Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing

Mainul Hoque1,4, Zhe Ji1,2,4, Dinghai Zheng1, Wenting Luo1,2, Wencheng Li1, Bei You1, Ji Yeon Park1, Ghassan Yehia3 & Bin Tian1,2

Alternative cleavage and polyadenylation (APA) generates diverse mRNA isoforms. We developed 3′ region extraction and deep sequencing (3′READS) to address mispriming issues that commonly plague poly(A) site (pA) identification, and we used the method to comprehensively map pAs in the mouse genome. Thorough annotation of gene 3′ ends revealed over 5,000 previously overlooked pAs (~8% of total) flanked by A-rich sequences, underscoring the necessity of using an accurate tool for pA mapping. About 79% of mRNA genes and 66% of long noncoding RNA genes undergo APA, but these two gene types have distinct usage patterns for pAs in introns and upstream exons. Quantitative analysis of APA isoforms by 3′READS indicated that promoter-distal pAs, regardless of intron or exon locations, become more abundant during embryonic development and cell differentiation and that upregulated isoforms have stronger pAs, suggesting global modulation of the 3′ end–processing activity in development and differentiation.

Cleavage and polyadenylation of nascent RNA is essential for maturation of almost all eukaryotic mRNAs and is coupled to termination of transcription1. The cleavage and polyadenylation site, or pA, is defined by surrounding cis elements2,3 including UGUA, AAUAAA or its variants (also known as the polyadenylation signal or PAS) and U-rich upstream elements as well as U-rich and GU-rich downstream elements. The 3′ end–processing machinery, composed of ~20 core factors and many associated factors4, is responsible for the cleavage and polyadenylation reaction.

Over half of human mRNA genes have multiple pAs, leading to mRNA isoforms that contain different coding sequences, 3′ UTRs or both5. APA can play a significant role in mRNA metabolism by controlling the length of the 3′ UTR and its encoded cis elements6,7. Dynamic regulation of 3′ UTRs by APA has been reported in different tissue types8,9; cell proliferation, differentiation and development10,11; cancer cell transformation12,13; and response to extracellular stimuli14. By contrast, pAs in introns and upstream exons have not been fully studied at the genomic level. In addition, the extent that APA regulates long noncoding RNAs (lncRNAs), which are increasingly found to play important roles in the cell15,16, is largely unknown.

Identification of pAs is typically based on the presence of a cDNA sequence corresponding to the poly(A) tail, which is generated by oligo(dT)-based reverse transcription17,18. However, oligo(dT) can also prime at internal A-rich sequences, which produce a string of sequenced As that is indistinguishable from a real poly(A) tail19. This internal priming problem is usually addressed computationally by eliminating putative pAs mapped to genomic A-rich regions, but the approach cannot guarantee full elimination of false positives caused by internal priming and can discard real pAs. In addition, some short (oligo(A)) tails are synthesized by noncanonical poly(A) polymerases, such as those involved in exosome-mediated RNA decay20. Although the length distribution of the oligo(A) tail is not yet clear, a recent study showed that the RNA species bound by yeast Trf4, a poly(A) polymerase in the TRAMP complex involved in nuclear RNA surveillance, have a median A-tail length of 5 nucleotides (nt)21. The same study found that about 10% of the population has oligo(A) tails of between 10 and 14 nt, making them potential targets for oligo(dT) priming and sequencing as false positives.

Here we present 3′READS, a high-throughput method to map pAs and quantitatively measure APA isoform expression. This method completely addresses the issue of internal priming and minimizes the complication of oligo(A) tails on degraded RNAs. Using 3′READS, we systematically mapped pAs in the mouse genome and found APA isoforms in ~79% of mRNA and ~66% of lncRNA genes. We uncovered over 5,000 real pAs located in genomic A-rich regions, which have hitherto been overlooked, underscoring the necessity of addressing the internal priming issue for comprehensive pA analysis. Using 3′READS, we studied APA in the progression of development and differentiation, and we found an overall downregulation of isoforms using promoter-proximal pAs, including those in the 3′-most exon as well as in introns and upstream exons. Expression and cis-element analyses indicated that this general trend of transcript lengthening is likely due to global modulation of 3′ end–processing activity.
RESULTS
Development of 3′READS

We reasoned that the internal priming problem could be solved by eliminating the use of oligo(dT) in reverse transcription and sequencing and that oligo(A) tailed-transcripts could be avoided by using conditions that discriminate between RNAs with long and short A tails. To this end, we developed 3′READS (Fig. 1a and Online Methods). Briefly, after fragmentation of RNA, we captured poly(A)-containing RNA fragments on magnetic beads coated with a special chimeric oligonucleotide that contained 45 thymidines (Ts) at the 5′ portion and 5 uridines (Us) at the 3′ portion, dubbed CU$_{5}$T$_{45}$. Its long sequence compared to that of a coated column commonly used for poly(A)$^+$/RNA selection. This column led to far fewer PASS reads (27%) and more reads mapped to A-rich or other regions (3.5-fold, Fig. 1c,d), supporting the effectiveness of using CU$_{5}$T$_{45}$ in distinguishing poly(A) tails from internal A-rich sequences. Notably, because reads lacking additional As after alignment are not used for pA identification in 3′READS, the issue of ‘internal priming’ essentially does not exist. In addition, because the five Us in the CU$_{5}$T$_{45}$ oligonucleotide can protect some As from digestion by RNase H through RNA:RNA base pairing, the eluted RNAs were more likely to have terminal As than those eluted from oligo(dT)$_{10–25}$-coated beads (Fig. 1c), making the resultant reads more usable for pA analysis.

To further evaluate the performance of 3′READS, we examined PASS reads mapped to ribosomal RNAs (rRNAs), small nucleolar RNAs (snoRNAs) and small nuclear RNAs (snRNAs), which are not polyadenylated. Read mapping to these RNAs would be due either to internal A-rich sequences or to the oligo(A) tail produced during their maturation or degradation. The CU$_{5}$T$_{45}$ oligonucleotide generated only 17% as many PASS reads mapped to rRNAs, snoRNAs or snRNAs as oligo(dT)$_{10–25}$ did (Fig. 1e).

Figure 1 | Mapping pAs by 3′READS. (a) Schematic of the 3′READS method. (b) Optimization of washing condition to enrich RNAs with long poly(A) tails. A60/A15 indicates the proportion of eluted A60 to A15 RNAs after RNase H treatment. (c) Reads generated using the CU$_{5}$T$_{45}$ oligonucleotide or oligo(dT)$_{10–25}$. Top, schematic showing alignment of a read to genomic DNA. The last aligned position (LAP) and the putative pA are indicated by arrows. Bottom, distribution of poly(A) site–supporting (PASS) reads used for pA identification (≥2 As immediately downstream of the LAP) or reads with <2 As immediately downstream of the LAP, where the LAP is either near a pA (≤24 nt) or further away (>24 nt). (d) Nucleotide profiles around the LAP (set to position 0) from CU$_{5}$T$_{45}$-generated reads (top) and oligo(dT)$_{10–25}$-generated reads (bottom). Reads whose LAP is >24 nt away from the pA are split into those flanked by A-rich or non–A-rich sequences, and their profiles are shown in addition to PASS read profiles. Percent of total reads are shown in each graph. An A-rich sequence is defined as ≥6 consecutive As or ≥7 As in a 10-nt window in the −10 to +10 nt region around the LAP. (e) Percentage of PASS reads assigned to rRNA, snRNA and snoRNA genes for data generated by CU$_{5}$T$_{45}$ or oligo(dT)$_{10–25}$. The ratio of the values is indicated.
We next compared 3′READS with PolyA-seq\textsuperscript{22} and PAS-Seq\textsuperscript{23}, two recent deep-sequencing methods for pA mapping that use oligo(dT) for reverse transcription. 3′READS generated fewer reads aligned to rRNAs, snRNAs or snRNAs, by a factor of \(\text{>10}\) (Supplementary Fig. 2a), indicating that 3′READS can substantially mitigate false positives caused by internal A-rich sequences and oligo(A) tails. Notably, over half of the original pAs (not filtered for internal priming) mapped by PolyA-seq were surrounded by A-rich sequences (Supplementary Fig. 2b), and the overall nucleotide profile of these sites was similar to that of the sites for non-PASS reads in this study (Fig. 1d). This indicates that internal priming is a serious issue for methods that use oligo(dT) for reverse transcription. We also compared our data with those of 3P-Seq, which does not use oligo(dT) for reverse transcription. 3′READS gave rise to 54% more usable reads for pA mapping than 3P-Seq\textsuperscript{24} (Supplementary Fig. 2c). This result is presumably due to the stringent washing condition and smaller number of sample processing steps used in 3′READS (Supplementary Discussion).

Using replicate samples, we found that 3′READS had good reproducibility, with Pearson correlation \(r \geq 0.95\) between replicates (Supplementary Fig. 3a). As expected, genes expressed at low levels had higher variations than those expressed at high levels. When different isoforms of a gene were combined, the 3′READS data were well correlated with those of RNA-seq (Pearson \(r = 0.89\)) (Supplementary Fig. 3b), indicating that 3′READS data are quantitative. Taken together, our data indicate that 3′READS is an accurate and efficient method for quantitative analysis of APA isoforms and gene expression.

Comprehensive mapping of pAs in the mouse genome

Using 3′READS, we set out to comprehensively map pAs in the mouse genome, which, despite the central role of mice as a model for mammalian biology, has poor pA annotations as compared to the human genome\textsuperscript{17}. We used RNA samples from (i) male and female whole bodies; (ii) embryos at 11, 15 and 17 days (d); (iii) brain and testis tissues at different postnatal stages; and (iv) over 11 cell lines, yielding \(\sim 54\) million PASS reads in total (Supplementary Table 2). We found \(\sim 25\%\) of the PASS reads were aligned to regions downstream of RefSeq-supported 3′ ends, indicating incomplete gene annotation by the RefSeq database (Fig. 2a,b). To address this issue, we used cDNAs and expressed sequence tags (ESTs) from the US National Center for Biotechnology Information databases, and strand-specific RNA-seq reads from the ENCODE project\textsuperscript{25} to connect the pAs mapped by 3′READS to RefSeq-defined genic regions (Online Methods). This step resulted in extension of the 3′ end for 9,612 genes with a median extension length of 307 nt (Fig. 2c). Our 3′READS data expanded the number of pAs currently annotated for mouse in the PolyA_DB 2 database by more than 2.5-fold (Fig. 2d). Consistent with our previous results\textsuperscript{5}, 42% of the pAs were associated with AAUAAA, 15% with AUUAAA and 22% with variants of A[A/U]UAAA, and 21% were not associated with any prominent PAS in the region from \(-40\) to \(-1\) nt relative to the cleavage site (Supplementary Fig. 4a).

In total, we examined 17,551 mRNA genes and 2,600 IncRNA genes in the mouse genome. When we required the relative abundance of an APA isoform to be above 5% in at least one sample, 78.5% of mRNA genes and 66.0% of IncRNA genes were found to undergo APA (Fig. 2e). On average, we found 4.0 pAs per mRNA gene and 2.6 pAs per IncRNA gene (Fig. 2e and Supplementary Fig. 4b). Data simulation indicated that our mouse pA collection for mRNA genes was near saturation with the RNA samples used in the study (Supplementary Fig. 4c). Overall, the pAs in mRNA and IncRNA genes were surrounded by similar \(cis\) elements (Supplementary Fig. 5 and Supplementary Table 3).

pAs in A-rich regions

Notably, 5,392 identified pAs (7.6% of total) were surrounded by genomic A-rich sequences, which would have been filtered out as internal priming candidates if a method employing oligo(dT) in reverse transcription had been used\textsuperscript{26}. Except for the A-rich sequences around the cleavage site, these pAs, named A-rich pAs for simplicity, had upstream A-rich and downstream U-rich peaks around the cleavage site similar to those of regular pAs (Fig. 3a). This is in stark contrast to the internal A-rich sequences that led to non-PASS reads (Fig. 1d). Notably, transcripts using A-rich pAs were generally more abundant than those using non–A-rich pAs (Fig. 3b), and A-rich pAs were more likely to be associated with AAUAAA than were non–A-rich pAs in both \(-40\)- to \(-11\)-nt and \(-10\)- to \(-1\)-nt regions (Fig. 3c).

We next wanted to validate the A-rich pAs identified by 3′READS. We reasoned that the surrounding regions of real pAs

\[\text{Figure 2} | \text{Mouse pAs identified by 3′READS. (a) Distribution of PASS reads in the mouse genome} \text{ (data from all samples are included). CDS, coding sequence; EST, expressed sequence tag.} \text{ (b) An example gene (Pde3a) showing PASS reads from 3′READS and RNA-seq reads (ENCODE project) used to assign pAs to the gene. RPM, reads per million.} \text{ (c) Histogram of the length of 3′ end extension for RefSeq mRNA genes (9,612 genes with extension >0 nt).} \text{ (d) Venn diagram comparing pAs in the PolyA_DB 2 database with those identified in this study.} \text{ (e) Percentage of mRNA or IncRNA genes considered to have APA at different isoform relative-abundance cutoffs.} \]
should exhibit stronger binding of cleavage and polyadenylation factors than random regions in the gene. To test this, we carried out cross-linking immunoprecipitation and high-throughput sequencing (CLIP-seq) using C2C12 cells and an antibody against the core cleavage and polyadenylation factor CstF64 (Supplementary Fig. 6 and Online Methods). A-rich and non-A-rich pAs had similar CstF64 bindings in their surrounding regions, and both types had much stronger (>13-fold) CstF64 association than randomly selected regions in genes (Fig. 3d). This result further confirms that the A-rich pAs we identified are genuine sites.

Alternative pAs in mRNA and lncRNA genes

We next examined alternative pAs in the mouse genome. pAs can be located in the 3′-most exon or upstream regions (Fig. 4a). pAs in the former group were further divided into the ‘single’ type when there was only one pA in the 3′-most exon or into the ‘first’, ‘middle’ or ‘last’ types according to their relative locations. pAs in upstream regions were grouped into the ‘intronic’ type, if there was RefSeq evidence indicating that the pA could be removed by splicing, or the ‘exonic’ type otherwise. As we did previously, we further separated intronic pAs into two subgroups: intronic pAs in skipped terminal exons or composite terminal exons (Fig. 4a).

Figure 3 | Comparison of pAs flanked by A-rich or non–A-rich sequences. (a) Nucleotide profile around the pAs identified in this study. (b) Relative abundance of isoforms using pAs flanked by A-rich sequences or other pAs (non–A-rich). Cumulative fraction curves are based on all genes and samples analyzed in this study. (c) Distribution of pAs sequences in the −40- to –1 nt region for A-rich and non–A-rich pAs. Table, frequencies of AAUAAA in −40- to –11 nt and −10- to –1 nt regions for A-rich and non–A-rich pAs. The difference in AAUAAA frequency in these two regions is significant \( P = 2.5 \times 10^{-14}, \chi^2 \) test. (d) Enrichment of CstF64 CLIP-seq reads around A-rich or non–A-rich pAs relative to randomly selected gene regions. Error bars, 90% confidence interval derived from 1,000× bootstrapping of data.

Figure 4 | APA of mouse mRNA and lncRNA genes. (a) Schematic of pA types; dashed lines indicate splicing. (b) Distribution of alternative pAs in different regions of mRNA or lncRNA genes. The \( P \) value for the difference in distribution between mRNA and lncRNA genes is 0 (\( \chi^2 \) test). (c) Relative abundance of APA isoforms using different types of pAs. The cumulative fraction curve is based on all genes and samples analyzed in this study. (d) Frequency of various PAS types for different types of pAs in mRNA and lncRNA genes. (e) Enrichment of CstF64 CLIP-seq reads around different types of pAs relative to randomly selected gene regions. Error bars, 90% confidence interval derived from 1,000× bootstrapping of data. (f) mRNA regions affected by alternative pAs. pAs were grouped by gene type (multi-exon or single exon) and pA location. mRNA regions were separated into 5′ UTR, coding sequence (CDS) and 3′ UTR. For intronic pAs, the mRNA region affected was defined by the exon immediately upstream of the pA. (g) Distribution of 3′ UTR length for genes without APA in the 3′ UTR (single) and genes with APA in the 3′ UTR (shortest and longest isoforms shown). (h) APA regulates conserved elements in lncRNAs. Conserved elements are derived from 30 mammalian species (Online Methods). In total, 599 and 391 lncRNA genes contain the first pA in the upstream region and 3′-most exon, respectively. Only isoforms with relative expression level >20% were analyzed.
We found that mRNA genes were more likely to have alternative pAs in the 3′-most exon, whereas lncRNA genes were more likely to have pAs in upstream regions (Fig. 4b): 70% of lncRNA genes with APA had intronic or upstream exonic pAs compared to 53% for mRNA genes with APA. This notion was further supported by relative expression levels of different APA isoforms (Fig. 4c): for mRNA genes, APA isoforms using 3′-most exon pAs were expressed at much higher levels than those using upstream-region pAs, whereas the difference between these isoform types was much smaller for lncRNA genes. The PAS usage pattern for different pA types in lncRNA genes was similar to that for mRNA genes (Fig. 4d). For example, the single and last pAs were more likely to be associated with AAUAAA than other types. Confirming the overall validity of identified pAs, all types of pAs had much stronger (>10-fold) CstF64 binding than randomly selected regions (Fig. 4e). However, pAs in different locations appeared to have distinct interactions with CstF64: single pAs in genes had the highest CstF64 binding, and pAs in composite terminal exons and in the middle of 3′-most exons had the lowest binding. Future analyses are needed to address underlying mechanisms for variation of CstF64 binding.

About one-third of all alternative pAs in multi-exon mRNA genes were in upstream regions, most of which (>97%) led to isoforms with different coding sequences (Fig. 4f). APA in the 3′-most exon resulted in about sevenfold difference in 3′ UTR length between the shortest and longest isoforms (medians of 249 nt and 1,773 nt for these isoforms, respectively; Fig. 4g). Therefore, APA can substantially affect the proteome and mRNA metabolism in the cell. To understand the significance of APA for lncRNAs, we examined pA locations relative to conserved elements of lncRNAs, assuming these elements are important for lncRNA functions. We found that ~45% of the conserved elements were downstream of the first pA when the site was located in an intron or upstream exon; ~15% were downstream when the first pA was in the 3′-most exon (Fig. 4h), suggesting that APA can play a notable role in regulation of lncRNA functions.

Regulation of APA in development and differentiation

We previously reported progressive lengthening of 3′ UTRs in mouse development and in cell differentiation by surveying about 500–1,800 genes using microarray data10. For a more systematic analysis of APA, we used 3′READS to characterize differentiation of C2C12 and 3T3-L1 cells, which represent myogenesis and adipogenesis, respectively (Fig. 5a). In addition, we compared whole embryos at 11 and 15 embryonic days. We first examined APA in the 3′-most exon (Fig. 5b). Consistent with our previous findings, genes with relatively upregulated distal pA isoforms significantly outnumbered those having relatively
upregulated proximal pA isoforms in 3T3-L1 differentiation, C2C12 differentiation and embryonic development (by 5.1-, 2.2- and 2.1-fold, respectively). In addition, significantly more APA events were consistently regulated in these processes than were oppositely regulated (Supplementary Fig. 7a). However, distinct APA events in each sample set could be clearly discerned. An example of consistent APA regulation in cell differentiation is shown in Supplementary Figure 8a. This result confirms our previous finding of general 3’ UTR lengthening in development and differentiation.

We next examined alternative pAs in upstream regions, which had not been examined by microarrays because of probe limitations. We first grouped together all isoforms using intron or upstream exon pAs for each gene and compared their change of abundance with that of isoforms using 3’-most exon pAs, which were also grouped together (Fig. 5c). Notably, we found that more genes had relatively upregulated 3’-most exon pA isoforms than relatively upregulated intron or upstream exon pA isoforms, by 5.6-, 4.0- and 4.2-fold for 3T3-L1 differentiation, C2C12 differentiation and embryonic development, respectively. Like APA in the 3’-most exon, both commonly and distinctly regulated APA events in these sample sets could be identified (Supplementary Fig. 7b). An example of consistent APA regulation in cell differentiation is shown in Supplementary Figure 8b. Together with the observations of APA in 3’-most exons, this result indicates that isoforms using promoter-distal pAs are generally upregulated in development and differentiation, regardless of intron or exon locations.

We next wanted to address whether isoforms regulated in development and differentiation share features other than their pA locations. We first examined isoform abundance in the whole-body mix and cell line mix samples. Isoforms that are relatively upregulated in development and differentiation tend to have higher expression levels in these samples than isoforms that are relatively downregulated (Fig. 5d), suggesting that isoforms with strong pAs are more likely to be upregulated than those with weak pAs. Consistent with this hypothesis, we found that 5-mers known to enhance cleavage and polyadenylation were enriched for regions around the pAs of upregulated isoforms (Fig. 5e), including AAUAAA in the −40- to −1-nt region, UGUA and U-rich elements in the −100- to −41-nt region, and UGUG elements in the +1- to +100-nt downstream region. Consistently, upregulated isoforms were more likely to have AAUAAA, as compared to other PAS types, than downregulated isoforms (Supplementary Fig. 9). Thus, we conclude that pA strength is a significant parameter in determining APA regulation in development and differentiation.

DISCUSSION
Numerous deep sequencing methods for pA analysis have recently been reported22,23,24,28–32. However, most of these methods use the oligo(dT) sequence for reverse transcription, opening the possibility of internal priming. For example, more than half of the original pAs mapped by PolyA-seq are flanked by A-rich sequences (Supplementary Fig. 2b), underscoring the severity of this problem. Direct RNA sequencing using the Helicos system28 does not require reverse transcription, but this method can also be affected by internal A-rich sequences because oligo(dT) is used to fill the poly(A) tail region before sequencing.

The key issue with internal priming is that it is impossible to determine whether the unaligned As in reads come from the real poly(A) tail or the oligo(dT) sequence in the primer. We found that even under a stringent condition, such as the one used in 3READS, RNAs with internal A-rich sequences can still bind oligo(dT) (Fig. 1c). In addition, RNA fragments not from genomic A-rich regions can also bind oligo(dT) (Fig. 1d), presumably owing to oligo(A) tails. Surprisingly, these two types of RNA species can account for ~17% and ~60% of the total reads generated from CU3T45 oligonucleotide and oligo(dT)10–25, respectively. Thus, for pA mapping, it is critically important (i) not to use oligo(dT) for priming in reverse transcription and (ii) to use unaligned As in reads for quality control. On the other hand, because 3’READS is not affected by the internal priming issue, we have been able to uncover nearly 8% of all mouse pAs that are located in genomic A-rich regions; they would have been removed by normal criteria of standard filtering methods to address internal priming26.

We found a global trend of upregulation of isoforms using promoter-distal pAs in development and differentiation. The regulation of alternative pAs in the 3’-most exon is consistent with the result we previously reported using microarray data33,34 and is in line with the notion that proliferating cells generally express short 3’ UTR isoforms11. The regulation of intron and upstream exon pAs in development and differentiation, however, has never been reported mainly because of technical limitations in using microarrays. Our result indicates that both intron/upstream exon pAs and 3’-most exon pAs follow the same global trend of regulation, leading to transcript lengthening. Several mechanisms may contribute to this phenomenon. First, our cis-element analysis indicated that, compared to downregulated isoforms, upregulated isoforms tend to have pAs with enhancing elements for pA usage, such as upstream AAUAAA, UGUA, and U-rich elements and downstream GU-rich elements. This suggests that the 3’ end processing activity weakens in development and differentiation. This is consistent with our previous result showing downregulation of mRNA expression in development and differentiation for many genes involved in 3’ end processing34. Importantly, this result readily explains why not all genes have 3’ UTR lengthening in these processes: it is the strength of pA rather than its location that is the primary determinant for regulation. Second, a recent study reported a significant role for U1 small nuclear ribonucleoprotein (snRNPs) in regulating transcript length35. Modulation of the activity of U1 snRNP relative to that of 3’ end processing in development and differentiation can contribute to greater usage of promoter-distal pAs. The detailed mechanisms and biological implications of this phenomenon need to be elucidated in future.

METHODS
Methods and any associated references are available in the online version of the paper.

Accession codes. NCBI Gene Expression Omnibus GSE42398.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS
We thank Y. Zhao and A. Antes for technical assistance, V. Jin for help with 3T3-L1 differentiation, C. MacDonald (Texas Tech University) for the anti-CstF64...
in the online version of the paper.

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details are available in the online version of the paper.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2288.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Colgan, D.F. & Manley, J.L. Mechanism and regulation of mRNA polyadenylation. Genes Dev. 11, 2755–2766 (1997).
2. Proudfoot, N.J. Ending the message: poly(A) signals then and now. Genes Dev. 25, 1770–1782 (2011).
3. Tian, B. & Graber, J.H. Signals for pre-mRNA cleavage and polyadenylation. Wiley Interdiscip. Rev. RNA 3, 385–396 (2011).
4. Shi, Y. et al. Molecular architecture of the human pre-mRNA 3′ processing complex. Mol. Cell 33, 365–376 (2009).
5. Tian, B., Hu, J., Zhang, H. & Lutz, C.S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. 33, 201–212 (2005).
6. Di Giambattista, D.C., Nishida, K. & Manley, J.L. Mechanisms and consequences of alternative polyadenylation. Mol. Cell 43, 853–866 (2011).
7. Lutz, C.S. & Moreira, A. Alternative mRNA polyadenylation in eukaryotes: an effective regulator of gene expression. Wiley Interdiscip. Rev. RNA 2, 23–31 (2011).
8. Zhang, H., Lee, J.Y. & Tian, B. Biased alternative polyadenylation in human tissues. Genome Biol. 6, R100 (2005).
9. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470–476 (2008).
10. Ji, Z., Lee, J.Y., Pan, Z., Jiang, B. & Tian, B. Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc. Natl. Acad. Sci. USA 106, 7028–7033 (2009).
11. Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A. & Burge, C.B. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. Science 320, 1643–1647 (2008).
12. Mayr, C. & Bartel, D.P. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncopgenes in cancer cells. Cell 138, 673–684 (2009).
13. Singh, P. et al. Global changes in processing of mRNA 3′ untranslated regions characterize clinically distinct cancer subtypes. Cancer Res. 69, 9422–9430 (2009).
14. Flavell, S.W. et al. Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. Neuron 60, 1022–1038 (2008).
15. Chen, L.L. & Carmichael, G.G. Long noncoding RNAs in mammalian cells: what, where, and why? Wiley Interdiscip. Rev. RNA 1, 2–21 (2010).
16. Wang, K.C. & Chang, H.Y. Molecular mechanisms of long noncoding RNAs. Mol. Cell 43, 904–914 (2011).
17. Lee, J.Y., Yeh, I., Park, J.Y. & Tian, B. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. Nucleic Acids Res. 35, D165–D168 (2007).
18. Brockman, J.M. et al. PACdb: PolyA Cleavage Site and 3′-UTR Database. Bioinformatics 21, 3691–3693 (2005).
19. Nam, D.K. et al. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. Proc. Natl. Acad. Sci. USA 99, 6152–6156 (2002).
20. Schmidt, M.J. & Norbury, C.J. Polyadenylation and beyond: emerging roles for noncanonical poly(A) polymerases. Wiley Interdiscip. Rev. RNA 1, 142–151 (2010).
21. Wlotzka, W., Kudla, G., Gronemann, S. & Tollner-De, D. The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. EMBO J. 30, 1790–1803 (2011).
22. Derti, A. et al. A quantitative atlas of polyadenylation in five mammals. Genome Res. 22, 1173–1183 (2012).
23. Shepard, P.J. et al. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA 17, 761–772 (2011).
24. Jan, C.H., Friedman, R.C., Ruby, J.G. & Bartel, D.P. Formation, regulation and evolution of Caenorhabditis elegans 3′UTRs. Nature 469, 97–101 (2011).
25. ENCODE Project Consortium. A user’s guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9, e1001046 (2011).
26. Lee, J.Y., Park, J.Y. & Tian, B. Identification of mRNA polyadenylation sites in genomes using cDNA sequences, expressed sequence tags, and Trace. Methods Mol. Biol. 419, 23–37 (2008).
27. Tian, B., Pan, Z. & Lee, J.Y. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. Genome Res. 17, 156–165 (2007).
28. Ozsolak, F. et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell 143, 1018–1029 (2010).
29. Mangone, M. et al. The landscape of C. elegans 3′ UTRs. Science 329, 432–435 (2010).
30. Fu, Y. et al. Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res. 21, 741–747 (2011).
31. Fox-Walsh, K., Davis-Turak, J., Zhou, Y., Li, H. & Fu, X.D. A multiplex RNA-seq strategy to profile poly(A+) RNA: application to analysis of transcription response and 3′ end formation. Genomics 98, 266–271 (2011).
32. Yoon, O.K. & Brem, R.B. Noncanonical transcript forms in yeast and their regulation during environmental stress. RNA 16, 1256–1267 (2010).
33. Ji, Z. et al. Transcriptional activity regulates alternative cleavage and polyadenylation. Mol. Syst. Biol. 7, 534 (2011).
34. Ji, Z. & Tian, B. Reprogramming of 3′ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. PLoS ONE 4, e8419 (2009).
35. Berg, M.G. et al. U1 snRNP determines mRNA length and regulates isofrom expression. Cell 150, 53–64 (2012).
Online Methods

Cell culture, tissue harvest and RNA samples. Mouse cell lines 2B75, CMT93, B16, F9 and C2c12 were cultured in DMEM with 10% FBS (FBS) and NIH3T3, 3T3-L1 and MC3T3-E1 cells were cultured in DMEM with 10% FCS (FCS). Differentiation of C2c12 and 3T3-L1 cells was carried out as previously described. Differentiated C2c12 and 3T3-L1 cells correspond to 4 d and 8 d after initiation of differentiation, respectively. Mouse whole-body tissue RNA sample was purchased from SABiosciences, and cell line mix 1 sample was purchased from Agilent. All mouse embryos and pups used in this study were derived from mating of FVB females and males. To obtain embryos, pregnant females were sacrificed by CO2 asphyxiation at 11, 15 and 17 d of pregnancy. Embryos were carefully dissected free of decidual and extraembryonic tissues. Postnatal pups were sacrificed at 3 weeks, 6 weeks and 9 weeks after birth. The whole brain and testes were cleared from tunica albuginea, and the seminiferous tubules were removed. All tissue samples were flash frozen in liquid nitrogen.

All animal work was conducted according to a protocol approved by the Institutional Animal Care and Use Committee (IACUC) at UMDNJ-New Jersey Medical School. Total RNA from cells and tissues was isolated using Trizol (Invitrogen) or the Qiagen kit. Binding of RNA with CU5T45 oligo–coated beads was carried out by washing with the low-salt buffer (10 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA) and the cleavage site as immediately downstream of the LAP. For 3′ end (reverse sequencing). Adaptor sequences and primer sequences are listed in Supplementary Table 1. cDNA libraries were sequenced on an Illumina Genome Analyzer GAIIx (72 nt, single read). RNA-seq was carried out using essentially the same protocol except that (i) the 3′ end region extraction step using the CU5T45 oligo was omitted and (ii) fragmented RNA was dephosphorylated at the 3′ end with shrimp alkaline phosphatase (Roche) before ligation with adaptors.

ClIP-seq. Our CLIP-seq method was largely based on the protocol used by Wang et al., with some minor modifications. Briefly, three 15-cm dishes of C2c12 cells were UV irradiated using Stratagene (Stratagene) at 254 nm with 2,000 × 100 µJ/cm² and then lysed on ice. The lysate was treated with 20 U/ml RNase T1 (Fermentas) at 25 °C for 10 min, which was followed by centrifugation at 14,000 g for 5 min at 4 °C. The supernatant was incubated for 2 h at 4 °C with mouse anti-CstF64 monoclonal antibody (gift from C. MacDonald, Texas Tech University) conjugated to magnetic Protein G Dynabeads (Invitrogen). Coimmunoprecipitated RNA fragments were dephosphorylated by calf intestinal alkaline phosphatase (NEB) and end labeled with [γ-32P]ATP by T4 polymerase (NEB). The RNA-protein complex was then resolved with bis-Tris–buffered SDS-PAGE and transferred to a nitrocellulose membrane, which was then exposed to X-ray film. The 85- to 150-kDa region was cut out, and the RNA from the cut-out membrane was isolated. cDNA library preparation was based on the protocol for Illumina’s Small RNA v1.5 kit. The libraries were sequenced on an Illumina Genome Analyzer GAIIx.

Identification of pA. For forward sequencing, the reads were aligned to the reference genome (mm9) and exon-exon junction database by Bowtie18, with the first 25 nt as seed and up to 2 mismatches allowed. The exon-exon junction database contained all possible exon-exon junction sequences in the genome9. Aligned reads were scored from 5′ to 3′ using the scheme: +1 for match and −2 for mismatch. The position in a read with the maximum score was considered as the last aligned position (LAP). The best hit for each read was chosen and was considered uniquely mapped if its score was greater than the second-best hit by at least 5. If a read contained ≥2 nongenomic As immediately after the LAP, we considered the read as a poly(A) site–supporting (PASS) read and the cleavage site as immediately downstream of the LAP. For data from reverse sequencing, we first trimmed the 5′ region of read, including the first four random nucleotides and subsequent continuous Ts. We then aligned the reads to the reference genome and exon-exon junction database by Bowtie using the first 36 nt, allowing up to 2 mismatches. For uniquely aligned reads, we compared the trimmed Ts with reference genome and exon-exon junction sequences. The reads with at least two nongenomic Ts were considered PASS reads. Because each pA can have multiple cleavage positions in a small window2, we merged cleavage positions into pAs: we first clustered together cleavage positions located within 24 nt from one another. If a cluster size was ≤24 nt, the position with the greatest number of PASS reads was used as the representative position for the pA. If a cluster was >24 nt, we first identified the cleavage site with the greatest number of PASS reads and reclustered reads located >24 nt from the position. This process was repeated until all pAs in the cluster were defined. To reduce false positives, we required a real pA to have (i) PASS

3′READS. Total RNA was subjected to one round of poly(A) selection using the Poly(A)Purist MAG kit (Ambion) according to the manufacturer’s protocol, which was followed by fragmentation using Ambion’s RNA fragmentation kit at 70 °C for 5 min. Poly(A)-containing RNA fragments were isolated using the CU5T45 oligo (Sigma), which was bound to the MyOne streptavidin C1 beads (Invitrogen) through biotin at its 5′ end. The oligo(dT)10–25-coated beads were from the Poly(A)Purist MAG kit. Binding of RNA with CU5T45 oligo–coated beads was carried out at room temperature for 1 h in the binding buffer (10 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA) and was followed by washing with the low-salt buffer (10 mM Tris-HCl, pH 7.5, 1 mM NaCl, 1 mM EDTA, 10% formamide). RNA bound to the CU5T45 oligo was digested with RNase H (5 U in 50 µl reaction volume) at 37 °C for 1 h, which also eluted RNA from the beads. Eluted RNA fragments were purified by phenol:chloroform extraction and ethanol precipitation followed by phosphorylation of the 5′ end with T4 kinase (NEB). Phosphorylated RNA was then purified by the RNeasy kit (Qiagen) and was sequentially ligated to a 5′-adenylated 3′ adaptor with the truncated T4 RNA ligase II (Bioo Scientific) and to a 5′ adaptor with T4 RNA ligase I (NEB). The resultant RNA was reverse transcribed to cDNA with Superscript III (Invitrogen), and the cDNA was amplified by 12 cycles of PCR with Phusion high-fidelity polymerase (NEB).

We designed adaptor sequences so that the RNA fragments could be sequenced from the 5′ end (forward sequencing) or from the 3′ end (reverse sequencing). Adaptor sequences and primer sequences are listed in Supplementary Table 1. cDNA libraries were sequenced on an Illumina Genome Analyzer GAIIx (72 nt, single read). RNA-seq was carried out using essentially the same protocol except that (i) the 3′ end region extraction step using the CU5T45 oligo was omitted and (ii) fragmented RNA was dephosphorylated at the 3′ end with shrimp alkaline phosphatase (Roche) before ligation with adaptors.
reads from more than one sample and (ii) ≥2 distinct PASS reads (defined by the number of As and the four random Ns) and ≥5% of all PASS reads for the same gene in at least one sample.

Extension of the 3′ end of genes. We used cDNA and EST data from NCBI and directional paired-end RNA-seq data from the ENCODE project\textsuperscript{25} to extend the 3′ ends defined by RefSeq. An extended region is between the 3′ end defined by RefSeq and pAs mapped by 3′READS and is covered by cDNA/EST sequences or RNA-seq reads without a gap greater than 40 nt. We also required that the 3′ UTR extension not exceed the transcription start site of the downstream gene. For genes located in an intron of another gene, the 3′ end extension was required not to go beyond the 3′ SS of the intron.

pAs flanked by A-rich sequences. An A-rich sequence around a pA was defined as ≥6 consecutive As or ≥7 As in a 10-nt window in the −10- to +10-nt region around the pA. pAs associated with A-rich sequences are typically filtered because they can be derived from internal priming when a primer containing oligo(dT) is used in reverse transcription\textsuperscript{17}.

APA analysis. The expression level of each APA isoform was indicated by the reads per million (RPM) value, which was calculated as the total number of PASS reads per million total uniquely mapped PASS reads for the sample. For analysis of APA regulation, Fisher’s exact test was used to examine whether the abundance of an APA isoform compared to that of other isoforms was significantly different between two samples.

IncRNA genes. IncRNAs were based on noncoding genes annotated in the RefSeq and Ensemble databases, excluding rRNAs, microRNAs, snoRNAs, snRNAs, tRNAs and those overlapping with mRNA genes on the same strand. We required IncRNAs to be longer than 200 nt. Conserved elements were obtained from the UCSC table browser (Euarchontoglires Conserved Elements for mm9)\textsuperscript{39} and were mapped to exonic regions of IncRNAs.

Identification of PAS. As previously described\textsuperscript{5}, we first selected the 6-mer with the highest occurrence in the −40- to −1-nt region upstream of all pAs. Once a 6-mer (PAS) was identified, all associated pAs were removed and the remaining pAs were searched for the next most prominent 6-mer. This process was repeated until the top ten most prominent PAS 6-mers were identified.

Cis-element analysis. We studied cis elements in four regions around the pA, i.e., −100 to −41 nt, −40 to −1 nt, +1 to +40 nt and +41 to +100 nt. As previously described\textsuperscript{40}, for each region, we calculated Z_{oe} for each 6-mer the difference between observed (o) and expected (e) occurrence frequencies: $Z_{oe} = (N_o(H) - N_e(H))/\text{SD}_{oe}(H)$, where $N_o(H)$ is the observed occurrence of 6-mer H, $N_e(H)$ is the expected occurrence based on the first-order Markov Chain model of the region, and $\text{SD}_{oe}(H)$ is the s.d. of $N_o(H) - N_e(H)$\textsuperscript{40}.

CLIP-seq data analysis. CLIP-seq reads were aligned to the mouse genome (mm9) using the program Novoalign (http://www.novocraft.com/). Identical reads in a sample were counted only once. The reads with deletion(s), which are caused by skipping of reverse transcriptase at the UV cross-linked nucleotide\textsuperscript{41}, were used for analysis, and the genome location corresponding to the deletion was used to indicate the read location. The enrichment score was the ratio of read density within 40 nt around the pA to that in randomly selected regions of the same size in the same gene set. A bootstrap resampling method was used to get the 90% confidence interval as described before\textsuperscript{33}. The Z score for 5-mers was calculated by $Z = N(P)/\text{SD}(P)$, where $N(P)$ is the occurrence of 5-mer P, and $\text{SD}(P)$ is the s.d. of $N(P)$.

36. Zhang, Y. et al. Adipose-specific deletion of autophagy-related gene 7 (atg7) in mice reveals a role in adipogenesis. Proc. Natl. Acad. Sci. USA 106, 19860–19865 (2009).
37. Wang, Z., Talley, J., Brieze, M., Turner, D. & Ule, J. CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. Methods 48, 287–293 (2009).
38. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).
39. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050 (2005).
40. Hu, J., Lutz, C.S., Wilusz, J. & Tian, B. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA 11, 1485–1493 (2005).
41. Zhang, C. & Darnell, R.B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat. Biotechnol. 29, 607–614 (2011).