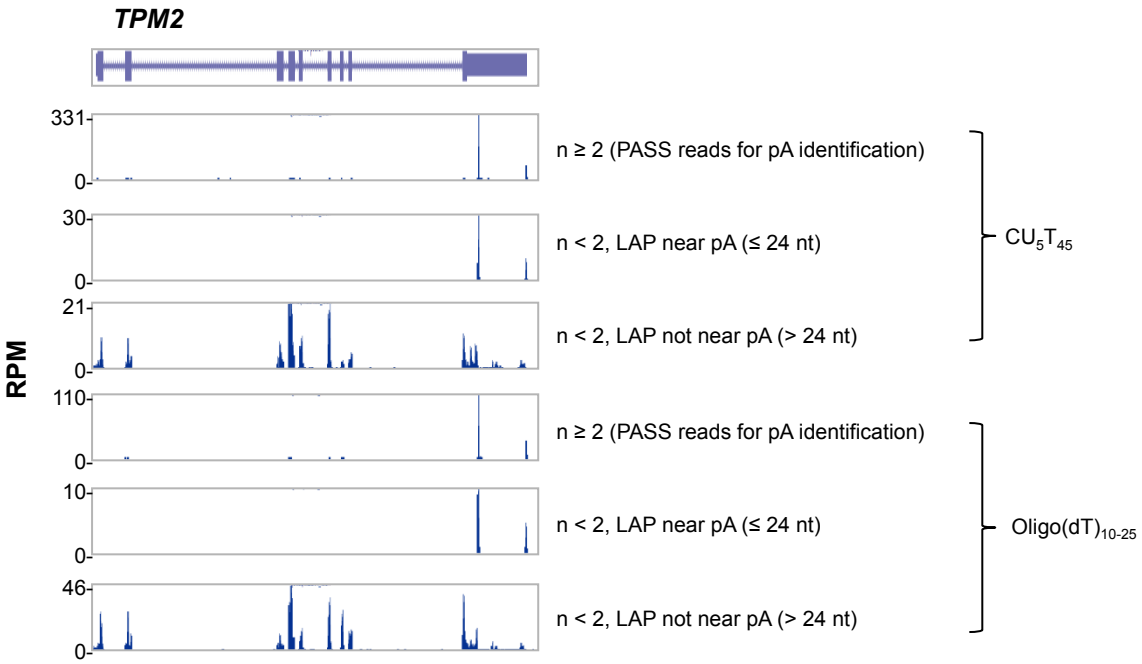
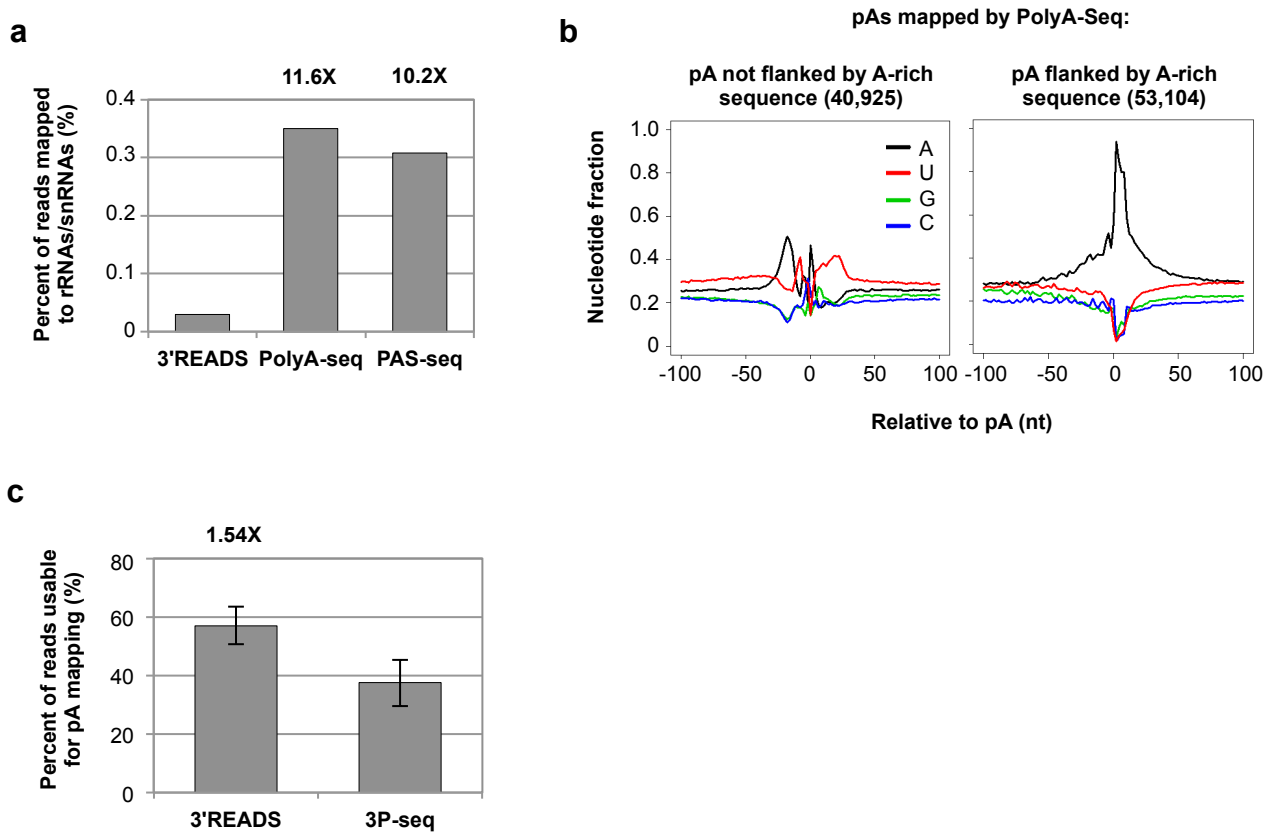
Supplementary Figure 1



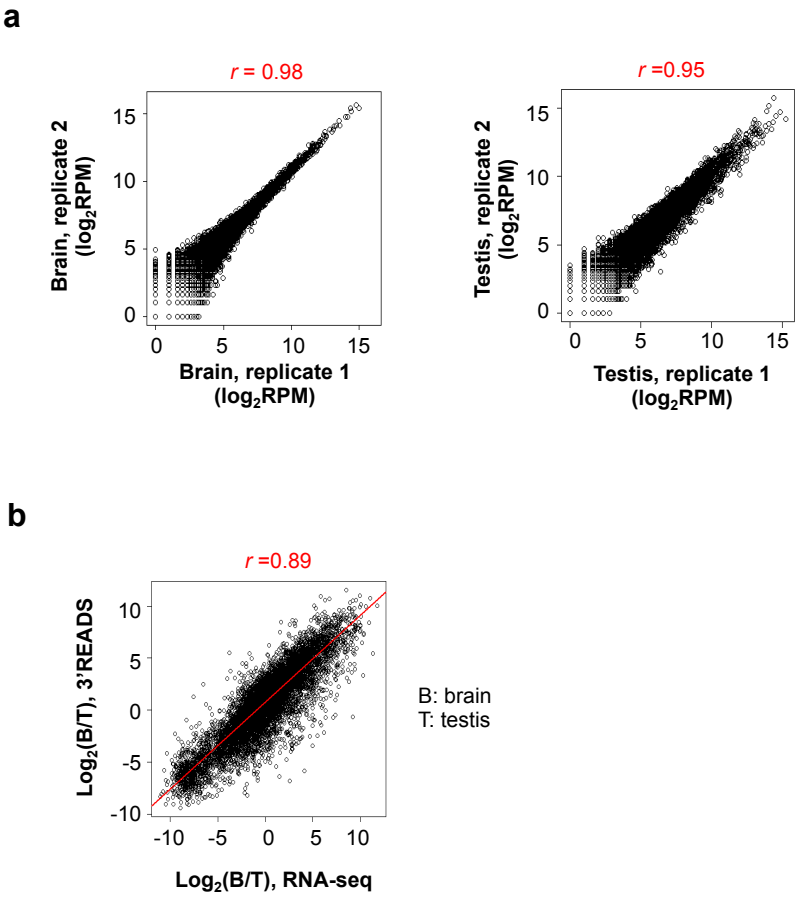
Supplementary Figure 1. 3'READS data. Example of three types of reads mapped to a gene (*TPM2*) using CU_5T_{45} oligo (top) or oligo$(dT)_{10\text{-}25}$ (bottom). PASS, polyA site supporting; LAP, last aligned position.

Supplementary Figure 2



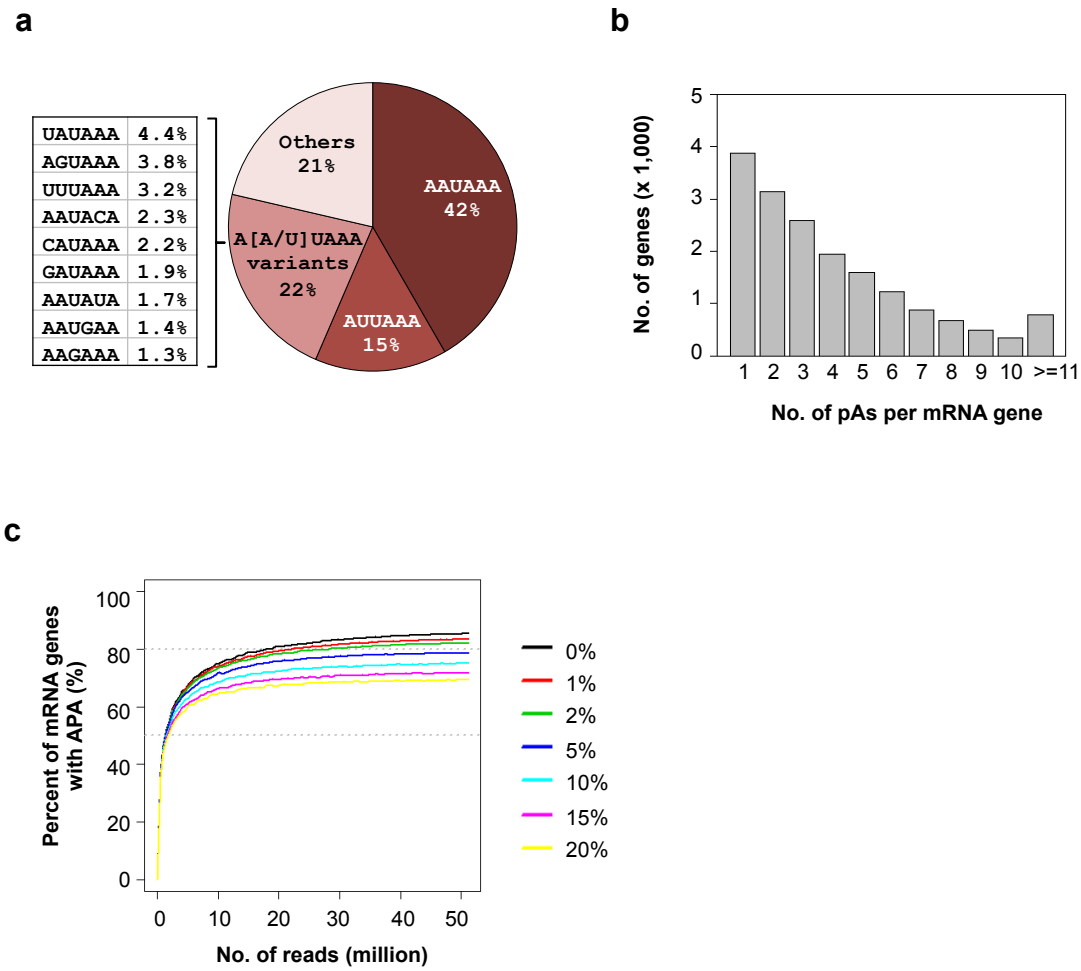
Supplementary Figure 2. Comparison of 3'READS with other methods. (a) Percent of reads assigned to rRNA, snoRNA, or snRNA genes, using 3'READS, PolyA-seq and PAS-seq. The ratio of the value for PolyA-seq or PAS-seq to that for 3'READS is indicated. **(b)** Nucleotide profiles around pAs mapped by PolyA-seq. pAs were divided into those not flanked by A-rich sequences (left) and those were (right). See Methods for identification of A-rich sequences. **(c)** Percent of reads useful for pA mapping for 3'READS and 3P-seq. The useful reads were defined as reads containing extra unaligned As/Ts that support the poly(A) tail. The ratio of value for 3'READS to that for 3P-seq is indicated.

Supplementary Figure 3



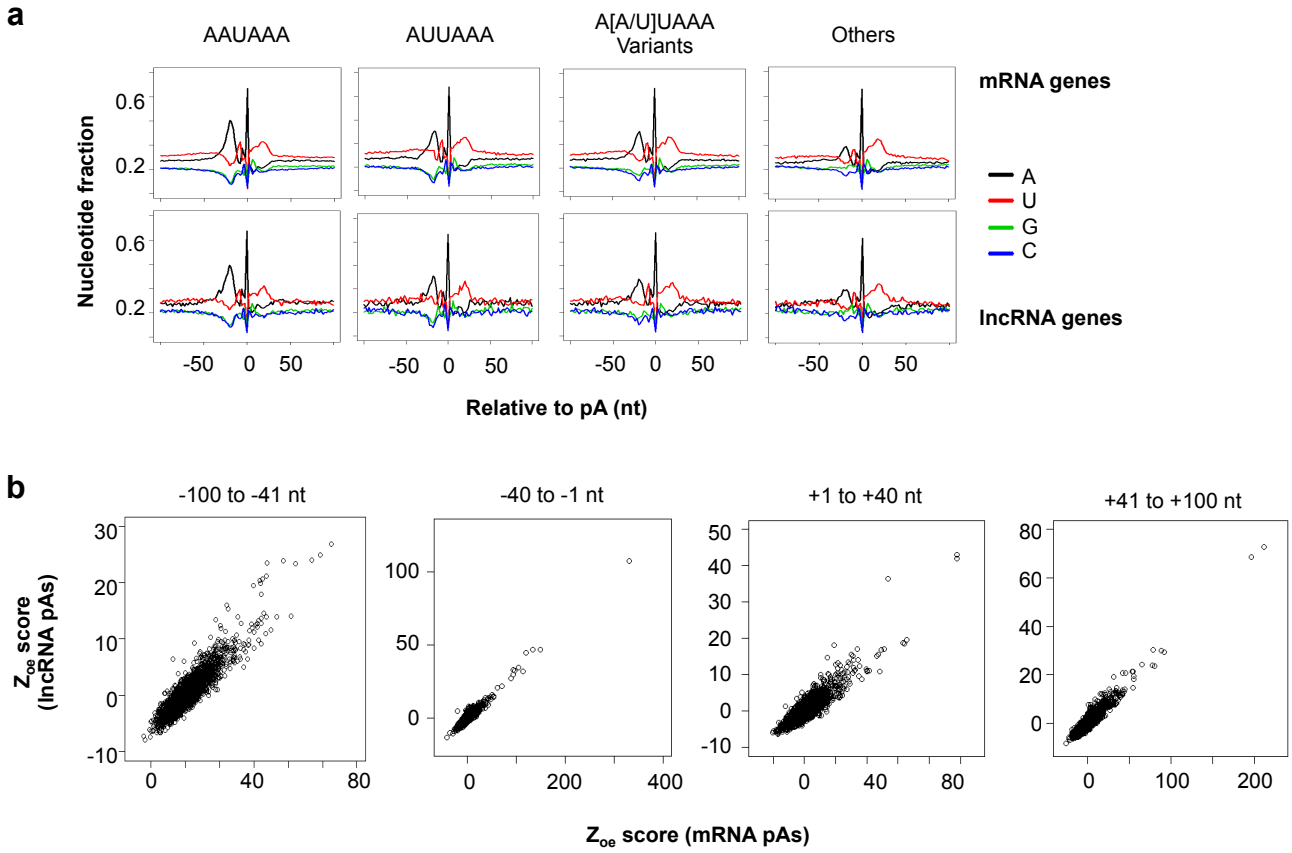
Supplementary Figure 3. 3'READS is reproducible and quantitative. (a) Comparison of 3'READS data from two brain (left) and testis (right) replicates. Reads were combined into genes. Each dot represents a gene. **(b)** Comparison of 3'READS data and RNA-seq data. Reads were combined into genes, and ratio of abundance was calculated for both 3'READS (y-axis) and RNA-seq (x-axis). Pearson correlation coefficient (*r*) is shown on each graph.

Supplementary Figure 4



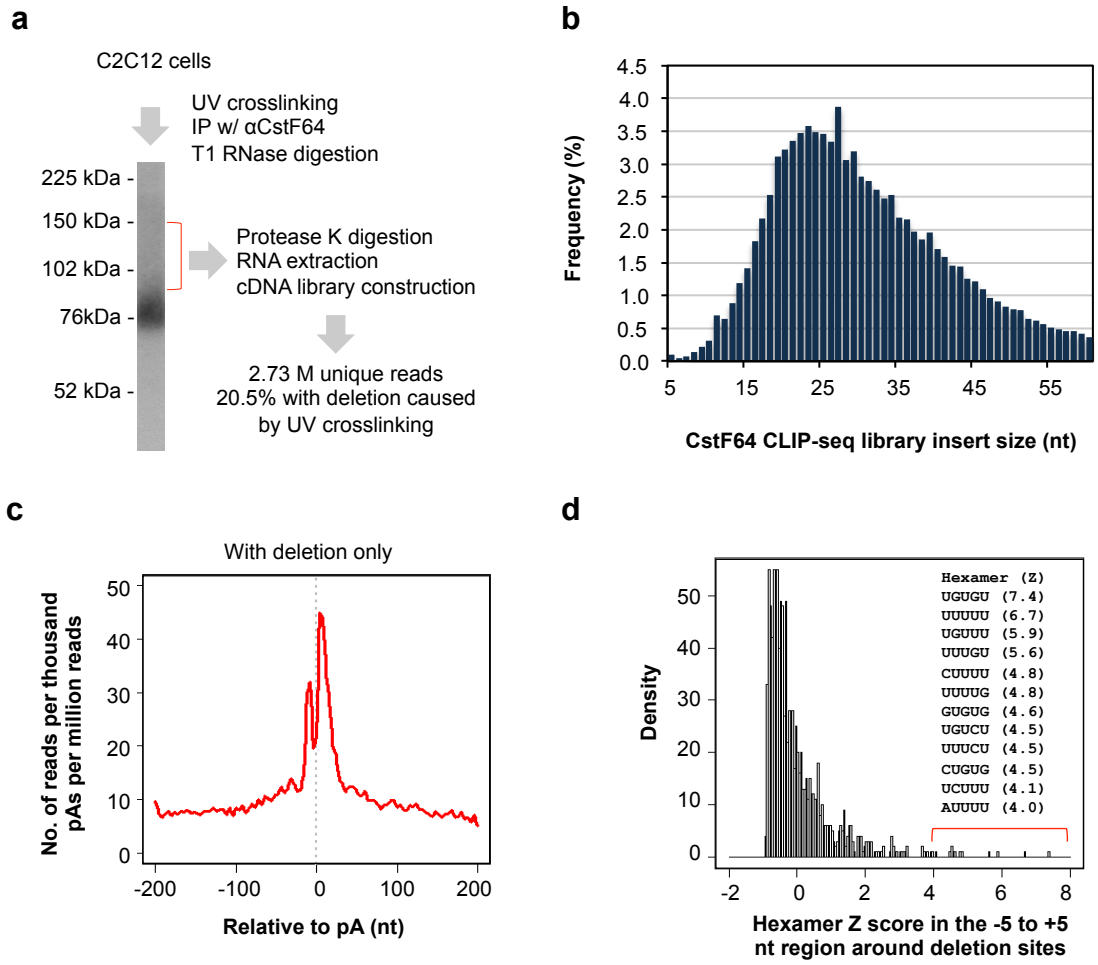
Supplementary Figure 4. APA in the mouse genome. (a) Distribution of PAS types for mouse pAs (see Methods for PAS identification). **(b)** Histogram of mRNA genes with different numbers of pAs. **(c)** Saturation curves showing sequencing depth vs. percent of APA genes identified**.** Different isoform relative abundance values were used as cutoffs, as indicated in the graph.

Supplementary Figure 5



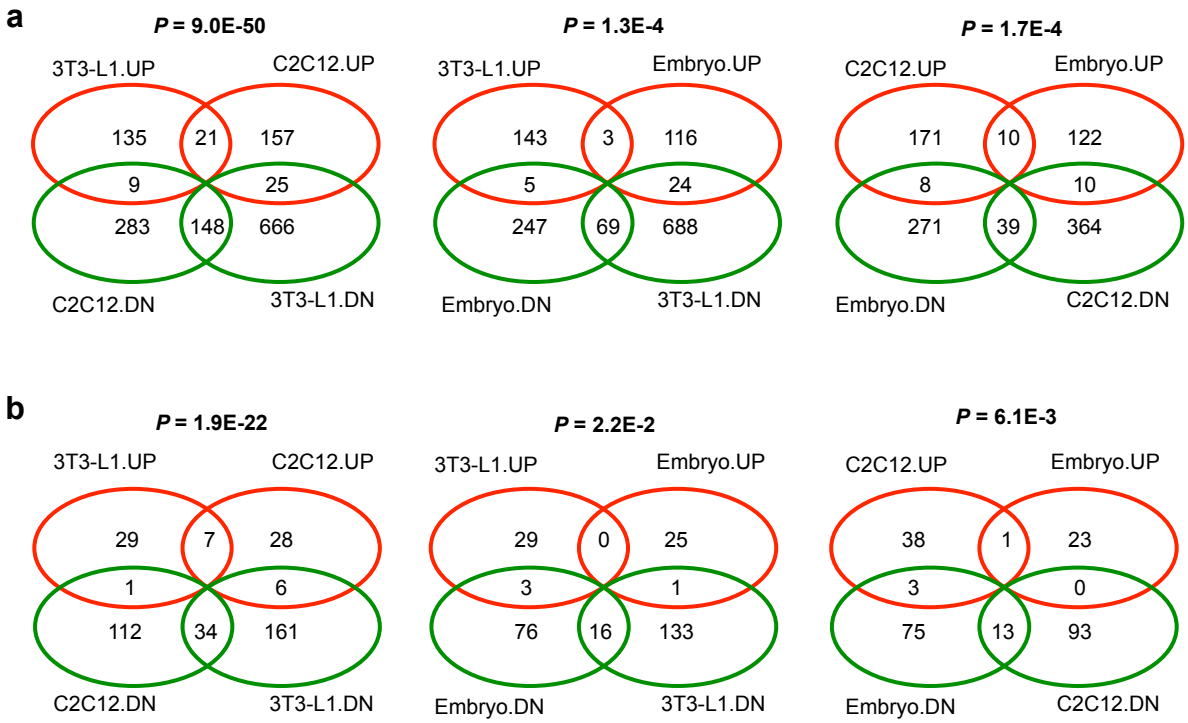
Supplementary Figure 5. Comparison of pAs in mRNA and lncRNA genes. (a) Nucleotide profiles around pAs in mRNA and lncRNA genes. pAs were grouped based on the PAS type. **(b)** Correlation of hexamer Z_{oe} scores in regions around the pA between pAs in mRNA and lncRNA genes. Four regions were analyzed, namely, -100 to -41 nt, -40 to -1 nt, +1 to +40 nt, and +41 to +100 nt, with the pA set at position 0. The Z_{oe} score represents the difference between observed and expected occurrences (see Methods for detail). Top overrepresented hexamers are also listed in Table S3.

Supplementary Figure 6



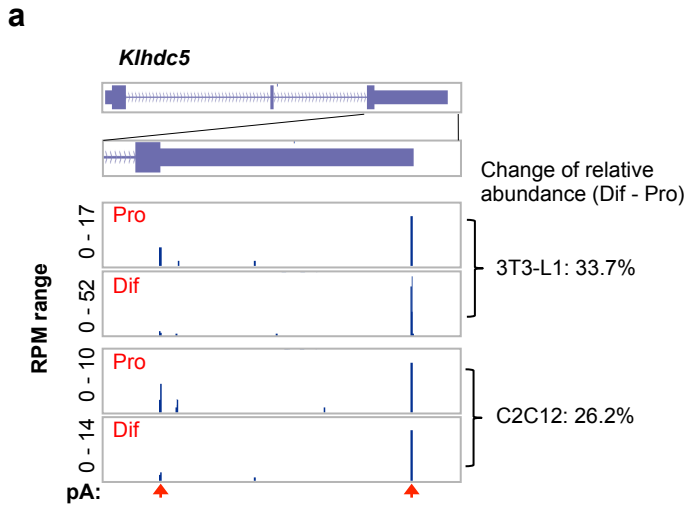
Supplementary Figure 6. CLIP-seq analysis of CstF64. (a) Schematic of the CstF64 CLIP-seq experiment. An autoradiograph of the membrane containing CstF64 with radioactively labeled RNA is shown. The region cut out for processing is indicated with a bracket. **(b)** Distribution of insert size of the CstF64 CLIP-seq library. Insert size was calculated from the reads, excluding the 3' adapter sequence. Since the read length is 70 nt and we use 10 nt for mapping with the 3' adapter sequence, the maximum insert size shown here is 60 nt. Those with insert size < 5 nt are not presented. **(c)** Distribution of CstF64 CLIP-seq reads around pAs. Only the reads with deletion(s) caused by UV crosslinking were used. **(d)** Hexamers enriched for genomic regions crosslinked to CstF64. The -5 to +5 nt region around deletion(s) was examined, which ensured that each hexamer in the region covered the deletion(s). Only reads mapped to the -100 to +100 nt region around the pA were used. Top 12 hexamers with respect to the Z score are shown. See Methods for calculation of the Z score.

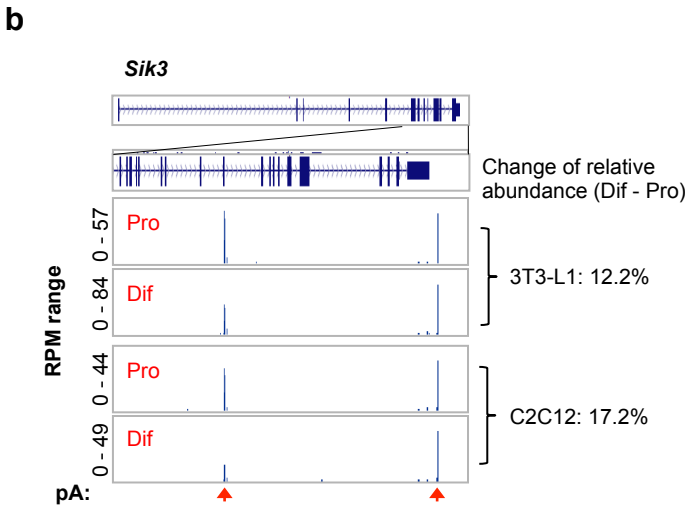
Supplementary Figure 7



Supplementary Figure 7. Comparison of regulated APA events in differentiation of 3T3-L1 and C2C12 cells and embryonic development. (a) Venn diagrams showing overlap of significantly regulated 3'-most exon pAs. **(b)** Venn diagrams showing overlap of significantly regulated upstream region (I/E) pAs. Upregulated (UP) and downregulated (DN) isoforms are those shown in Figures 5B and 5C. Only pAs in genes expressed in both comparing conditions were used. 3T3-L1.UP, C2C12.UP, and Embryo.UP are pAs whose isoforms are relatively upregulated in 3T3-L1 differentiation, C2C12 differentiation, and embryonic development, respectively. 3T3-L1.DN, C2C12.DN, and Embryo.DN are pAs whose isoforms are relatively downregulated in 3T3-L1 differentiation, C2C12 differentiation, and embryonic development, respectively. P-values indicating significance of overlap (Fisher's Exact test) are shown.

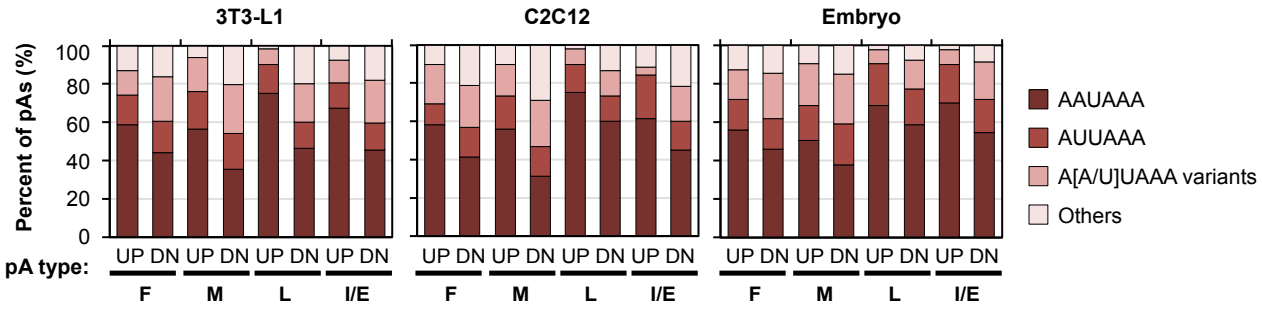
Supplementary Figure 8





Supplementary Figure 8. Examples of APA in cell differentiation. (a) An example gene (*Klhdc5*) showing relatively upregulated distal pA isoform in the 3'-most exon (vs. the proximal pA isoform) during differentiation. The PASS read distributions in proliferating and differentiated cells are shown. *Klhdc5* expresses multiple 3'-most exon pA isoforms. The two most abundant isoforms were compared, and are indicated by arrows. **(b)** An example gene (*Sik3*) showing relatively upregulated 3'-most exon pA isoforms (vs. upstream region pA isoforms) during differentiation. The two most abundant isoforms, one using an intronic pA and the other 3'-most exon pA, are indicated by arrows. The difference in relative abundance between differentiated cells and proliferating cells (Dif-Pro) is indicated.

Supplementary Figure 9



Supplementary Figure 9. PAS distribution for the pAs whose isoforms were significantly regulated in differentiation of 3T3-L1 and C2C12 cells and embryonic development. Upregulated (UP) and downregulated (DN) isoforms are those shown in Figure 5D.

Supplementary Table 1. Adapters and primers used in this study.

3' adapters	Forward sequencing: 5'-rApp/TGGAATTCTCGGGTGCCAAGG/3ddC Reverse sequencing: 5'-rApp/NNNNGATCGTCGGACTGTAGAACTCTGAAC/3ddC[1]
5' adapters	Forward sequencing: 5'-CCUUGGCACCCGAGAAUUCCA Reverse sequencing: 5'-GUUCAGAGUUCUACAGUCCGACGAUC
RT primers	Forward sequencing: 5'-CTAGCAGCCTGACATCTTGAGACTTG Reverse sequencing: 5'-GCCTTGGCACCCGAGAATTCCA
PCR primers	5'-AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA 5'-CAAGCAGAAGACGGCATACGAGAT[CGTGAT]GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA[2]

[1]Ns are random nucleotides used 1) to facilitate separation of clusters on the flow cell by Illumina software (Ts at the beginning of the read can cause problems); and 2) to distinguish different RNA fragments and eliminate redundant reads caused by PCR.
[2]The nucleotides in bracket are the index used for multiplexing sequencing.

Supplementary Table 2. Samples used in this study.

Sample	# of PASS reads
Whole body tissue mix (SABiosciences), several male and female Balb/c mice, whole bodies without fur.	3,692,214
Cell line mix 1 (Agilent), 11 cell lines	6,738,355
Cell line mix 2 (Tib75, CMT93, B16, F9, MC3T3-E1, N2A, NIH3T3)	5,972,921
Embryos at 11, 15, and 17 days	7,309,767
Proliferating and differentiated 3T3-L1 cells	6,631,037
Proliferating and differentiated C2C12 cells	11,703,170
Brain (postnatal 3 weeks, 6 weeks, 9 weeks)	3,905,248
Testis (postnatal 3 weeks, 6 weeks, 9 weeks)	8,795,777
Total	**54,748,489**

Supplementary Table 3. Top hexamers overrepresented in different regions around the pA.

Sub-region	Hexamer	mRNA	lncRNA
-100 to -41 nt	AAAAAA	92	25
	UAUAUA	89	25
	AUAUAU	84	25
	UUUUUU	67	24
	UAUUUU	69	15
	AAUAAA	61	22
	GUGUGU	54	24
	CACACA	54	16
	UGUAUA	57	13
	AAAUAA	49	20
-40 to -1 nt	AAUAAA	358	107
	AUAAAG	156	43
	AAAUAA	149	48
	AUAAAU	135	48
	AUAAAA	127	32
	AUUAAA	115	34
	UAAAUA	107	36
	AUAAAC	104	36
	CAAUAA	97	28
	UAAUAA	97	26
+1 to +40 nt	UGUGUG	111	50
	GUGUGU	110	49
	AAAAAA	65	39
	UUUAUU	73	24
	UUAUUU	72	23
	UUUUUU	69	19
	UAUAUA	58	18
	AUAUAU	56	19
	UAUUUA	54	18
	UAUUUU	55	15
+41 to +100 nt	GUGUGU	235	83
	UGUGUG	217	78
	ACACAC	104	34
	UUUUUU	104	28
	CACACA	94	33
	UAUAUA	96	20
	AUAUAU	92	23
	AAAAAA	76	28
	UCUCUC	61	28
	UUUAUU	62	21

Top 10 most enriched hexamers based on Z_{oe} in regions near the pA. Hexamers are sorted by the average Z_{oe} score for both mRNA and lncRNA pAs. Values shown in the table are Z_{oe} scores (see Methods for calculation of Z_{oe}).

Supplementary Discussion

Comparison with 3P-seq. The 3P-seq method developed by the Bartel group[1] uses splint ligation to ensure that only the RNAs with 3' terminal As are captured and sequenced, which elegantly addresses the internal priming issue. However, the RNase T1 digestion and multiple steps of ligation and reverse transcription in 3P-seq not only require substantial efforts for optimization of experimental condition but can also introduce noise of various kinds. By contrast, 3'READS has fewer steps and is much easier to implement. In addition, 3'READS uses a washing condition that maximally separates long and short A-tailed RNA-species, which can minimize the complication of oligo(A) tails. 3P-seq, however, does not address this issue. As such, 3'READS generates 54% more reads usable for pA mapping than 3P-seq (Supplementary Figure 2c).

Extent of APA in the mouse genome. We found that ~79% of mRNA genes and ~66% of lncRNA genes in the mouse genome have APA. These numbers are significantly higher than what we previously reported using cDNAs/ESTs[2]. Because of biased usage of pAs in tissues[3, 4] and in development[5], we used samples representing diverse tissue types and development stages. In addition, we specifically included brain and testis samples because of their unique trends of pA usage compared to other tissue types[3, 6]. Simulation of data indicated that our pA mapping is saturated for the samples used in this study (Supplementary Figure 4c). Nevertheless, some pAs may still escape our identification if they are used in certain cell types or under particular conditions not analyzed in this study. Future work will need to explore special biological settings to completely catalog pAs in the mouse genome.

Intron/upstream exon pAs. Interestingly, we found that about half of the genes with APA have intron/upstream exon pAs. Given their impact on the CDS of mRNA, their regulation can significantly affect the proteome in the cell. Notably, for mRNA genes, the abundance of isoforms using intron/upstream exon pAs is much lower than that of isoforms using 3'-most exon pAs (Figure 4c). However, in general, the PAS types of intron/upstream exon pAs and their binding to CstF64 are not substantially different than those of 3'-most exon pAs (Figures 4d and 4e). This suggests that the expression of intron/upstream exon pA isoforms is inhibited by mechanisms other than 3' end processing. One possibility is that splicing can suppress the usage of these pAs despite their recognition by cleavage/polyadenylation factors, leading to low efficiency of 3' end processing. Another possibility is that intron/upstream exon pA isoforms may be subject to rapid RNA decay. This is particularly relevant to pAs in CDS exons which can lead to unstable transcripts due to lack of an in-frame stop codon[7]. Further studies are needed to examine these mechanisms. On the other hand, lncRNA genes tend to have relatively higher expression of isoforms using intron/upstream exon pAs. How their APA regulation differs from that of mRNA genes and how usage of alternative pAs impacts lncRNA functions are to be further explored.

CLIP-seq. Consistent with the known properties of CstF64, our CLIP-seq reads were highly enriched around pAs (Supplementary Figure 6C), and GU-rich and U-rich hexamers were significantly overrepresented in regions with deletions (Supplementary Figure 6d). Interestingly, while reads were mostly mapped to the immediately downstream region of pA, a peak in the immediately upstream region was also discernable (Supplementary Figure 6c), suggesting CstF64 may also bind sequences upstream of the pA. However, co-immunoprecipitation of other cleavage/polyadenylation factor(s) binding to the region cannot be completely ruled out.

Supplementary References

1. 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).
2. 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).
3. Zhang, H., Lee, J.Y. & Tian, B. Biased alternative polyadenylation in human tissues. *Genome Biol* **6:R100** (2005).
4. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476 (2008).
5. 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 U S A* **106**, 7028-7033 (2009).
6. Liu, D. et al. Systematic variation in mRNA 3'-processing signals during mouse spermatogenesis. *Nucleic Acids Res* **35**, 234-246 (2007).
7. van Hoof, A., Frischmeyer, P.A., Dietz, H.C. & Parker, R. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science* **295**, 2262-2264 (2002).