Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing

Bin Tian, Zhenhua Pan and Ju Youn Lee

Genome Res. 2007 17: 156-165 originally published online January 8, 2007
Access the most recent version at doi:10.1101/gr.5532707

Letter

Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing

Bin Tian,[1] Zhenhua Pan, and Ju Youn Lee

Department of Biochemistry and Molecular Biology, New Jersey Medical School, University of Medicine and Dentistry of New Jersey, Newark, New Jersey 07101, USA

mRNA polyadenylation and pre-mRNA splicing are two essential steps for the maturation of most human mRNAs. Studies have shown that some genes generate mRNA variants involving both alternative polyadenylation and alternative splicing. Polyadenylation in introns can lead to conversion of an internal exon to a 3′ terminal exon, which is termed composite terminal exon, or usage of a 3′ terminal exon that is otherwise skipped, which is termed skipped terminal exon. Using cDNA/EST and genome sequences, we identified polyadenylation sites in introns for all currently known human genes. We found that ~20% human genes have at least one intronic polyadenylation event that can potentially lead to mRNA variants, most of which encode different protein products. The conservation of human intronic poly(A) sites in mouse and rat genomes is lower than that of poly(A) sites in 3′-most exons. Quantitative analysis of a number of mRNA variants generated by intronic poly(A) sites suggests that the intronic polyadenylation activity can vary under different cellular conditions for most genes. Furthermore, we found that weak 5′ splice site and large intron size are the determining factors controlling the usage of composite terminal exon poly(A) sites, whereas skipped terminal exon poly(A) sites tend to be associated with strong polyadenylation signals. Thus, our data indicate that dynamic interplay between polyadenylation and splicing leads to widespread polyadenylation in introns and contributes to the complexity of transcriptome in the cell.

[Supplemental material is available online at www.genome.org.]

Maturation of mRNA involves multiple steps of processing, including capping, splicing, and polyadenylation (Proudfoot et al. 2002). Splicing and polyadenylation are responsible for removing introns and adding poly(A) tails, respectively. Essential signals for splicing out an intron from a pre-mRNA include *cis* elements at the 5′ splice site (5′ss), at the 3′ splice site (3′ss), and at the branchpoint site in the intron (Burge et al. 1999). Splicing involves two steps: First, the 5′ss is attacked by the 2′OH of an adenosine at the branchpoint, resulting in a 5′ exon with a free 3′OH and a lariat consisting of the intron and 3′ exon; second, the 3′OH of the 5′ exon is joined with the 3′ss of the 3′ exon via a transesterification reaction and the lariat is released. An array of proteins and RNAs are involved in the splicing reaction, including several small nuclear RNAs (snRNAs) and their associated proteins that form small nuclear ribonucleoproteins (snRNPs). Regulators of splicing include various SR (serine- and arginine-rich) proteins and hnRNP (heterogeneous nuclear ribonucleoproteins) proteins (Jurica and Moore 2003). In addition, a number of *cis* elements located in both exons and introns play enhancing or repressing roles in splicing (Ladd and Cooper 2002). Alternative splicing occurs in ~40%–60% human genes, contributing to the functional complexity of the human genome (Modrek and Lee 2002).

mRNA polyadenylation is a two-step reaction (Colgan and Manley 1997; Edmonds 2002), involving a specific endonucleolytic cleavage at the polyadenylation site (poly(A) site) and subsequent polymerization of an adenosine tail. Proteins that participate in the polyadenylation reaction in mammals include cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), cleavage factors CF I and CF II, and poly(A) polymerase (PAP). The sequence surrounding the poly(A) site, called the poly(A) region herein, contains various *cis* elements for polyadenylation, including the upstream polyadenylation signal (PAS), such as AAUAAA, AUUAAA and other hexamer variants (Beaudoing et al. 2000; Tian et al. 2005), and downstream U-rich and GU-rich elements (Zhao et al. 1999). In addition, a number of auxiliary elements have been suggested or shown to play a role in regulating polyadenylation (Hu et al. 2005 and references therein). Recently, another type of RNA polyadenylation process has been identified in eukaryotic cells, which involves a different set of proteins and has been implicated in RNA degradation in the nucleus (LaCava et al. 2005; West et al. 2006). Over half of the human genes have multiple poly(A) sites, potentially resulting in transcripts encoding distinct protein products and/or possessing variable 3′ untranslated regions (3′ UTRs) (Tian et al. 2005; Yan and Marr 2005).

Growing lines of evidence indicate that splicing and polyadenylation are coupled events that take place cotranscriptionally (Proudfoot et al. 2002). First, a number of human genes have mRNA variants whose production involves both alternative splicing and alternative polyadenylation. For example, the immunoglobulin M (*IgM*) heavy chain gene has a poly(A) site located in an intron, leading to mRNA variants differentially expressed in different stages of B cell development. Both polyadenylation and splicing activities have been found to be responsible for the regulation of alternative transcripts of the *IgM* heavy chain gene (Edwalds-Gilbert and Milcarek 1995; Takagaki et al. 1996; Bruce et al.

[1]**Corresponding author.**
E-mail btian@umdnj.edu; fax (973) 972-5594.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5532707.

17:156–165 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

2003). Another extensively studied case is the calcitonin/calcitonin gene-related peptide gene (*CALCA*), where the usage of an intronic poly(A) site is regulated by the splicing factor SRp20 in a tissue-specific manner to make mRNA variants (Lou et al. 1998). In addition, a number of studies have shown that perturbation of splicing can affect polyadenylation, and vice versa (Niwa and Berget 1991; Nesic and Maquat 1994; Cooke et al. 1999). Second, several splicing factors have been shown to regulate polyadenylation, such as U1A, U1 70, PTB, SRp20, and p54nrb (Gunderson et al. 1994; Lutz et al. 1996; Lou et al. 1998; Moreira et al. 1998; Liang and Lutz 2006), and a number of protein–protein interactions have been reported between polyadenylation and splicing factors, such as U1 snRNP with CF Im (Awasthi and Alwine 2003), CPSF with U2 snRNP (Kyburz et al. 2006), and U2AF 65 with CF Im (Millevoi et al. 2006), suggesting mechanistic interplay between these two processes. Third, both splicing and polyadenylation factors interact extensively with the C-terminal domain (CTD) of RNA polymerase II (Hirose and Manley 2000; Proudfoot et al. 2002; Kaneko and Manley 2005), which plays critical roles at various stages of transcription, i.e., initiation, elongation, and termination, suggesting temporal and spatial coordination between splicing and polyadenylation.

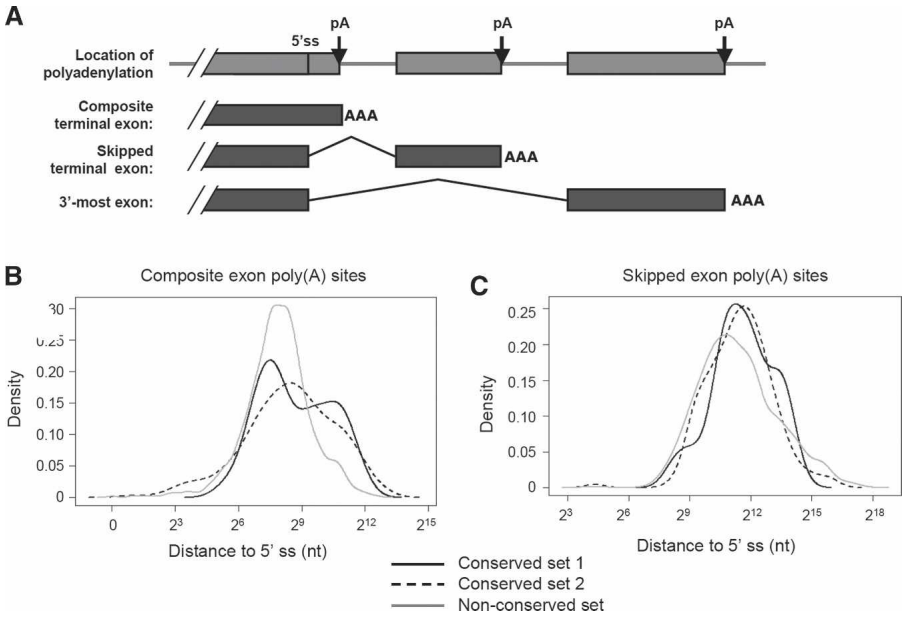

Figure 1. Intronic poly(A) sites in human genes. (*A*) Schematic of poly(A) sites located in different types of exons, i.e., composite terminal exon, skipped terminal exon, and 3′-most exon. 5′ss indicates 5′ splice site; pA, poly(A) site. Exons are shown as boxes. Splicing is indicated by an angled line. (*B*) Distance between 5′ss and composite exon poly(A) sites. Median values are 295 nt, 355 nt, and 238 nt for poly(A) sites in the conserved set 1, conserved set 2, and nonconserved set, respectively. (*C*) Distance between 5′ss and skipped exon poly(A) sites. Median values are 3445 nt, 2997 nt, and 2320 nt for poly(A) sites in the conserved set 1, conserved set 2, and nonconserved set, respectively. As indicated, solid black lines are for poly(A) sites in the conserved set 1, dotted black lines are for poly(A) sites in the conserved set 2, and solid gray lines are for poly(A) sites in the nonconserved set.

In the present work, we systematically examined polyadenylation events in introns, similar to those in the *IgM* heavy chain and *CALCA* genes. We define an intronic poly(A) site as a site that is located upstream of the 3′-most exon of a gene, and is spliced out in some transcripts of the gene. Intronic poly(A) sites can be divided into two types (Fig. 1A): one whose usage leads to conversion of an internal exon to a 3′ terminal exon, such as in the case of *IgM* heavy chain gene, and the other whose usage leads to inclusion of an otherwise skipped exon, such as in the case of *CALCA*. Exons associated with these types are termed composite terminal exons and skipped terminal exons, respectively, as reported in Edwalds-Gilbert et al. (1997) and Zhao et al. (1999). For simplicity, we call the poly(A) sites associated with these two types of exons composite exon poly(A) sites and skipped exon poly(A) sites, respectively. Using cDNA/EST and genome sequences, we found that ~20% of human genes have at least one intron containing poly(A) sites. More composite exon poly(A) sites were found than skipped exon poly(A) sites. Intronic polyadenylation can lead to different mRNA and protein products. Using human versus mouse and human versus rat whole genome alignments, we found that 10% of human intronic poly(A) sites are conserved in rodent genomes, whereas 51% of 3′-most poly(A) sites and 29% of other poly(A) sites (not 3′-most poly(A) sites) in 3′-most exons are conserved. Conserved intronic poly(A) sites are associated with stronger polyadenylation signals than nonconserved ones. Quantitative analysis of the expression of mRNA variants generated by intronic polyadenylation suggests that the intronic polyadenylation activity varies between cell types for most genes. Furthermore, we found that weak 5′ splice site and large intron size are the determining factors for the usage of composite exon poly(A) sites, whereas skipped exon poly(A) sites are associated with strong polyadenylation signals. Taken together, our data indicate that dynamic interplay between splicing and polyadenylation leads to widespread intronic polyadenylation events in the human genome and contributes to the complexity of the transcriptome in the cell.

Results

We have previously found that over half of the human genes have alternative polyadenylation products (Tian et al. 2005). Interestingly, a large number of poly(A) sites are located upstream of the 3′-most exon, including introns and internal exons. Alternative poly(A) sites in the 3′-most exon can lead to variable 3′ UTRs that contain different *cis* elements for mRNA metabolism, such as AU-rich elements and miRNA target sequences (Farh et al. 2005; Khabar et al. 2005). However, little is known about polyadenylation in introns on a global level. In the present study, we set out to address what percentage of human genes contain intronic poly(A) sites; how many of them are conserved in other mammalian genomes, such as mouse and rat genomes; what are the roles of intronic polyadenylation in modulating gene functions; and what are the characteristics of intronic poly(A) sites and introns that harbor them.

Widespread mRNA polyadenylation in human introns

To identify intronic poly(A) sites, we first aligned human cDNA/ESTs with human genome sequences and mapped all

poly(A) sites on the human genome (for details, see Methods). We then used NCBI RefSeq (Pruitt and Maglott 2001) and UCSC KnownGene sequences (Hsu et al. 2006) as mRNA templates and identified poly(A) sites in their introns using cDNA/ESTs. To eliminate spurious transcripts and genes or transcription units located entirely in an intron, we required that cDNA/ESTs ending at a poly(A) site in an intron of a template sequence must overlap with the template sequence by at least 32 nt. Each unique intron-poly(A) site pair is an intronic polyadenylation event, which is defined by the 5′ss, 3′ss, and poly(A) site on a given chromosome. The poly(A) site for each event was classified as composite exon poly(A) site or skipped exon poly(A) site according to the sequences of its supporting cDNA/ESTs. When there were equal numbers of cDNA/ESTs supporting both forms, an event was classified as "both." As summarized in Table 1, of 16,610 human genes we surveyed, we identified 4625 intronic poly(A) sites and 5088 intronic polyadenylation events in 3344 genes. Thus, 20% of human genes have at least one intronic polyadenylation event. Some introns contain more than one poly(A) site, and some poly(A) sites are situated in different introns (different 5′ss and 3′ss). There are more polyadenylation events resulting in composite terminal exons than skipped terminal exons. Most introns containing poly(A) sites are flanked by exons containing coding sequences (CDS) (Table 1).

To address how many human intronic poly(A) sites are conserved in other species, we mapped all poly(A) sites in the mouse and rat genomes, and identified intronic poly(A) sites in these two species by the same method described above. We then used the human versus mouse and human versus rat whole genome alignments and identified human versus mouse and human versus rat orthologous poly(A) site pairs (for details, see Methods). For each orthologous poly(A) site pair, we required that (1) the human and mouse/rat sites are located within 24 nt in the human and mouse/rat genome alignment, and (2) they are nearest to one another in a reciprocal manner, i.e., the mouse/rat poly(A) site is the nearest one to the human poly(A) site on the mouse/rat genome and the human site is the nearest one to the mouse/rat site on the human genome. In sum, we identified 449 human intronic poly(A) sites that have orthologous sites in mouse or rat genomes, among which 135 intronic poly(A) sites have their orthologous sites also classified as intronic poly(A) sites in mouse or rat, and 314 poly(A) sites have orthologous sites in mouse or rat, but the sites in these two species are not classified as intronic poly(A) sites. The former group is called conserved set 1; the latter, conserved set 2. The main reason for poly(A) sites being in the conserved set 2 is that much fewer cDNA/EST and template sequences are available for mapping mouse and rat intronic poly(A) sites. For example, the well-known skipped exon poly(A) site in *CALCA* is in the conserved set 2. On the other hand, poly(A) sites belonging to the conserved set 1 can be frequently utilized poly(A) sites, or their corresponding mRNAs are highly expressed, which can make them easier to be detected in mouse and rat genomes than those in the conserved set 2 (for further discussion, see below). Thus, 90% of human intronic poly(A) sites (4176) do not have orthologous sites in the mouse or rat genomes, which constitute a nonconserved set. To assess the degree of conservation, we applied the same mapping method to 31,512 poly(A) sites located in the 3′-most exons of the RefSeq and KnownGene mRNA sequences. We also classified these sites into 3′-most poly(A) sites (16,107 in total) and other sites in the 3′-most exons (15,405 in total). Using the same method for identifying orthologous poly(A) sites, we found that 51% (8141) of 3′-most poly(A) sites and 29% (4474) of other poly(A) sites in 3′-most exons have orthologous sites in mouse or rat genomes. Thus, the conservation of intronic poly(A) sites is much less than that of sites located in 3′-most exons. We discuss this low conservation in the Discussion. All intronic polyadenylation events are shown in Supplemental Tables 1 through 3.

Table 1. Intronic poly(A) sites in human genes

	Conserved set 1	Conserved set 2	Nonconserved set
Genes (total unique: 3344)	121	279	3123
Intronic poly(A) sites (total: 4625)	135	314	4176
Intronic polyadenylation events[a] (total: 5088)	159	349	4580
Terminal exon type[b]			
Composite	93	176	3202
Skipped	65	168	1340
Both	1	5	38
Affected region			
5′ UTR	3	6	179
CDS	143	274	4147
3′ UTR	13	69	254

Conserved set 1 contains human intronic poly(A) sites whose orthologous sites in mouse or rat genomes were also found to be in introns. Conserved set 2 contains human poly(A) sites that have orthologous sites in mouse or rat genomes, but the orthologous sites were not classified as intronic poly(A) sites due to lack of supporting evidence (for details, see Methods). Nonconserved set contains human intronic poly(A) sites that do not have orthologous sites identified in mouse or rat genomes.
[a]An intronic polyadenylation event is defined as a poly(A) site in a particular intron (5′ss + 3′ss) of a RefSeq or KnownGene mRNA.
[b]Terminal exon type is the terminal exon resulting from an intronic polyadenylation event. "Both" indicates that there exist equal numbers of cDNA/EST evidences supporting composite and skipped terminal exons.

As shown in Figure 1B, most composite exon poly(A) sites are located 63–958 nt (10th–90th percentiles) from the 5′ss of an intron. However, poly(A) sites in the conserved sets appear to have bimodal distributions. As expected, skipped exon poly(A) sites are located farther away from the 5′ss than composite exon poly(A) sites, mostly ranging from 554–16,286 nt (10th–90th percentiles). No significant differences can be discerned among different groups for the distance between poly(A) site and 3′ss (Supplemental Fig. 1). For both composite and skipped exon poly(A) sites, 3′ terminal exons generated by poly(A) sites in the conserved sets are larger than those generated by sites in the nonconserved set, but smaller than 3′-most exons (Supplemental Fig. 2).

Intronic polyadenylation activity varies in different cell lines

Of the intronic poly(A) sites in conserved set 1, some are already known, such as two sites in *PAP* (Zhao and Manley 1996), one site in *CSTF3* (Pan et al. 2006), and several sites listed in (Edwalds-Gilbert et al. 1997), including alpha-tropomyosin, (2′–5′) oligoadenylate synthetase, etc. Other sites have not been previously reported and appear to have the potential to significantly regulate functions of gene products. For example, the cyclin C gene has a skipped exon poly(A) site, resulting in a protein isoform containing a poor PEST motif (proline [P], glutamic acid [E], serine [S], threonine [T] domain) at its C terminus, whereas the isoform derived from using poly(A) sites in the 3′-most exon has a strong PEST motif (Supplemental Fig. 3). Since the PEST motif is responsible for protein stability, the half-lives of these two

protein isoforms are presumably different. The cyclin C protein interacts with CDK8, which regulates transcription via phosphorylation of the CTD of Pol II (Hengartner et al. 1998) and several other transcription factors (Akoulitchev et al. 2000; Liu et al. 2004). It also plays a role in cell's entry into the cell cycle from G0 phase by interacting with CDK3 (Ren and Rollins 2004). Interestingly, the cyclin C/CDK8 complex itself has been shown to regulate protein turnover of the Notch receptor intracellular domain (ICD) by phosphorylation of its PEST domain (Fryer et al. 2004). Whether cyclin C can regulate its own turnover rate and whether alternative polyadenylation can play a role in modulating its activity are to be further explored in the future. We also found that there are two composite exon poly(A) sites in the *C20orf67* gene (previously known as *PCIF1*; see Supplemental Table 1), whose product has been shown to interact with the phosphorylated CTD of Pol II (Fan et al. 2003). The poly(A) sites are located upstream of the exon containing the start codon (Supplemental Fig. 4), which represents one of the few cases that intronic polyadenylation leads to production of noncoding RNAs, thereby potentially shutting down protein expression.

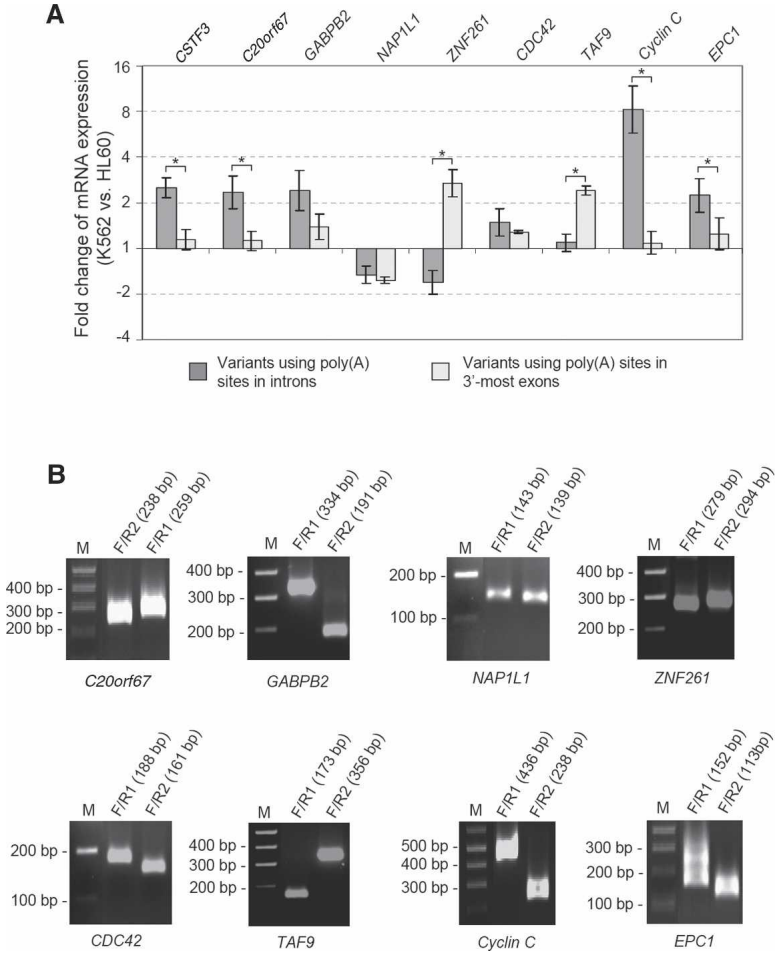

Figure 2. Intronic polyadenylation activity varies between cell lines. (*A*) QPCR results of nine genes that contain intronic poly(A) sites. For each gene, two sets of primers were used to detect the mRNA variant(s) generated by intronic polyadenylation and the mRNA variant(s) generated by polyadenylation in the 3′-most exon. For each variant type, the mRNA expression level (QPCR value) from K562 cells was compared with that from HL60 cells. For each gene, fold changes of intronic polyadenylation variants and 3′-most exon variants were compared, and those significantly different (*P*-value < 0.05, *t*-test) are indicated by asterisks. The result is based on two experiments, each with samples in duplicate. Error bar is SD. (*B*) PCR products using mRNAs from human K562 cells. M indicates molecular marker; F/R1, products by primers F and R1; and F/R2, products by primers F and R2 (for primer sequences and their targeted regions, see Supplemental Fig. 4). The expected molecular weight based on supporting cDNA/ESTs for each PCR product is indicated *above* each lane.

We set out to validate some of the intronic polyadenylation events, and to address whether the intronic polyadenylation activity varies under different cell conditions. To this end, we used two human myeloid leukemia cell lines, K562 and HL60, which represent two stages of myeloid development, with K562 being undifferentiated blast cells and HL60 being at the promyelocyte stage of maturation (Koeffler and Golde 1980). We selected nine genes that have conserved intronic polyadenylation events in rodents, including *CSTF3*, *C20orf67*, *GABPB2* (GA binding protein transcription factor, β subunit 2), *NAP1L1* (nucleosome assembly protein 1-like 1), *ZNF261* (zinc finger protein 261, currently known as *ZMYM3*), *CDC42* (cell division cycle 42), *TAF9* (TATA box binding protein-associated factor, 32 kDa), cyclin C, and *EPC1* (enhancer of polycomb homolog 1). To the best of our knowledge, most of the intronic polyadenylation events have not been reported so far, except for *CSTF3* (Pan et al. 2006) and *GABPB2* (Gugneja et al. 1995). Using mRNAs from K562 and HL60 cells and primer sets that distinguish mRNA variants generated by intronic polyadenylation (or intronic polyadenylation variants) from mRNA variants generated by polyadenylation in 3′-most exons (or 3′-most exon variants, see Supplemental Figure 4 for primer designs and sequences), we confirmed the usage of intronic poly(A) sites for all nine genes (Fig. 2). Furthermore, using quantitative PCR (QPCR), we compared intronic polyadenylation variants with 3′-most exon variants with respect to their change of mRNA expression in the two cell lines. As shown in Figure 2A, for most genes, different variants are expressed differently in K562 cells versus HL60 cells. Several genes, including *CSTF3*, *C20orf67*, cyclin C, and *EPC1*, have up-regulated expression in K562 for both variants, but the intronic polyadenylation variants appear to be up-regulated to a greater extent than the 3′-most exon variants (*P*-values < 0.05, *t*-tests). On the other hand, *ZNF261* and *TAF9* have the opposite relationship between the two types of variants. While mRNA stability may be a factor that influences the steady-state level for some of the variants, the overall trend from these data suggests that the relative frequency of intronic polyadenylation events as opposed to polyadenylation in 3′-most exons may vary under different conditions for most genes. This is consistent with our previous bioinformatic finding that alternative poly(A) sites are utilized differently in different tissues (Zhang et al. 2005). However, a more systematic validation approach is needed to confirm this notion for a larger number of genes.

Composite exon poly(A) sites are associated with weak 5′ss and large intron size

We then wanted to examine whether certain features of intron are different between introns with poly(A) sites and introns without poly(A) sites. We first analyzed 5′ss, 3′ss, and intron size for introns containing composite exon poly(A) sites. For 5′ss and 3′ss, we generated consensus sequences to build position-specific scoring matrices (PSSMs) using all GT-AG type introns of RefSeq and KnownGene sequences (181,669 introns in total). PSSMs for 5′ss and 3′ss were then used to score each 5′ss and 3′ss. A high score indicates a good similarity to the consensus and, presumably, a stronger signal for promoting splicing. We focused on the –3 to +6 nt region at the 5′ss and the –22 to +2 nt region at the 3′ss. Figure 3, A and B, shows distributions of 5′ss and 3′ss scores for different groups of introns, including introns without poly(A) sites and introns containing composite exon poly(A) sites in the conserved set 1, conserved set 2, and nonconserved set. We found that introns with poly(A) sites in all three sets have lower 5′ ss scores than introns without poly(A) sites (P-values $<1 \times 10^{-5}$, Wilcoxon tests), but similar 3′ ss scores were observed for all groups. This observation was also confirmed by a modified Kolmogorov-Smirnov test (or mKS test) (Mootha et al. 2003), which used a randomization scheme to get the probability (termed E-value) that the difference between two groups of introns is due to random chance (for details, see Methods). As shown in Figure 3, A and B, introns containing composite exon poly(A) sites have significantly lower 5′ss scores (E-value = 0) than do introns without poly(A) site, but they have similar 3′ss scores (E-value = 0.084). Furthermore, we examined potential base pairs between the 5′ss sequence and U1 snRNA sequence for different groups of introns. Indeed, both ΔG and number of potential base pairs suggest introns containing composite exon poly(A) sites have weaker 5′ss than do introns without poly(A) sites (Supplemental Fig. 5A).

We then compared the intron size among different groups of introns (Fig. 3C). We found that introns with poly(A) sites are significantly larger than are introns without poly(A) sites by Wilcoxon tests (P-values $<1 \times 10^{-8}$) and the mKS test (E-value = 0, Figure 3C). Since

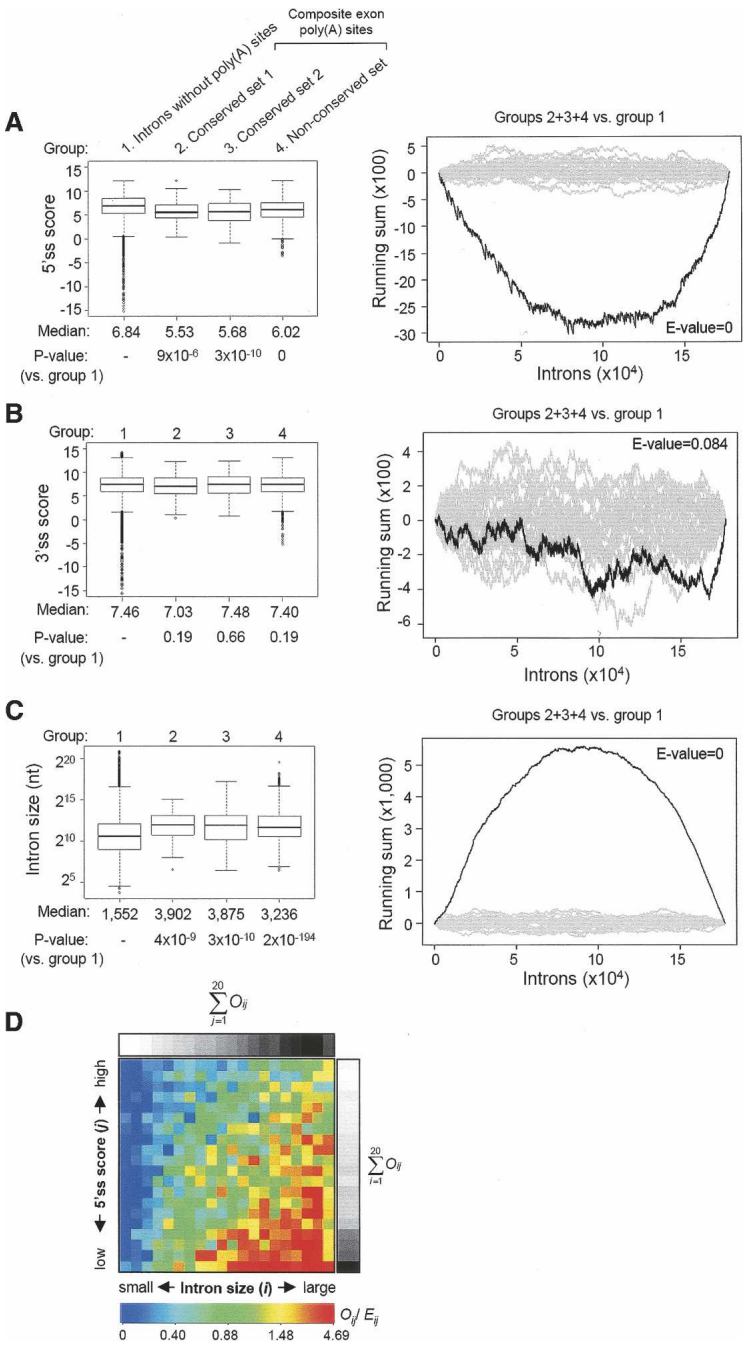

Figure 3. Characteristics of introns containing composite exon poly(A) sites. (*A*) Boxplots of 5′ss scores for four groups of introns (*left*) and a mKS test result (*right*) comparing introns without poly(A) sites with introns with composite exon poly(A) sites with respect to 5′ss scores. (*B*) As in *A* except that 3′ss scores are plotted and compared. (*C*) As in *A* except that intron sizes are plotted and compared. For boxplots, median values and *P*-values from the Wilcoxon tests comparing each group with group 1 are shown. For mKS tests, the E-values are expected values as described in Methods. The E-values for *A* and *B* represent the probability of getting smaller values in groups 2 + 3 + 4 than in group 1 by random chance, and the E-value for *C* represents the probability of getting higher values in groups 2 + 3 + 4 than in group 1 by random chance. In each graph, the black line is the running sum of the real data, and the gray lines are 25 randomly selected running sums from 1000 randomized data. (*D*) Intron distribution map for introns with composite exon poly(A) sites. *x*-axis is intron size (*i*) from small to large, and *Y*-axis is 5′ss score (*j*) from low to high, as indicated in the graph. The ratios of observed values to expected ones (O_{ij}/E_{ij}) are shown in a heatmap, where colors are used to represent values according to the color scale under the graph. The row sum $\sum_{i=1}^{20} Oij$ and column sum $\sum_{j=1}^{20} Oij$ are also shown in grayscale bars presented *next to* and *above* the graph, respectively, with black representing the highest value and white representing the lowest value.

the group of introns without poly(A) sites contains a population of small introns (<512 nt) (Supplemental Fig. 6A), we carried out a mKS test only using introns >512 nt. As shown in Supplemental Figure 6B, the introns with composite exon poly(A) sites are still significantly larger than introns without poly(A) sites (E-value = 0).

To examine whether introns with composite poly(A) sites are associated with both weak 5′ss and large intron size at the same time, we developed an intron distribution map, which indicates how a group of introns are distributed against all introns with respect to 5′ss score and intron size. First, a 20 × 20 grid was constructed, and the value range for each cell was defined by all RefSeq and KnownGene introns. We then put introns with composite poly(A) sites into the grid based on their 5′ss score and intron size. The number of introns with composite poly(A) sites in each cell, called observed value, was compared with the expected value, which was calculated based on the distribution of all introns. If the distribution of an intron group is similar to that of all introns, the ratios of observed values to expected values for all cells should be close to one. If an intron group has a different distribution than all introns, introns from the group are overrepresented in cells with ratios larger than one and underrepresented in cells with ratios less than one. As shown in Figure 3D, high ratios are mostly located in the lower right part of the grid, indicating that introns with composite poly(A) sites are associated with both weak 5′ss and large intron size. Taken together, these data demonstrate that poly(A) sites in composite exons are associated with weaker 5′ss and larger intron size than, but similar 3′ss to, other introns. Since weak 5′ss and large intron size would increase the time to splice out an intron, this result indicates that there exists a dynamic competition between splicing and polyadenylation when a composite exon poly(A) site is encountered in an intron.

Characteristics of introns containing skipped exon poly(A) sites

To delineate the characteristics of introns containing skipped exon poly(A) sites, we first examined their 5′ss scores. In contrast to introns containing composite exon poly(A) sites, introns containing skipped exon poly(A) sites have higher 5′ss scores than introns without poly(A) sites (Fig. 4A). We then examined the 3′ss of introns upstream of the skipped exons (termed upstream introns) and introns containing the skipped terminal exons (termed full introns), as depicted in Figure 4B. We found that 3′ss scores of upstream introns in general are lower than those of introns without poly(A) sites, but full intron 3′ss scores are similar to those of introns without poly(A) sites (Fig. 4C; mKS tests shown in Supplemental Fig. 7). However, comparison of upstream and downstream 3′ss scores corresponding to the same 5′ss did not reveal systematic differences (Fig. 4D), indicating that other factors may play regulatory roles in the usage of skipped exon poly(A) sites. In addition, upstream introns appear to be slightly larger than introns without poly(A) sites, and as expected, the full introns are significantly larger than are introns without poly(A) sites (Fig. 4E). Thus, unlike composite exon poly(A) sites, which are usually situated in large introns with

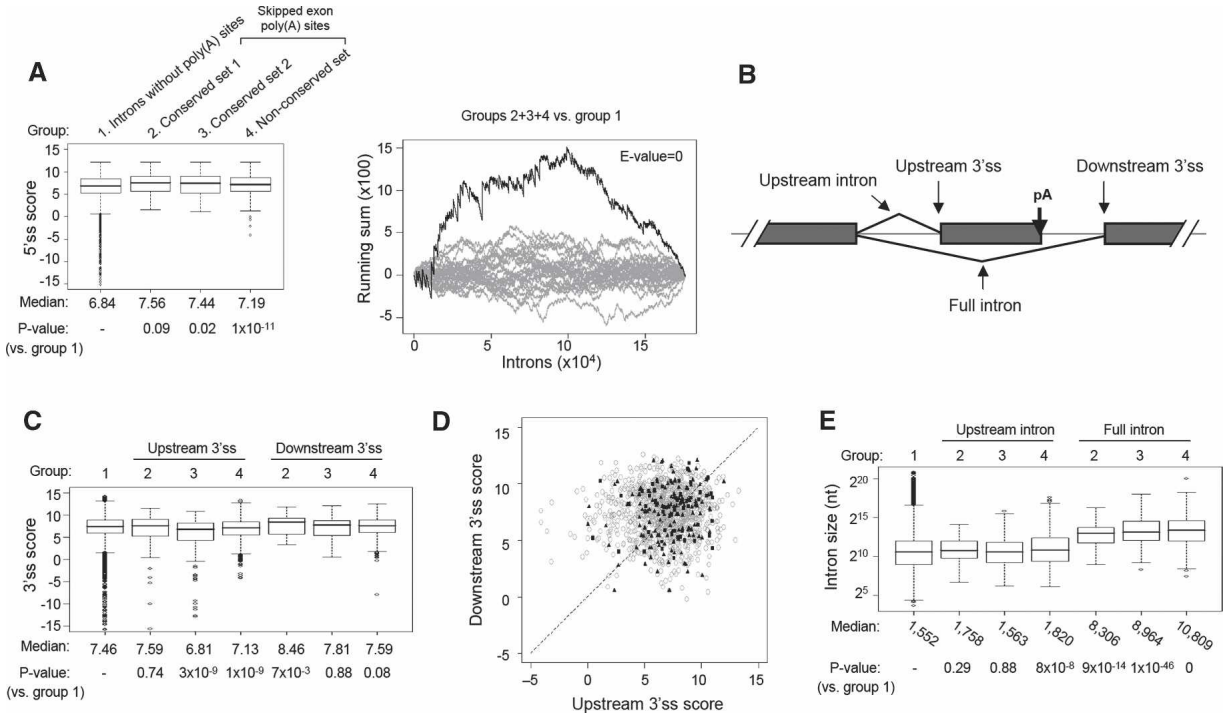

Figure 4. Characteristics of introns containing skipped exon poly(A) sites. (*A*) Boxplots of 5′ss scores for 4 groups of introns (*left*) and a mKS test result (*right*) comparing introns without poly(A) sites with introns with skipped exon poly(A) sites with respect to 5′ss scores. The E-value represents the probability of getting higher values in groups 2 + 3 + 4 than in group 1 by random chance. (*B*) Schematic of a skipped terminal exon in an intron. (*C*) Boxplots of 3′ss scores for four groups of introns. Both upstream 3′ss and downstream 3′ss (indicated in *B*) are shown. (*D*) Scatterplot of upstream 3′ss scores (*x*-axis) and downstream 3′ss scores (*y*-axis). Each dot represents a skipped terminal exon with an upstream 3′ss and a downstream 3′ss for the same 5′ss. Solid squares are for poly(A) sites in the conserved set 1; solid triangles, for poly(A) sites in the conserved set 2; and gray circles, for poly(A) sites in the nonconserved set. (*E*) Boxplots of intron size for four groups of introns. Both upstream introns and full introns are shown. For boxplots, median values and *P*-values from the Wilcoxon tests comparing each group with group 1 are shown.

weak 5′ss, no obvious intron characteristics can explain the usage of skipped exon poly(A) sites.

Distinct polyadenylation signals for different groups of intronic poly(A) sites

We then examined the frequency of usage of PAS hexamers for different groups of intronic poly(A) sites. PAS hexamers can be classified into four types: AAUAAA, AUUAAA, other PAS (11 variants as described in Methods), and no PAS. As demonstrated in many biochemical assays, the strength of a PAS hexamer for promoting polyadenylation is in the order AAUAAA > AUUAAA > other PAS > no PAS. We examined four types of poly(A) sites: 3′-most exon poly(A) sites, other poly(A) sites in 3′-most exons, composite exon poly(A) sites, and skipped exon poly(A) sites. For each type, we analyzed conserved and nonconserved sites separately. As shown in Figure 5, we found that conserved poly(A) sites are associated with stronger signals than are nonconserved ones for every type. 3′-Most poly(A) sites are associated with stronger signals than are other poly(A) sites in the 3′-most exon, as reported before (Tian et al. 2005).

Most conserved intronic poly(A) site groups are similar to 3′-most poly(A) sites, with stronger signals than other poly(A) sites in 3′-most exons. Interestingly, skipped exon poly(A) sites in the conserved set 1, conserved set 2, and nonconserved set are all associated with stronger signals than are composite exon poly(A) sites in the respective groups, indicating that skipped exon poly(A) sites are usually very strong. In fact, the skipped exon poly(A) sites in the conserved set 1 utilize AAUAAA to the greatest extent among all groups. To further examine sequences surrounding the intronic poly(A) sites in greater detail, we generated scores for both the upstream region (−40 to –3 nt) and downstream region (+3 to +40 nt) of a poly(A) site. A score was calculated by comparing a poly(A) region sequence, i.e., −40 to −3 nt or +3 to +40 nt, to the consensus sequence of the region, which is represented as a PSSM. As shown in Supplemental Figure 8, upstream scores and downstream scores show similar trends with respect to the difference between conserved and nonconserved groups, and are consistent with the result of PAS hexamers. Skipped exon poly(A) sites are associated with stronger signals in the upstream region than other poly(A) site types in comparisons among conserved site groups as well as comparisons among nonconserved site groups. Thus, the strength of polyadenylation signal varies among different types of intronic poly(A) sites, suggesting distinct mechanisms in their usage.

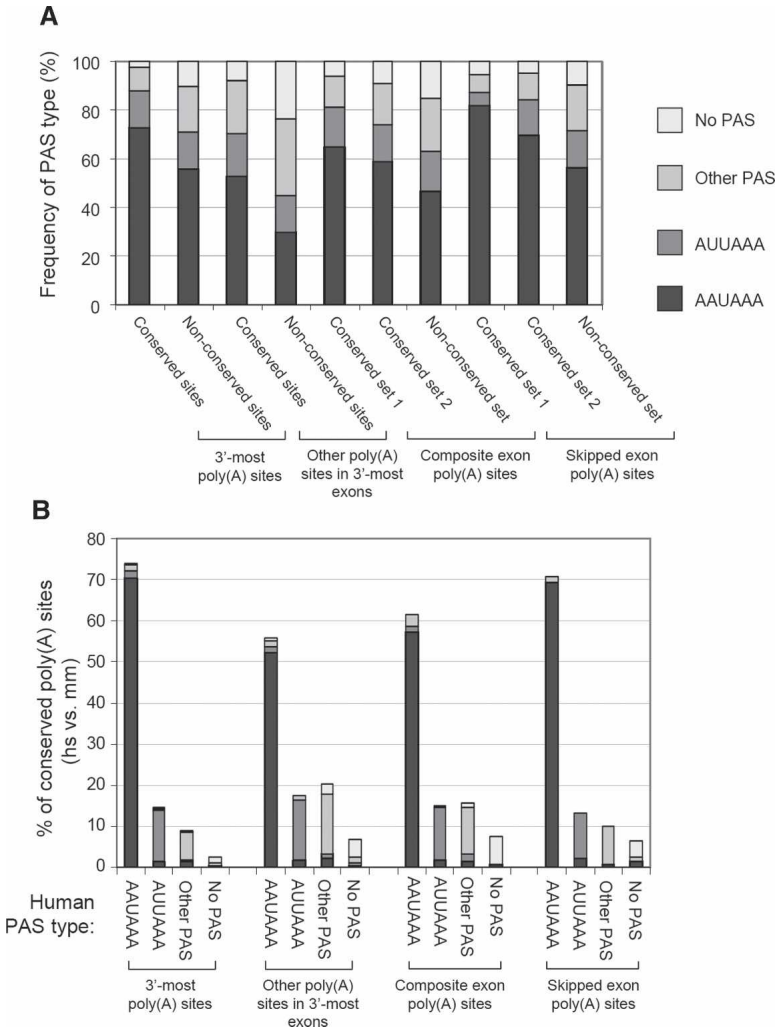

Figure 5. PAS hexamer frequency and conservation for different types of poly(A) site. (*A*) Frequency of four types of PAS hexamers in 10 groups of poly(A) site. Poly(A) site types are indicated at the *bottom* of the graph. The −40 to −1 nt region was used for identifying PAS hexamers. Other PAS corresponds to any one of the 11 variants of AAUAAA (for details, see Methods), and no PAS indicates that no PAS hexamers can be found in the –40 to –1 nt region. (*B*) Conservation of PAS type between human and mouse orthologous poly(A) site pairs. Conserved sets were combined for composite exon poly(A) sites and skipped exon poly(A) sites. Each bar represents the percentage of conserved poly(A) sites having a given PAS type (indicated below the bar) in a human poly(A) site group (indicated at the *bottom* of the graph). Thus, the sum of four bars for a poly(A) site group is one. Each bar contains four areas, representing the frequency of four PAS types for the corresponding mouse sites.

To understand how different types of poly(A) sites are under selective pressure through evolution, we analyzed the conservation of PAS type between human and mouse and between human and rat orthologous poly(A) site pairs. As shown in Figure 5B and Supplemental Figure 9, PAS type is well conserved for all types of poly(A) sites. For human and mouse orthologous poly(A) site pairs, 91% of 3′-most poly(A) sites, 86% of other sites in 3′-most exons, 88% of composite exon poly(A) sites, and 93% of skipped exon poly(A) sites have the identical PAS type, i.e., AAUAAA versus AAUAAA, AUUAAA versus AUUAAA, other PAS versus other PAS, and no PAS versus no PAS. Human and rat orthologous poly(A) sites have similar numbers. The difference between different poly(A) site types is attributed to the percentage of AAUAAA, which is the most conserved PAS type. However, AUUAAA, other PAS, and no PAS also appear to be conserved

through evolution, as their conservation patterns are significantly different than random. Thus, conserved poly(A) sites located in different parts of a gene are under similar selective pressures with respect to the polyadenylation signal, suggesting that they have comparable importance for gene regulation.

Discussion

We found that ~20% of human genes can have mRNA polyadenylation events in introns, resulting in either creation of composite terminal exons or usage of skipped terminal exons. Most of the intronic polyadenylation events can potentially lead to mRNA variants encoding different proteins, indicating that alternative polyadenylation can significantly contribute to the complexity of the proteome in the cell. The conservation of intronic polyadenylation in mouse or rat genomes is lower than that of poly(A) sites in 3′-most exons. While problems in genome sequences and genome alignments can lead to false negatives in orthologous site mapping, the low conservation can be mainly ascribed to two reasons: First, fewer cDNA/ESTs are available for poly(A) site mapping in mouse and rat genomes (Tian et al. 2005; Lee et al. 2007), and mouse and rat genes are usually less well annotated than human ones, resulting in fewer RefSeq and KnownGene sequences that can be used as templates for finding intronic polyadenylation events. In fact, the main reason we combined mouse and rat data for finding orthologous poly(A) sites for human ones is that fewer poly(A) sites and intronic polyadenylation events were found in mouse and rat genomes. This problem can be further exacerbated by the fact that mouse and rat cDNA/ESTs are derived from a much smaller range of tissue/cell types than human ones, considering the tissue/cell-specific regulation of alternative polyadenylation (Zhang et al. 2005). The fact that some human poly(A) sites have orthologous sites in mouse and/or rat genomes but the sites in these two species are not classified as intronic poly(A) sites (those in the conserved set 2) is consistent with this notion. Thus, the total number of human intronic poly(A) sites is likely to be greater than we report here, as more cDNA/ESTs and RefSeq and KnownGenes become available.

Second, variation in genome sequence that leads to species-specific polyadenylation signals is probable for some nonconserved intronic poly(A) sites. While in general polyadenylation signals have been found to be selected against on the sense strand of genes (Glusman et al. 2006) as they can terminate transcription, they may arise in recent history of evolution. Consistent with this notion, nonconserved intronic poly(A) sites are associated with weaker signals than are conserved ones, with respect to PAS hexamers and upstream and downstream sequences. On a similar note, both conserved and nonconserved alternative splicing events between human and mouse genomes have been reported (Thanaraj et al. 2003; Yeo et al. 2005). Thus, different species can have distinct alternative mRNA processing patterns, potentially contributing to the speciation of organisms. It is to be further explored in wet laboratory settings how alternative polyadenylation can contribute to different gene products and, therefore, function in different species.

We found that introns containing composite exon poly(A) sites tend to have weak 5′ss and large size, indicating that the usage of this group of poly(A) sites may be mainly governed by the timing of splicing and polyadenylation. Presumably, weak 5′ss and large intron size would require longer time to splice out an intron, creating a time window for polyadenylation in introns. Interestingly, the 3′ss does not seem significant, indicating that polyadenylation in introns may take place before the completion of transcription of an intron. Conversely, these data suggest that protein factors that enhance splicing, particularly those functioning at 5′ss, can inhibit intronic polyadenylation. This notion is in accord with the fact that several factors in the U1 snRNP, which is directly involved in binding to 5′ss during splicing, have inhibitory effects on polyadenylation, including U1A and U1 70K (Gunderson et al. 1994; Gunderson et al. 1998). On a similar note, we previously found high expression of U1A in human brain tissues, in which decreased usage of poly(A) sites upstream to 3′-most exons was also observed (Zhang et al. 2005). In addition, our result suggests that suboptimal splicing activity in the cell can potentially lead to enhanced usage of composite exon poly(A) sites. These events would be less conserved for reasons described above. Supporting this notion, composite exon poly(A) sites account for 70% of the sites in the nonconserved set but 50%–60% of the sites in the conserved sets (Table 1).

Compared with intronic polyadenylation events involving composite terminal exons, the interaction between splicing and polyadenylation may be distinct when skipped terminal exons are utilized. No characteristics of introns can explain the usage of skipped exon poly(A) sites, indicating that other factors, such as exonic and intronic enhancer/silencers, may govern their usage (Goren et al. 2006; Wang et al. 2006), and their regulation may be similar to exon skipping events. However, skipped exon poly(A) sites tend to be associated with strong poly(A) signals. Conceivably, strong polyadenylation signals are utilized by skipped terminal exons to overcome splicing during transcription and mRNA processing. In this sense, it is to be explored whether machine learning models can be constructed to accurately predict intronic polyadenylation events (Cheng et al. 2006), and to distinguish composite exon poly(A) sites from skipped exon poly(A) sites using both intron and poly(A) site parameters.

Methods

Poly(A) site mapping

Human, mouse, and rat poly(A) sites were identified as described in Tian et al. (2005). Briefly, human, mouse, and rat cDNA/EST (NCBI, August 2005 versions) sequences were aligned with their genomes (UCSC; hg17 for human, mm5 for mouse, and rn3 for rat) by BLAT (Kent 2002). Dangling poly(A) tails (>8 nt) of the aligned cDNA/ESTs were used to find poly(A) sites. Sites located in A-rich regions, i.e., six or more consecutive As or seven or more As in a 10-nt window in the –10 to +10 nt region surrounding the site were considered as internal priming candidates and were not used in this study. cDNA/ESTs without poly(A) tails were also used if their 3′ ends were located within 24 nt from a site supported by poly(A/T)-tailed cDNA/ESTs. The orientation of a cDNA/EST on the genome was inferred by its splicing sites as previously described (Tian et al. 2005). Poly(A) sites located in introns of NCBI RefSeq (August 2005 versions) or UCSC KnownGene (March 2006 versions) sequences were identified. We required at least 32 nt overlap between the cDNA/ESTs supporting the intronic poly(A) sites and the RefSeq and KnownGene sequences. All poly(A) sites can be queried in the PolyA_DB 2 database (Lee et al. 2007).

Identification of orthologous poly(A) sites

Orthologous poly(A) sites were identified by using UCSC human versus mouse (hg17 vs. mm5), mouse versus human (mm5 vs. hg17), human versus rat (hg17 vs. rn3), and rat versus human

(rn3 vs. hg17) whole genome alignments (axtNet files) (Schwartz et al. 2003). A pair of human and mouse/rat poly(A) sites were considered orthologous when (1) the human and mouse/rat sites are located within 24 nt in the human and mouse/rat genome alignment; and (2) they are nearest to one another in a reciprocal manner, i.e., the mouse/rat poly(A) site is the nearest one to the human poly(A) site using hg17 versus mm5 or hg17 versus rn3, and the human one is the nearest to the mouse/rat one using mm5 versus hg17 or rn3 versus hg17.

Splice site scores

We used 5′ss and 3′ss of 181,669 GT-AG type introns of human RefSeq and KnownGene sequences to build PSSMs. For 5′ss, we used –3 to +6 nt surrounding the 5′ss, with 3 nt in the exon and 6 nt in the intron; for 3′ss, we used –22 to +2 nt surrounding the 3′ss, with 22 nt in the intron and 2 nt in the exon. Each entry in a PSSM was calculated by $M_{ij} = \log_2(f_{ij}/g_i)$, where f_{ij} is the frequency of nucleotide i at position j, and g_i is the frequency of nucleotide i in the whole region. The score for each individual sequence was calculated by $S = \sum_j m_{i,j}$, where $m_{i,j}$ is the score of nucleotide i at position j in the PSSM. 5′ss sequences were also scored by their ability to hybridize with U1 snRNA. We used the sequence 5′-ACTTACCTG of U1 snRNA to form duplex structures with 5′ss sequences using the RNAduplex function of ViennaRNA (Hofacker 2003).

PAS hexamers and poly(A) region scores

The −40 to −1 nt region (poly(A) site was set at position 0) was used to identify PAS hexamers, including AAUAAA, AUUAAA, and 11 single nucleotide variants, including UAUAAA, AGUAAA, AAGAAA, AAUAUA, AAUACA, CAUAAA, GAUAAA, AAUGAA, UUUAAA, ACUAAA, and AAUAGA (Beaudoing et al. 2000; Tian et al. 2005). We used all human poly(A) regions in the PolyA_DB 2 database (Lee et al. 2007) to generate PSSMs for the –40 to −3 nt and +3 to +40 nt regions. The –2 to +2 nt region was not used, as this region could not be unambiguously resolved by alignment tools (data not shown). The score for each poly(A) region was calculated as described above.

Statistical analyses

Wilcoxon rank sum tests and mKS tests were carried out in program R (http://www.r-project.org). *t*-tests were carried out in Microsoft Excel. We followed what was described in Mootha et al. (2003) for the mKS test. Briefly, given a set of values N containing n entries and another set M containing m entries, the following method was used to assess whether values in M were significantly higher or lower than those in N. N and M were first combined, and the combined set (M+N) was then ordered from high value to low value and a running sum was computed across all entries starting at the highest value. A value of $v1$ was added to the running sum if the entry was from N, and otherwise $v2$ was added, where $v1 = \sqrt{(m/n)}$, and $v2 = -\sqrt{(n/m)}$. Thus, the overall sum was zero. The maximum and minimum values, O_{max} and O_{min} respectively, of the running sum were used as empirical statistics and can be considered as observed values. To obtain their significance, we randomly selected m entries from (M+N), and calculated the maximum and minimum values, E_{max} and E_{min} respectively, which were considered as expected values. The process was repeated 1000 times. The probability for rejecting the null hypothesis that M contains larger values than N was the fraction of 1000 E_{max} that were higher than O_{max}. The probability for rejecting the null hypothesis that M contains smaller values than N was the fraction of 1000 E_{min} that were smaller than O_{min}. These probabilities were called E-values in this study.

Intron distribution map

Intron distribution maps were used to analyze the distribution of a subset of introns (M) in the all intron set (N) with respect to 5′ss score and intron size. Maps were constructed first by creating a 20 × 20 grid with intron size as x-axis and 5′ss score as y-axis. For each axis, we used 20 quantile values of all introns to define the range for each cell. Introns in M and N were put into cells according to their 5′ss score and intron size values. The number of introns from M in each cell was considered as observed value O_{ij}. The expected value for each cell was calculated by $E_{ij} = m^*c_{ij}/n$, where m is the number of introns in M, n is the number of introns in N, and c_{ij} is the number of introns from N in cell C_{ij}. The ratios of observed value to expected value (O_{ij}/E_{ij}) were plotted in a heatmap. The sums of O_{ij} in both rows ($\sum_{i=1}^{20} Oij$) and columns ($\sum_{j=1}^{20} Oij$) were also calculated and presented in heatmaps.

Cell lines and QPCR

Human K562 cells were maintained in the Dulbecco's Modified Eagles Medium (DMEM) supplemented with 10% fetal bovine serum (FBS). Human HL60 cells were maintained in RPMI-1640 with 10% FBS. All media were also supplemented with 50 U/mL penicillin and 50 µg/mL streptomycin. Total cellular mRNAs were extracted by RNeasy Kit (Qiagen) according to manufacturer's protocol. mRNAs were reverse-transcribed by M-MLV reverse transcriptase using oligo-dT_{15}. Real-time QPCR was carried out using the 7500 Real-time PCR system (Applied Biosystems) with Syber-Green I as dye. Primers for different genes are provided in Supplemental Figure 4. All primers were obtained from the Molecular Resource Facility at UMDNJ. QPCR values of all transcripts were normalized to those of Cyclophilin A (*CYPH*) transcripts from the same cell. Primers for *CYPH* were 5′-ATGGTCAACCC CACCGTGT and 5′-AATCCTTTCTCTCCAGTGCTCAG. After 40 cycles of amplification, products were run on a 2% agarose gel, and stained with ethidium bromide. All samples were run in duplicate, and all experiments were conducted twice.

Acknowledgments

We thank James Manley for helpful discussion; Carolyn Suzuki, Michael Tsai, Tsafi Pe'ery, Carol S. Lutz, and Jun Hu for critical reading of the manuscript; and Anita Antes for assistance with cell culture. B.T. was supported by the Foundation of the University of Medicine and Dentistry of New Jersey.

References

Akoulitchev, S., Chuikov, S., and Reinberg, D. 2000. TFIIH is negatively regulated by cdk8-containing mediator complexes. *Nature* **407:** 102–106.

Awasthi, S. and Alwine, J.C. 2003. Association of polyadenylation cleavage factor I with U1 snRNP. *RNA* **9:** 1400–1409.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10:** 1001–1010.

Bruce, S.R., Dingle, R.W., and Peterson, M.L. 2003. B-cell and plasma-cell splicing differences: A potential role in regulated immunoglobulin RNA processing. *RNA* **9:** 1264–1273.

Burge, C.B., Tuschl, T., and Sharp, P.A. 1999. Splicing of precursors to mRNAs by the spliceosomes. In *The RNA world* (eds. R.F. Gesteland et al.), pp. 525–560. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Cheng, Y., Miura, R.M., and Tian, B. 2006. Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics* **22:** 2320–2325.

Colgan, D.F. and Manley, J.L. 1997. Mechanism and regulation of mRNA polyadenylation. *Genes & Dev.* **11:** 2755–2766.

Cooke, C., Hans, H., and Alwine, J.C. 1999. Utilization of splicing elements and polyadenylation signal elements in the coupling of polyadenylation and last-intron removal. *Mol. Cell. Biol.*

19: 4971–4979.
Edmonds, M. 2002. A history of poly A sequences: From formation to factors to function. *Prog. Nucleic Acid Res. Mol. Biol.* **71:** 285–389.
Edwalds-Gilbert, G. and Milcarek, C. 1995. Regulation of poly(A) site use during mouse B-cell development involves a change in the binding of a general polyadenylation factor in a B-cell stage-specific manner. *Mol. Cell. Biol.* **15:** 6420–6429.
Edwalds-Gilbert, G., Veraldi, K.L., and Milcarek, C. 1997. Alternative poly(A) site selection in complex transcription units: Means to an end? *Nucleic Acids Res.* **25:** 2547–2561.
Fan, H., Sakuraba, K., Komuro, A., Kato, S., Harada, F., and Hirose, Y. 2003. PCIF1, a novel human WW domain-containing protein, interacts with the phosphorylated RNA polymerase II. *Biochem. Biophys. Res. Commun.* **301:** 378–385.
Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. 2005. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310:** 1817–1821.
Fryer, C.J., White, J.B., and Jones, K.A. 2004. Mastermind recruits CycC:CDK8 to phosphorylate the Notch ICD and coordinate activation with turnover. *Mol. Cell* **16:** 509–520.
Glusman, G., Qin, S., El-Gewely, M.R., Siegel, A.F., Roach, J.C., Hood, L., and Smit, A.F. 2006. A third approach to gene prediction suggests thousands of additional human transcribed regions. *PLoS Comput. Biol.* **2:** e18.
Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. 2006. Comparative analysis identifies exonic splicing regulatory sequences. The complex definition of enhancers and silencers. *Mol. Cell* **22:** 769–781.
Gugneja, S., Virbasius, J.V., and Scarpulla, R.C. 1995. Four structurally distinct, non-DNA-binding subunits of human nuclear respiratory factor 2 share a conserved transcriptional activation domain. *Mol. Cell. Biol.* **15:** 102–111.
Gunderson, S.I., Beyer, K., Martin, G., Keller, W., Boelens, W.C., and Mattaj, L.W. 1994. The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. *Cell* **76:** 531–541.
Gunderson, S.I., Polycarpou-Schwarz, M., and Mattaj, I.W. 1998. U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. *Mol. Cell* **1:** 255–264.
Hengartner, C.J., Myer, V.E., Liao, S.M., Wilson, C.J., Koh, S.S., and Young, R.A. 1998. Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent kinases. *Mol. Cell* **2:** 43–53.
Hirose, Y. and Manley, J.L. 2000. RNA polymerase II and the integration of nuclear events. *Genes & Dev.* **14:** 1415–1429.
Hochsmann, M., Toller, T., Giegerich, R., and Kurtz, S. 2003. Local similarity in RNA secondary structures. *Proc. IEEE Comput. Soc. Bioinform. Conf.* **2003:** 159–168.
Hofacker, I.L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31:** 3429–3431.
Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. 2006. The UCSC known genes. *Bioinformatics* **22:** 1036–1046.
Hu, J., Lutz, C.S., Wilusz, J., and Tian, B. 2005. Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation. *RNA* **11:** 1485–1493.
Jurica, M.S. and Moore, M.J. 2003. Pre-mRNA splicing: Awash in a sea of proteins. *Mol. Cell* **12:** 5–14.
Kaneko, S. and Manley, J.L. 2005. The mammalian RNA polymerase II C-terminal domain interacts with RNA to suppress transcription-coupled 3′ end formation. *Mol. Cell* **20:** 91–103.
Kent, W.J. 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* **12:** 656–664.
Khabar, K.S., Bakheet, T., and Williams, B.R. 2005. AU-rich transient response transcripts in the human genome: Expressed sequence tag clustering and gene discovery approach. *Genomics* **85:** 165–175.
Koeffler, H.P. and Golde, D.W. 1980. Human myeloid leukemia cell lines: A review. *Blood* **56:** 344–350.
Kyburz, A., Friedlein, A., Langen, H., and Keller, W. 2006. Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3′ end processing and splicing. *Mol. Cell* **23:** 195–205.
LaCava, J., Houseley, J., Saveanu, C., Petfalski, E., Thompson, E., Jacquier, A., and Tollervey, D. 2005. RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121:** 713–724.
Ladd, A.N. and Cooper, T.A. 2002. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol.* **3:** reviews0008.
Lee, J.Y., Yeh, I., Park, J.Y., and Tian, B. 2007. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.* (in press).
Liang, S. and Lutz, C.S. 2006. p54nrb is a component of the snRNP-free U1A (SF-A) complex that promotes pre-mRNA cleavage during polyadenylation. *RNA* **12:** 111–121.
Liu, Y., Kung, C., Fishburn, J., Ansari, A.Z., Shokat, K.M., and Hahn, S. 2004. Two cyclin-dependent kinases promote RNA polymerase II transcription and formation of the scaffold complex. *Mol. Cell. Biol.* **24:** 1721–1735.
Lou, H., Neugebauer, K.M., Gagel, R.F., and Berget, S.M. 1998. Regulation of alternative polyadenylation by U1 snRNPs and SRp20. *Mol. Cell. Biol.* **18:** 4977–4985.
Lutz, C.S., Murthy, K.G., Schek, N., O'Connor, J.P., Manley, J.L., and Alwine, J.C. 1996. Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro. *Genes & Dev.* **10:** 325–337.
Millevoi, S., Loulergue, C., Dettwiler, S., Karaa, S.Z., Keller, W., Antoniou, M., and Vagner, S. 2006. An interaction between U2AF 65 and CF I(m) links the splicing and 3′ end processing machineries. *EMBO J.* **25:** 4854–4864.
Modrek, B. and Lee, C. 2002. A genomic view of alternative splicing. *Nat. Genet.* **30:** 13–19.
Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. 2003. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34:** 267–273.
Moreira, A., Takagaki, Y., Brackenridge, S., Wollerton, M., Manley, J.L., and Proudfoot, N.J. 1998. The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3′ end formation by two distinct mechanisms. *Genes & Dev.* **12:** 2522–2534.
Nesic, D. and Maquat, L.E. 1994. Upstream introns influence the efficiency of final intron removal and RNA 3′-end formation. *Genes & Dev.* **8:** 363–375.
Niwa, M. and Berget, S.M. 1991. Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns. *Genes & Dev.* **5:** 2086–2095.
Pan, Z., Zhang, H., Hague, L.K., Lee, J.Y., Lutz, C.S., and Tian, B. 2006. An intronic polyadenylation site in human and mouse CstF-77 genes suggests an evolutionarily conserved regulatory mechanism. *Gene* **366:** 325–334.
Proudfoot, N.J., Furger, A., and Dye, M.J. 2002. Integrating mRNA processing with transcription. *Cell* **108:** 501–512.
Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.
Ren, S. and Rollins, B.J. 2004. Cyclin C/cdk3 promotes Rb-dependent G0 exit. *Cell* **117:** 239–251.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human–mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.
Takagaki, Y., Seipelt, R.L., Peterson, M.L., and Manley, J.L. 1996. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* **87:** 941–952.
Thanaraj, T.A., Clark, F., and Muilu, J. 2003. Conservation of human alternative splice events in mouse. *Nucleic Acids Res.* **31:** 2544–2552.
Tian, B., Hu, J., Zhang, H., and Lutz, C.S. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33:** 201–212.
Wang, Z., Xiao, X., Van Nostrand, E., and Burge, C.B. 2006. General and specific functions of exonic splicing silencers in splicing control. *Mol. Cell* **23:** 61–70.
West, S., Gromak, N., Norbury, C.J., and Proudfoot, N.J. 2006. Adenylation and exosome-mediated degradation of cotranscriptionally cleaved pre-messenger RNA in human cells. *Mol. Cell* **21:** 437–443.
Yan, J. and Marr, T.G. 2005. Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res.* **15:** 369–375.
Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T., and Burge, C.B. 2005. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl. Acad. Sci.* **102:** 2850–2855.
Zhang, H., Lee, J.Y., and Tian, B. 2005. Biased alternative polyadenylation in human tissues. *Genome Biol.* **6:** R100.
Zhao, W. and Manley, J.L. 1996. Complex alternative RNA processing generates an unexpected diversity of poly(A) polymerase isoforms. *Mol. Cell. Biol.* **16:** 2378–2386.
Zhao, J., Hyman, L., and Moore, C. 1999. Formation of mRNA 3′ ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol. Mol. Biol. Rev.* **63:** 405–445.

Received May 21, 2006; accepted in revised form November 20, 2006.