Long Conserved Fragments Upstream of Mammalian Polyadenylation Sites

Eric S. Ho†, and Samuel I. Gunderson*

Department of Molecular Biology and Biochemistry, Nelson Laboratories, Rutgers University, Piscataway, New Jersey

†Present address: Department of Molecular Genetics, Microbiology and Immunology, University of Medicine and Dentistry of New Jersey-Robert Wood Johnson Medical School, Piscataway, New Jersey

*Corresponding author: Email: gunderson@biology.rutgers.edu.

Accepted: 27 May 2011

Abstract

Polyadenylation is a cotranscriptional nuclear RNA processing event involving endonucleolytic cleavage of the nascent, emerging pre-messenger RNA (pre-mRNA) from the RNA polymerase, immediately followed by the polymerization of adenine ribonucleotides, called the poly(A) tail, to the cleaved 3′ end of the polyadenylation site (PAS). This apparently simple molecular processing step has been discovered to be connected to transcription and splicing therefore increasing its potential for regulation of gene expression. Here, through a bioinformatic analysis of *cis*-PAS–regulatory elements in mammals that includes taking advantage of multiple evolutionary time scales, we find unexpected selection pressure much further upstream, up to 200 nt, from the PAS than previously thought. Strikingly, close to 3,000 long (30–500 nt) noncoding conserved fragments (CFs) were discovered in the PAS flanking region of three remotely related mammalian species, human, mouse, and cow. When an even more remote transitional mammal, platypus, was included, still over a thousand CFs were found in the proximity of the PAS. Even though the biological function of these CFs remains unknown, their considerable sizes makes them unlikely to serve as protein recognition sites, which are typically $\leq$15 nt. By harnessing genome wide DNaseI hypersensitivity data, we have discovered that the presence of CFs correlates with chromatin accessibility. Our study is important in highlighting novel experimental targets, which may provide new understanding about the regulatory aspects of polyadenylation.

Key words: polyadenylation, *cis*-regulatory elements, bioinformatics, conserved fragments.

Introduction

The majority of vertebrate protein-coding messenger RNA precursors (pre-mRNAs) undergo required posttranscriptional modifications namely 5′ capping, splicing, and polyadenylation, in the nucleus before being exported to the cytoplasm. Collectively, these are often called posttranscriptional processing events even though these three processes are actually orchestrated cooperatively during transcription. RNA processing serves vital biological functions and is thought to facilitate diversity. Intriguingly, polyadenylation is the only pre-mRNA modification out of the three that is preserved in all domains of life (Sarkar 1997; Portnoy and Schuster 2006), that is, prokaryotes, archaea, and eukaryotes. Despite its ancient origin, polyadenylation has become more sophisticated during the course of evolution (ca. 3 billion years). Here, we use the more complex mammalian species as a model to investigate the possible regulatory elements of polyadenylation.

All eukaryotic protein-coding mRNAs are polyadenylated except histones. Polyadenylation consists of two sequential enzymatic reactions, that is, the endonucleolytic cleavage of nascent pre-mRNA emerging from the transcription complex, immediately followed by the polymerization of adenosine nucleotides to the cleaved 3′ end of the pre-mRNA molecule. The location of the endonucleolytic cleavage site, namely the polyadenylation site (PAS), is specific despite the fact that ~54% of human and ~32% of mouse genes are found to possess more than one PAS (Tian et al. 2005). The polyadenosine nucleotides polymerized at the 3′ end of the mRNA is commonly known as the poly(A) tail. The typical length of the poly(A) tail in mammals is 200–250 nt long, whereas lower organisms tend to have a shorter poly(A) tail, for example, about 70 nt in yeast, 10–20 nt in *Escherichia coli* (Karnik et al. 1987; Taljanidisz et al. 1987). Polyadenylation is a nontemplate driven process, in contrast to transcription and DNA replication. It takes place in the nucleus, however, not without exception as poly(A) tail lengthening and shortening is known to occur in the cytoplasm as best evidenced by examples from *Xenopus* oocyte maturation and early embryogenesis (Piqué et al. 2008).

It is well known that vertebrate PAS activation requires a large protein complex and two distinct sequence elements, the first being a highly conserved hexanucleotide, called the poly(A) signal located 10–30 nt upstream of the PAS. The two most prevalent forms of poly(A) signal in vertebrates are AAUAAA and AUUAAA, collectively called the canonical poly(A) signal. According to our unpublished data and that of others (Beaudoing et al. 2000; Tian et al. 2005), AAUAAA and AUUAAA are found in approximately 66% and 16% of mammalian genes, respectively. The second sequence element, called the downstream sequence element (DSE), begins ~15 nt downstream of the PAS. Unlike the poly(A) signal, it has a quite degenerate consensus sequence enriched in uracil (U) and guanine (G) but not simple $(GU)_n$ repeats as reported previously using SELEX and NMR studies (Takagaki and Manley 1997; Perez Canadillas and Varani 2003; Salisbury et al. 2006). Due to its nucleotide bias, this downstream polyadenylation element is often named the U/GU-rich region. In addition, experimental data indicated that cleavage and polyadenylation occur deterministically at a fixed location (±10 nt) between the poly(A) signal and the U/GU-rich region. A recent computational study of PAS downstream sequences from various metazoans suggested that DSEs exhibit a 5′ to 3′ transition from UG-rich to U-rich (Salisbury et al. 2006), an observation consistent with our recent work (Ho et al. 2009).

Both upstream and downstream *cis*-polyadenylation elements have been studied experimentally and bioinformatically. Bioinformatic analysis discovered the enrichment of certain hexamers upstream, up to 100 nt, in human (Hu et al. 2005) or downstream, up to 60 nt, of PASs in metazons (Salisbury et al. 2006). Through experimental studies, various functions have been attributed to other *cis*-regulatory elements including but not limited to, the inhibition of polyadenylation through a U-rich region downstream of the PAS (Zhu et al. 2007), stabilization of the polyadenylation complex by U-rich elements upstream of the PAS (Danckwardt et al. 2004, 2006, 2007; Kaufmann et al. 2004), alteration of polyadenylation by U/GU-rich elements downstream of the PAS (Liu et al. 2008), alteration of the cleavage step through proximal and distal G-rich elements downstream of the PAS (Phillips et al. 2004; Dalziel et al. 2007), and U1A autoregulation through U1A binding a polyadenylation inhibition element (PIE) (Boelens et al. 1993; Gunderson et al. 1994, 1997). So far, these studies emphasized the presence of short (≤15 nt) *cis*-regulatory elements flanking up to 100 nt upstream the PASs. Furthermore, other related studies found highly conserved regions (HCRs) in noncoding sequences (Duret et al. 1993; Duret and Bucher 1997), but no preferred locations were reported. Ten of the HCRs have been tested in a viral-based reporter gene assay (Spicher et al. 1998). Five HCRs were shown to affect mRNA stability and two affected translation efficiency. All these studies have largely ignored the possibility that highly conserved elements could be effecting 3′ end processing (Siepel et al. 2005).

Here, we undertake a *trans*-mammalian (human, mouse, cow, platypus) analysis of sequences flanking the PAS and report the following new findings: 1) there is selection pressure not only for the highly conserved and short poly(A) signal but also in the farther upstream region (up to 200 nt) of the PAS, 2) there is a prevalence of long (>30 nt up to 500 nt) conserved fragments (CFs) up to 200 nt upstream of the PAS in distant mammalian species, and 3) there is evidence for a role of CFs in chromatin architecture.

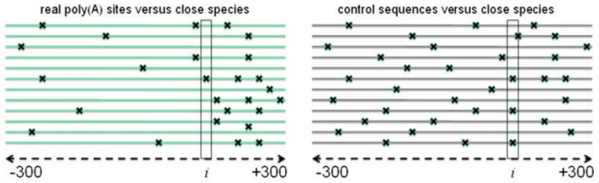
Materials and Methods

EST-Based PASs

Human and mouse expressed sequence tags (ESTs) obtained from the NCBI Refseq database (ftp://ftp.ncbi.nih.gov/refseq/release/) were used to compile a set of reliable PASs. Only polyadenylated ESTs, that is, either beginning with six or more T's or ending with six or more A's, were selected. Those polyadenylated ESTs were then mapped to the respective genome in order to make sure that the T/A-tracks of the ESTs did not originate from the genome and to determine the direction of transcription. By using this method, 17,090 and 8,779 human and mouse PASs were collected, respectively. Detailed procedures can be found in Supplement B (Supplementary Material online).

Selection Pressure in PAS Flanking Regions

We used the substitution rate of the orthologous PAS flanking regions among different organisms to measure the degree of selection pressure. Unfortunately, noncoding regions such as the 3′ untranslated regions (UTRs), which embody the PASs, are generally not conserved among remote species, making sequence alignment unfeasible. Furthermore, nucleotide sequence comparison often suffers from the homoplasy effect, that is, a given recent substitution is reverted to its ancestral form over a long evolutionary time unless the interested region is subjected to high selection pressure. To overcome this issue, the approach to harness highly similar genomes between close species genomes was adopted to examine the existence of selection pressure flanking the PAS. Three pairs of close species were used: namely human–chimpanzee, human–rhesus (rhesus macaque), and mouse–rat. The percentages of genome identity between human–chimpanzee and


real poly(A) sites versus close species
control sequences versus close species
-300
i
+300
-300
i
+300

$$mismatch_i = \frac{mismatch\_poly(A)_i}{mismatch\_control_i}$$

Fig. 1.—Mismatch ratio. Green lines on the left illustrate 600-nt real PAS flanking regions. Gray lines on the right represent control sequences. Cross symbols represent a mismatch detected between a close species pair. The mismatch ratio is computed for each position, denoted by *i*, across all sequences.

human–rhesus are 99% (Chimpanzee Sequencing and Analysis Consortium 2005) and 94% (Rhesus Macaque Genome Sequencing Consortium 2007), respectively. For the mouse–rat pair, only 8–10% of substitutions were accumulated because they diverged ~12 to 24 Ma (Rat Genome Sequencing Project Consortium 2004).

If a given genomic region is subjected to neutral selection, one would expect random substitutions to be distributed evenly along that region; otherwise, they are either localized or depleted in that region. Based upon this intuition, the following procedure was devised to examine the extent of selection pressure flanking the PAS.

1. Obtain 17,080 human and 8,799 mouse PASs from our EST-based PAS database (described in the Supplement B, Supplementary Material online).
2. Consider regions [−300, +300] (hereafter, in convention [−M, +N], M denotes nucleotides upstream and N denotes nucleotides downstream of the PAS).
3. Identify homologous PASs in close species pairs human–chimpanzee, human–rhesus, and mouse–rat using NCBI-BlastN (Camacho et al. 2009).
4. Remove sequences from genes with 3′ UTRs < 500 nt or that contain coding regions as found in some genes with multiple PASs.
5. Compile a control data set that is of the same length and number as the PAS sequences from step 1. Three control sequences were prepared from random locations in the intergenic region, open reading frame (ORF) (spliced), and intronic region with canonical poly(A) signals.
6. Examine the mismatch ratio (explained below and in fig. 1) for each position among homologous pairs in [−300, +300] of the PAS.

Using the above procedure, 16,835, 16,759, and 8,604 pairs of homologous PASs were found between human–chimpanzee, human–rhesus, and mouse–rat, respectively. For both real and control result sets, the number of mismatches was counted between each pair of close species for each position along the [−300, +300] region. Then, the two mismatch counts were combined into a ratio per position as shown in figure 1. (Note: the mismatch ratio was set to undefined during plotting if the number of mismatches in control sequences was zero. Because large numbers of PAS regions were considered, this situation was found to occur only in the first and last three positions at either end, thereby not affecting the overall analysis.) The mismatch ratio reflects the comparative substitution rate in PAS flanking regions versus control sequences. A value close to 1, >1, and <1 indicates neutral, higher, and lower substitution rates, respectively, in the PAS flanking regions as compared with the control. The choice of control sequences was based on the assumption that intergenic sequence is subjected to the least selection pressure, whereas the strongest pressure is in the ORF. The comparison of the PAS flanking region with these two extremes enables us to understand the magnitude of selection pressure. Besides the PAS flanking region, other types of genomic sequences, such as 5′ splice sites, parts of the 3′ UTR far from the PAS and introns, were included in this study in order to confirm the validity of this method. The degree and the extent of conservation of the region flanking the PAS were examined by plotting the mismatch ratio for these two pairs of close species (fig. 1).

Conserved Fragments

Four evolutionarily remote mammalian species were chosen in this analysis namely human, mouse, cow, and platypus. Gene homologous information (based on proteins) of human, mouse, and cow was obtained from the NCBI HomoloGene database (HomoloGene 2009). As the genome of platypus was completed only recently, little expression data is available to obtain its homologous information with other species. To circumvent this, human PAS flanking sequences were used to search against the platypus genome in order to identify homologous regions in platypus. Because two different ways were used to obtain the homologous information, the four mammalian species were divided into two homologous groups, namely HMC, which was composed of human, mouse, and cow and HMCP, which contained all four species.

To explore the conservation of the region that spans [−500, +500] while avoiding the influence of the ORF, genes possessing 3′ UTRs shorter than 500 nt or regions overlapping with the ORF were dropped from the analysis. Low complexity and repeat fragments were masked off from the sequences using RepeatMasker (Smit et al.). The multiple sequence alignment tool T-COFFEE (Notredame et al. 2000) was then used to align the PAS flanking regions for each orthologous group. A score value, in the range of 0–100, was returned from each alignment process, where 0 and 100 represents no and perfect alignment, respectively. Based on the alignment report, CFs were extracted from each orthologous gene group and duplicated fragments were eliminated for those genes having multiple closely spaced PASs. Altogether, 10,765 and 5,362 orthologous genes from groups HMC and HMCP were aligned, respectively. A 15-nt sliding window was used to scan the alignment, base by base. For a given gene, a "good" alignment was defined to be ≤3 mismatches (80% identity) and overlapping good windows were then stitched together to form the final CF for that gene.

For control purposes, we selected two other genomic regions as controls, one was the 500-nt region downstream of PAS and the other was the 500-nt region at the 5′ most of the 3′ UTR given its length was at least 1,000 nt. The minimum length requirement ensured the CFs, which were found in the control region, did not overlap with those in the flanking region of the PAS. We named the former control the downstream control and the latter the 5′ control. The same CF searching procedure was used to identify CFs in the two control regions in the HMC group.

Chromatin Structure of PAS

We investigated the openness of chromatin in PAS flanking regions through DNaseI hypersensitivity (HS) studies in four different human cell lines, one normal (H1-hESC) and three transformed (K562, HeLa S3, and GM12878). No equivalent mouse data were used in our analysis as there has been insufficient analysis of this type using mouse cell lines. All data were downloaded from the ENCODE project (Rosenbloom et al. 2010) hosted in the UCSC Genome Browser (Karolchik et al. 2003) website (http://genome.ucsc.edu/ENCODE/). Two independent HS data sets were obtained, one was produced by ENCODE Open Chromatin Map from the Crawford/Duke, Leib/UNC, and Lyer/UT-Austin labs (Crawford, Davis, et al. 2006; Crawford, Holt, et al. 2006; Boyle et al. 2008) and the other from the University of Washington DNaseI HS by Digital DNaseI (Sabo et al. 2004).

In all data sets mentioned above, high-resolution raw sequence data were mapped to 16,730 human PAS flanking sequences having no overlap with the ORF along regions [−500, +50] for each cell line. As a result, each sequence is associated with the number of DNaseI cut sites. However, only nonzero data were included in our analysis as zero data have an ambiguous interpretation of being either nondetectable or insensitive to DNaseI. We split these mapped results into three groups according to the size of the CF namely 11,669 without CF, 4,522 with short CF (≥30 and <200 nt), and 534 with long CF (≥200 nt).

In order to evaluate any statistically significant differences among these three groups in terms of HS mapping, we employed the resampling technique followed by a two-sample *t*-test. In each run, we performed 2,000 rounds of sampling. The average DNaseI cut sites in each round was computed by sampling equal numbers of PASs from the two compared groups; the sample size was set to 50% of the smaller group. The distributions of these 2,000 pairs of sample averages from the two compared groups were confirmed to exhibit normality using Q–Q normal plot (through qqnorm function in R) that justified the use of *t*-test for our purpose.

Results

Selection Pressure on the Farther Region 200 nt Upstream of the PAS

The mismatch ratios between flanking PASs for human–rhesus and mouse–rat pairs are illustrated in figure 2. Because human and chimpanzee diverged more recently at only ∼6 Ma as compared with ∼12 to 24 and 25 Ma for mouse–rat and human–rhesus, respectively (Rat Genome Sequencing Project Consortium 2004; Rhesus Macaque Genome Sequencing Consortium 2007), the plot of mismatch ratios for the human–chimpanzee pairs experiences substantial random fluctuations, therefore in the main text, we will present data from human–rhesus and mouse–rat pairs only, whereas the mismatch ratio plots for human–chimpanzee pair can be found in Supplement C (Supplementary Material online).

In figure 2, the blue line represents the mismatch ratio between the real PAS and the intergenic control sequence, likewise for the green line except that the control is changed to the ORF. The gray line represents the comparison between the two types of control sequences, that is, ORF/3′ UTR versus intergenic.

For the human–rhesus pair (fig. 2*A*), the mismatch ratio of real PAS sequence versus intergenic sequence (blue line) is <1 for the entire region indicating a stronger selection pressure in the PAS sequences than in the intergenic sequences. However, the experienced selection pressure is weaker than the pressure to preserve the ORF (green line) except for the region ∼30 nt upstream of the PAS, which is the preferred location of the poly(A) signal. A similar pattern is observed between the mouse–rat comparison as shown in figure 2*B*. In addition, the region upstream of the poly(A) signal not only experienced a stronger selection pressure than the region downstream but also covered a wider region as indicated by the drastic drop of the mismatch ratio after ∼50 nt from the PAS as shown in figure 2*A* and *B*. This asymmetrical pressure is not caused by any possible uneven selection pressure in the two types of control sequences along the considered region because the mismatch ratio line (gray line) for ORF versus intergenic stays at a steady level (∼0.5) across the entire region.

In order to determine the range of the selection pressure on the upstream region starting from the poly(A) signal, the first 600 nt of 3′ UTR, equivalent to the 5′ part of the 3′ UTR, was chosen as a control rather than the ORF as the PAS upstream flanking region is actually part of the 3′ UTR. Differences in selection pressure between the 5′ part of the 3′ UTR and the 3′ UTR near the PAS can thus be ascribed to the PAS. As shown in figure 2*C* and *D*, when

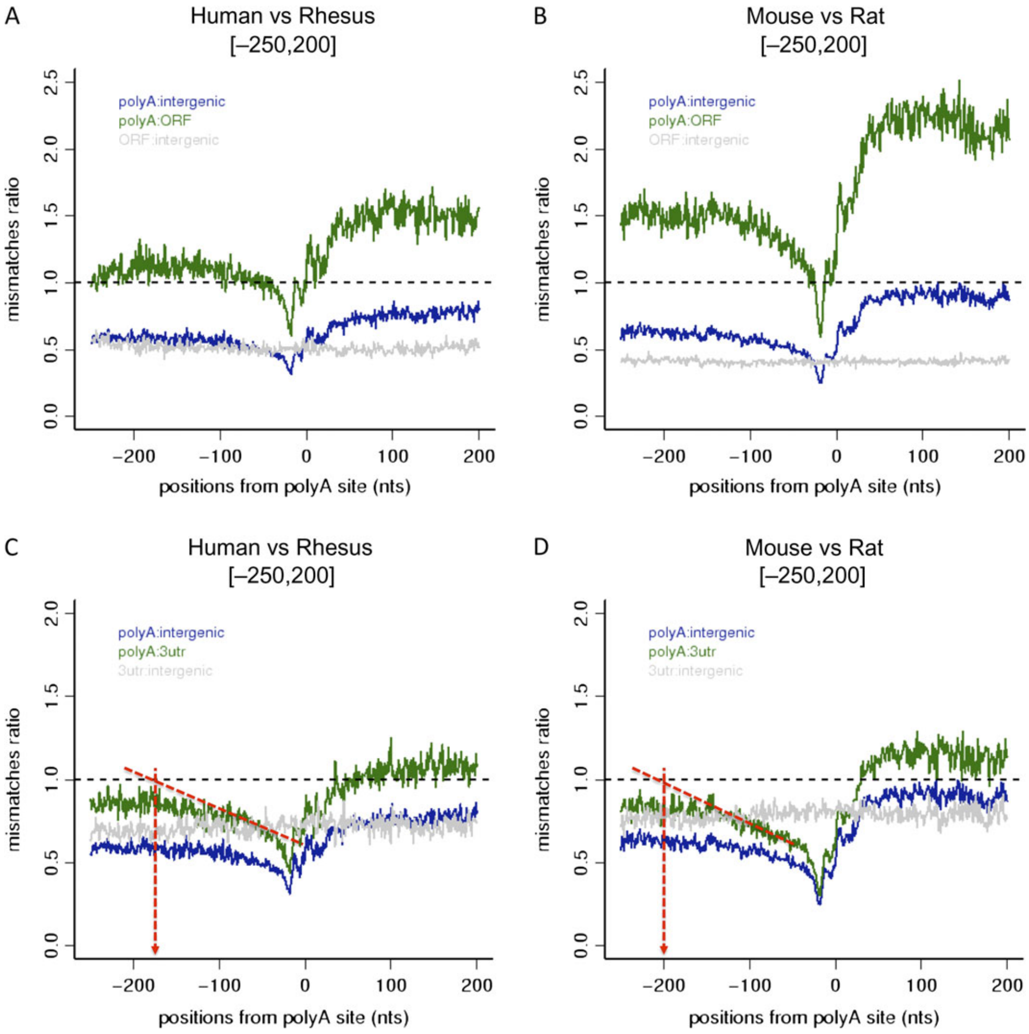


FIG. 2.—Selection pressure by close species comparison. Shown is the mismatch ratio in the PAS flanking region between close species. (*A* and *B*) The mismatch ratio variation for region [−250, +200] and (*C* and *D*) the PAS flanking region versus 3′ UTR.

the 5′ part of the 3′ UTR is taken as the control, the mismatch ratio (green) line asymptotically approaches 1 in the upstream direction and becomes flat by ~200 nt upstream of the PAS. The horizontal mismatch ratio plot (gray line in fig. 2*C* and *D*) between the 3′ UTR and the intergenic region is similar to that of the ORF versus intergenic in figure 2*A* and *B* indicating the 3′ UTR does not exhibit uneven selection pressure across the considered region. The data also indicate 3′ UTRs experience a lower substitution rate than intergenic sequences, which is in agreement with prior studies that many expression-related regulatory elements are located in the 3′ UTR (Xie et al. 2005) but with less clear positional preference.

Although the above close species analysis supports the existence of selection pressure flanking the PAS, it is prudent to do several types of control analysis to rule out alternative explanations such as artifacts inherent in the computation methods and alternative biological mechanisms. One possible artifact is the NCBI-Blast algorithm favors alignment of sequences in the middle of an alignment over sequences near the two ends. To refute this possibility, we repeated the same analysis as in figure 2*A* and *B* except the region of interest was shifted upstream or downstream by 200 nt. The pattern in these plots remains largely unchanged except it is shifted to the left or right (supplementary figs. E1–E4, Supplementary Material online). Hence, alignment bias can be ruled out in this study. To examine whether the selection pressure pattern depends on proximal repeats of PAS, only the single PAS genes were selected to produce the plot. We have found that the same pattern persists in both close species pairs (supplementary figs. E5 and E6, Supplementary Material online).

Another possible reason for the selection pressure pattern may be due to the presence of the highly conserved poly(A) signal AWUAAA (W = A or U). To examine this, a set of 600-nt long intronic sequences (17,080 from human, 8,799 from mouse) with AWUAAA positioned ~270 nt from the 5′ end was randomly sampled. We dub this the pseudo-PAS sequence set and more details on its assembly can be found in Supplement D (Supplementary Material online). Analysis of this sequence data set can be found in supplementary figs. E7 and E8 (Supplementary Material online). Results clearly showed that these sequences had no selection pressure pattern as the mismatch ratio is close to 1 when compared with intergenic sequences. Thus, the highly conserved hexanucleotide by itself failed to reproduce the same asymmetrical pattern exhibited by the real PAS flanking region. Moreover, if the distinct mismatch ratio pattern were simply caused by the highly conserved poly(A) signal, then, figure 2*A*–*D* should show a symmetric pattern too, however, nothing of that is observed

The same analysis was also applied to the 5′ splice site (5′ss) region found in the first exon as it is well documented that 5′ss recognition is facilitated by the presence of short sequence elements located immediately upstream of the 5′ss (Fairbrother et al. 2002; Wang et al. 2004). These sequence elements, commonly known as exonic splicing enhancers, are targets of serine-rich proteins (SR proteins) (Graveley 2000). Because 5′ss splicing enhancers are essential for pre-mRNA processing, they must be subjected to positive selection pressure. The mismatch ratio has the lowest value just upstream of the 5′ss and then rises abruptly immediately after the exon–intron junction in the 5′ to 3′ direction. Plots can be found in supplementary figs. E9 and E10, Supplementary Material online.

Finally, 30% and 38% of human and mouse genes were found either to overlap or be close (<1,000 nt separation) to a neighboring gene. To examine whether such close proximity or overlap with a gene influences this analysis, such genes were removed from the initial data set leaving 12,195 and 5,553 pairs of human–rhesus and mouse–rat homologous PAS regions (supplementary figs. E11 and E12, Supplementary Material online). There is no observable difference in the variation of mismatch ratio with respect to the unfiltered sequences (fig. 2*A* and *B*). Thus this battery of analysis has confirmed the presence of selection pressure on sequences within 0–200 nt upstream of the PAS.

Percentage of Alignment of PAS Flanking Regions Among Remote Mammals

The close species comparison presented above revealed the presence of selection pressure 200 nt upstream of the PAS, supporting the existence of other nonrepetitive *cis*-elements. Although previous attempts in identifying *cis*-acting PAS elements were successful in capturing the enrichment of short and fixed-size sequence motifs, such attempts largely neglected the hunt for evolutionarily conserved gene-specific elements. In order to identify the sequences preserved by this selection pressure, we switched the evolutionary time scale from close to distant mammalian species for this task. Four mammalian species were selected namely human, mouse, cow, and platypus.

The multiple alignment program T-COFFEE was used to align 10,765 and 5,362 orthologous gene groups in HMC and HMCP, respectively. The relationship between the fraction of alignment by position was plotted separately by alignment score as shown in figure 3. Two alignment score thresholds were used namely 50 and 70. Empirically, an alignment score above 50 generally indicates the presence of long fragments (>30 nt). Note that higher alignment scores are often associated with longer and/or multiple CFs.

Red and blue lines denote high and low scoring groups, respectively. Each line represents the variation in fraction of genes containing the same nucleotide as human along the flanking region of PAS. In total, 5,261 of 10,765 genes or 49% were found to achieve higher than 50 alignment score in the HMC group (fig. 3*A*). In the HMCP group, 2,668 of 5,362 genes or 50%, similar to the HMC group were found to exceed a 50 alignment score. When a more stringent threshold, 70, was adopted, the number of genes dropped to 2,160 (20%) for the HMC group and the HMCP group dropped even more to 629 genes (12%). But raising the threshold resulted in a higher fraction of alignment (compare fig. 3*A* and *C* or between *B* and *D*).

Not surprisingly, for both high and low scoring groups, the best alignment was attained at around 21 nt upstream from the PAS, which is the preferred location of the poly(A) signal. The peak occurred at 31 nt instead of 21 nt upstream in the HMCP group with threshold 70 (fig. 3*D*), the fractions of alignment between them differ by 3 percentage points only. The trend of the plot resembles that of the close species comparison method where selection pressure is asymmetrical, that is, higher in strength and range in the upstream than the downstream region. However, the degree of alignment seems to extend farther than 200 nt upstream for a subset of high scoring genes as revealed in figure 3*C* and *D*. In total, 1,080 of 2,160 orthologous HMC-group genes show a high degree of alignment but not necessarily in one continuous stretch, for up to 400 nt upstream. This observation provides an intriguing opportunity to look into the conservation of the noncoding sequence of each gene.

Identification of CFs

The two independent methods, close and remote species comparisons, presented here suggest the conservation pressure is prominent upstream rather than downstream of the PAS, thus the rest of the analysis will concentrate on the upstream region only. Based on the multiple alignment results, CFs were extracted from genes with alignment scores > 50, longer than 30 nt, and limited to one fragment per nonoverlapping gene. Altogether, 2,987 and 1,130 nonredundant conserved upstream fragments were discovered in the HMC and HMCP groups, respectively. In the two control

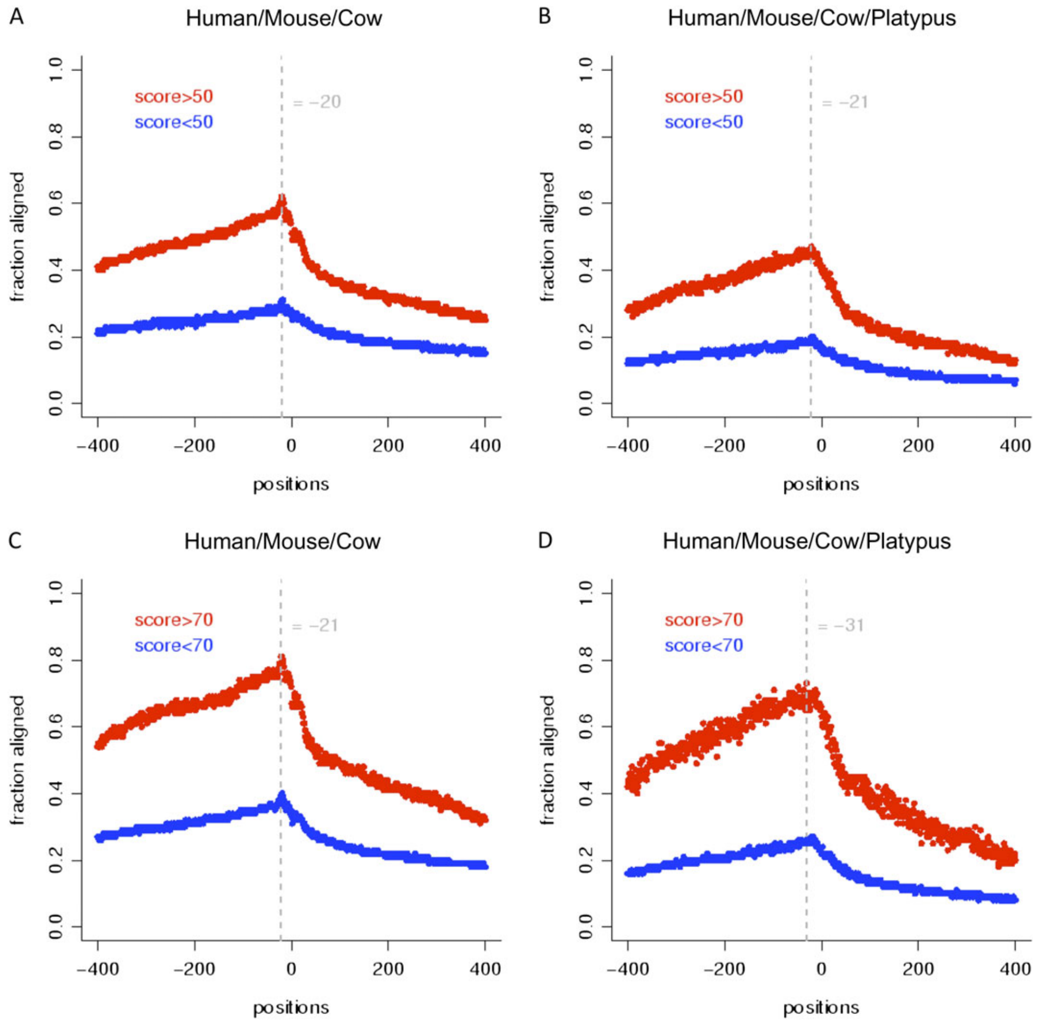


Fig. 3.—Fraction of alignment along the PAS flanking positions. Red and blue lines denote high and low scoring groups, respectively. (*A*) HMC group with threshold 50, (*B*) HMCP with threshold 50, (*C*) HMC with threshold 70, and (*D*) HMCP with threshold 70.

groups, the threshold was set to 50. CFs were found five and two times more in the upstream PAS flanking region than in the downstream, and the 5′ control regions, respectively. (More details can be found in Supplement F, Supplementary Material online.) The distribution of their lengths is shown in figure 4 where almost two-thirds of the CFs were between 30 and 100 nt long in the HMC group. Several CFs were found to be as long as 400–500 nt (fig. 4*A* and *B*). As expected, smaller numbers of CFs were found in the HMCP group, however, both groups exhibit similar distribution (fig. 4*A* and *B*).

Distance of CF From PAS

To explore the CF to PAS distance (based on 3′ end of CF), the relationship between fragment length and proximity to the PAS was examined. Figure 5 displays the distribution of the distance of these human CFs from the PAS in both the HMC and the HMCP groups. Almost half of the CFs were found to reside within 20 nt from the PAS in the HMC group (fig. 5*A*), and the remaining CFs were uniformly distributed along the upstream region, suggesting there is no particular relation between the size of the CF and proximity to the PAS. A consistent picture is found in both the HMC and the HMCP groups (fig. 5*C*). Furthermore, the length of CFs that were found within 20 nt from the PAS were analyzed as shown in figure 5*B* and *D*. Their distribution closely resembles the overall distribution of CFs where the majority of them were between 30 and 100 nt long.

Examples of CFs

A sample of alignments and CFs for three genes will be illustrated namely polypyrimidine tract binding protein 2 (PTBP2), FBJ murine osteosarcoma viral oncogene homolog (FOS), and oligodendrocyte transcription factor 1 (OLIG1). These three genes manifest different degrees of conservation near the PAS.

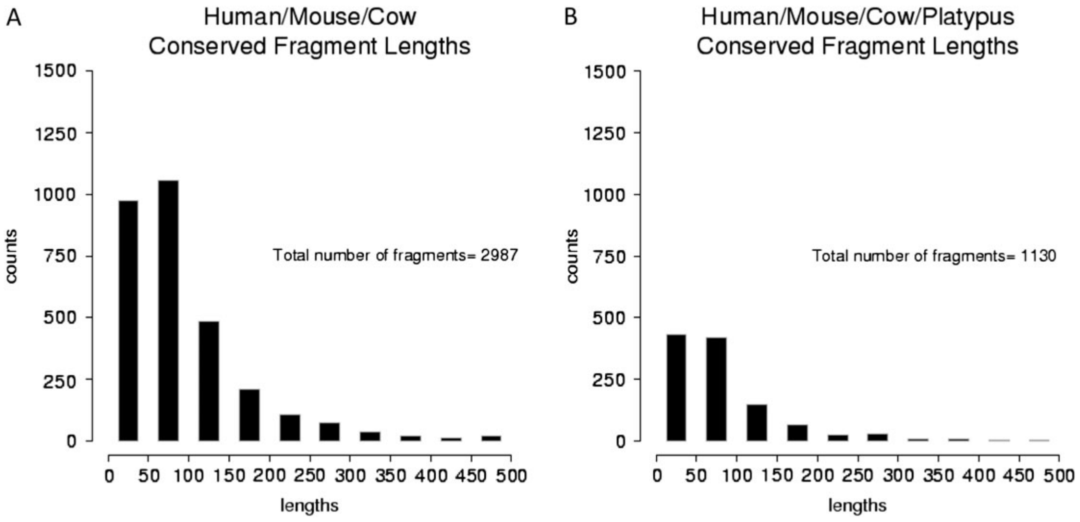


Fig. 4.—Distribution of length of human conserved upstream fragments. (*A*) In HMC group and (*B*) in HMCP group.

All alignments can be found in Supplement G (Supplementary Material online). PTBP2 and FOS are extreme examples as they contain 400 to nearly 500-nt long CFs starting from the PAS in the 5′ direction. PTBP2 is reported to control the assembly of other splicing regulatory proteins and binds to intronic polypyrimidine tracts during splicing. PTBP2 is similar to PTBP1 except for the fact that it is abundant mainly in brain. It is evident there is a continuous stretch of CFs among human, mouse, and cow including the poly(A) signal with the CFs being rich in A and T but not of low complexity as repeated and low complexity regions were removed before alignment. The degree of conservation is amazing, as it is even higher than the coding sequence.

Another example is FOS, which is a well-studied oncogene that regulates cell proliferation, differentiation, and transformation. The total conserved region of FOS, excluding the repeat masked fragment, is about 400 nt.

Not all CFs discovered here include the poly(A) signal like PTBP2 and FOS. For instance, a 34-nt long CF was found to locate ~100 nt upstream from the PAS. OLIG1 is a transcription factor in oligodendrocyte development (Lu et al. 2001) that plays a role in remyelination after injury (Labombarda et al. 2009). The small sample of genes discussed here is very limited, suggestive of a regulatory function yet to be discovered. Especially, the region of conservation of OLIG1 between human and mouse expands significantly. A full list of alignments of the upstream region among the four mammalian species can be found in Supplement H (Supplementary Material online).

CFs Are Gene Specific

To examine whether our collection of CFs share sequence similarity, an exhaustive pairwise comparison was performed among CFs in order to cluster them into groups by subsequence similarity. However, no significant similarity was found among them, which confirmed one previous study (Spicher et al. 1998), except for three pairs of genes namely MORF4L1/MORF4L2, RPL27AP6/RPL27A, and TUBA3C/TUBA4A. Each pair shares about a 100 nt long highly similar fragment. For these pairs, their similarity is more likely due to gene duplication rather than sharing a common regulatory binding site in the 3′ UTR because the proteins encoded by these genes also exhibit a high degree (77–97%) of identity.

Biological Function of CFs

At present, the only long CF that has been studied experimentally is that of the U1A gene. An approximately 53-nt long CF, called the PIE, is conserved among mammalian U1A genes (alignment in Supplement I, Supplementary Material online) and binds two molecules of the U1A protein upstream of PIE is a shorter (11 nt) conserved 5′ss-like sequence that was shown to bind the U1 snRNP splicing factor (Guan et al. 2007). The collective action of the PIE and 5′ss-like sequence is to repress the PAS as part of a negative autoregulatory feedback system. Thus, U1A is an example of a CF composed of smaller individual binding sites.

Chromatin Structure of CFs

In spite of the U1A example above, it seems very unlikely that most of these long CFs are target sites of RNA binding proteins as known sites are usually ≤15 nt, Hence, we speculate that they may serve a role in chromatin modeling. One approach to explore this aspect is through DNaseI HS mapping, which is an accurate method to identify genomic regulatory regions such as promoters, enhancers, and silencers upstream of transcription start sites. Two previous studies were conducted to perform genome wide mapping of DNaseI HS sites in an array of human tissues (Sabo et al. 2004; Boyle et al. 2008). However, current HS mapping studies mostly emphasized transcription regulatory elements.

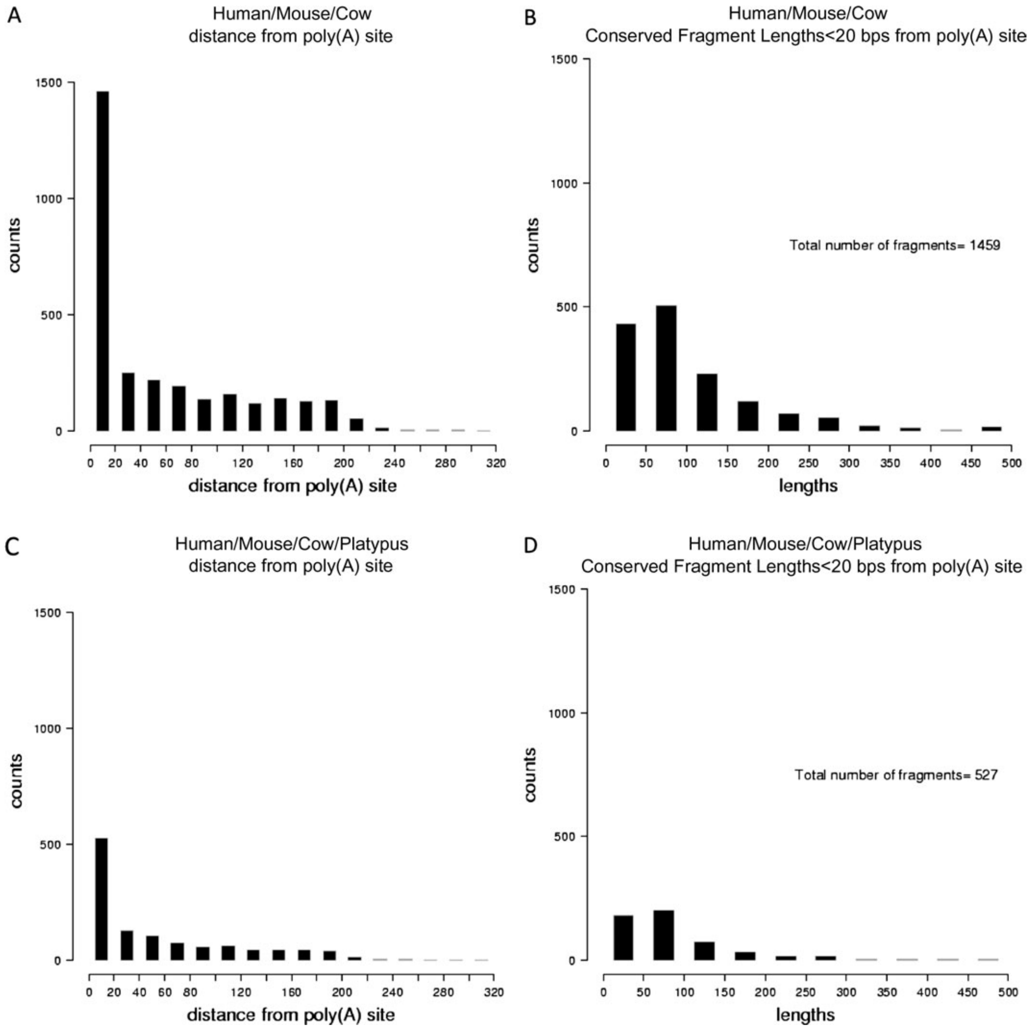


Fig. 5.—Distance of human CFs (based on 3′ end of CF) from the PAS. (*A*) Distance of CF from PAS in the HMC group, (*B*) length of CF $<$ 20 nt from the PAS in the HMC group, (*C*) distance of CF from the PAS in the HMCP group, and (*D*) length of CF $<$ 20 nt from the PAS in the HMCP group.

Only one recent bioinformatic study has reported a depletion of nucleosomes near PASs (Spies et al. 2009) where it was suggested that such depletion is unlikely to be related to expression even if its function is largely unknown. Hence, understanding of the chromatin structure at the 3′ end is still very limited. To examine this issue more carefully, we conducted a comprehensive HS mapping along the PAS flanking regions in human so as to elucidate the possible impact of CFs to the accessibility of chromatin near the PASs.

Data from two independent DNaseI HS data sets were mapped to 16,730 human PAS flanking regions [−500, +50] in four tissues. We chose these two data sets because they were performed independently at two different institutions, used different protocols, and analyzed four cell types, HeLa S3, K562, Hu ESC, and GM12878 each with distinct features, three being transformed cells and one being a primary human embryonic stem cell. Furthermore, all four cell types are from distinct tissues and ESCs pluripotent cells that are not yet committed to a differentiation pathway. These PAS flanking regions were further split into three groups namely no CF, short CF (<200 nt), and long CF (≥200 nt). We also examined other threshold lengths, such as 100 nt, but no significant differences were seen in the outcome analysis. Results in figure 6*A* and *B* show that the presence of CFs makes the region less accommodative to DNaseI endonucleoytic cleavage (red bar vs. blue/purple bars), and the differences were statistically significant according to pairwise *t*-test (*P* values were in the range of $\sim 10^{-13}$ to 10^{-16}). Furthermore, 7 (except K562(1), K562(2), and HeLa S3(1) from the University of Washington's data set) of 11 samples exhibit size dependent chromatin accessibility, indicating that PAS flanking regions

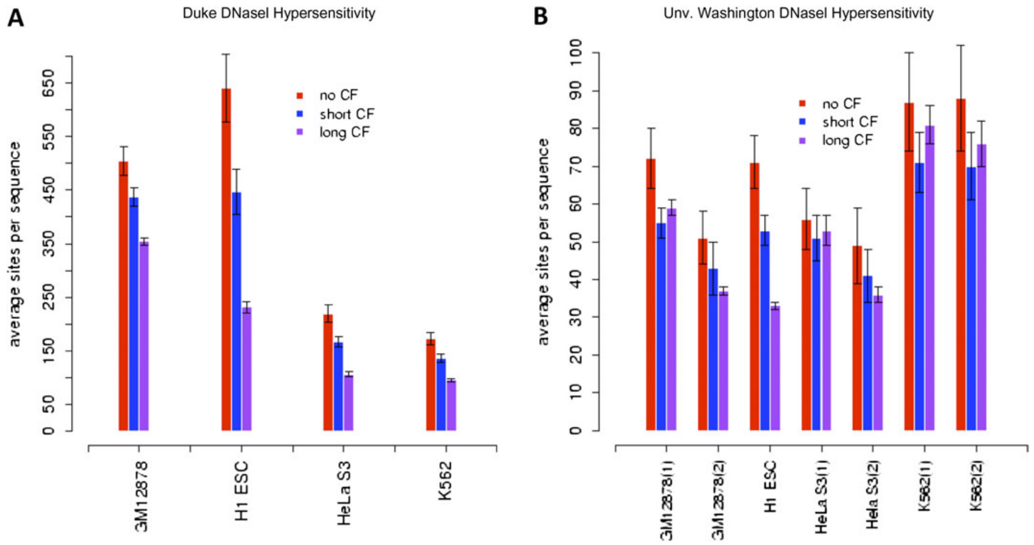


Fig. 6.—Comparison of chromatin structure of human PAS flanking regions [−500, +50] with and without CFs. Vertical axis represents the average number of DNaseI sites per sequence. "no CF" (red), "short CF" (blue), and "long CF" (purple) denote PAS sequence without CF, with CF < 200 nt, and CF > 200 nt, respectively. (*A*) DNaseI HS data obtained from Duke/University of Utah/University of North Carolina. (*B*) Similar DNaseI HS data set obtained from University of Washington. The bracketed number after the tissue labels GM12878, HeLa S3, and K562 represent replicate.

with shorter CFs were cut more frequently by DNaseI than those regions containing longer CFs (blue bar vs. purple bar in fig. 6*A* and *B*).

Discussion

Asymmetrical Selection Pressure Flanking PAS

Results show that close species comparison is capable of revealing the radiation of asymmetrical selection pressure from the poly(A) signal. Such a finding reveals that the upstream region involved in polyadenylation is longer than reported previously (Legendre and Gautheret 2003; Hu et al. 2005; Tian et al. 2005). Even though the requirement of the upstream poly(A) signal and downstream U/GU-rich region are well established, the asymmetrical selection pressure present in up to 200 nt upstream of the PAS suggests the existence of other unknown *cis*-elements in the upstream region that may involve signaling the arrival of PAS to the transcription complex.

Unlike 5′ss sequences, a sharp fall in the mismatch ratio is not observed in the upstream region (supplementary figs. E9 and E10, Supplementary Material online). Three possible explanations may account for the lack of a sharp fall. First, the upstream binding factor(s) is flexible in acting at a distance. Such action-at-a-distance is common for RNA-based regulation and often derives from secondary and tertiary folding patterns of the RNA itself. Second, the selection pressure for the region [−200, −100] is gene specific rather than basal and thus can only be seen when comparing orthologous genes as done here. Third, unlike frameshift mutations caused by mis-splicing, no severe drawback would be expected if cleavage occurs at a slightly (±5 nt) different position. According to previous studies (Legendre and Gautheret 2003; Hu et al. 2005; Tian et al. 2005), one characteristic of the upstream region is the gradual elevation of uracil composition in the 5′ to 3′ direction in the region [−100, −30]. The maximum increment is about 5% which happens immediately 5′ of the poly(A) signal. One study asserted that a stronger PAS possesses higher uracil content than a weaker one (Hu et al. 2005). However, we have found that the entire set of human and mouse 3′ UTRs, except the region 50 nt immediately after the stop codon and the last 100 nt at the 3′ end, is evenly enriched with uracil (~29%) and adenine (~27%) (Supplement J, Supplementary Material online). A similar observation has also been reported in diverse species (Graber et al. 1999). If the polyadenylation machinery solely relies on a uracil-rich signal, false signals in the 3′ UTR should appear more frequent than the real one. Even taking the two canonical poly(A) signals into account to enhance specificity, such an idea helps little to improve the recognition of PAS as poly(A) signals occur ubiquitously. Close to 3.4 and 2.2 million canonical poly(A) signals were found in human and mouse introns, respectively. Examination of the region [−500, +500] in those intronic sequences show they contain 30% A and T, which is similar to the 3′ UTR in terms of nucleotide composition. Hence, it is likely that additional gene-specific *cis*-elements are preserved by nature near the PAS.

Besides, it is intriguing to notice that even in the absence of good alignment in the low scoring plots (blue) as shown in figure 3, these plots still exhibit an asymmetrical pattern between the upstream and downstream regions around the PAS, suggesting the possibility of degenerate and/or short sequence elements in the upstream region.

Widespread Prevalence of CFs

Assuming 2,500 human genes means that 5–12% (1,130 in HMCP and 2,987 in HMC) of all mammalian genes carry a CF near the PAS. Such a large proportion can hardly be accounted for by chance only, especially as we had eliminated overlapping genes as discussed previously in the Results. As shown in figure 4, large numbers of the CFs are longer than the well-studied AU-rich, U-rich, G-rich, and C-rich regions, which regulate mRNA stability within their target proteins suggesting these CFs utilize a novel mode of regulation.

The approach discussed here complements previous work to search for overrepresented short and fixed-length *cis*-elements of polyadenylation (Graber et al. 1999; Hu et al. 2005; Hutchins et al. 2008). Previous work may be predisposed with the model that these *cis*-elements are binding targets of one or two factors. But the long CFs reported here are likely to play a role other than RNA protein recognition sites as they are much longer than known binding sites. Previous work identified long (>50 nt) and at least 70% conserved sequences in the noncoding regions among metazoans (Duret et al. 1993; Duret and Bucher 1997), and these sequences can be retrieved from the ACUTS database. But no analysis has been done to indicate their location bias near to the PAS as what we have shown here. A recent study has shown nucleosome depletion at around the [−100, +100] region (Spies et al. 2009). Double-stranded homopolymeric stretches of deoxyadenosine (10–20 nt) (Segal and Widom 2009), poly(A) signal and T-rich content are suggested for the diminishing of nucleosomes for both high and low usage PAS. Another important insight comes from the study of ultraconserved elements (UCEs). By comparing human, mouse, and rat genomes, 481 identical genomic segments longer than 200 nt were found, and they are also highly conserved in chicken and dog (Bejerano et al. 2004). Some of them function as long-range enhancers (Pennacchio et al. 2006), driving development (Woolfe et al. 2005), regulating splicing (Lareau et al. 2007; Ni et al. 2007), and epigenetic modification (Bernstein et al. 2006; Lee et al. 2006). At present, only one report demonstrated that the deletion of a subset of UCEs, postulated to be enhancers, could yield viable mice (Ahituv et al. 2007). Even though the CFs discovered here cannot be considered as ultraconserved, their conservation among distant mammalian species is so high and long that it is perplexing if they happen by pure chance during the course of evolution. Readers interested in the widespread of conservation among mammals may check the UCSC conservation track (Karolchik et al. 2003).

Possible Functions of CFs

What may be the possible roles of these CFs? It is well established that the presence of a highly conserved poly(A) signal at ~20 nt upstream and a U/GU-rich region at ~15 nt downstream from the PAS is sufficient to cause the polyadenylation machinery to cleave the nascent pre-mRNA from the transcription complex. Two efficiency elements located upstream of the PAS have been reported to promote polyadenylation. One was found in PAPOLA and PAPOLG genes (Venkataraman et al. 2005) with sequence consensus UGUAN. The other was A-rich sequence in the intronless MC4R gene (Nunes et al. 2010). Many of the CFs reported here are located less than 20 nt from the PAS (fig. 5) and they lack significant sequence similarity except for the three probably duplicated genes. These observations indicate that most genes with CFs do not regulate by common factor.

Half of the CFs were found closer than 20 nt upstream of the PAS, suggesting that they may be correlated to polyadenylation activity, otherwise there is no reason to support their biased proximity to the PAS. However, even with such positional preference, one cannot exclude the possibility that these CFs are required by other biological processes, such as mRNA stability and translation regulation as one in vitro study has shown these functionalities in 7 of the 10 selected HCRs (Spicher et al. 1998). We speculate that these CFs may serve as gene-specific promoter elements, as PAS and promoters are known to influence each other. Even though CFs longer than 100 nt are unusual, one should not overlook the rest of the 30–100 nt long CFs as multiple RNA protein recognition sites could comprise a CF as in the case of the U1A gene's CF.

CFs and Chromatin Structure

Besides serving as protein binding targets, we have also investigated the chromatin modeling role of these CFs. According to our study, CFs correlate with a more compact chromatin structure though we are unsure about their impact on expression, regulation, and efficiency of polyadenylation. Moreover, we have considered the methylation aspect of the PAS flanking region especially for the trimethylation of histone H3 Lysine 36 (H3K36me3) as it has been reported to be relevant to transcription termination (Lian et al. 2008). In addition, one similar study (Kolasinska-Zwierz et al. 2009) has shown that H3K36me3 chromatin marks are preferentially found in exonic rather than intronic sequences in *Caenorhabitis elegans* and such a methylation pattern is found to be conserved in human and mouse, indicating that H3K36me3 is of biological importance, likely, for mediating splicing. There are a number of established examples of the interactions of splicing factors with the polyadenylation complex (Lutz et al. 1996; Gunderson et al. 1998; Shi et al. 2009). However, the current embargo-free genome wide H3K36me3 data are still inadequate to reconstruct a consistent picture about CFs, methylation, and 3′ end processing. Thus, more data of this kind and CF mutagenesis studies are needed in the future in order to elucidate the interplay between chromatin structure and polyadenylation.

Supplementary Material

Supplementary figures B, C.1, C.2, E1–E12, G1–G4, I, and J are available at *Genome Biology and Evolution online* (http://gbe.oxfordjournals.org/).

Acknowledgments

This work is supported by National Institutes of Health (NIH) grant GM57286 to S.I.G. and NIH K12 GM093854-01 to E.S.H.

Literature Cited

Ahituv N, et al. 2007. Deletion of ultraconserved elements yields viable mice. PLoS Biol. 5(9):e234.

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. Genome Res. 10(7):1001–1010.

Bejerano G, et al. 2004. Ultraconserved elements in the human genome. Science. 304(5675):1321–1325.

Bernstein BE, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125(2):315–326.

Boelens WC, et al. 1993. The human U1 snRNP-specific U1A protein inhibits polyadenylation of its own pre-mRNA. Cell 72(6):881–892.

Boyle AP, et al. 2008. High-resolution mapping and characterization of open chromatin across the genome. Cell 132(2):311–322.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437(7055):69–87.

Crawford GE, et al. 2006. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat Methods. 3(7):503–509.

Crawford GE, et al. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 16(1):123–131.

Dalziel M, Nunes NM, Furger A. 2007. Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly(A) site of the intronless melanocortin receptor 1 gene are critical for efficient 3′ end processing. Mol Cell Biol. 27(5):1568–1580.

Danckwardt S, et al. 2004. The prothrombin 3′end formation signal reveals a unique architecture that is sensitive to thrombophilic gain-of-function mutations. Blood 104(2):428–435.

Danckwardt S, et al. 2006. The prothrombin 20209 C. T mutation in Jewish-Moroccan Caucasians: molecular analysis of gain-of-function of 3′ end processing. J Thromb Haemost. 4(5):1078–1085.

Danckwardt S, et al. 2007. Splicing factors stimulate polyadenylation via USEs at non-canonical 3′ end formation signals. EMBO J. 26(11):2658–2669.

Duret L, Bucher P. 1997. Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol. 7(3):399–406.

Duret L, Dorkeld F, Gautier C. 1993. Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. Nucleic Acids Res. 21(10):2315–2322.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. Science. 297(5583):1007–1013.

Graber JH, Cantor CR, Mohr SC, Smith TF. 1999. In silico detection of control signals: mRNA 3′-end-processing sequences in diverse species. Proc Natl Acad Sci U S A. 96(24):14055–14060.

Graveley BR. 2000. Sorting out the complexity of SR protein functions. RNA 6(9):1197–1211.

Guan F, et al. 2007. A bipartite U1 site represses U1A expression by synergizing with PIE to inhibit nuclear polyadenylation. RNA. 13(12):2129–2140.

Gunderson SI, et al. 1994. The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. Cell 76(3):531–541.

Gunderson SI, Polycarpou-Schwarz M, Mattaj IW. 1998. U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. Mol Cell. 1(2):255–264.

Gunderson SI, Vagner S, Polycarpou-Schwarz M, Mattaj IW. 1997. Involvement of the carboxyl terminus of vertebrate poly(A) polymerase in U1A autoregulation and in the coupling of splicing and polyadenylation. Genes Dev. 11(6):761–773.

Ho ES, Jakubowski CD, Gunderson SI. 2009. iTriplet, a rule-based nucleic acid sequence motif finder. Algorithms Mol Biol. 4:14.

HomoloGene. 2009. NCBI HomoloGene database build 63. Available from: ftp://ftp.ncbi.nih.gov/pub/HomoloGene/build63/.

Hu J, Lutz CS, Wilusz J, Tian B. 2005. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA 11(10):1485–93.

Hutchins LN, Murphy SM, Singh P, Graber JH. 2008. Position-dependent motif characterization using non-negative matrix factorization. Bioinformatics 24(23):2684–2690.

Karnik P, Taljanidisz J, Sasvari-Szekely M, Sarkar N. 1987. 3′-terminal polyadenylate sequences of Escherichia coli tryptophan synthetase alpha-subunit messenger RNA. J Mol Biol. 196(2):347–354.

Karolchik D, et al. 2003. The UCSC Genome Browser Database. Nucleic Acids Res. 31(1):51–54.

Kaufmann I, Martin G, Friedlein A, Langen H, Keller W. 2004. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. EMBO J. 23(3):616–626.

Kolasinska-Zwierz P, et al. 2009. Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet. 41(3):376–381.

Labombarda F, et al. 2009. Effects of progesterone on oligodendrocyte progenitors, oligodendrocyte transcription factors, and myelin proteins following spinal cord injury. Glia. 57(8):884–897.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature 446(7138):926–929.

Lee TI, et al. 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125(2):301–313.

Lian Z, et al. 2008. A genomic analysis of RNA polymerase II modification and chromatin architecture related to 3′ end RNA polyadenylation. Genome Res. 18(8):1224–1237.

Legendre M, Gautheret D. 2003. Sequence determinants in human polyadenylation site selection. BMC Genomics. 4(1):7.

Liu D, Fritz DT, Rogers MB, Shatkin AJ. 2008. Species-specific cis-regulatory elements in the 3′-untranslated region direct alternative polyadenylation of bone morphogenetic protein 2 mRNA. J Biol Chem. 283(42):28010–28019.

Lu QR, Cai L, Rowitch D, Cepko CL, Stiles CD. 2001. Ectopic expression of Olig1 promotes oligodendrocyte formation and reduces neuronal survival in developing mouse cortex. Nat Neurosci. 4(10):973–974.

Lutz CS, et al. 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(3):325–337.

Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, Ares M Jr. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. 21(6):708–718.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302(1):205–217.

Nunes NM, Li W, Tian B, Furger A. 2010. A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. EMBO J. 29(9):1523–1536.

Pennacchio LA, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444(7118):499–502.

Perez Canadillas JM, Varani G. 2003. Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. EMBO J. 22(11):2821–2830.

Phillips C, Pachikara N, Gunderson SI. 2004. U1A inhibits cleavage at the immunoglobulin M heavy-chain secretory poly(A) site by binding between the two downstream GU-rich regions. Mol Cell Biol. 24(14):6162–6171.

Piqué M, López JM, Foissac S, Guigó R, Méndez R. 2008. A combinatorial code for CPE-mediated translational control. Cell 132(3):434–448.

Portnoy V, Schuster G. 2006. RNA polyadenylation and degradation in different Archaea; roles of the exosome and RNase R. Nucleic Acids Res. 34(20):5923–5931.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521.

Rhesus Macaque Genome Sequencing Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316(5822):222–234.

Rosenbloom KR, et al. 2010. ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Res. 38(Database issue):D620–D625.

Sabo PJ, et al. 2004. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. Proc Natl Acad Sci U S A. 101(13):4537–4542.

Salisbury J, Hutchison KW, Graber JH. 2006. A multispecies comparison of the metazoan 3′-processing downstream elements and the CstF-64 RNA recognition motif. BMC Genomics 7:55.

Sarkar N. 1997. Polyadenylation of mRNA in prokaryotes. Annu Rev Biochem. 66:173–197.

Segal E, Widom J. 2009. Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr Opin Struct Biol. 19(1):65–71.

Shi Y, et al. 2009. Molecular architecture of the human pre-mRNA 3′ processing complex. Mol Cell. 33(3):365–376.

Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15(8):1034–1050.

Smit AFA, Hubley R, Green P. RepeatMasker at http://repeatmasker.org

Spicher A, et al. 1998. Highly conserved RNA sequences that are sensors of environmental stress. Mol Cell Biol. 18(12):7371–7382.

Spies N, Nielsen CB, Padgett RA, Burge CB. 2009. Biased chromatin signatures around polyadenylation sites and exons. Mol Cell. 36(2):245–254.

Takagaki Y, Manley JL. 1997. RNA recognition by the human polyadenylation factor CstF. Mol Cell Biol. 17(7):3907–3914.

Taljanidisz J, Karnik P, Sarkar N. 1987. Messenger ribonucleic acid for the lipoprotein of the Escherichia coli outer membrane is polyadenylated. J Mol Biol. 193(3):507–515.

Tian B, Hu J, Zhang H, Lutz CS. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. 33(1):201–212.

Venkataraman K, Brown KM, Gilmartin GM. 2005. Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. Genes Dev. 19(11):1315–1327.

Wang Z, et al. 2004. Systematic identification and analysis of exonic splicing silencers. Cell 119(6):831–845.

Woolfe A, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3(1):e7.

Xie X, et al. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature. 434(7031):338–345.

Zhu H, Zhou HL, Hasman RA, Lou H. 2007. Hu proteins regulate polyadenylation by blocking sites containing U-rich sequences. J Biol Chem. 282(4):2203–2210.

Associate editor: Bill Martin