the splicing pattern of exons flanked by introns of substantially lower GC content. We were able to do that by altering intron length in four minigenes: *MYH1* and *TTN* with short flanking introns and *CA3* and *THSD7B* with long flanking introns. It is important to note that all intron modifications were performed without altering the original splice sites. We lengthened the introns of the *MYH1* and *TTN* minigenes by an insertion of either 500 nucleotides (nt) or 2,000 nt long segments. Original intron lengths were ~100 nt. The GC content level of the inserted segments

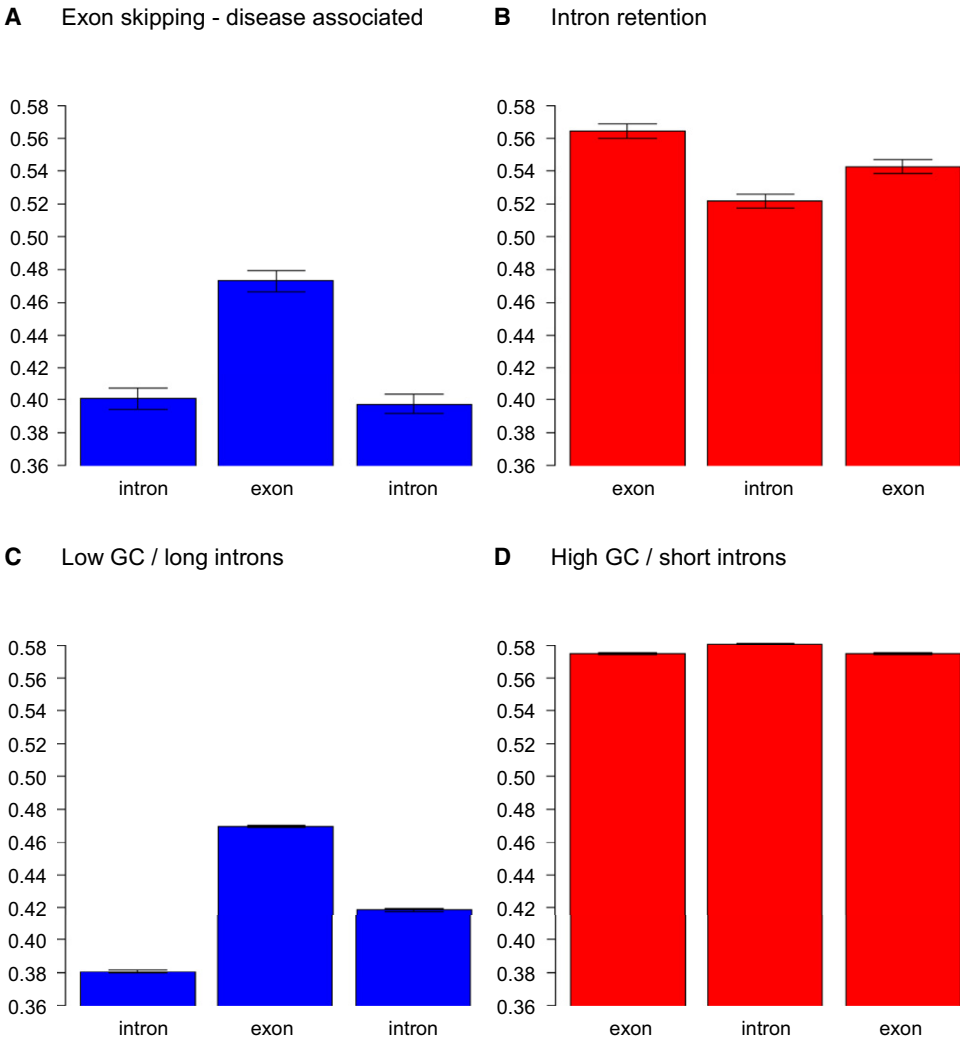


Figure 2. Connection between GC Content and Splicing Unit Recognition

The GC content (y axis) of (A) the surrounding genomic area of exons that are skipped due to disease associated mutations and (B) the surrounding genomic area of alternatively retained introns. For comparison we performed a similar analysis for (C) exons flanked by introns that were defined as part of the "low GC" content group and (D) introns and flanking exons that were part of the "high GC" content group. See Experimental Procedures for a description of sequence selection. See Figure S2.

was similar to that of the original introns. We shortened each intron of the *CA3* and *THSD7B* minigenes from approximately 2 kb to 150–450 nt by deleting central intronic segments in order to avoid impairing the effects of regulatory sequences as their frequency is higher near the splice sites (Fairbrother et al., 2004; Majewski and Ott, 2002; Sorek and Ast, 2003; Sugnet et al., 2006; Yeo et al., 2007). Nevertheless, this does not exclude the possibility that in some cases we eliminated regulatory sequences.

The wild-type (WT) *MYH1* and *TTN* minigenes exhibited inclusion of the middle exon (Figures 3A and 3B, lane 1, respectively). Lengthening the introns by 500 nt or by 2,000 nt did not alter the splicing pattern (Figures 3A and 3B, lanes 2 and 3). Interestingly, it seems that considerable differential GC content between the introns and the central exon allows lengthening of the introns even 20-fold without impairing recognition and splicing of the exon.

The WT *CA3* and *THSD7B* minigenes exhibit inclusion of the central exon (Figures 3C and 3D, lane 1, respectively). After shortening the downstream intron to 150 nt the *CA3* exon remained included (Figure 3C, lane 2) and the *THSD7B* exon was mostly included as well (Figure 3D, lane 2). Shortening of the *CA3* and *THSD7B* upstream intron to 300 nt and 450 nt, respectively, did not alter the splicing pattern, and the exon remained included (Figures 3C and 3D, lane 3; see also Extended Results and Figure S3). Finally, shortening of both introns did not change the splicing pattern of the *CA3* exon and the *THSD7B* central exon was mostly included (Figures 3C and 3D, lane 4). Our results demonstrate that shortening of long introns of low GC content minigenes without altering their GC content retains

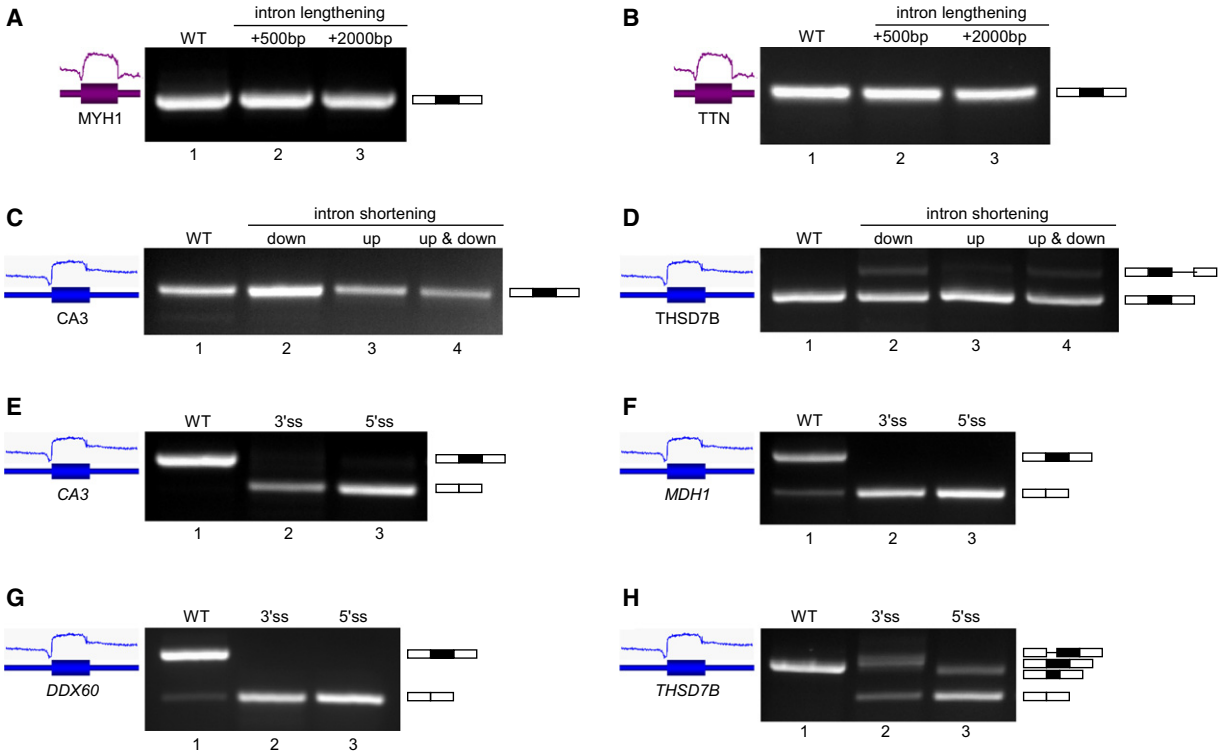


Figure 3. Manipulation of Intron Length and Splice-Site Scores in Minigenes with Exon-Intron Differential GC Content

Minigenes were introduced into 293T cells by transfection. Total RNA was extracted 48 hr after transfection, mRNA was reverse transcribed and splicing products were separated on a 1.5% agarose gel.

(A and B) Effect of lengthening of (A) *MYH1* and (B) *TTN* introns. Lane 1, WT minigenes; lane 2, both introns extended with segments of 500 nt; lane 3, both introns extended with segments of 2,000 nt.

(C and D) Effect of shortening of (C) *THSD7B* and (D) *CA3* introns. Lane 1, WT minigene; lane 2, downstream intron shortened to 150 nt; lane 3, upstream intron shortened to 300 nt in C and 450 nt in D; lane 4, shortening of both the upstream and the downstream introns to the same lengths examined in lanes 2 and 3.

(E–H) Effect of abolishing the splice-site signals of the (E) *CA3*, (F) *MDH1*, (G) *DDX60*, and (H) *THSD7B* minigenes. Lane 1, WT minigenes; lane 2, abolishing the 3′ splice site; lane 3 abolishing the 5′ splice site. The mRNA products are shown on the right of each panel; boxes define exons, lines represent retained introns. The group each minigene belongs to is shown on the left.

See Figure S3.

inclusion of the central exon. In general, these results demonstrate that exon-intron differential GC content compensates for intron length manipulations and contributes to exon selection.

Splice-Site Mutations Adjacent to the Exon-Intron GC Content Differential Lead to Exon Skipping

Our results imply that it is possible to provide a priori predictions for the effect mutations would have on the splicing activity of exons flanked by long introns of lower GC content (Figures 2A and 2C). To evaluate this we abolished the 3′ or the 5′ splice-site signals of the *CA3*, *MDH1*, *DDX60*, and *THSD7B* minigenes that exhibit exon-intron differential GC content (Figures 3E–3H). The WT exons were all recognized and constitutively or alternatively spliced (Figures 3E–3H, lane 1). In accordance with our bioinformatic findings, impairing the 3′ splice site or 5′ splice site led to full skipping of the CA3, MDH1, and DDX60 exons (Figures 3E–3G, lanes 2 and 3). Abolishing the splice-site signals of the THSD7B minigene caused exon skipping as well the selection of cryptic splice sites (Figure 3H, lanes 2 and 3). It is of interest to note that the 23 nt segment that was included due to the cryptic 3′ splice-site selection (Figure 3H, lane 2) has high GC content (52.2%) compared to the GC content of the intron (37%). Hence, the high GC content platform within a low GC environment remains. Our results demonstrate that just as mutations adjacent to an exon-intron GC content differential lead to disease-associated exon skipping events, so do mutations of the same sort in the splice sites of exons flanked by introns of lower GC content. This constitutes an additional indication that the exon-intron GC content differential influences the way by which the splicing machinery identifies exons.

The Relationship between the Exon-Intron Differential GC Content and the Inclusion Level of Alternative Exons

Since we have shown that the exon-intron GC content differential is associated with exon selection, we were also interested in examining the relationship between exon inclusion levels and the size of the exon-intron GC content differential. In this analysis, we used the abovementioned intron retention data set as well as an exon skipping data set that include tissue-specific readings across 48 human tissues and cell lines (Experimental

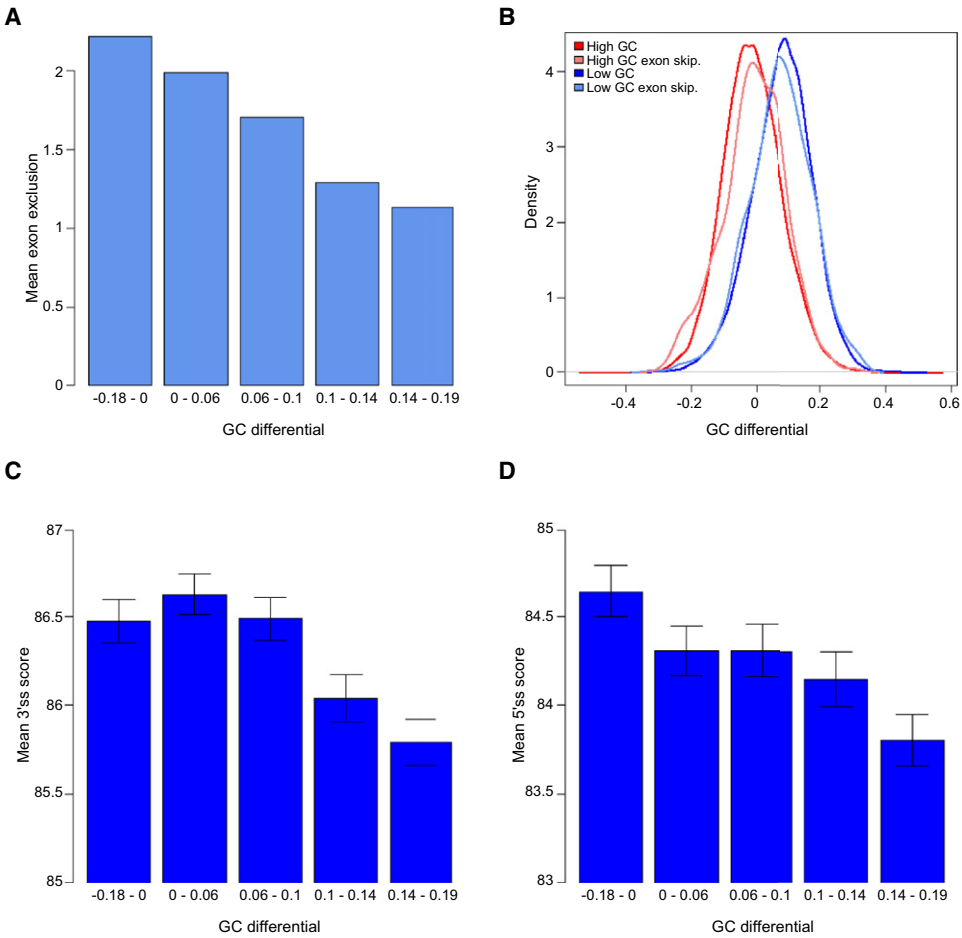


Figure 4. The Relationship between the Exon-Intron Differential GC Content and the Inclusion Level of Alternative Exons

(A) The mean exclusion level (y axis; represented by the mean change in percent of splice form composition across the tissues and cell lines) of low GC skipped exons was plotted against the exon-intron GC differential (x axis).

(B) Exon-intron GC differential distribution of four groups: "low GC," "high GC," low GC skipped exons, and high GC skipped exons. The x axis represents the GC differential and the y axis represents the density.

(C and D) The mean splice-site scores (y axis) of the "low GC" group were plotted against the exon-intron GC differential (x axis) for both the (C) 3′ splice site and the (D) 5′ splice site. The confidence intervals represent the standard error of the mean of each bin.

See Figure S4.

Procedures). In agreement with our previous findings, more alternative splicing events of the exon skipping type were found within the group exhibiting exon-intron differential GC content rather than the group with uniformly high GC content across exons and their flanking short introns (599 versus 360 events, respectively). On the other hand, intron retention events were almost exclusively found within the group with similar GC content in both introns and exons (292 events) rather than the group with differential GC content (only one event). These results suggest a possible connection to the manner in which the spliceosome recognizes the primary splicing unit (e.g., either the exon or the intron) in different genomic architectures. Moreover, examination of the 599 exon skipping events found in the group exhibiting exon-intron differential GC content revealed that as the GC differential increased so did the percentage of transcripts that included the alternatively spliced exon. In essence, a positive relationship between the size of the GC differential and exon inclusion level was identified (Figure 4A). Next, we plotted the GC content distributions of both low GC (599 cases) and high GC (360 cases) alternatively skipped exons against their corresponding GC content groups. This was performed in order to ensure that the GC content distributions of the low and high GC skipped exons resemble that of their corresponding groups, so we could treat them as their valid representatives. Indeed, our results show that the low GC skipped exons resemble the low GC content group and that the high GC skipped exons are similar to the high GC content group (see Figure 4B).

These results suggest that the GC content difference between the exon and its surrounding introns may affect the measure in which an exon is included in the mature RNA transcript. We examined the relationship between the splice-site signal strength (calculated using the Shapiro and Senapathy algorithm;

Shapiro and Senapathy, 1987) and the GC difference between the exon and its surrounding introns. Strengths at both the 3′ and 5′ splice sites weakened as the exon-intron GC content differential became higher (Figures 4C and 4D). This suggests that the elevation of the exon-intron GC content differential may compensate for the weaker splice-site signals.

Furthermore, GC-rich splicing enhancers were previously shown to be distributed differently in human introns according to intron length (Yeo et al., 2004). We examined the prevalence of these sequences in both exon-intron GC architectures while controlling for GC content (Extended Experimental Procedures). Our results revealed them to be similarly overabundant in both GC groups (Figure S4), suggesting they do not contribute to the differences observed between the group displaying exon-intron differential GC content and the group without it. Finally, all of the aforementioned results do not only indicate that exon-intron differential GC content impacts exon selection, but rather that the size of the GC content differential has an effect on exon inclusion levels.

Generating Exon-Intron GC Content Differential Rescues Splicing

The abovementioned results indicate that differential GC content impacts exon selection. To further examine this we decided to use a minigene, in which the central exon and its flanking introns do not exhibit differential GC content (from the high GC content group), and to facilitate such a differential in order to test its effect on splicing. As a control, we took low GC content minigenes with exon-intron differential GC content (from the ancestral group) and a high splicing efficacy, and abolished the GC differential. We chose one minigene, *PLXNB1*, from the high GC content group and two minigenes, *MYH1* and *TTN*, from the ancestral group for this evaluation (all three minigenes include short introns; the scores of the gene splice sites are indicated in Table S2). Strikingly, when we generated a GC content differential by replacing the introns of *PLXNB1* with those of *MYH1* or *TTN*, a significant change in splicing pattern was observed. The central exon in WT *PLXNB1* is fully skipped (Figures 5A and 5B, lane 1), and the intron replacement has led in both cases to exon inclusion as well as to the retention of the upstream intron (Figures 5A and 5B, lane 2). When we replaced the flanking introns of *MYH1* and *TTN* with those of *PLXNB1* to eliminate the exon-intron differential GC content, we found that the splicing pattern changed from a fully included central exon in the WT minigenes to a fully skipped central exon in the chimeric minigenes (Figures 5C and 5D). In conclusion, we have shown that transforming a minigene without an exon-intron GC content differential into one that has a GC differential rescues splicing of an otherwise skipped exon; in contrast, loss of exon-intron GC content differential inhibited exon recognition. These results indicate that the GC content differential between exons and their flanking introns is a major component in the exon selection process.

Increase and Decrease of Exon-Intron Differential GC Content Affect Exon Selection

Our previous finding has demonstrated that generating exon-intron differential GC content leads to exon inclusion. In addition, our bioinformatic analysis (displayed in Figure 4A) indicates that the higher the exon-intron GC differential the more included the exon. This led us to examine the effect distinctive exon-intron differential GC content levels have on exon inclusion. To this end, we systematically modified the GC content of the middle exon of several minigenes and examined the outcome on the splicing activity. We used the *MDH1*, *DDX60*, *CA3*, and *THSD7B* minigenes, which are derived from the group of exons flanked by long introns of substantially lower GC content. We replaced the middle exon of each minigene with one of two exons of the same length while maintaining the original splice sites. One exon was of lower GC content and the other of higher GC content (see Table S3). The WT *MDH1* and *DDX60* minigenes were alternatively spliced (Figures 5E and 5F, lane 1). Replacing the middle exon of each minigene with another of higher GC content resulted in an elevation of the inclusion level to full exon inclusion (Figures 5E and 5F, lane 2), whereas the replacement with a lower GC content exon led to full exon skipping in the *MDH1* minigene (Figure 5E, lane 3) and to a higher level of exon skipping in the *DDX60* minigene (Figure 5F, lane 3). In the same vein, the WT *CA3* exon was fully included (Figure 5G, lane 1), and decreasing the differential GC content led to alternative splicing (Figure 5G, lane 3), whereas its increase kept the exon fully included (Figure 5G, lane 2). Finally, The WT *THSD7B* exon was fully included (Figure 5H, lane 1), and, in accordance with our other results, the lower GC content exon was fully skipped (Figure 5H, lane 3). The higher GC content exon replacement resulted in exon inclusion as well as the selection a cryptic 3′ splice site (Figure 5H, lane 2). Interestingly, the 23 nt segment included due to the cryptic 3′ splice site is the same one of elevated GC content previously described (Figure 3H, lane 2). It maintains the high GC content platform within a low GC environment. These results demonstrate that increasing the differential GC content between the exon and the flanking introns improves exon recognition, and elevates exon inclusion level, whereas decreasing it enhances exon skipping. Our findings indicate to the importance of the exon-intron GC content differential in the correct identification of exons by the splicing machinery.

Nucleosome Occupancy Differs According to Exon-Intron GC Content Architecture

To further investigate how the exon-intron differential GC content may contribute to exon recognition we examined exon-intron GC content architecture from an epigenetic perspective. Previous studies have shown higher level of nucleosome occupancy in exons compared to the flanking intron sequences. This suggests that exon recognition may be facilitated by chromatin organization (Kolasinska-Zwierz et al., 2009; Kornblihtt et al., 2009; Schwartz et al., 2009; Tilgner et al., 2009). Moreover, nucleosomes tend to bind GC-rich, rather than AT-rich sequences (Prendergast and Semple, 2011; Tillo and Hughes, 2009). It therefore appears that the GC topography of introns and exons may drive nucleosomes to bind exons preferentially, ensuring their identification by the splicing machinery (Kornblihtt et al., 2009; Schwartz and Ast, 2010). We set out to examine the nucleosome occupancy in both the high and low GC content groups by analyzing genome-wide

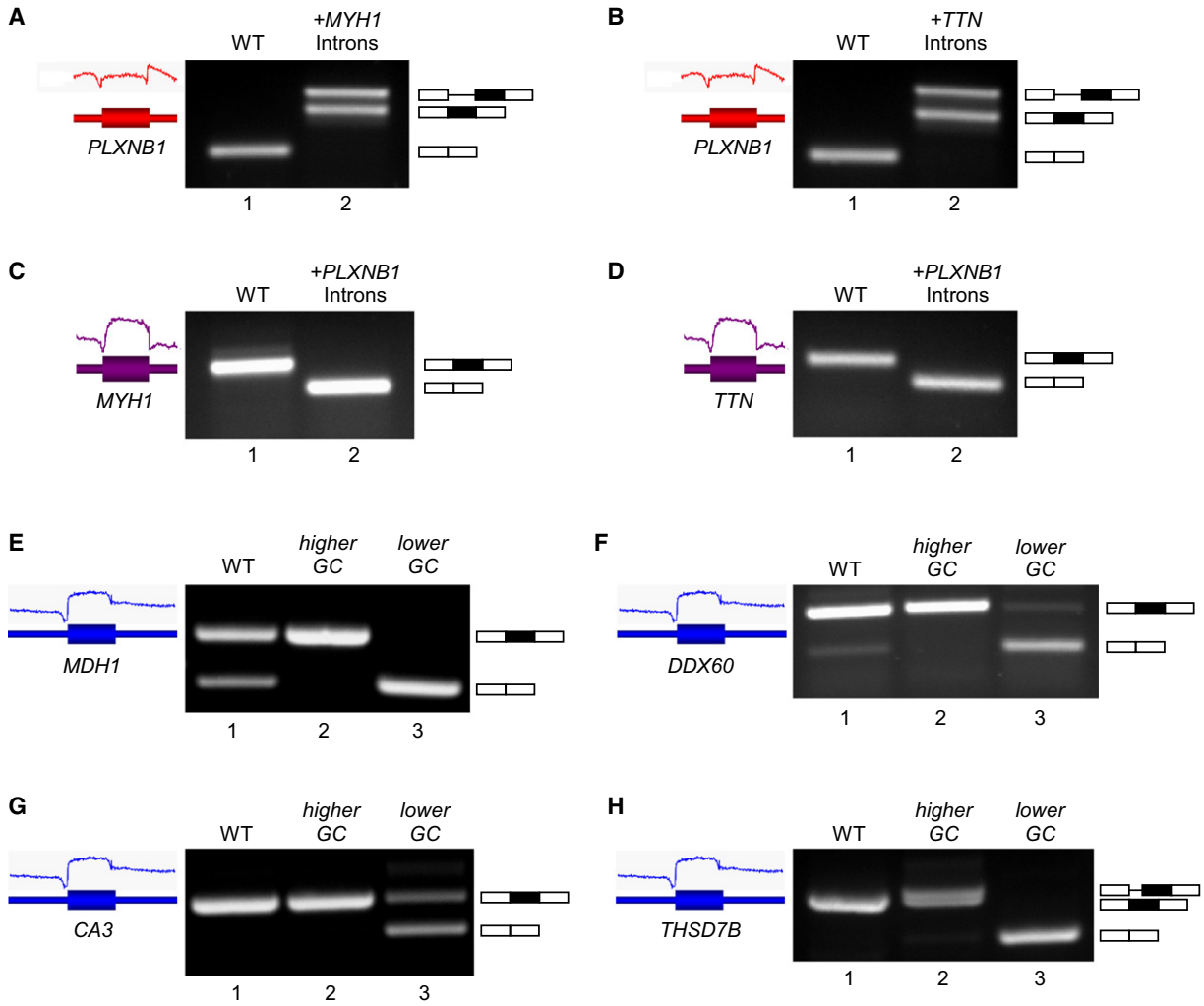


Figure 5. The Effect of Differential GC Content on Exon Inclusion

Plasmids were introduced into 293T cells by transfection. Total RNA was extracted 48 hr after transfection, mRNA was reverse transcribed and splicing products were separated on a 1.5% agarose gel.

(A and B) Replacement of *PLXNB1* introns with those from *MYH1* (A) or *TTN* (B). Lane 1, WT *PLXNB1*; lane 2, PLXNB1 with *MYH1* or *TTN* flanking introns.

(C) Replacement of *MYH1* introns with those of *PLXNB1*. Lane 1, WT *MYH1*; lane 2, *MYH1* with *PLXNB1* flanking introns.

(D) Replacement of *TTN* introns with those of *PLXNB1*. Lane 1, WT *TTN*; lane 2, *TTN* with *PLXNB1* flanking introns.

(E–H) Replacement of WT exons (lane 1) with exons of higher (lane 2) and lower (lane 3) GC content for the following minigenes: (E) *MDH1*, (F) *DDX60*, (G) *CA3*, and (H) *THSD7B*. The mRNA products are shown on the right of each panel; boxes define exons, lines represent retained introns. The group each minigene belongs to is shown on the left.

nucleosome-positioning in human T cells (Experimental Procedures). Our results revealed a different spatial distribution of the nucleosome occupancy along the exon-intron structure of these groups. In high GC content introns, there was higher nucleosome occupancy than in low GC content introns. Interestingly, exons of the high GC content group had only a slight elevation in nucleosome levels compared to their flanking intron, whereas exons in the low GC group had substantially higher nucleosome occupancy than did their flanking introns. Even more surprising is the fact that in spite of their GC content level the low GC exons exhibited higher nucleosome occupancy than the high GC exons group (Figure 6A).

We experimentally validated these results by measuring nucleosome occupancy of representative genes from the high and low GC content groups: *PLXNB1* and *THSD7B*, respectively. For each gene, we tested the endogenous nucleosome occupancy of the exon, which was previously used for our analyses, and its flanking introns (see Extended Experimental Procedures). Even though both internal exons exhibited significantly higher nucleosome occupancy levels than flanking introns (Mann-Whitney test $p < 0.01$), we found that *THSD7B*, the gene with the lower GC content, had exon-intron differential nucleosome occupancy two hundred times higher than that of the *PLXNB1* gene (Figure 6B).

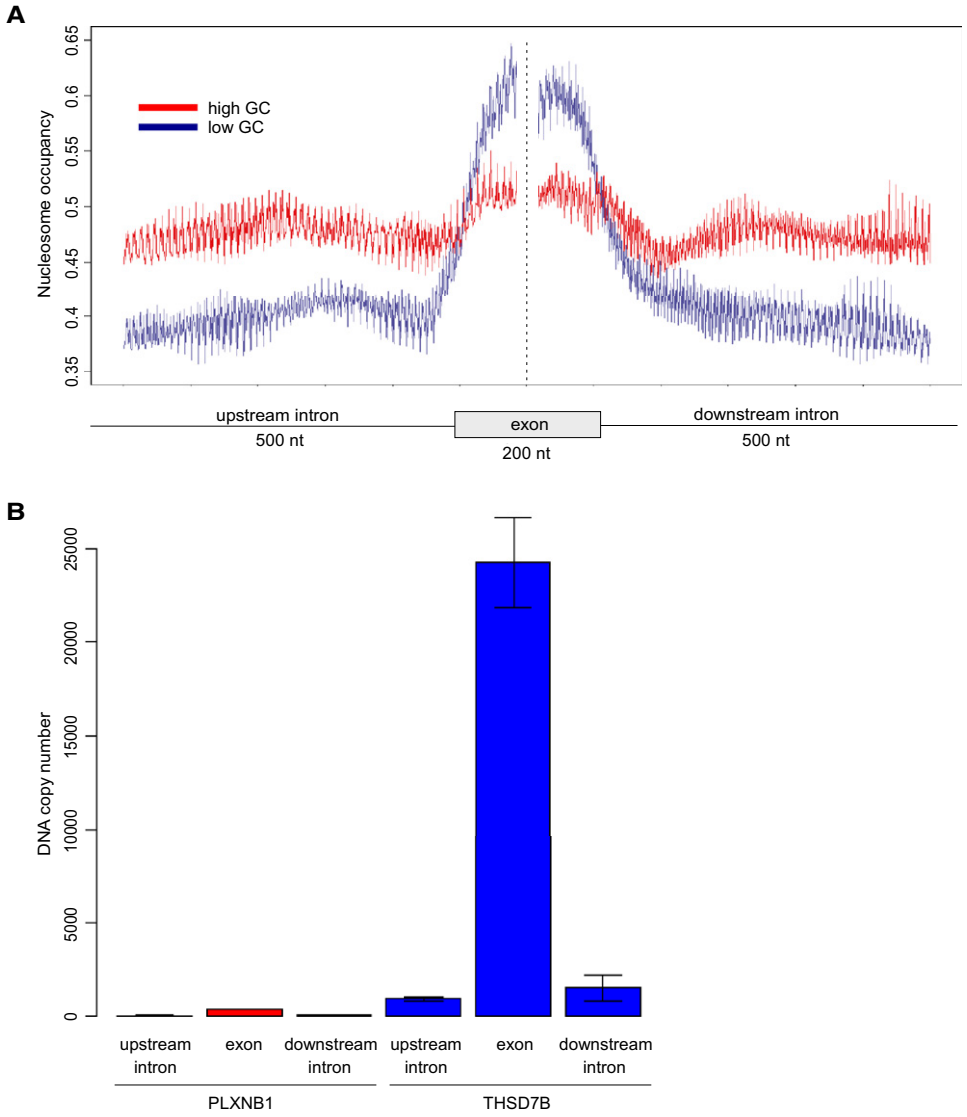


Figure 6. Nucleosome Occupancy Differs According to Exon-Intron GC Content Architecture

(A) The spatial distribution of nucleosome occupancy in human T cells was examined for low GC content genes with a significant exon-intron GC content differential (blue) and the high GC content gene without such a differential (red). The x axis represents the examined structure containing exons flanked by 500 bp from each side. The y axis represents the fraction of nucleosome occupancy in each position.

(B) Nuclei from 293T cells were extracted, following MNase digestion. Mononucleosomal DNA was extracted from an agarose gel and subjected to absolute real-time PCR analysis with primers to the exons and their flanking introns. Data are presented as DNA copy number. Real-time PCR experiments were performed in triplicate; results shown are mean values ±SD.

DISCUSSION

In this study, we examined gene structure in GC-poor versus GC-rich regions. Mammalian and avian genomes, but not invertebrates and zebrafish, exhibit a negative correlation between the GC content of exons and the length of their flanking introns. The exons in low GC content regions are flanked by introns with markedly lower GC content, a phenomenon absent in exons in high GC content regions. Reconstruction of the ancestral state revealed a plausible genomic structure of exons surrounded by short introns that significantly differ in their GC content level. This genomic structure has evolved into two distinct groups of exons: (1) regions displaying elevated GC content containing exons flanked by short introns and (2) low GC content regions containing exons flanked by long introns with a significantly lower GC content than the exons. These two groups were noticeable in our analysis in the chicken genome and present in all of the examined mammals (Figure 1D). Presumably, this is the result of the two different ways by which the splicing machinery recognizes the primary splicing unit. Two models for the mechanism of exon and intron selection were previously suggested: intron and exon definition. In intron definition, the

splicing machinery recognizes an intronic unit and places the basal splicing machinery across introns. As this type of recognition is presumably impaired in long introns, the introns recognized by intron definition are probably under evolutionary selection to remain short. Indeed, in lower eukaryotes, where introns are generally short and can be directly recognized by the splicing machinery, intron definition is probably the dominant mode of splicing. On the other hand, in higher eukaryotes, such as vertebrates, where most introns are long, the splicing machinery had to adapt to identify short exons among the long introns via the exon definition mechanism (Ast, 2004; Berget, 1995; Hertel, 2008; Keren et al., 2010; Niu, 2008; Ram and Ast, 2007).

Our results suggest that GC content differences are associated with two different types of alternative splicing: intron retention and exon skipping. In general, we expect an exon to be skipped where the splicing machinery fails to recognize it, namely, where exon definition is the dominant mechanism. We analyzed introns containing disease-associated mutations that cause an adjacent exon to be skipped. We examined these events since their causing a known disease proves they occur and implies these exons' functional importance. The skipped exons have significantly higher GC content than the flanking introns. This result resembles our finding concerning exons embedded within low GC content regions (Figures 2A and 2C). Our results therefore imply a connection between a significant exon-intron difference in GC content and exon recognition via the exon definition mechanism. On the other hand, examination of intron retention events in numerous organs and cell lines revealed minor difference in GC content between the neighboring exons and the retained intron (Figures 2B and 2D). The fact that we did not find a considerable GC content difference in the intron retention cases along with the fact that intron retention events are associated with short introns (Berget, 1995; Talerico and Berget, 1994) suggest that these events are probably the result of the splicing machinery's difficulty in recognizing these introns via the intron definition mechanism.

Since exons in low GC content regions are flanked by long introns, they are presumably spliced via exon definition. We hypothesized that the difference observed in GC content between the exons and their flanking introns (Figure 1D) helps "mark" these exons for recognition by the splicing machinery. This marking is less important for the recognition of exons flanked by short introns, as they are arguably spliced via intron definition. Our experimental results demonstrate that the lengthening of short introns surrounding exons of substantially higher GC content does not change the splicing pattern of those exons, and they remain included in the mRNA (Figures 3A and 3B). We found these results interesting since previous studies have shown that, in accordance with the intron definition model, expansion of small introns in yeast or *Drosophila* causes loss of splicing, cryptic splicing, or intron retention (Guo et al., 1993; Talerico and Berget, 1994). Since the intron expansion presumably lead to a switch of splice-site recognition from cross-intron interactions to cross-exon interactions (Fox-Walsh et al., 2005), we propose that the exon-intron GC content differential compensates for the intron expansion. Reciprocally, shortening of long introns flanking exons of considerably higher GC content did not inhibit exon selection (Figures 3C and 3D). Of further indication of the role the exon-intron GC content differential has in exon definition is the fact the splice-site mutations adjacent to it led almost exclusively to exon skipping (Figures 3E–3H).

In addition to the general association between GC content architecture and splicing unit recognition, we were also able to identify a positive correlation between the exon-intron GC content difference and the inclusion level of alternatively spliced cassette exons; specifically, the higher the difference the more the exon is included (Figure 4A). We find this phenomenon even more interesting since exon inclusion levels increase despite a decrease in 5′ and 3′ splice-site signal scores (Figures 4C and 4D). This further stresses the importance of the difference in GC content between exons and introns and indicates that the magnitude of that difference may play a role in fine-tuning exon inclusion levels.

To illustrate the importance of the GC topography in exon selection, we replaced the introns of the high GC content group (i.e., short introns with a high GC content level similar to that of the central exon) with introns of lower GC content. In this manner, we created an exon-intron GC content differential that was not present in the WT minigene. This changed the splicing pattern from exon skipping in the WT minigene to products that included the central exon (Figures 5A and 5B). A reciprocal change—abolishing the exon-intron GC content differential by replacing low GC content introns with high GC content introns—led to an opposite result. In this case, we observed a change from exon inclusion in the WT construct to exon skipping (Figures 5C and 5D). Although the effect the exon-intron GC content difference has in a short intronic environment is unclear, the fact that we did not change the splicing mode (according to intron length; Talerico and Berget, 1994), but did abolish the exon-intron GC content difference, may suggest its involvement in the intron definition mechanism as well.

To further illustrate the effect of the exon-intron GC differential, we designed an experiment meant to provide a wider understanding concerning the effect distinct levels of differential GC content have on exon splicing. We replaced exons from the low GC content group (i.e., those with long introns with lower GC content compared to the central exon) with exons with higher or lower GC content. Our findings revealed a positive relationship between the exon-intron GC content differential and the inclusion level of the exon (Figures 5E–5H). Increasing the differential GC content of the exon increased exon inclusion while decreasing it elevated exon skipping levels.

We do not assume that core spliceosomal components are more likely to associate with GC-rich genomic areas. We do show, however, that epigenetic markers, which impact splicing, are affected by GC content. It has been shown that nucleosome occupancy is higher in exons than introns. Thus, exon recognition may be facilitated by chromatin structure (de Almeida et al., 2011; Kim et al., 2011; Kolasinska-Zwierz et al., 2009; Kornblihtt et al., 2009; Schwartz et al., 2009; Tilgner et al., 2009). Moreover, nucleosomes tend to bind GC-rich sequences rather than AT-rich sequences (Prendergast and Semple, 2011; Tillo and Hughes, 2009). It therefore appears that the GC topography of introns and exons may serve as a force that drives nucleosomes to bind exons preferentially, thus ensuring their identification by the splicing machinery (Kornblihtt et al., 2009;

Schwartz and Ast, 2010). Our results add another layer concerning exon "marking" by nucleosome occupancy. We suggest that the variable differential GC content in exons flanked by long introns may be associated with exon recognition via exonic nucleosome enrichment. This hypothesis was supported by our results (see Figure 6) and those of previous studies indicating nucleosome enrichment in exons flanked by long introns compared to exons flanked by short introns (Spies et al., 2009). However, our analysis differs from that of Spies et al. (2009) since intron definition was previously suggested to be limited to intron length of 300 bp (Berget, 1995; Hertel, 2008), and therefore Spies et al.'s categorization of short introns (500–1,000 bp) is irrelevant to this mode of splicing recognition, and cannot be applied to our analysis. Moreover, their categorization of exons flanked by short introns (500–1,000 bp) and long introns (>5 kb) resulted in smaller intron data sets—approximately 35% of human introns—compared to roughly 50% in our analysis. Finally, our study has added an extra layer of information regarding differential GC content correlation with nucleosome occupancy in these specific gene structures.

Furthermore, we postulate that the observed changes in splicing pattern brought by manipulations of GC differential size in a minigene system (Figures 5E–5H) could be mediated by nucleosomal changes. In that respect, it is of interest to note that previous works have already shown a connection between the splicing mechanism and chromatin structure using a minigene system in order to recapitulate endogenous gene behavior (Nogues et al., 2002; Schor et al., 2009; Subtil-Rodríguez and Reyes, 2010).

Overall, the results we present show that during evolution changes in genomic GC content architecture had a fundamental impact on the splicing machinery, arguably resulting in a separation into two modes of splicing unit recognition—intron and exon definition. These two recognition modes could represent two exon groups: one contains considerable differential GC content compared to their flanking introns while the other does not. The exon group exhibiting differential GC content compared to flanking introns, which remained similar to the ancestral state in that respect, is better recognized by the splicing machinery. This presumably allowed the expansion of the flanking introns without losing exon recognition via the exon definition model. This strongly supports the hypothesis that alternative splicing became prevalent in mammalian and mostly human genomes due to intron expansion, as expansion beyond a certain length led to suboptimized recognition of internal exons and therefore to exon skipping, the most abundant form of alternative splicing. Thus, we have identified a determinant for exon recognition, as our results indicate that GC content architecture has an influential effect on the identification of exons embedded between long introns, arguably via exon definition.

EXPERIMENTAL PROCEDURES

Analysis of GC Content in Exons and Their Flanking Introns

For "genomic data set compilation," see Extended Experimental Procedures. We used the annotation provided by the University of California, Santa Cruz (UCSC) tables to identify the flanking introns of each exon. For each exon, we then extracted the adjacent 150 nucleotides (nt) of the upstream and downstream introns. For the upstream intron sequence, we discarded 20 nt from the 3′ end as these are part of the splice-site signal. For the same reason, we also discarded the first 6 nt of the downstream introns and the first 2 and the last 3 nt of the exons. For both the upstream and downstream introns we also removed bases that were part of the preceding or succeeding exon's splice-site signal, if relevant. We then defined exons as flanked by "short" introns if both the upstream and downstream introns were shorter than their respective 25th percentile values. Similarly, we defined exons as flanked by "long" introns if both the upstream and downstream introns were longer than the 75th percentile of both the upstream and the downstream introns. Exons that did not fit in either of these two categories were discarded.

Mutation and Alternative Splicing Data Sets Compilation

Multiple sources were used to generate a comprehensive mutation data set. We downloaded 478 human mutations that were collected and analyzed previously by Krawczak et al. (2007). This data set was comprised of single base-pair substitutions located within the splice-sites of 38 different human genes. Additional mutations were collected from the "Aberrant Splicing Database" which includes an extensive collection of mammalian genetic disease mutations (Nakai and Sakamoto, 1994) and from published literature (Desmet et al., 2009; Roca et al., 2003). Next, we excluded redundant mutations and redundant exons/introns that contained them (for example, introns that contained more than one mutation—leading to identical exon skipping events—were used once). We collected a total of 199 disease-associated mutations that lead to exon skipping. In addition, we constructed two data sets: one of intron retention and one of exon skipping events. These data sets were based on previously published microarray data for 24,426 alternative splicing events across 48 tissues and cell lines (Castle et al., 2008). For each event, these data sets contain 48 values indicative of the change in relative abundance of the splice forms (as the change in percent of splice form composition) across the tissues and cell lines. The intron retention data set includes 584 intron retention events found to be significantly different in at least one tissue or cell line compared to a mixed tissue pool. The exon skipping data set includes significantly different events in at least one tissue or cell line that were intersected with the data sets of exons flanked by long or short introns.

Construction of Nucleosome Occupancy Maps for Exons with Long and Short Flanking Introns

The mapping of nucleosome occupancy was done using the MNase-Seq data of active T cells obtained from Schones et al. (2008). In order to map nucleosome positioning upon the exon-intron genomic regions, we constructed a script that enabled us to cast a given value (per genomic position) onto a set of coordinates, resulting in a per-nucleotide map of values. We used this script to cast the obtained nucleosome positioning data onto the coordinates of exons and flanking intronic sequences (600 nt upstream and downstream) of both groups of exons ("low GC" and "high GC").

Minigene Choice Criteria and Cloning

We cloned genomic sequence from three representative groups: (1) exons with high GC content flanked by short introns with similar GC content; intron length was restricted to 250 nt; exon length was restricted to 185 nt; exon and intron GC content is of at least 65%; (2) exons with low GC content flanked by long introns with significantly lower GC content; intron length was restricted to 1,000–2,500 nt; exon length was restricted to 185 nt; intron GC content less than 40% and at least 8% under that of the exon; (3) exons with low GC content flanked by short introns with significantly lower GC content (representing the ancestral gene architecture); intron length was restricted to 250 nt; exon length was restricted to 185 nt; intron GC content less than 40% and at least 8% under that of the exon. For further procedures pertaining to the minigenes see Extended Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Results, Extended Experimental Procedures, four figures, and four tables and can be found with this article online at doi:10.1016/j.celrep.2012.03.013.

ACKNOWLEDGMENTS

This work was supported by a grant from the Israel Science Foundation (ISF 61/09); Joint Germany-Israeli Research Program (ca-139); Deutsche-Israel Project (DIP MI-1317); and Israel Cancer Research Foundation (ICRF). D.H. and S.G. are fellows of the Edmond J. Safra Bioinformatics Center at Tel-Aviv University. D.B is a fellow of the Converging Technologies Program of the Israeli Council for Higher Education. T.P is supported by a grant from the Israel Science Foundation (878/09).

Received: December 27, 2011
Revised: March 7, 2012
Accepted: March 30, 2012
Published online: May 3, 2012

REFERENCES

Ast, G. (2004). How did alternative splicing evolve? Nat. Rev. Genet. *5*, 773–782.

Barbosa-Morais, N.L., Carmo-Fonseca, M., and Aparício, S. (2006). Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. Genome Res. *16*, 66–77.

Berget, S.M. (1995). Exon recognition in vertebrate splicing. J. Biol. Chem. *270*, 2411–2414.

Bernardi, G. (1993). The vertebrate genome: isochores and evolution. Mol. Biol. Evol. *10*, 186–204.

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. Gene *241*, 3–17.

Bernardi, G. (2001). Misunderstandings about isochores. Part 1. Gene *276*, 3–13.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. Science *228*, 953–958.

Blencowe, B.J. (2000). Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem. Sci. *25*, 106–110.

Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q., and Krainer, A.R. (2003). ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res. *31*, 3568–3571.

Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., and Johnson, J.M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat. Genet. *40*, 1416–1425.

Collins, L., and Penny, D. (2005). Complex spliceosomal organization ancestral to extant eukaryotes. Mol. Biol. Evol. *22*, 1053–1066.

Cooper, T.A., and Mattox, W. (1997). The regulation of splice-site selection, and its role in human disease. Am. J. Hum. Genet. *61*, 259–266.

Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell *136*, 777–793.

Cross, S.H., Clark, V.H., Simmen, M.W., Bickmore, W.A., Maroon, H., Langford, C.F., Carter, N.P., and Bird, A.P. (2000). CpG island libraries from human chromosomes 18 and 22: landmarks for novel genes. Mamm. Genome *11*, 373–383.

de Almeida, S.F., Grosso, A.R., Koch, F., Fenouil, R., Carvalho, S., Andrade, J., Levezinho, H., Gut, M., Eick, D., Gut, I., et al. (2011). Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. Nat. Struct. Mol. Biol. *18*, 977–983.

Desmet, F.O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. *37*, e67.

Deutsch, M., and Long, M. (1999). Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res. *27*, 3219–3228.

Eisenberg, E., and Levanon, E.Y. (2003). Human housekeeping genes are compact. Trends Genet. *19*, 362–365.

Fairbrother, W.G., Holste, D., Burge, C.B., and Sharp, P.A. (2004). Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. *2*, E268.

Fox-Walsh, K.L., Dou, Y., Lam, B.J., Hung, S.P., Baldi, P.F., and Hertel, K.J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc. Natl. Acad. Sci. USA *102*, 16176–16181.

Gelfman, S., Burstein, D., Penn, O., Savchenko, A., Amit, M., Schwartz, S., Pupko, T., and Ast, G. (2012). Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. Genome Res. *22*, 35–50.

Guo, M., Lo, P.C., and Mount, S.M. (1993). Species-specific signals for the splicing of a short Drosophila intron in vitro. Mol. Cell. Biol. *13*, 1104–1118.

Hastings, M.L., and Krainer, A.R. (2001). Pre-mRNA splicing in the new millennium. Curr. Opin. Cell Biol. *13*, 302–309.

Hertel, K.J. (2008). Combinatorial control of exon recognition. J. Biol. Chem. *283*, 1211–1215.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nat. Rev. Genet. *11*, 345–355.

Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. *35*, 125–131.

Kim, S., Kim, H., Fong, N., Erickson, B., and Bentley, D.L. (2011). Pre-mRNA splicing is a determinant of histone H3K36 methylation. Proc. Natl. Acad. Sci. USA *108*, 13564–13569.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S., and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. Nat. Genet. *41*, 376–381.

Kornblihtt, A.R., Schor, I.E., Allo, M., and Blencowe, B.J. (2009). When chromatin meets splicing. Nat. Struct. Mol. Biol. *16*, 902–903.

Krawczak, M., Reiss, J., and Cooper, D.N. (1992). The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum. Genet. *90*, 41–54.

Krawczak, M., Thomas, N.S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., and Cooper, D.N. (2007). Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum. Mutat. *28*, 150–158.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al; International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lercher, M.J., Urrutia, A.O., Pavlícek, A., and Hurst, L.D. (2003). A unification of mosaic structures in the human genome. Hum. Mol. Genet. *12*, 2411–2415.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. Science *302*, 1401–1404.

Majewski, J., and Ott, J. (2002). Distribution and characterization of regulatory elements in the human genome. Genome Res. *12*, 1827–1836.

Nakai, K., and Sakamoto, H. (1994). Construction of a novel database containing aberrant splicing mutations of mammalian genes. Gene *141*, 171–177.

Niu, D.K. (2008). Exon definition as a potential negative force against intron losses in evolution. Biol. Direct *3*, 46.

Nogues, G., Kadener, S., Cramer, P., Bentley, D., and Kornblihtt, A.R. (2002). Transcriptional activators differ in their abilities to control alternative splicing. J. Biol. Chem. *277*, 43110–43114.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Prendergast, J.G., and Semple, C.A. (2011). Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. Genome Res. *21*, 1777–1787.

Ram, O., and Ast, G. (2007). SR proteins: a foot on the exon before the transition from intron to exon definition. Trends Genet. *23*, 5–7.

Roca, X., Sachidanandam, R., and Krainer, A.R. (2003). Intrinsic differences between authentic and cryptic 5′ splice sites. Nucleic Acids Res. *31*, 6321–6333.

Roy, S.W., and Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. Nat. Rev. Genet. *7*, 211–221.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887–898.

Schor, I.E., Rascovan, N., Pelisch, F., Alló, M., and Kornblihtt, A.R. (2009). Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. Proc. Natl. Acad. Sci. USA *106*, 4325–4330.

Schwartz, S., and Ast, G. (2010). Chromatin density and splicing destiny: on the cross-talk between chromatin structure and splicing. EMBO J. *29*, 1629–1636.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol. *16*, 990–995.

Schwartz, S.H., Silva, J., Burstein, D., Pupko, T., Eyras, E., and Ast, G. (2008). Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. *18*, 88–103.

Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. *15*, 7155–7174.

Sorek, R., and Ast, G. (2003). Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res. *13*, 1631–1637.

Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased chromatin signatures around polyadenylation sites and exons. Mol. Cell *36*, 245–254.

Sterner, D.A., Carlo, T., and Berget, S.M. (1996). Architectural limits on split genes. Proc. Natl. Acad. Sci. USA *93*, 15081–15085.

Subtil-Rodríguez, A., and Reyes, J.C. (2010). BRG1 helps RNA polymerase II to overcome a nucleosomal barrier during elongation, in vivo. EMBO Rep. *11*, 751–757.

Sugnet, C.W., Srinivasan, K., Clark, T.A., O'Brien, G., Cline, M.S., Wang, H., Williams, A., Kulp, D., Blume, J.E., Haussler, D., and Ares, M., Jr. (2006). Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput. Biol. *2*, e4.

Talerico, M., and Berget, S.M. (1994). Intron definition in splicing of small Drosophila introns. Mol. Cell. Biol. *14*, 3434–3445.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. Nat. Struct. Mol. Biol. *16*, 996–1001.

Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics *10*, 442.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Yandell, M., Mungall, C.J., Smith, C., Prochnik, S., Kaminker, J., Hartzell, G., Lewis, S., and Rubin, G.M. (2006). Large-scale trends in the evolution of gene structures within 11 animal genomes. PLoS Comput. Biol. *2*, e15.

Yeo, G., Hoon, S., Venkatesh, B., and Burge, C.B. (2004). Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc. Natl. Acad. Sci. USA *101*, 15700–15705.

Yeo, G.W., Van Nostrand, E.L., and Liang, T.Y. (2007). Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. PLoS Genet. *3*, e85.

Zhang, C., and Darnell, R.B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat. Biotechnol. *29*, 607–614.