PAN-TISSUE ANALYSIS OF APA REGULATION IN HYBRID MICE

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by
Yisheng Li

Berlin, 2019

1st Reviewer: Prof. Dr. Wei Chen
2nd Reviewer: Prof. Dr. Florian Heyd

Date of defense: 21.11.2019

Acknowledgement

Here I would like to express my sincere gratitude to my supervisors Prof. Dr. Wei Chen and Prof. Dr. Florian Heyd, without whom I could never complete this thesis. Their patience, motivation, enthusiasm, and immense knowledge encouraged me to continue my journey in biology study. I still remember that Prof. Dr. Wei Chen guided me to start the sequencing work in tRNA isoforms with great patience for the first time when I joined his lab without any idea in sequencing and bioinformatics. I spent around a half-year until the first library was successfully prepared. Since then, I start to fond the work and my study in bioinformatics. The next year, I was so glad that my first study proposal got accepted by Prof. Dr. Florian Heyd. He encouraged me to pursue the unknown and do researches with my own ideas and to be motivated by an aspiration for science instead of the requirements of the studying field. In four years, their help and advice have always accompanied me for research and thesis writing. I could not have imagined having better advisors and mentors for my Ph.D. study.

Besides my advisors, I would like to express my gratitude to Dr. Bernhard Schaefke and Dr. Xudong Zou for their encouragement, insightful comments, hard questions and inspiring discussion in my study. My sincere thanks also go to Dr. Bernhard Schaefke who spent a lot of time and effort in my thesis revising and helped me to translate the thesis.

Additionally, I would like to thank my colleagues, Dr. Bin Zhang and Dr. Meisheng Xiao for sharing their experience and knowledge in the field of alternative polyadenylation (APA) and for teaching me all protocols related to the study, Dr. Qingsong Gao and Dr. Xi Wang for their help in teaching me the computing skills. Also, I thank my fellow lab-mates in Southern University of Science and Technology: Min Zhang and Yuhao He for their hard-working and insightful discussion. We worked together and learned from each other in many projects. They are also my friends who support me spiritually. It was fantastic to have the opportunity to acquaint and work together with these wonderful people during these 4 years of study.

Last but by no means least, I would like to express my appreciation to my family: my parents Ganlin Luo and Bei Li, who raised me in the first place and supported me along the way. With a special mention to my wife Yunting Cai, she is the one who has provided me with moral and emotional support in my life and she is also the source of my motivation. I appreciate her supports for my career and her help in my daily life. She was always keen to know what I was

doing and how I was proceeding. Whenever a significant momentous was reached, we shared screams of joy.

Very special gratitude goes out to all down at the China Scholarship Council (CSC) Graduate Scholarships for helping and providing the funding for my study and the work.

Thanks for all your encouragement!

Zusammenfassung

Alternative Polyadenylierung (APA), die sowohl durch *cis*-Elemente als auch durch *trans*-Faktoren reguliert wird, ist in Eukaryoten weit verbreitet und wird als ein wichtiger Mechanismus der Genregulation angesehen. Durch APA können die Länge der 3'-UTR eines mRNS-Transkripts und die darin enthaltenen *cis*-regulatorischen Elemente so geändert werden, dass es seine Stabilität, Translationseffizienz, Export aus dem Zellkern und die Lokalisierung der mRNS oder des translatierten Proteins beeinflusst. Falls eine Polyadenylierungsstelle (PAS) vor dem Stoppcodon verwendet wird, kann APA auch zu verschiedenen Protein-Isoformen mit unterschiedlichen Eigenschaften führen. Zahlreiche Studien deuten darauf hin, dass die durch APA hervorgerufenen Längenänderungen der 3'-UTR eine wichtige Rolle bei der onkogenen Transformation, der Pluripotenz, der Lymphozytenaktivierung, der neuronalen Stimulation sowie bei der embryonalen Entwicklung und Differenzierung spielen könnten. In den meisten Fällen ist jedoch die Funktion von APA, die in Hochdurchsatz-Sequenzierungsexperimenten beobachtet wurde, weiterhin unklar. APA als molekularer Mechanismus ist ein Phänotyp auf niedriger Ebene in der Hierarchie biologischer Organisation und hat möglicherweise nur sehr begrenzte Auswirkungen auf die Fitness des Organismus. Daher schlugen einige Forscher die "Fehlerhypothese" vor, wonach in den meisten Fällen APA auf fehlende molekulare Präzision zurückzuführen ist, und die APA-Varianz eines Gens innerhalb eines Gewebes oder APA-Unterschiede zwischen Organen im Allgemeinen neutral oder sogar schädlich und ohne biologische Funktion sind. In ähnlicher Weise wurde vermutet, dass die APA-Divergenz zwischen den Arten weitgehend nicht adaptiv ist. Dieses Szenario würde mit der (nahezu-)neutralen Theorie der molekularen Evolution übereinstimmen, die vorhersagt, dass Gene unter schwächerem Selektionsdruck schneller neutrale (oder leicht schädliche) Mutationen akkumulieren als solche unter stärkerer negativer Selektion.

Um die allgemeine und gewebeabhängige Funktion und Regulation von APA und ihre Evolution in Säugetieren zu erforschen, verwendeten wir 3'-mRNS-Sequenzierung für mehrere Gewebe der weiblichen Nachkommen aus einer F1-Kreuzung zwischen der Hausmaus (*Mus musculus*), vertreten durch die Labormaus C57BL/6J, und der algerischen Maus (*Mus spretus*), vertreten durch die Inzuchtlinie SPRET/EiJ. Wir haben die Faktoren analysiert, die die APA-Diversität regulieren, um die Vorhersagen der Fehlerhypothese zu überprüfen. In dieser Studie wurden alle annotierten PASs in neun Geweben der F1-Hybridmaus quantifiziert und verschiedene Merkmale von Genen mit einer einzigen PAS und Multi-PAS-Genen (Genen mit APA) umfassend charakterisiert. Als nächstes überprüften wir den Effekt der relativen PAS-Position auf die PAS-Stärke und untersuchten den funktionellen Unterschied zwischen PASs des Rangs 1 (Haupt-PASs) und des Rangs 2 in verschiedenen Gengruppen. Die allel-spezifische Quantifizierung der PAS-Verwendung ermöglichte es uns, die Gene mit Unterschieden in APA-Mustern oder mRNS-Leveln zwischen den Spezies zu identifizieren und verschiedene Evolutionsmuster in APA und Genexpression aufzudecken.

Unsere Ergebnisse zeigen, dass die APA-Muster von Multi-PAS-Genen im Allgemeinen mit der Fehlerhypothese übereinstimmen und dass die APA-Diversität eines Gens innerhalb oder zwischen Geweben größtenteils auf molekulare Fehler aufgrund schwacher *cis*-Regulation zurückgeführt werden kann. Die meisten Gene haben nur eine optimale PAS, deren Verwendungshäufigkeit mit den Expressionsniveaus der mRNS in den Geweben korreliert. Zusätzlich wird die Beziehung zwischen dem Genexpressionslevel und der relativen PAS-Verwendungshäufigkeit auch durch die PAS-Lokalisierung beeinflusst und könnte eine direkte funktionelle Beziehung zwischen der Transkriptionsregulation und der PAS-Wahl widerspiegeln. Obwohl für die meisten Gene nur die Verwendung einer einzigen PAS (der Haupt-PAS) erwünscht zu sein scheint, haben einige Gene auch alternative PASs, die offenbar eine funktionelle Bedeutung haben. Sie sind stark konserviert und können mit den Haupt-PASs um die Polyadenylierungsmaschinerie konkurrieren. Außerdem fanden wir eine kleinere Anzahl von Genen, die stark organspezifische APA-Muster aufweisen. In diesen Genen unterliegt die PAS-Verwendung einer intensiven *trans*-Regulation und ist normalerweise für die C57BL/6J- und SPRET/EiJ-Allele in der F1-Hybridmaus ähnlich, was auf eine evolutionäre Konservierung hinweist. Im Gegensatz dazu existieren viele Unterschiede in PAS-Verwendungsmustern zwischen den beiden Allelen in Genen mit niedrigem Expressionsniveau und unter schwachem Selektionsdruck.

Abstract

Alternative polyadenylation (APA), which is regulated by both *cis*-elements and *trans*-factors, is widespread across all eukaryotic species and is recognized as a major mechanism of gene regulation. It could change the 3'UTR of an mRNA transcript affecting its stability, translation efficiency, nuclear export and mRNA or translated protein localization, or, if an exonic/intronic polyadenylation site (PAS) upstream of the stop codon is used, it could affect a gene's coding region to produce different protein isoforms with distinct properties. Accumulating evidence suggests that global APA-mediated 3'UTR length change might play an important role in oncogenic transformation, pluripotency, lymphocyte activation, neuronal stimulation and in embryonic development and differentiation. However, recent studies found limited effects of 3'UTRs in most genes compared to other regulatory elements located in 5'UTRs or coding sequence. APA as a molecular trait is a low-level phenotype in the hierarchy of biological organization, and might only exert very limited effects on organismal fitness. Therefore, some researchers proposed the “error hypothesis”, stating that most observed APA is noise and that APA diversity within and between tissues is generally neutral or deleterious, and not functional. Similarly, it has been suggested that APA divergence between species is largely non-adaptive. This scenario would be consistent with the (nearly) neutral theory of molecular evolution, which predicts that genes under relaxed selective constraints accumulate neutral (or slightly deleterious) changes at a faster rate than those under stronger purifying selection.

In order to clarify the general and tissue-dependent function and regulation of APA and its evolution in mammals, we applied 3'mRNA sequencing for multiple tissues of an F1 hybrid between the C57BL/6J (*Mus musculus*) and SPRET/EiJ (*Mus spretus*) mouse strains. We analyzed the factors regulating APA diversity and addressed the question whether APA is generally non-adaptive as proposed by the error hypothesis. In this study, we quantified all annotated PASs in nine tissues of the F1 hybrid mouse and comprehensively characterized different features of single-PAS genes and multi-PAS genes. Next, we checked the positional effects on PAS strength and discussed the functional difference between rank 1 and rank 2 PASs among distinct gene groups. By quantifying PAS usage in each allele, we studied the genes with divergent major PAS expression level and dN/dS ratio difference, and unveiled different evolutionary patterns between APA patterns and gene expression (mRNA levels).

We found that in general APA of multi-PAS genes is consistent with the error hypothesis, and that most APA diversity within and between tissues appears to reflect noise, resulting from molecular error due to weak *cis*-regulation. However, we did not find different selective constraint in dN/dS between genes with high and with low APA diversity, but found strong correlation between mRNA abundance and APA accuracy. The minor and major relative PAS usage is also affected by PAS position. In addition to most major PAS, many minor PASs appear to have functional importance. They are highly conserved and can compete with the major PASs. Last, we found a small fraction of genes exhibits strongly tissue-regulated APA patterns. In these genes, PAS usage is under intensive *trans*-regulation between the C57BL/6J and SPRET/EiJ alleles in the F1 hybrid mouse. Whereas many divergent PASs exist between the two alleles in genes with low expression level and under relax selective constraints, comparing these with genes showing allelic mRNA transcript level differences, we unveiled different evolutionary patterns between APA and gene expression.

Directory

Introduction

Cleavage and polyadenylation

When a gene is transcribed by RNA Polymerase II (Pol II) in Eukaryotes, mRNA cleavage and polyadenylation are essential steps to free the pre-mature mRNA from its DNA template and add a poly(A) tail to its 3'end. The terminal site of genome where the poly(A) tail is added is commonly referred to as polyadenylation site (PAS). This process was discovered back to 40 years ago[1–3] and has since been regarded as an important process to regulate gene expression[4]. In brief, a hexamer signal such as AAUAAA, AUUAAA or one of its variants located upstream of the mRNA's cleavage site is first recognized by the cleavage and polyadenylation specificity factor (CPSF). The factor CPSF-73 (Ysh1 in yeast) preferentially cuts the CA dinucleotide at the cleavage site in preference, forming a 3' mRNA end of the transcript. A poly(A) tail is then synthesized by untemplated poly(A) polymerase (PAP), which process is also facilitated by CPSF and PABPN1 bound at the transcript end.

Only after polyadenylation, mature mRNA can be transported out of the nucleus into the cytoplasm. The poly(A) tail dependent exportation was first discovered in Saccharomyces *cerevisiae* by demonstrating that any mutation in a *cis*-element or a *trans*-factor causing malfunction of mRNA cleavage and polyadenylation would result in mRNA retention in the nucleus[5–7], similar as any malfunction in 5' cap addition or failure to remove introns by splicing. Transcripts with a poly(A) tail can first interact with the TREX (TRanscription-Export) complex and get transported to the Mex67:Mtr2 complex to form an export-competent messenger ribonuclear particles (mRNP) which is then transmitted to the cytoplasm. Factors such as Nab2 coordinating the nuclear steps and DEAD-box helicase Dbp5 detaching the mRNA from the Mex67:Mtr2 complex in the cytoplasm are also essential for this process[8–10].

Besides being a crucial step of RNA transcription maturation and nuclear export, forming the poly(A) tail on an mRNA's 3' untranslated region (3'UTR) is also known to be important in gene expression regulation. First of all, poly(A)-binding proteins (PABPs) can regulate mRNA stability in the cytoplasm. Several studies showed that binding of PABPC (cytoplasmic PABP) at the poly(A) tail inhibits uridylation of the tail and protects mRNA from degradation by the exosome[11,12]. On the other hand, PABPCs can interact with the complexes Pan2–Pan3 and Ccr4–Not causing deadenylation or contribute to microRNA-mediated repression, resulting in mRNA decay[13]. Although it remains unknown how the direction in which a PABPC would affects mRNA stability is determined, in most cases, PABPs functions as protectors to reduce mRNA degradation speed[13]. In embryonic development, the expression level of some PABPs is stage dependent, e.g. embryonic poly(A)-binding protein (ePABP) from *Xenopus* egg with specific affinity in mRNA's AU-rich elements (ARE), indicating its role in UTR motif mediated regulation[12]. Besides mRNA stability, a closed-loop formed by PABP and EIF4G for mRNA is proposed as a canonical model to facilitate translation initiation and further stabilize mRNA during translation[14].

Interestingly, polyadenylation does not only happen inside the nucleus. For poly(A) tail elongation, polyadenylation can also occur in the cytoplasm by cytoplasmic poly(A) polymerase GLD-2, which is recruited by cytoplasmic polyadenylation element binding proteins (CPEB) and CPSF bound to cytoplasmic polyadenylation elements (CPEs) and polyadenylation site (PAS) respectively. For post-transcriptional mRNA cleavage, a terminator 5′-phosphate-dependent exonuclease (TEX) treatment sequencing test validated that mRNA cleavage can happen in the cytoplasm and produce a shortened mRNA with newly added poly(A) tail and an uncapped downstream 3' UTR region[15]. Though key proteins involved in this process remain to be determined, CPEB might help in initiating PAS signal recognition for this post-transcriptional cleavage. Importantly, polyadenylation in the cytoplasm can make it possible for CPEB to regulate the mRNA's stability and translation efficiency of mRNAs, which plays an important role in controlling oogenesis, cell division, cellular senescence and even synaptic plasticity in neurons[16]. Malfunction of CPEB or erroneous cytoplasmic polyadenylation also contribute to tumor development and chronic liver disease due to dysregulation of angiogenesis problem[17,18].

Alternative polyadenylation

Another intriguing aspect of polyadenylation is the fact that the cleavage and polyadenylation site of one gene is not necessarily fixed. This RNA processing that generates distinct PASs for one gene is termed alternative polyadenylation (APA). It was first observed in Adenovirus type 2 (Ad2) where five different 3' terminal structures of the transcript were identified[19,20]. APA widens the choices of mRNA produced for a single gene, adding to the number of possible isoforms introduced by alternative splicing and alternative transcription start sites (TSS). Thanks to advanced techniques in 3'mRNA sequencing such as poly(A)-position profiling sequencing (3P-seq) and 3' region extraction and deep sequencing (3'READS)[21,22], increasing numbers of PASs are detected and annotated to genes for several species. As recorded in current PAS databases, there are 50% of all genes undergo APA in flies[23], 69% genes in zebrafish[23], 70% in mice[22] and 79% in human[24]. On average, each of these species has 2.6 PASs, 2.8 PASs, 3.8 PASs and 4.1 PASs per gene respectively[25]. While some PASs located in an intron or an exon's CDS region change the open reading frame (ORF) and produce distinct protein isoforms, the majority of PASs is located in the 3'UTR of annotated genes and affects the length of the 3'UTR. The latter APA can produce mRNA isoforms harboring different *cis*-elements, which could recruit various RNA binding proteins (RBPs), microRNAs (miRNA) and lnRNA to regulate transcript stability, translational efficiency, RNA localization and translated protein localization[25].

Functions of alternative polyadenylation

Currently, among the above factors, the effect of 3'UTR lengths on miRNA targeting is the best studied. miRNAs are an approximately 21 nt non-coding RNAs which can complementarily bind to 3'UTR with its 2-8 nt seeding region, in some cases also to an mRNA's CDS region. They function by repressing translation initiation and triggering mRNA deadenylation and decay[26]. By producing RNA transcripts with different 3'UTRs, APA can regulate a gene's miRNA-mRNA interactions and consequently the levels of its protein product. In cancer cells, for example, ATP binding cassette subfamily G member 2 (ABCG2) can be overexpressed and increase the cell lines' drug resistance by switching to the mRNA isoform with shorter 3'UTR lacking the hsa-miR-519c binding site[27]. APA of the Heat shock protein 70 (Hsp70) transcript can produce a shorter transcript without the binding site for miR-378* to bypass miR-378*-mediated suppression under the condition of an ischemic or heat shock stimulus[28]. Extensive bioinformatics studies have found that shorter 3'UTR isoforms are preferred in embryonic tissues, cancer cells and cells undergoing proliferation compared to differentiated cells[23,29–32]. Also a global analysis conducted by the Tian lab in different human and murine tissues and cell lines found a negative correlation between gene expression level and UTR length according to RUD scores (Relative expression of isoforms Using Distal polyA sites)[33]. These studies indicated that mRNA's UTR length might be related to cell status through controlling the miRNA binding sites of different mRNA isoforms. An intuitive thought is that longer 3'UTR increases miRNA binding probability and leads to repression of mRNA translation and reduced mRNA stability. However, a study in miR-17/92 cluster's role in upregulating cell's proliferative stage revealed that only miRNA targeting near the poly(A) tail functions effectively on mRNA regulation, whereas longer 3'UTR can provide a buffering place for miRNA to compete with each other in regulating transcript fate[34,35]. Additionally, a decrease in miRNA enrichment with increasing 3'UTR length was found in an evolutionary search of miRNA density in different species by Bartel Lab, suggesting a sophisticate role in 3'UTR's length of miRNA targeting regulation[36]. Therefore, the choice and binding efficiency of miRNA on transcripts becomes one of key factors in mediating gene expression as well as cell status during stimulus or development.

Similarly, many RBPs bound to distinct *cis*-elements in the 3’-UTR could determine mRNA's stability. Some well-studied RBPs binding to the AU-rich elements (AREs) such as AU-rich binding factor-1 (AUF1), CUG-binding protein (CUG-BP) and KH splicing regulatory protein (KSRP) destabilize the mRNA by recruiting the exosome or other RNA degrading enzymes. For examples, Staufen 1 (STAU1) can be recruited to double-stranded RNA when an Alu element is (imperfectly) paired with a lnRNA, and initiate mRNA decay[37]. TIA-1 and TIA-1-related protein (TIAR) can silence mRNA translation[38–40]. In contrast, proteins like the embryonic lethal abnormal vision (ELAV) proteins (HuR and HuD) can compete with these RNA binding proteins and stabilize mRNA[41]. Other KH-domain RNA-binding proteins bind at pyrimidine-rich element to prevent mRNA's deadenylation[42], and PUF protein binding is involved in the CCR4–NOT deadenylase in Saccharomyces *cerevisiae*[43].

Regulation of mRNA processing by RBPs is of great functional importance, so that errors in the interaction between 3'UTRs and the binding RBPs are often related to severe disease. E.g., in systemic lupus erythematosus (SLE) patients, malfunctional of the proximal PAS in IFN-regulatory factor 5 (IRF5) caused by a single nucleotide polymorphism (SNP) results in the majority of the mRNA transcripts to be produced in a longer isoform containing an ARE. This type of mRNA degrades quickly, leading to a reduction in the total IRF5 protein level in patients[44]. Study also showed that loss of muscleblind-like (MBNL) protein, a RBP which can regulate both splicing and polyadenylation, disturbs cellular APA patterns globally, which matches the failure of postnatal APA pattern development in the mouse polyCUG model and human myotonic dystrophy (DM)[44]. Similar as found in miRNA interaction with mRNA's 3'UTR, many cross-linking immunoprecipitation (CLIP) works showed the enrichment of RBP toward 3' end of 3'UTR, suggesting RBP nearby poly(A) tail can efficiently regulate mRNA fate[45,46], although longer mRNA isoforms have a higher chance to bind RBP or lnRNA in their destabilizing elements. However, in mouse fibroblast, evidences from 3P-seq, polysome profiling and ribosome footprint profiling data also showed limited influence in stability and translation efficiency of most mRNAs regardless of their choices among different length isoforms[47].

Another important role of RBPs binding to distinct mRNA isoforms is to control their localization in the cell. This mechanism occurs in two steps: mRNA nuclear exportation and mRNA cytoplasmic localization. Early in 2012 Sarah *et al.* reported differential transcript isoform enrichment in nucleus and cytoplasm[48]. A detailed study conducted by Jonathan *et al.* found that isoforms with longer 3'UTR are retained in the nucleus. Further work in HEK293 cells showed that about 10% of mRNA APA isoform exhibit different abundance in cytoplasm vs. nucleus which indicates some mRNA can be subject to different regulation[49]. However, till recently only a fewer *cis*-elements such as inverted

Alu-repeats are known to be involving mRNA nuclear retention[50], how mRNA isoform is selectively regulated and exported from nucleus is largely unclear.

Compared to mRNA's nuclear retention, subcellular localization of mRNA in cytoplasm is better characterized. Study of oocytes revealed a ~50nt *bicoid* localization element (BLE1) guides *bicoid* mRNA's transportation to the anterior pole of Drosophila oocytes[51]. In *Xenopus* oocytes, 340 nucleotides of the 3'UTR containing VM1 (YYUCU) and E2 (WYCAC) elements help *Vg1* RNA localize to vegetal hemisphere[52]. Other *cis*-element such as E3 in yeast *ASH1*, a 54nt sequence "zipcode" in chick *β-actin* and a 94nt sequence in *CamKIIα* of mature neuron dendrites all play an essential role for RNA localization in the appropriate subcellular component[53–55]. Except for many *cis*-elements directly regulating mRNA transportation, *cis*-element that controls mRNA degradation also can make mRNA exhibit a distinct spatial localization. For example, *nanos* and *Hsp83* mRNA in the *Drosophila* embryo are unstable and vulnerable to Smaug triggered CCR4/Not deadenylase but not in the posterior polar plasm and the pole cells[56,57].

Regulated by APA, RNA can switch its 3'UTR with different localization signal to produce protein in distinct desired subcellular component. To be explicit, in hippocampal neurons, there are two kinds of brain-derived neurotrophic factor (BDNF) transcripts. One is with longer 3'UTR, and the other with a shorter 3'UTR. The long 3'UTR isoform is localized in dendrites and is translated to meet the local demands in neurons, whereas the short one is only enriched in somata[58]. Furthermore, APA regulated localization is not only limited in controlling mRNA itself. mRNAs of CD-47 and PD-L1 with AU-rich elements transcribed on their 3'UTR can be aggregated in ER subdomain by TIS granules (TIS11B formed membraneless organelle) where their 3'UTRs can mediate the interaction between SET and its translated protein, making their protein products transported to membrane[59]. This mechanism is especially important for some secretory proteins.

On the other hand, if PAS locates in gene's coding region, it will lead to two kinds of results: 1. PAS located in alternative last exon will generate new isoform with different ORF to enrich protein diversity. 2. PAS located in intron or CDS can repress gene translation. The first scenario is prevalent in blood-derived immune cells. This type of APA can specifically regulate B-cell development to adjust different cellular environments[60]. Take *CALCA* and immunoglobulin M (lgM) as examples. If *CALCA* mRNA is terminated at proximal PAS, it encodes hormones calcitonin, whereas another protein calcitonin gene-related peptide (CGRP) is produced by *CALCA* mRNA choosing distal PAS[61]. The hormones calcitonin is enriched in thyroid, whereas CGRP express more prevalently in hypothalamus. During B cell activation, membrane-bound IgM's heavy chain can also switched by using proximal PAS to function as secreted protein[62]. However, in the second scenario, mRNA is terminated early during transcription and may not possess an in-frame stop codon. This premature PAS can result in mRNA degradation by a non-stop decay pathway[60]. In some case like truncated cleavage stimulation factor 77 (Cstf-77) mRNA isoform, which has a conserved in-frame stop codon after intronic termination, its truncated mRNA isoform can still be translated into protein but without functional C-terminal. Because Cstf-77 itself also play an important part in mediating APA, this shortened protein is often viewed as a negative feedback product to repress high expression levels of full-length functional CSTF77 protein[63].

cis-elements in polyadenylation

Because of the potentially important role of APA in gene regulation, it is crucial to understand the mechanisms determining PAS choice during transcription. Currently, various of *cis*-elements in PAS flanking regions and *trans*-factors binding to PASs or affecting polymerase are known to determining PAS cleavage strengthAmong the *cis*-elements in the PAS region, the most common and well-studied sequences are the canonical polyadenylation signals, the AAUAAA/AUUAAA hexamers and their close variants. In mammals, different experimental validations of signal strength using point mutations in these hexamers in various genes showed a similar order, which starts with AAUAAA as the strongest PAS cleavage signal, followed by AUUAAA (~80% of the processing efficiency of AAUAAA), AGUAAA (~20%) and other variants[64–66]. However, PASs annotated by 3'READS in the Tian Lab showed that 28.6% of PAS regions in human and 35.9% in mouse contain no canonical hexamer or any of its variants[67]. Their cleavage strength could be contributed by some other shared features in PASs' flanking: 1) An A-rich sequence at the cleavage site (6.1% in human, 5.4% in mouse), 2) a secondary structure such as a stem loop to enhance cleavage, 3) G-rich (sometimes G-quadruplex structure) and U-rich (UAUA/UGUA) elements upstream of the PAS, 4) U-rich and GU-rich elements downstream[68,69]. Additionally, *cis*-elements recognized by splicing factors such as MNBL and SPSF can also function to mediate PAS strength[70]. The most efficient repressing element, U1 anti-sense sequence, was well studied decades ago and found to inhibit polyadenylation intensively by repressing PAS interaction with PAPα[71]. Therefore, reduced levels of U1 snRNA can often lead to increased usage of PAS near the 5' splicing site and globally shorten the mRNA's 3'UTR[72].

Trans-factors in regulating the alternative polyadenylation

If we apply a 'first come, first served' model to APA, we would expect that the major function of distal PASs is to guarantee transcription termination when proximal PASs are not recognized properly. However, in some cases distal PASs can be highly used in genes with strong proximal PAS signals. In fact, PASs located in different regions in a gene can compete with each other through *trans*-factor regulation. For example, the core polyadenylation complex cleavage factor I (CFI) has two subunits CFI25 or CFI68. When these two subunits are downregulated, an increase in proximal PAS occurs[73,74]. This suggests that these two subunits might function as PAS repressor and reduce the proximal PAS usage so that the distal one can be selected in multi-PAS genes. Besides PAS repressing proteins, *trans*-factors recognizing the core PAS signal can also affect the choice of PAS. Fip1 regulates APA depending on its interaction with CPSF and PAS distance, and can promote stem cell self-renewal. It can enhance the weaker proximal PAS usage of genes which have proximal and distal PASs separated by long genomic distances, but can also repress proximal PAS usage in genes with two close PASs[30]. Fip1 also cooperates with CFI in controlling splicing interference in 3' mRNA processing[75]. Other core factors currently found in PAS regulation are CSTF77 and CPSF30. Knockdown of CSTF77 or CPSF30 in oxidative stress can induce global APA variation [76,77]. Some RBPs such as Nuclear poly(A)-binding protein 1 (PABPN1) can also regulate APA, although their roles in APA regulation is not clear[77,78].

Tissue specific APA patterns

Because the expression level of *trans*-factors regulating a gene's APA pattern can vary between tissues, different isoforms generated by APA can exhibit tissue-specific expression to fulfill distinct functional requirements in different tissues. In *C. elegans*, DAPK-1/DAP kinase, SSUP-72/SSU72, PINN-1/PIN1 peptidyl-prolyl isomerase and SYDN-1 work together to regulate expression of the long isoform of *unc-44* (UNC-44L) in neurons. Interruption of their interaction would cause ectopic expression of UNC-44L in the epidermis and disrupt epidermal morphology[79]. In its body muscle tissue, *rack-1* and *tct-1* are expressed in their shorter 3'UTR isoforms to evade *miR-50* and *miR-85* targeting, which can increase protein expression needed for proper muscle function. In *Drosophila*, neurons predominantly express distal APA isoforms whereas testis is strongly biased towards short 3'UTR isoforms[80]. In mammals, the *CALCA* gene produces the hormone calcitonin prevalently in thyroid, but a different product, CGRP, in hypothalamus, by utilizing its distal PAS as mentioned before[61]. Similar global tissue-specific APA patterns were documented in zebrafish and human[23,24].

Additionally, components of the pre-mRNA 3′ end processing complex also can function in a tissue specific manner. For example, τCstF-64 (*Cstf2t*) is expressed in germ line cells and brain. Knockout of this factor delays accumulation of the testis-specific histone (H1fnt) in the germ line and causes infertility in male mice, but also increases memory retention in spatial learning tasks for female mice[81,82]. CstF-64 expression level in mouse primary B cells can regulated IgM heavy chain switch, increasing of which can lead to produce IgM in the secreted form (µs) instead of the membrane-bound form (µm)[83].

Some RBPs regulating PAS cleavage strength regulation showed a special regulatory role in neurons where genes often have longer 3'UTR to fulfill different functions compared to other cell types. Nova, a neuron-specific splicing factor binding to the YCAY motif, can enhance the usage of distal PASs. Knockout of Nova can globally increase the utilization of proximal PAS[84]. Embryonic-lethal abnormal visual (ELAV) in *Drosophila melanogaster* or HUR in mammal can repress proximal PAS usage or selectively block PAS with U-rich sequences, causing 3' UTR extensions during neural development[85,86]. Its repression mechanism is unknown, but might relate to ELAV's interaction with polII during transcription. TAR DNA-binding protein 43 (TDP43) and FUS are responsible for the neurodegenerative disease amyotrophic lateral sclerosis (ALS) and also influence APA. Evidence shows that FUS can interact with CPSF and CstF complex and bind to the carboxy-terminal domain (CTD) of the largest subunit of Pol II which affects the transcription. Interestingly, these findings indicate transcription itself can influence polyadenylation as well.

Regulation of transcription, splicing and PAS choice

The transcription factor MAZ specific G-rich sequence can lead to Pol II pausing and enhance upstream PAS cleavage efficiency[87]. On the other hand, transcription elongation factors such as SPT5, TFIIS and RPB2 can promote distal PAS usage. When these factors are mutated, cells exhibit slower elongation rates for transcription and produce shorter and early terminated nascent mRNA[88]. Recently, Liu's work in the *RpII215^{C4}* mutant *Drosophila melanogaster*, which possesses a mutant RNA polymerase II (RNAPII) with a decreased elongation rate, showed that there is a tissue-specific impact of the transcription elongation rate. They observe that a slower elongation rate can cause proximal or weaker PAS utilization in most tissues, but that this factor seems not to affect isoform expression in neurons in which APA appears to be strongly regulated by a different tissue-specific mechanism[89].

In addition, since splicing and polyadenylation are tightly connected to each other during transcription, factors involved in alternative splicing can also regulate APA. For intronic PASs, their interaction directly determines the usage of the PAS located. Plenty of evidence showed that inhibition of splicing factors could lead to strong increase of the intronic PAS usage[77]. For PASs located in other region, splicing and polyadenylation factors can couple with mutual reinforcement. For instance, U1 snRNP-A protein (U1A) can interact with CPSF160 to enhance SV40 polyadenylation efficiency[90]. CPSF and the U2 snRNP are cooperated in pre-mRNA 3' end processing and splicing so that removal of CPSF100 *in vitro* or impairment of U2 snRNP binding inhibits both splicing and 3' end cleavage[91]. Similar factors in promoting mRNA exportation via interacting with Nuclear export factor 1 (NXF1), SRSF3 and SRSF7 can function oppositely for splicing and polyadenylation. The first one will decrease exon inclusion and decrease the 3'UTR length, whereas the other one promotes exon inclusion and leads to an extension of 3' UTRs[92].

Evolution of APA

Apparently, the role of APA in regulating RNA isoform expression and thereby affecting mRNA stability, localization, translation and even protein localization indicates its great potential for regulating gene expression in different conditions or in different cell types. Since changes of gene expression are associated with morphological, physiological, or behavioral differences across species and could function as a driving force for pleiotropic phenotype differences[93], APA might also play an essential role in shaping species- or lineage-specific phenotypes, adaptive organismic traits and therefore evolutionary fitness.

According to conservation studies in mammals, most of the *trans*-factors involved in polyadenylation are well conserved between human and yeast with only a few exceptions such as CFI in mammals and Hrp1 in yeast[94]. In contrast, cis-elements differ significantly in the sequences as well as locations between metazoan and yeast[95]. But if limited in mammals, genes' PASs have similar polyadenylation signals enriched in PAS flanking region and exhibit good conservation among species such as chicken, mouse, rat and human[96]. These conserved sites are also found with higher processing efficiency than non-conserved sites[97]. Especially, polymorphism analysis in polyadenylation signals of PASs indicated that PASs located in distal 3'UTR are under strong selection and are more conserved in mammals compared to proximal PAS[98]. Recently, a comprehensive study of APA conservation[67] showed that globally over 80% mRNAs have at least one conserved PAS and more than half of all genes possess multiple conserved PASs to construct similar APA patterns in genes in mammals. These genes with multiple conserved PASs exhibit high average gene expression levels as well as tissue specificity in brain and muscle.

However, in a recent report Xu and Zhang propose the “error hypothesis” stating that most genes have only one optimal polyadenylation site and that APA is caused largely by deleterious polyadenylation errors[99]. They show a consistent trend that genes with lower expression levels exhibit higher APA diversity in multiple tissues from five mammals. They also claim that most APA is deleterious and under negative selection, based on the low PAS number observed compared to the larger number of pseudo-PASs in the complementary strand.

According to the neutral theory of molecular evolution, most intraspecific polymorphisms and interspecific divergences in DNA sequences are effectively neutral, and their patterns across different genes reflect the relative strengths of purifying selection and genetic drift[100]. In extension of the neutral theory of molecular evolution, Zhang proposes that the fraction of fitness-related phenotypic traits can be different on various levels in the hierarchy of biological organization. Higher level phenotypes such as organismal traits are more likely to be adaptive or beneficial, whereas the majority of molecular traits such as APA belong to effectively neutral categories[101]. Consistently, observation in murine and human naive and activated T cells failed to show any correlation between 3'UTR length and corresponding changes in gene expression[102]. Also, limited effects of 3'UTR in mRNA's stability and translational efficiency are detected in most genes of mouse fibroblast. Although mRNA isoforms with distal PASs generally have significantly lower stability and greater translational efficiency, different choices between PASs in the 3'UTR only account for slight effects in posttranscriptional regulation when compared to the attributes from other elements existing in the 5'UTR or coding sequence[47].

Approach in studying *cis*- and *trans*- regulation in APA evolution

Since changes in *cis*-regulatory elements and/or *trans*-acting factors in conjunction result in APA divergence, it is impossible to tell which one causes differences in usage of specific PASs simply by studying APA patterns in a single strain or by just comparing two species with each other. Therefore, a hybrid model together with the parental strains are often applied to effectively distinguish influence from *cis*-elements and *trans*-factors. The underlying logic is the following: the APA divergence between two parental strains is due to both *cis*- and *trans*- differences, while in their F1 hybrid only *cis*- differences affect allelic APA divergence, because both alleles share the same *trans*- regulatory environments. After acquiring the divergence only caused by *cis*- difference in the F1 hybrid, the *trans*- effect on parental divergent APA can be inferred by comparing the differences between the two parental strains to those between alleles in the F1 hybrid. Applying F1 hybrids studies in Saccharomyces yeasts[100], Drosophila[103] and mice[104] have unveiled different strengths between *cis*-elements and *trans*-factors in allele-specific gene expression and splicing regulation.

In our lab, we use a hybrid system consisting of two evolutionarily diverged mouse strains (*Mus musculus* C57BL/6J and *Mus spretus* SPRET/EiJ) and their F1 hybrid to study both alternative splicing and alternative polyadenylation[105,106]. These two parental strains have about 1.5 million years of divergent evolutionary history, resulting in more than 35.4 million SNPs and 4.5 million insertions and deletions (indels) between their genomes[107]. Their significant genomic differences, on average one SNP or indel per 75bp, enable us to a great extent to unambiguously assign the allelic reads from mRNA-seq or 3'mRNA seq data. Hence, the allelic PAS usage and gene expression levels can be precisely quantified. Our previous global APA divergence analysis in F1 hybrid fibroblasts found that among 24,721 PASs there are 3,747 (15.2%) PASs with significant usage differences between the two strains' alleles, 1876 (50.1%) of which exhibit *cis*- dominant divergence in contrast to only 572 (15.3%) *trans*-divergence. In all five different types of PASs, classified according to their annotated position, *cis*-contributions dominate. Similar findings were observed also in alternative splicing[106]. The much higher prevalence of *cis*-divergence indicates that evolutionary variants with local effects for both APA and alternative splicing in RNA processing were more likely to be tolerated than those with potentially pleiotropic consequences.

In order to clarify the general and tissue-dependent function as well as regulation of APA and its evolution in mammals, we applied 3'mRNA sequencing for multiple tissues of an F1 hybrid between the C57BL/6J (*Mus musculus*) and SPRET/EiJ (*Mus spretus*) mouse strains. We analyzed the factors regulating APA diversity and addressed the question whether APA is generally non-adaptive as proposed by the error hypothesis. In this study, we quantified all annotated PASs in nine tissues of the F1 hybrid mouse and comprehensively characterized different features of single-PAS genes and multi-PAS genes. Next, we checked the positional effects on PAS strength and discussed the functional difference between rank 1 and rank 2 PASs among distinct gene groups. We found that while in general APA of multi-PAS genes is consistent with the error hypothesis, some minor PASs appears to be functional. Additionally, a small fraction of genes is regulated in a strongly tissue-dependent manner and possesses functional minor PASs. By quantifying PAS usage in each allele, we studied the genes with divergent isoform expression levels and their dN/dS ratios. Comparing these with genes showing allelic mRNA transcript level differences, we unveiled different evolutionary patterns between APA and gene expression.

Results

Quantification of alternative polyadenylation patterns across tissues in F1 hybrid mice

To study APA regulation across different tissues, we quantified the PAS usage in eight organs (cortex, cerebellum, heart, muscle, lung, liver, kidney, spleen) as well as embryonic stem cells (ESC) from an F1 hybrid cross between the C57BL/6J and SPRET/EiJ mouse strains using Lexogen 3' mRNA Rev sequencing, a technique only targeting the 3' ends of mRNAs with poly(A) tails (Methods). For each tissue, we sequenced three replicates with an average of about 40 million reads per sample, of which approximately 70% uniquely mapped to at least one of the reference genomes (Methods and Table 1). To assure all reads were derived from the mRNA's 3' end, we further required mapped reads to be located within 24 nucleotides away from the cleavage position of a PAS annotated in the polyA_DB3 database[108]. After filtering, on average, 17.15 million reads (about half of the uniquely mapped reads) could be assigned to one of the annotated PASs (hereafter termed “PAS reads”). Among these reads, an average of 8.14 million reads (47.5%) could be unambiguously assigned to either the SPRET/EiJ or the C57BL/6J genome (see Methods and Table 1), and were later used for allelic PAS usage analysis.

We then counted the number of PAS reads for each PAS and the total number of PAS reads mapped to each Refseq protein-coding gene. A gene was considered to be expressed in a tissue only if it was covered by at least 20 PAS reads in each of the three replicates. In order to avoid sampling error in PAS identification caused by differences in sequencing depth between genes, we confirmed all results with those of a downsampled data set, randomly picking 20 reads from each gene to analyze the gene's PAS usage (Methods). In total, 13,369 protein coding genes were expressed in at least one tissue (Table 2). Among them, 1,859 (13.90%) genes were single-PAS genes, expressing only one identical PAS across the nine tissues (2,133, 15.95% genes in the downsampled data). 77,213 PASs are detected in nine tissues (67,047 PAS in downsampled data, see Table 3), nearly half of which has higher than 5% usage in at least one tissues. On average, there are 5.8 PAS per gene of which 4.1 PAS located in 3'UTR per gene. These numbers are similar in different tissues. After downsampling, there are approximately 3.6 PAS and 2.8 PAS in the 3'UTR per gene. The PAS with the maximum average usage across all tissues in a multi-PAS gene was defined as a gene's "major PAS". Most of the multi-PAS genes' major PASs (10,896, 94.67%) are located in the 3'UTR, whereas only few genes (614, 5.33%) have their major PAS located upstream of the annotated last stop codon. Among the major PASs located in the 3'UTR, 1139 (10.45%) are the only PAS in the 3'UTR (3'UTR(S)), 2,314 (21.24%) are located in the most proximal position to the stop codon (3'UTR(F)), 4,201 (38.56%) in the most distal position (3'UTR(L)), and 3,242 (29.75%) in between (3'UTR(M)).

Table 1. Sequences reads number for nine tissues with three replicates

Tissues	Sequencing reads	Unique mapping reads to F1 genome	Reads mapped to annotated PAS	Reads assigned to alleles	Reads mapped to protein coding genes	Mapped reads assigned to BL6	Mapped reads assigned to SPR
ESC_1	38,310,730	23,751,837	15,947,315	7,412,455	12,816,562	3,638,874	3,421,311
ESC_2	35,091,726	22,394,965	15,005,588	7,498,364	12,025,935	3,677,664	3,459,803
ESC_3	21,681,581	15,338,107	10,446,957	4,914,733	8,311,962	2,415,183	2,272,770
Cerebellum_1	54,505,336	36,195,485	16,255,602	6,886,854	12,451,770	3,319,919	3,228,518
Cerebellum_2	62,398,509	41,470,979	20,299,148	9,556,173	16,082,283	4,604,615	4,466,850
Cerebellum_3	49,281,118	32,405,466	15,724,700	7,052,146	12,151,824	3,460,752	3,244,753
Cortex_1	51,514,545	29,840,400	15,199,678	6,325,032	11,615,472	3,044,764	2,968,090
Cortex_2	47,556,684	31,571,326	14,930,980	6,899,155	11,571,755	3,317,859	3,204,280
Cortex_3	60,813,146	41,415,226	21,981,871	10,146,001	17,427,532	4,998,186	4,635,994
Heart_1	39,198,588	27,717,264	18,155,775	9,303,786	15,193,531	4,396,249	4,549,960
Heart_2	37,603,240	25,237,822	16,492,023	7,774,114	13,326,095	3,700,408	3,775,168
Heart_3	41,424,038	28,660,969	17,738,042	8,490,478	14,262,853	4,125,355	4,030,999
Kidney_1	35,937,583	26,829,182	16,433,110	8,653,710	13,500,139	4,169,138	4,106,340
Kidney_2	37,532,445	26,426,207	16,212,675	8,665,554	13,119,943	4,152,133	4,116,348
Kidney_3	42,986,429	32,883,284	19,859,440	10,412,048	16,434,209	5,069,306	4,886,733
Liver_1	35,273,651	24,650,841	16,809,143	8,539,118	13,437,240	4,313,275	3,976,835
Liver_2	34,539,447	23,674,845	16,465,203	8,270,815	13,036,233	4,045,911	3,981,507
Liver_3	38,616,805	27,977,797	18,834,656	9,362,280	15,227,974	4,567,090	4,522,081
Lung_1	39,569,519	30,034,230	17,944,172	7,402,410	13,268,641	3,683,285	3,382,032
Lung_2	42,860,286	31,327,479	18,004,532	7,768,580	13,104,953	3,830,607	3,584,761
Lung_3	44,752,188	32,989,887	19,089,545	8,401,752	13,981,643	4,223,550	3,790,506
Muscle_1	33,885,070	25,768,338	17,804,093	8,893,782	13,448,180	4,508,899	4,183,531
Muscle_2	35,618,364	28,209,331	18,196,292	9,550,882	14,782,023	4,794,354	4,563,239
Muscle_3	37,610,304	28,760,316	19,131,365	9,624,880	14,665,449	5,031,716	4,365,346
Spleen_1	45,270,912	27,529,476	15,088,769	6,540,080	11,396,432	3,058,529	2,954,623
Spleen_2	45,438,308	33,121,258	18,709,988	8,252,102	14,098,293	3,831,834	3,701,635
Spleen_3	42,932,761	30,359,884	16,237,077	7,305,359	12,521,011	3,496,496	3,239,201
Average	41,933,456	29,131,193	17,148,064	8,144,542	13,454,072	3,980,591	3,800,489

Table 2. Quantification of PASs for protein coding genes across tissues.

Tissues	PAS # (PAS # with usage >=5%)	Average PAS # per gene (median)	Average PAS # in UTR per gene (median)	# of expressed genes in tissue	# of multi-PAS genes (# of genes with only one PAS detected in tissue)	# of single PAS genes*
ESC	52,037 (19,966)	5.7 (5)	4.1 (3)	9,227	8,409 (240)	799
Cerebellum	65,108 (23,541)	5.9 (5)	4.3 (3)	11,052	10,010 (157)	1,014
Cortex	64,634 (23,171)	5.9 (5)	4.3 (3)	10,903	9,916 (142)	959
Heart	58,276 (20,743)	5.8 (5)	4.3 (3)	10,007	9,144 (220)	840
Kidney	62,826 (22,604)	5.9 (5)	4.3 (3)	10,675	9,593 (131)	1,055
Liver	53,149 (19,746)	5.7 (5)	4.2 (3)	9,330	8,489 (213)	821
Lung	61,793 (23,539)	5.6 (4)	4.1 (3)	11,022	9,865 (215)	1,129
Muscle	48,994 (18,788)	5.6 (5)	4.1 (3)	8,750	8,044 (238)	685
Spleen	60,828 (22,670)	5.8 (5)	4.2 (3)	10,556	9,491 (203)	1,039
Total	76,970 (28,354)	5.8 (4)	4.1 (3)	13,369	11,510	1,859

*Single-PAS genes are those for which only one identical PAS is detected in all tissues in which the gene is expressed

Table 3. Quantification of PASs for protein coding genes across tissues (Downsampled data).

Tissues	PAS #	Average PAS # per gene (median)	Average PAS # in UTR per gene (median)	# of expressed genes in tissue	# of multi-PAS genes (# of genes with only one PAS detected in tissue)	# of single PAS genes*
ESC	33,259	3.6 (4)	2.8 (2)	9,204	8,218 (851)	986
Cerebellum	39,171	3.6 (4)	2.8 (2)	11,018	9,797 (945)	1,221
Cortex	38,778	3.6 (4)	2.8 (2)	10,872	9,696 (987)	1,176
Heart	34,991	3.5 (4)	2.8 (2)	9,980	8,953 (1058)	1,027
Kidney	38,021	3.6 (4)	2.8 (2)	10,641	9,396 (934)	1,245
Liver	33,245	3.6 (4)	2.8 (2)	9,307	8,306 (854)	1,001
Lung	38,781	3.5 (4)	2.7 (2)	10,991	9,657 (975)	1,334
Muscle	30,956	3.5 (4)	2.7 (2)	8,729	7,868 (886)	861
Spleen	37,939	3.6 (4)	2.8 (2)	10,525	9,286 (870)	1,239
Total	66,826	5.0 (4)	3.6 (3)	13,369	11,236	2,133

*Single-PAS genes are those for which only one identical PAS is detected in all tissues in which the gene is expressed

Single PAS genes and multiple PAS genes

As previous studies indicate that single-PAS and multi-PAS genes are two distinct classes of genes[109,110], we further examined their different features in our data. Consistent with previous findings, single-PAS genes have on average shorter UTR lengths (Fig. 1A) and are expressed in fewer tissues (Fig. 1B, Table 4-5). These features remain when sequencing depth is controlled (see Methods and Fig. 2A). Compared to the major PAS of multi-PAS gene, single PAS gene's PAS upstream regions (50nt upstreaming of the cleavage site) exhibit lower sequence conservation in Glires (rodents and lagomorphs) PhastCons score (Fig. 1C). It possibly indicates a relaxation of selective constraints in this region. We also check the dN/dS (Mouse vs. Rat) for protein selective restraint between single-PAS and multi-PAS genes. The dN/dS ratio is significantly higher in single-PAS gene. Even after downsampling, 6 of 9 tissues still possesses significant higher dN/dS rate in single-PAS genes, indicating a protein selective restraint exists for single-PAS gene. When comparing gene expression, the original data shows a different result as in downsampled data (data not shown), indicating a strong bias in detection of single-PAS genes by lower sequencing depth. It is expected since increasing sequencing depth would increase the chance to detected a gene's minor PAS. Due to different features and sequencing bias for single PAS genes, we excluded them in our further analysis for APA diversity of genes.

Table 4. Expression breadth vs. number of PASs

# Tissues	T1	T2	T3	T4	T5	T6	T7	T8	T9
PAS #									
1	439	331	183	111	107	72	87	125	404
2	246	218	166	124	94	104	118	176	598
3	126	176	136	95	83	70	98	134	679
4	54	119	78	74	61	74	73	166	698
5	22	77	67	60	57	51	77	159	679
6	9	41	45	38	40	41	63	111	613
>=7	16	108	119	105	125	91	187	415	3,256
Fraction of single-PAS genes	48%	31%	23%	18%	19%	14%	12%	10%	6%

Tx: x is the number of tissues in which the gene is expressed.

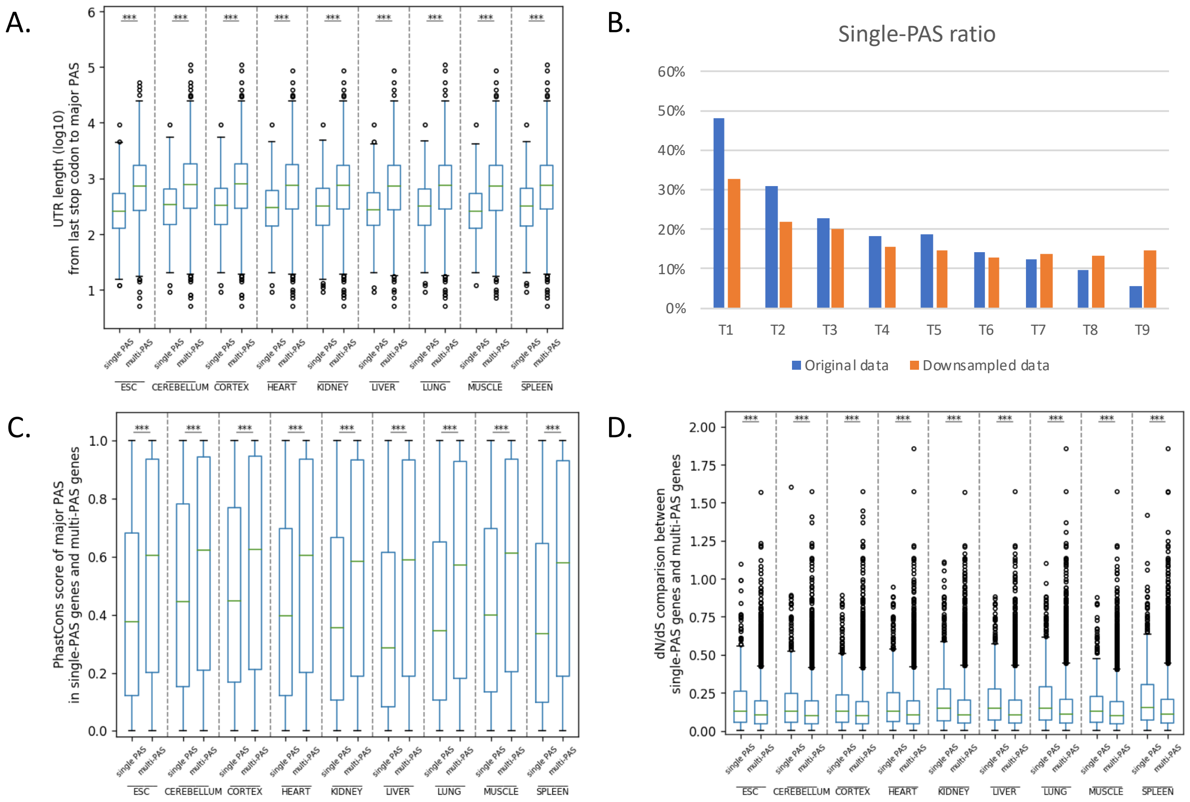


Figure 1. Different features between single-PAS genes and multi-PAS genes. A) 3'UTR length (distance from gene last annotated stop codon to major PAS) comparison between single-PAS genes and multi-PAS genes. B) Expression breadth comparison between single-PAS genes and multi-PAS genes. Tx: x indicates the number of tissues in which the gene is expressed. C) PhastCons score of major PAS region (-50nt, 0nt) comparison between single-PAS genes and multi-PAS genes for each tissue. D) Comparison in dN/dS (Mouse vs. Rat) between single-PAS genes and multi-PAS genes for each tissue. Mann-Whitney U test was used to determine the statistical significance (*** indicates $p<0.001$; ** $p<0.01$; * $p<0.05$)

Table 5. Expression breadth vs. number of PASs (Downsampled).

# Tissues	T1	T2	T3	T4	T5	T6	T7	T8	T9
PAS #									
1	494	352	163	100	96	88	118	179	578
2	347	385	148	134	124	116	143	210	667
3	252	263	133	103	122	104	122	227	639
4	141	192	106	92	80	97	124	189	529
5	94	123	94	70	73	76	96	151	441
6	68	85	44	47	41	58	70	105	307
>=7	106	193	116	98	118	144	189	289	793
Fraction of single-PAS genes	33%	22%	20%	16%	15%	13%	14%	13%	15%

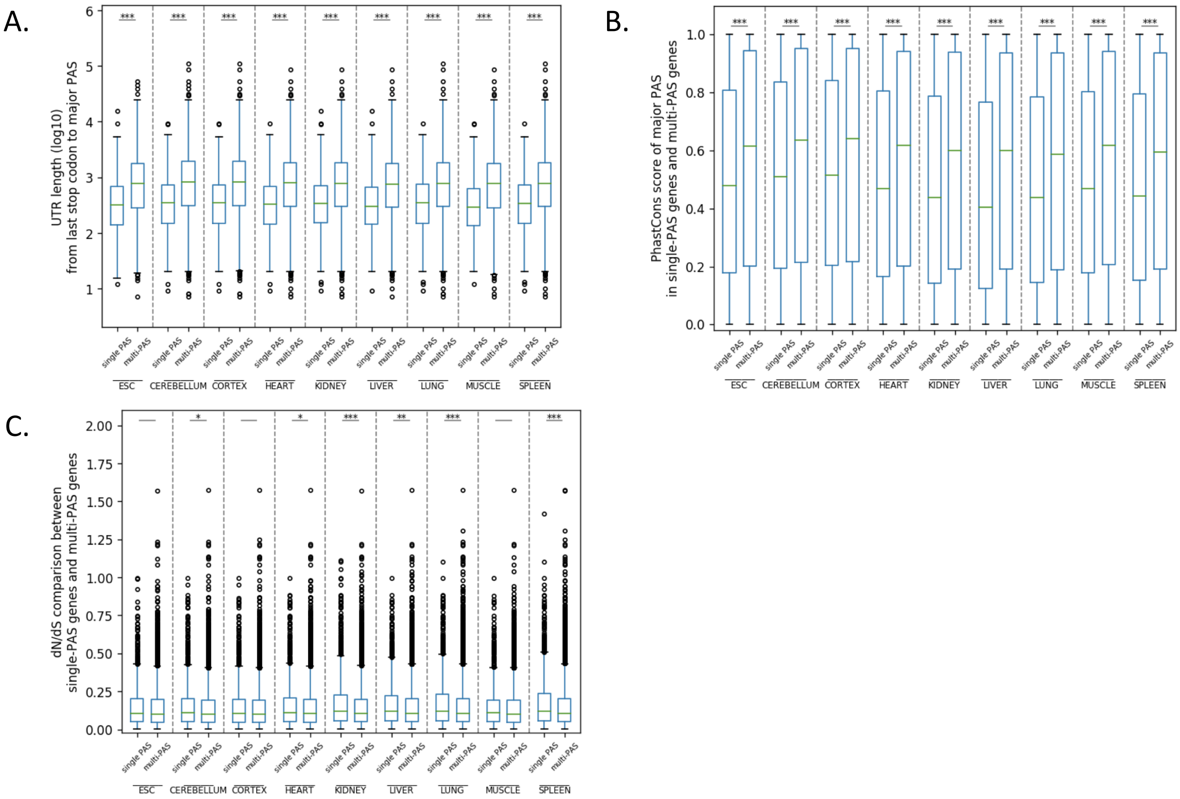


Figure 2. Different features between Single-PAS genes and multi-PAS genes in Downsampled data. A) 3'UTR length (distance from gene last annotated stop codon to major PAS) comparison between single-PAS genes and multi-PAS genes. B) PhastCons score of major PAS region (-50nt, 0nt) comparison between single-PAS genes and multi-PAS genes for each tissue. C) Comparison in dN/dS (Mouse vs. Rat) between single-PAS genes and multi-PAS genes for each tissue. Mann-Whitney U test was used to determine the statistical significance (*** indicates p<0.001; ** p<0.01; * p<0.05)

APA diversity within tissues

To test error hypothesis that genes with lower expression levels are expected to have higher APA diversity than highly expressed genes, we classified the multi-PAS genes according to a gene's Shannon index, which measures the entropy of the alternative PAS usage reflecting both the number of different PASs and the evenness of the usage distribution across these sites (see Methods), or according to the usage of their dominant PAS, for each tissue separately. We then compared the distribution of mRNA expression for the different gene groups in the respective tissue. As shown in Fig. 3A-B, in Liver (same results in other 8 tissues), most multi-PAS genes with higher entropy and lower dominant PAS usage expressed at lower level. Significantly negative correlations between gene's diversity (increase of gene's Shannon index or decrease of gene's dominant PAS usage) and gene expression level were detected in all tissues except spleen (Fig. 3C-D, Table 6). To avoid sequencing depth influence on PAS usage calculation, we applied downsampled dataset to recalculate PAS usage and Shannon index, and found same trend that genes with lower APA diversity have higher gene expression level (Fig. 4A-D, Table 7).

Table 6. Spearman correlation between gene expression level and APA diversity for multi-PAS genes. Genes are divided into high and low APA diverse genes according to their dominant PAS usage (DP) (less than 90% or at least 90%).

Dominant PAS usage and gene expression level	Spearman-rho			p-value		
	ALL	DP*<0.9	DP>=0.9	ALL	DP<0.9	DP>=0.9
ESC	0.15	0.04	0.10	<2.2e-16	1.5e-03	5.8e-08
Cerebellum	0.10	0.01	0.12	<2.2e-16	0.63	2.2e-12
Cortex	0.11	0.02	0.11	<2.2e-16	0.09	4.6e-12
Heart	0.14	0.02	0.12	<2.2e-16	0.21	2.8e-13
Kidney	0.13	0.02	0.14	<2.2e-16	0.06	<2.2e-16
Liver	0.12	0.01	0.09	<2.2e-16	0.30	2.9e-07
Lung	0.09	0.00	0.10	<2.2e-16	0.77	8.0e-09
Muscle	0.18	0.08	0.15	<2.2e-16	7.9e-08	4.0e-16
Spleen	0.08	-0.02	0.06	2.3e-14	0.22	2.4e-04
Shannon index and gene expression level	Spearman-rho			p-value		
	ALL	DP<0.9	DP>=0.9	ALL	DP<0.9	DP>=0.9
ESC	-0.13	-0.01	-0.06	<2.2e-16	0.57	1.0e-03
Cerebellum	-0.07	0.04	-0.07	9.6e-13	1.92E-03	5.4e-06
Cortex	-0.09	0.02	-0.07	<2.2e-16	0.16	6.7e-06
Heart	-0.13	0.00	-0.09	<2.2e-16	0.91	5.2e-08
Kidney	-0.12	0.00	-0.11	<2.2e-16	0.77	1.3e-10
Liver	-0.11	0.00	-0.06	<2.2e-16	0.85	2.4e-04
Lung	-0.07	0.06	-0.05	5.8e-11	1.01E-05	1.2e-03
Muscle	-0.17	-0.06	-0.12	<2.2e-16	8.60E-06	9.5e-11
Spleen	-0.06	0.05	-0.03	6.7e-09	1.59E-04	0.11

*DP: Dominant PAS usage

Table 7. Spearman correlation between gene expression level and APA diversity for multi-PAS genes in downsampled data.

Dominant PAS usage and gene expression	Spearman-rho			p-value		
	ALL	DP<0.9	DP>=0.9	ALL	DP<0.9	DP>=0.9
ESC	0.14	0.05	0.10	<2.2e-16	5.3e-04	3.5e-08
Cerebellum	0.08	0.00	0.06	1.2e-14	0.82	1.1e-04
Cortex	0.09	0.02	0.10	<2.2e-16	0.14	1.6e-09
Heart	0.12	0.02	0.11	<2.2e-16	0.07	4.8e-10
Kidney	0.11	0.04	0.11	<2.2e-16	4.2e-03	3.0e-10
Liver	0.10	0.01	0.08	<2.2e-16	0.64	3.8e-06
Lung	0.08	-0.01	0.05	4.2e-14	0.43	1.5e-03
Muscle	0.16	0.06	0.10	<2.2e-16	7.7e-06	7.7e-08
Spleen	0.07	0.01	0.10	4.8e-11	0.70	1.1e-08
Shannon index and gene expression level	Spearman-rho			p-value		
	ALL	DP<0.9	DP>=0.9	ALL	DP<0.9	DP>=0.9
ESC	-0.13	-0.03	-0.09	<2.2e-16	0.01	2.2e-06
Cerebellum	-0.07	0.02	-0.05	1.3e-10	0.06	2.3e-03
Cortex	-0.09	-0.01	-0.09	<2.2e-16	0.60	6.6e-08
Heart	-0.12	-0.02	-0.10	<2.2e-16	0.10	4.0e-09
Kidney	-0.10	-0.04	-0.10	<2.2e-16	5.7e-03	1.5e-08
Liver	-0.11	-0.02	-0.08	<2.2e-16	0.10	1.3e-05
Lung	-0.06	0.04	-0.04	1.1e-09	1.3e-03	0.03
Muscle	-0.16	-0.07	-0.09	<2.2e-16	2.79E-06	2.7e-07
Spleen	-0.06	0.00	-0.08	1.6e-09	0.75	1.2e-06

Additionally, we used average gene expression across all tissues to represent the gene's expression rank in mouse for the spearman correlation with APA diversity and observed more consistent and strong negative correlations between gene expression and APA diversity in all tissues (Table 8-9). This difference was most pronounced between genes with a dominant PAS usage above or equal to 90% and those below that level. Within the latter group, the positive correlation between gene expression and dominant PAS usage was still observable in most tissues, but much weaker, whereas the negative correlation between gene expression and the Shannon index largely disappeared and even reverted in some tissues.

Table 8. APA diversity in each tissue strongly correlates with average expression level across tissues.

Dominant PAS usage and average gene expression	Spearman-rho			p-value		
	ALL	DP<0.9	DP>=0.9	ALL	DP<0.9	DP>=0.9
ESC	0.16	0.03	0.19	<2.2e-16	0.04	<2.2e-16
Cerebellum	0.17	0.04	0.21	<2.2e-16	2.7e-03	<2.2e-16
Cortex	0.18	0.05	0.21	<2.2e-16	1.9e-04	<2.2e-16
Heart	0.15	0.02	0.18	<2.2e-16	0.23	<2.2e-16
Kidney	0.17	0.03	0.21	<2.2e-16	0.04	<2.2e-16
Liver	0.14	0.00	0.17	<2.2e-16	0.97	<2.2e-16
Lung	0.16	0.02	0.18	<2.2e-16	0.05	<2.2e-16
Muscle	0.14	0.04	0.16	<2.2e-16	1.8e-03	<2.2e-16
Spleen	0.15	0.01	0.17	<2.2e-16	0.44	<2.2e-16
Shannon index and average gene expression level	Spearman-rho			p-value		
	ALL	DP<0.9	DP>=0.9	ALL	DP<0.9	DP>=0.9
ESC	-0.15	-0.01	-0.16	<2.2e-16	0.38	<2.2e-16
Cerebellum	-0.16	-0.02	-0.17	<2.2e-16	0.07	<2.2e-16
Cortex	-0.17	-0.03	-0.18	<2.2e-16	0.02	<2.2e-16
Heart	-0.15	-0.01	-0.16	<2.2e-16	0.33	<2.2e-16
Kidney	-0.16	-0.01	-0.18	<2.2e-16	0.32	<2.2e-16
Liver	-0.14	0.01	-0.14	<2.2e-16	0.70	3.3e-16
Lung	-0.14	0.01	-0.14	<2.2e-16	0.36	<2.2e-16
Muscle	-0.13	-0.03	-0.14	<2.2e-16	0.03	2.0e-14
Spleen	-0.13	0.02	-0.14	<2.2e-16	0.15	<2.2e-16

Table 9. Gene APA diversity strongly correlated with average expression level in each tissue in downsampled data.

Dominant PAS usage and average gene expression	Spearman-rho			p-value		
	ALL	DP<0.9	DP>=0.9	ALL	DP<0.9	DP>=0.9
ESC	0.13	0.04	0.13	<2.2e-16	2.8e-03	9.3e-13
Cerebellum	0.14	0.04	0.12	<2.2e-16	2.6e-03	4.0e-13
Cortex	0.16	0.04	0.14	<2.2e-16	5.9e-04	6.3e-18
Heart	0.12	0.02	0.12	<2.2e-16	0.26	3.8e-13
Kidney	0.14	0.04	0.16	<2.2e-16	9.2e-04	<2.2e-16
Liver	0.11	-0.01	0.10	<2.2e-16	0.55	5.3e-09
Lung	0.12	0.02	0.08	<2.2e-16	0.14	1.9e-06
Muscle	0.11	0.04	0.10	<2.2e-16	6.3e-03	7.0e-08
Spleen	0.12	0.03	0.14	<2.2e-16	0.02	2.3e-16
Shannon index and average gene expression level	Spearman-rho			p-value		
	ALL	DP<0.9	DP>=0.9	ALL	DP<0.9	DP>=0.9
ESC	-0.13	-0.04	-0.12	<2.2e-16	3.8e-03	1.3e-11
Cerebellum	-0.14	-0.04	-0.11	<2.2e-16	5.1e-03	2.2e-11
Cortex	-0.16	-0.05	-0.13	<2.2e-16	6.0e-05	8.0e-16
Heart	-0.13	-0.03	-0.12	<2.2e-16	0.05	5.9e-12
Kidney	-0.15	-0.05	-0.15	<2.2e-16	2.0e-04	<2.2e-16
Liver	-0.12	-0.01	-0.10	<2.2e-16	0.44	3.7e-08
Lung	-0.12	0.00	-0.07	<2.2e-16	0.98	4.6e-05
Muscle	-0.11	-0.05	-0.09	<2.2e-16	1.6e-03	3.1e-07
Spleen	-0.12	-0.02	-0.13	<2.2e-16	0.15	1.0e-13

Additionally, higher APA diversity within each tissue could be caused either by higher variability in APA between cells or substructures (due to higher diversity of cell types, more extensive spatiotemporal regulation or due to higher polyadenylation noise) or by consistent high entropy APA patterns across all the cells of a given tissue. While higher APA variability between cells could be reflected in increased differences between biological replicates, this would not be predicted in the case of consistent and tightly regulated APA diversity. As expected in the former case, we observe a positive correlation between a gene's entropy and its APA variability among three biological replicates (Fig. 3E, Table 10), further supporting the error hypothesis. Similar as mentioned above, genes with dominant PAS usage above 90% exhibit strong correlations between gene expression level and gene's APA diversity (dominant PAS usage and Shannon index) as well as APA variability among replicates in all tissues (Fig. 3E, Table 10).

Table 10. Correlation between APA variability in replicates and gene Shannon index.

APA variability in replicates and gene Shannon index	Spearman-rho			p-value		
	ALL	DP<0.9	DP>=0.9	ALL	DP<0.9	DP>=0.9
ESC	0.64	0.21	0.63	<2.2e-16	<2.2e-16	<2.2e-16
Cerebellum	0.41	0.13	0.35	<2.2e-16	<2.2e-16	<2.2e-16
Cortex	0.51	0.18	0.38	<2.2e-16	<2.2e-16	<2.2e-16
Heart	0.42	0.15	0.37	<2.2e-16	<2.2e-16	<2.2e-16
Kidney	0.40	0.15	0.28	<2.2e-16	<2.2e-16	<2.2e-16
Liver	0.45	0.17	0.41	<2.2e-16	<2.2e-16	<2.2e-16
Lung	0.57	0.19	0.56	<2.2e-16	<2.2e-16	<2.2e-16
Muscle	0.64	0.22	0.58	<2.2e-16	<2.2e-16	<2.2e-16
Spleen	0.29	0.12	0.23	<2.2e-16	<2.2e-16	<2.2e-16

Because APA variability is another index for APA noise within different tissues, we checked the correlation between APA variability among replicates and gene expression levels in each tissue. Due to different homogeneities of tissues, only in ESC, liver and lung exist strong negative correlations (Fig. 3F, Table 11). Additionally, we compared the dN/dS for genes with different APA diversity, but did not find a strong negative correlation between gene's dN/dS and APA diversity (Shannon index) or variability in replicates, which indicating that gene's APA diversity is not directly related to selective constraints on the protein coding sequence (Table 12).

Table 11. Correlation between gene expression and APA variability in replicates

APA variability in replicates and gene expression level	Spearman-rho			p-value		
	ALL	DP<0.9	DP>=0.9	ALL	DP<0.9	DP>=0.9
ESC	-0.16	-0.10	-0.12	<2.2e-16	6.1e-14	1.4e-10
Cerebellum	-0.04	-0.03	-0.04	4.9e-05	7.6e-03	0.02
Cortex	0.02	0.07	0.01	0.05	2.2e-08	0.61
Heart	-0.09	-0.10	-0.04	<2.2e-16	7.0e-14	0.02
Kidney	-0.05	-0.03	-0.05	3.6e-07	0.01	5.8e-03
Liver	-0.20	-0.23	-0.14	<2.2e-16	<2.2e-16	4.0e-16
Lung	-0.22	-0.25	-0.15	<2.2e-16	<2.2e-16	<2.2e-16
Muscle	-0.14	-0.09	-0.08	<2.2e-16	1.3e-10	6.1e-06
Spleen	0.03	0.03	0.04	2.7e-03	0.01	0.02

APA variability in replicates and average gene expression level	Spearman-rho			p-value		
	ALL	DP<0.9	DP>=0.9	ALL	DP<0.9	DP>=0.9
ESC	-0.20	-0.14	-0.20	<2.2e-16	<2.2e-16	<2.2e-16
Cerebellum	-0.09	-0.04	-0.11	<2.2e-16	2.2e-03	6.7e-11
Cortex	-0.06	0.02	-0.05	1.4e-08	0.06	2.0e-03
Heart	-0.10	-0.08	-0.08	<2.2e-16	1.5e-08	9.9e-07
Kidney	-0.10	-0.06	-0.07	<2.2e-16	5.0e-07	3.2e-05
Liver	-0.20	-0.19	-0.18	<2.2e-16	<2.2e-16	<2.2e-16
Lung	-0.25	-0.23	-0.19	<2.2e-16	<2.2e-16	<2.2e-16
Muscle	-0.09	-0.01	-0.09	1.5e-15	0.35	7.9e-07
Spleen	-0.01	0.03	0.00	0.54	0.04	0.78

Table 12. No significant or strong correlation between dN/dS and APA diversity or variability within tissue separately.

dN/dS correlation with APA diversity or variability	Spearman-rho		p-value	
	Shannon index	APA variability	Shannon index	APA variability
ESC	-0.02	0.02	0.11	0.03
Cerebellum	0.02	0.00	0.05	0.72
Cortex	0.02	-0.01	0.02	0.39
Heart	-0.01	0.00	0.21	0.96
Kidney	-0.01	0.01	0.19	0.48
Liver	-0.03	0.02	0.01	0.04
Lung	-0.02	0.03	0.09	1.7e-03
Muscle	-0.03	-0.02	0.02	0.07
Spleen	-0.02	-0.04	0.09	1.7e-04

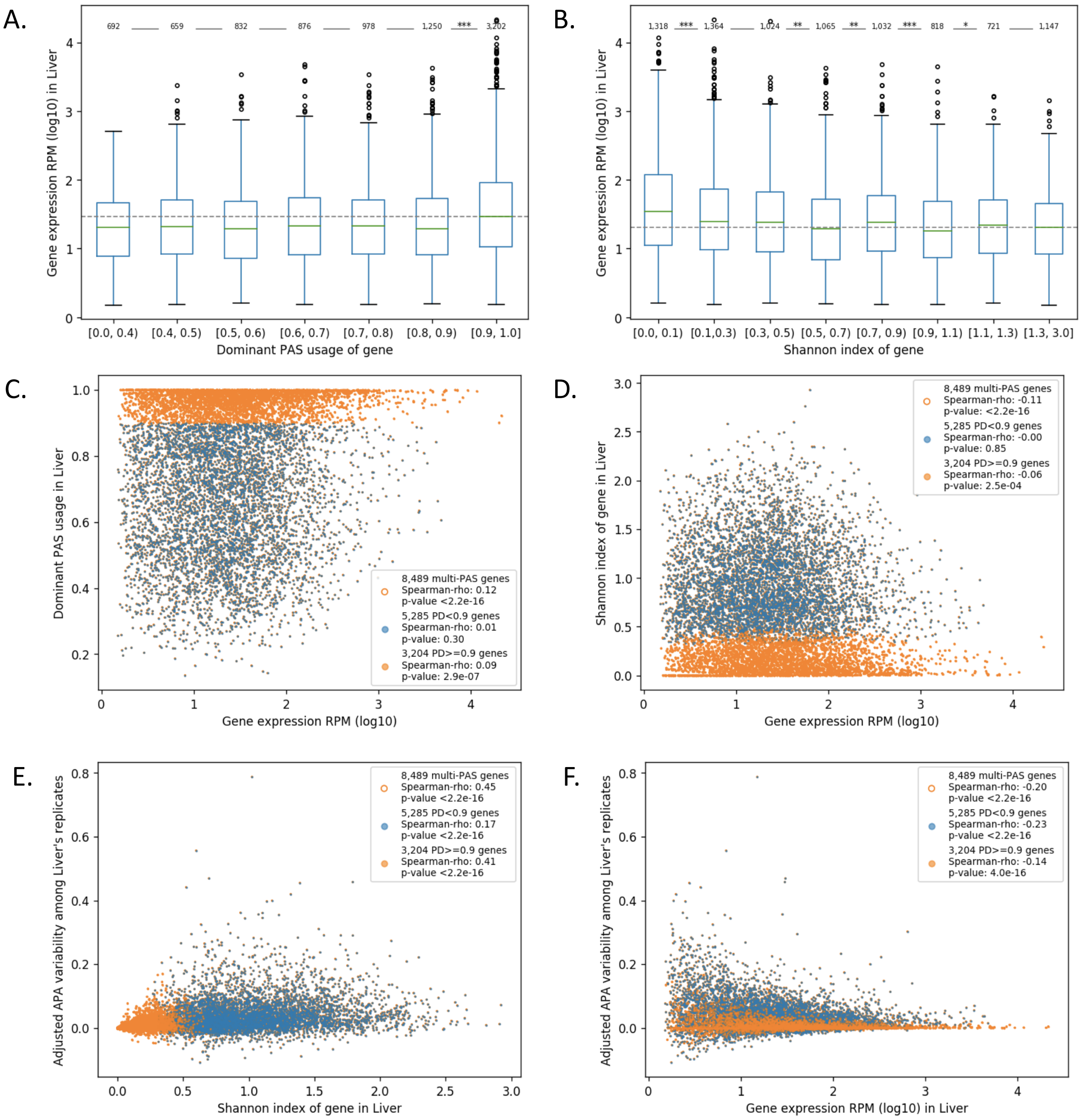


Figure 3. Gene with high APA diversity are lowly expressed. A) Gene expression level comparison among genes with different range of dominant PAS usage in liver. B) Gene expression level comparison among genes with different Shannon index in liver. C) Positive spearman correlation between gene expression level and gene's dominant PAS usage in liver. D) Negative spearman correlation between gene expression level and gene's Shannon index. E) Correlation between Shannon index and APA variability among replicates in liver. F) Correlation between gene expression level and APA variability among replicates in liver. Mann-Whitney U test was used to determine the statistical significance (*** indicates p<0.001; ** p<0.01; * p<0.05)

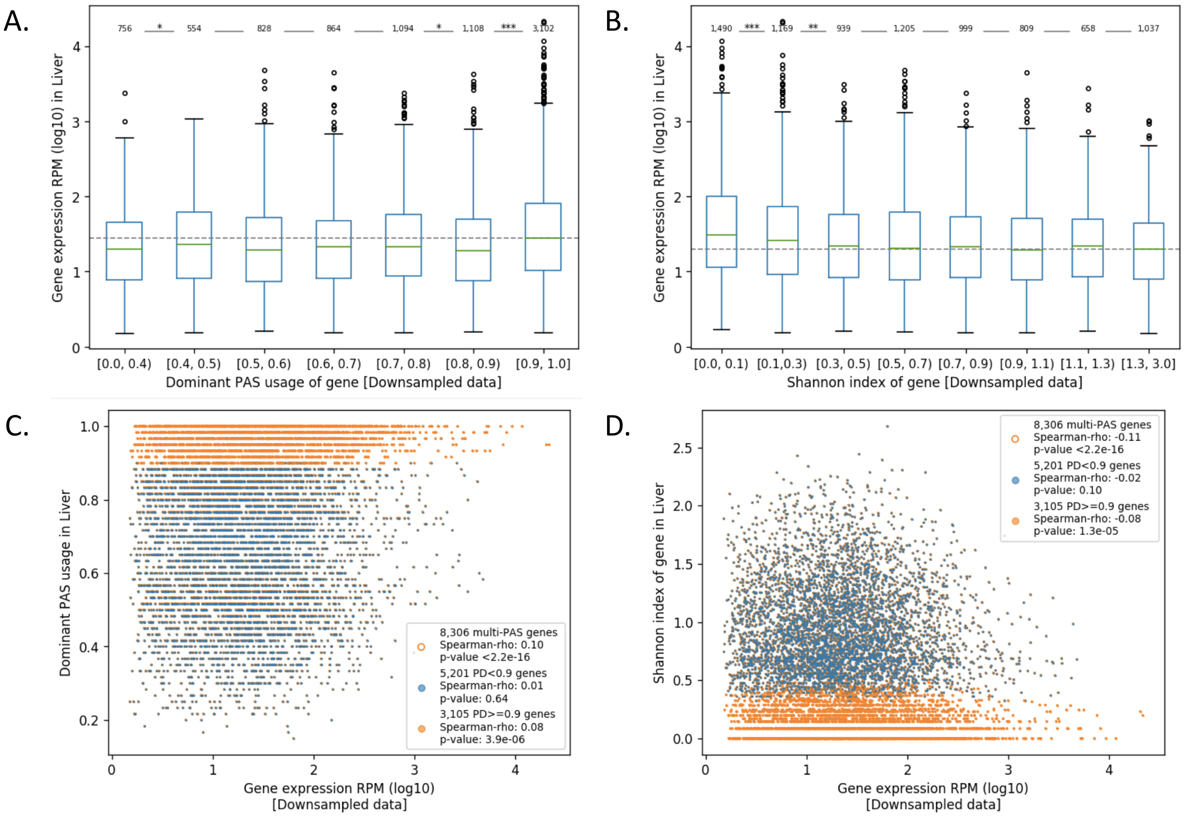


Figure 4. Gene with high APA diversity are lowly expressed in downsampled data A) Gene expression level comparison among genes with different range of dominant PAS usage in liver. B) Gene expression level comparison among genes with different Shannon index in liver. C) Positive spearman correlation between gene expression level and gene's dominant PAS usage in liver. D) Negative spearman correlation between gene expression level and gene's Shannon index. Mann-Whitney U test was used to determine the statistical significance (*** indicates p<0.001; ** p<0.01; * p<0.05)

APA diversity across tissues

Under the neutral hypothesis, we would expect genes with lower average expression levels to experience lower levels of purifying selection affecting APA, and therefore to exhibit higher polyadenylation diversity across tissues than highly expressed genes. To test this hypothesis, we assigned a “maximum switch score” for each gene by calculating the maximum pairwise APA difference between tissues in which the gene is expressed (see Methods). As predicted, we found that multi-PAS genes with higher maximum switch scores have slightly lower average expression levels and that a strong negative correlation between gene average expression level and maximum switch score exist in all genes especially for genes with maximum switch score not larger than 10% across all tissue pairs (Fig. 5A-B).

Next, we addressed the question whether the magnitude of APA differences between tissues (switch score) is related to the APA diversity within tissues and the variability between biological replicates of the same tissue. As expected under the error hypothesis, we found that genes with larger differences between tissues also have higher within-tissue entropy values and higher APA variability across biological replicates (Fig. 5C-D, Table 13-14).

Table 13. Gene's maximum switch score positively correlates with APA diversity and APA variability in each tissue.

Shannon index and maximum switch score	Spearman-rho			p-value		
	ALL	MSC*>0.1	MSC<=0.1	ALL	MSC>0.1	MSC<=0.1
ESC	0.71	0.27	0.71	<2.2e-16	<2.2e-16	<2.2e-16
Cerebellum	0.69	0.27	0.63	<2.2e-16	<2.2e-16	<2.2e-16
Cortex	0.70	0.28	0.63	<2.2e-16	<2.2e-16	<2.2e-16
Heart	0.74	0.32	0.74	<2.2e-16	<2.2e-16	<2.2e-16
Kidney	0.73	0.30	0.72	<2.2e-16	<2.2e-16	<2.2e-16
Liver	0.72	0.29	0.72	<2.2e-16	<2.2e-16	<2.2e-16
Lung	0.73	0.31	0.69	<2.2e-16	<2.2e-16	<2.2e-16
Muscle	0.72	0.29	0.73	<2.2e-16	<2.2e-16	<2.2e-16
Spleen	0.73	0.29	0.71	<2.2e-16	<2.2e-16	<2.2e-16
Gene's APA variability and maximum switch score	Spearman-rho			p-value		
	ALL	MSC>0.1	MSC<=0.1	ALL	MSC>0.1	MSC<=0.1
ESC	0.58	0.30	0.47	<2.2e-16	<2.2e-16	<2.2e-16
Cerebellum	0.34	0.11	0.24	<2.2e-16	<2.2e-16	<2.2e-16
Cortex	0.44	0.20	0.30	<2.2e-16	<2.2e-16	<2.2e-16
Heart	0.35	0.12	0.31	<2.2e-16	<2.2e-16	<2.2e-16
Kidney	0.36	0.18	0.24	<2.2e-16	<2.2e-16	<2.2e-16
Liver	0.37	0.13	0.33	<2.2e-16	<2.2e-16	<2.2e-16
Lung	0.47	0.18	0.40	<2.2e-16	<2.2e-16	<2.2e-16
Muscle	0.54	0.21	0.45	<2.2e-16	<2.2e-16	<2.2e-16
Spleen	0.26	0.10	0.16	<2.2e-16	2.4e-14	<2.2e-16

*MSC: Maximum Switch Score of gene

Table 14. Positive correlation between average APA variability and switch score in pairwise tissues.

Spearman-rho\ p-value	ESC	Cere-bellum	Cortex	Heart	Kidney	Liver	Lung	Muscle	Spleen
ESC	-	0.59	0.63	0.60	0.60	0.59	0.63	0.73	0.58
Cerebellum	<2.2e-16	-	0.43	0.44	0.42	0.47	0.54	0.63	0.42
Cortex	<2.2e-16	<2.2e-16	-	0.52	0.47	0.52	0.56	0.67	0.49
Heart	<2.2e-16	<2.2e-16	<2.2e-16	-	0.41	0.44	0.51	0.65	0.40
Kidney	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	-	0.41	0.49	0.64	0.41
Liver	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	-	0.49	0.63	0.40
Lung	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	-	0.68	0.48
Muscle	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	-	0.59
Spleen	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	-

Under the neutral scenario (larger APA differences between tissues are found in genes under relaxed selective constraints), we would also expect a positive correlation between maximum switch score across tissues and dN/dS ratios. However, as finding for within-tissue APA diversity (Table 12), this was not the case (Fig. 5E), indicating that selective constraints on APA regulation are largely independent of those on protein sequence evolution and that the relationship between low average expression levels and larger APA differences between tissues might be mediated by a mechanistic relationship between transcription/mRNA abundance and polyadenylation accuracy. This possibility is supported by the observation that generally a gene's expression level across tissues is positively correlated with its major PAS usage (as an estimate of polyadenylation accuracy) and negatively correlated with the variability between replicates in each tissue (Fig. 5F).

It is also important to note that if molecular error of "noisy" APA is purely result from weak *cis*-regulation in a gene, then it could be observed that a gene's APA difference in pairwise tissue comparison exhibits random variation. Therefore, globally all genes' APA variability in pairwise tissue comparison would only be correlated with average gene expression level as in a single tissue, but not with gene expression level difference between tissues. In fact, we found predominantly positive correlations between gene expression differences among tissue pairs and differences in major PAS usage (the tissue with higher gene expression tended to have higher major PAS usage or higher APA accuracy), and generally pairwise PAS usage differences were more strongly correlated with expression differences between two tissues than with the average gene expression levels of the tissue pair (Table 15-16), indicating a strong mechanistic relationship between transcription/mRNA abundance and polyadenylation accuracy.

Table 15. Spearman correlation between gene expression fold change and major PAS usage difference in tissue pairs.

Spearman-rho\ p-value	ESC	Cerebellum	Cortex	Heart	Kidney	Liver	Lung	Muscle	Spleen
ESC	-	0.18	0.18	0.11	0.14	0.11	0.13	0.20	0.10
Cerebellum	<2.2e-16	-	0.07	0.16	0.13	0.13	0.11	0.22	0.12
Cortex	<2.2e-16	<2.2e-16	-	0.16	0.13	0.15	0.11	0.21	0.12
Heart	<2.2e-16	<2.2e-16	<2.2e-16	-	0.11	0.09	0.13	0.25	0.06
Kidney	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	-	0.12	0.12	0.22	0.09
Liver	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	-	0.11	0.21	0.08
Lung	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	-	0.24	0.08
Muscle	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	-	0.21
Spleen	<2.2e-16	<2.2e-16	<2.2e-16	6.7e-09	<2.2e-16	3.8e-13	2.6e-15	<2.2e-16	-

Table 16. Spearman correlation between average gene expression fold change and major PAS usage difference in tissue pairs.

Spearman-rho\ p-value	ESC	Cerebellum	Cortex	Heart	Kidney	Liver	Lung	Muscle	Spleen
ESC	-	0.03	0.03	0.09	0.06	0.08	0.07	0.06	0.08
Cerebellum	0.02	-	-0.01	0.04	0.00	0.04	0.04	0.00	0.03
Cortex	0.01	0.36	-	0.05	0.02	0.05	0.05	-0.01	0.03
Heart	1.2e-14	8.7e-04	1.5e-05	-	-0.03	0.02	-0.01	-0.03	0.00
Kidney	2.4e-07	0.85	0.09	4.9e-03	-	0.06	0.02	0.00	0.02
Liver	7.1e-13	6.8e-04	1.8e-05	0.05	1.5e-08	-	-0.03	-0.05	-0.04
Lung	3.4e-11	3.5e-04	1.0e-05	0.37	0.14	0.01	-	-0.01	0.01
Muscle	3.3e-07	0.68	0.53	0.01	0.84	5.0e-06	0.38	-	0.00
Spleen	1.8e-12	3.1e-03	1.3e-03	0.78	0.02	6.8e-04	0.53	0.75	-

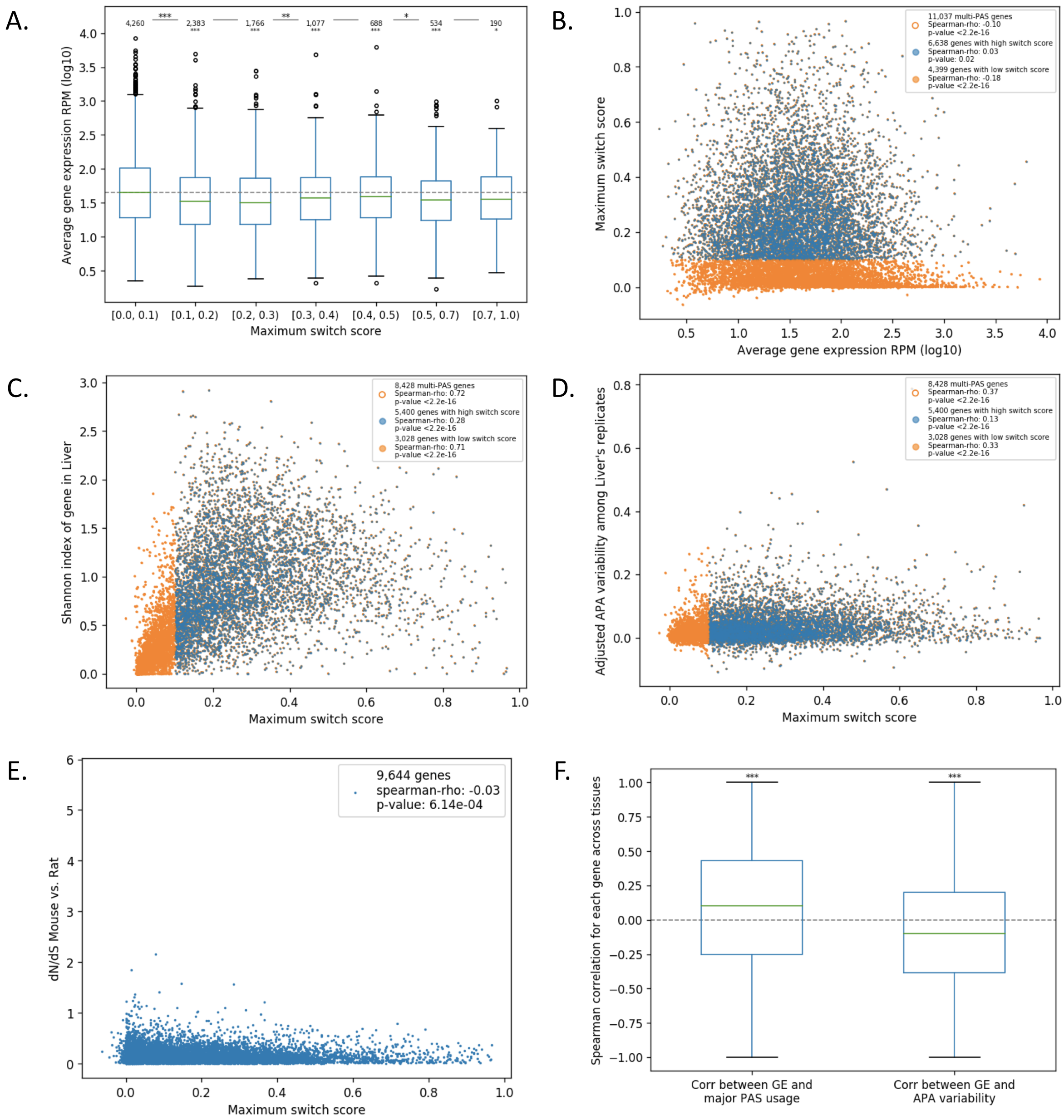


Figure 5. Switch score between tissues correlates with APA diversity and gene expression. A) Average gene expression comparison among genes with different maximum switch score across all tissue pair. Mann-Whitney U test was used to determine the statistical significance (*** indicates p<0.001; ** p<0.01; * p<0.05). B) Negative correlation between gene expression and maximum switch score T2more genes. C) Positive correlation between maximum switch score and gene's Shannon index in liver. D) Positive correlation between maximum switch score and adjusted APA variability among liver replicates. E) Non-positive correlation between dN/dS (mouse vs. rat) and maximum switch score for multi-PAS genes. F) Correlation between gene expression level and major PAS usage or APA variability for each gene across tissue. Binomial test was used to determine whether positive or negative correlated genes are significantly more than the random expectation of 50% (*** indicates p<0.001; ** p<0.01; * p<0.05).

Positional effects on correlations between PAS usage and gene expression levels

As previous studies have shown that PAS choice of a gene depends on its PAS position relative to other PASs (for a review see[111]), and that gene expression levels and minor PAS usage exhibit different correlations depending on the minor PAS location relative to the dominant PAS[112], we further investigated the effects of PAS location on PAS usage within tissues. In general, we found strong negative correlations between gene expression levels and usage of all PASs located upstream of the stop codon (leading to the expression of truncated proteins or alternative protein isoforms) and strong positive correlations between gene expression levels and 3'UTR major PAS usage for each tissue (Table 17). Next, we checked genes' major PAS usage in different localization and found that major PASs affecting coding region have lowest average usage (Fig. 6A). Also, these genes have lower average gene expression level than others except genes with their major located in single 3'UTR PAS (3'UTR(S)) (Fig. 6B).

Table 17. Correlation between gene expression and PAS usage for PASs with different locations.

PAS usage and gene expression	Spearman-rho				p-value			
	All PAS		Major PAS		All PAS		Major PAS	
	CDS	UTR	CDS	UTR	CDS	UTR	CDS	UTR
ESC	-0.29	0.09	0.03	0.16	<2.2e-16	<2.2e-16	0.54	<2.2e-16
Cerebellum	-0.20	0.02	-0.12	0.10	<2.2e-16	0.10	8.2e-03	<2.2e-16
Cortex	-0.23	0.04	-0.10	0.11	<2.2e-16	1.3e-05	0.02	<2.2e-16
Heart	-0.22	0.05	-0.01	0.14	<2.2e-16	9.2e-07	0.81	<2.2e-16
Kidney	-0.25	0.06	-0.02	0.13	<2.2e-16	1.0e-08	0.61	<2.2e-16
Liver	-0.22	0.08	0.03	0.12	<2.2e-16	3.1e-13	0.59	<2.2e-16
Lung	-0.18	0.01	-0.08	0.10	<2.2e-16	0.48	0.07	<2.2e-16
Muscle	-0.18	0.07	0.09	0.19	<2.2e-16	4.3e-09	0.09	<2.2e-16
Spleen	-0.15	0.00	0.00	0.08	<2.2e-16	0.81	0.95	8.5e-15

If we only consider gene with exactly 3 PASs in the 3'UTR (with or without additional minor PASs in the region upstream of the stop codon) and the major PAS located in the middle of 3'UTR PAS (3'UTR(M)) (608 genes), their minor proximal and minor distal PAS usage in the 3'UTR show negative correlations with gene expression of similar magnitude (Table 18). This suggests that the stronger negative correlation found for proximal minor PAS usage in a previous study[112] was mainly due to the inclusion of PASs upstream of the stop codon. Additionally, we selected all genes with exactly 3 PASs in 3'UTR, and compared their major usage correlation according to different 3'UTR positions: 491 genes with major PAS located in first 3'UTR PAS (3'UTR(F)), 608 in middle 3'UTR PAS (3'UTR(M)), 688 in last 3'UTR PAS (3'UTR(L)). Similar as observed in previous study[33], for genes with 3'UTR(F) major PAS, their major PAS usages are more strongly positively correlated with gene expression than genes with major PAS located elsewhere (Table 19), indicating that globally highly expressed genes with major PAS in 3'UTR(F) have higher ratio of mRNA with accurate APA (proximal one) than other genes.

Table 18. Spearman correlation between PAS usage and gene expression for genes with exact 3 UTR PAS and major PAS located in middle

PAS usage and gene expression	Spearman-rho			p-value		
	3'UTR(F)	3'UTR(M)	3'UTR(L)	3'UTR(F)	3'UTR(M)	3'UTR(L)
ESC	-0.17	0.24	-0.18	3.5e-04	4.2e-07	1.7e-04
Cerebellum	-0.26	0.30	-0.32	3.7e-09	1.6e-12	1.7e-13
Cortex	-0.22	0.29	-0.30	4.3e-07	4.5e-11	9.5e-12
Heart	-0.24	0.37	-0.36	1.0e-07	<2.2e-16	1.7e-15
Kidney	-0.36	0.34	-0.29	<2.2e-16	6.7e-15	5.2e-11
Liver	-0.17	0.24	-0.19	4.4e-04	3.1e-07	9.2e-05
Lung	-0.21	0.29	-0.29	1.6e-06	2.7e-11	2.0e-11
Muscle	-0.16	0.34	-0.26	9.6e-04	1.4e-12	6.0e-08
Spleen	-0.24	0.24	-0.25	5.5e-08	6.1e-08	1.4e-08

Table 19. Spearman correlation between major PAS usage and gene expression for genes with exact 3 UTR PASs

Major PAS usage and gene expression	Spearman-rho			p-value		
	3'UTR(F) 491 genes	3'UTR(M) 608 genes	3'UTR(L) 688 genes	3'UTR(F)	3'UTR(M)	3'UTR(L)
ESC	0.29	0.20	0.33	4.3e-08	4.2e-07	2.4e-13
Cerebellum	0.34	0.25	0.26	1.6e-12	1.6e-12	6.0e-10
Cortex	0.34	0.23	0.28	1.6e-12	4.5e-11	1.0e-11
Heart	0.34	0.31	0.26	1.7e-11	<2.2e-16	3.0e-09
Kidney	0.39	0.25	0.21	1.0e-15	6.7e-15	1.3e-06
Liver	0.28	0.14	0.14	5.3e-08	3.1e-07	2.1e-03
Lung	0.36	0.29	0.28	9.3e-14	2.7e-11	2.0e-11
Muscle	0.32	0.28	0.23	1.9e-09	1.4e-12	3.6e-06
Spleen	0.32	0.19	0.22	1.1e-10	6.1e-08	1.5e-07

To investigate the effects of PAS location on PAS usage across tissues, we calculated the Spearman's rank-order correlation between gene's major PAS usage and gene's expression across tissues for each gene expressed in at least three tissues. For genes with major PAS located in 3'UTR(M), 3'UTR(L) and 3'UTR(S), genes with positive correlation significantly outnumber the non-positive correlated genes (Fig. 4C). However, for the genes with major PAS affecting CDS, there are smaller number of positive correlated genes than that of non-positive correlated gene (252 vs. 268). Only slightly larger number of genes with major PAS in 3'UTR(F) PAS is found (1032 vs. 984). It suggests that for individual gene its distal major PAS usage or single major PAS in 3'UTR is more strongly coupled to transcriptional activity than proximal major PAS or major PAS affecting CDS in gene.

According to error hypothesis, genes' APA diversity would remain low in genes with high expression level in each tissue. We further calculated the correlation between gene expression and APA diversity across tissues for each gene expressed in at least three tissues. Significantly larger numbers of genes with gene expression negatively correlated Shannon index were found in genes having major PAS located in 3'UTR(M), 3'UTR(L) and 3'UTR(S) (Fig. 4D). Similarly, as findings in PAS location effects on PAS usage across tissues, there are not significantly more genes exhibiting negative correlations than expected by chance for genes with major PASs affecting CDS or genes with 3'UTR(F) major PASs (245 negatively correlated genes vs. 275 non-negatively correlated genes in genes with PASs affecting CDS, and 1,013 vs 1,003 in genes with 3'UTR(F) major PASs). Because APA variabilities among tissues' replicates are another index for analyzing the relation between APA noise and gene expression level. As expected, more genes with APA variability negatively correlated with gene expression were observed except for genes with major PAS located in 3'UTR(S) (Fig. 4E). Additionally, dN/dS comparison among genes with differently located major PASs shows that genes with major PAS located in 3'UTR(M) and 3'UTR(L) are under stronger selective constraint during evolution (Fig. 4F). These results indicate significantly different effects of PAS localization in APA regulation by gene transcription/mRNA abundance.

To further illustrate location effects in 3'UTR for minor PAS, we again focused on genes with exactly 3 PASs in their 3'-UTRs and with major PAS located in middle. We found that most of genes' expression levels across tissues positively correlate with 3'UTR(M) major PASs as well as usage of 3'UTR(L) minor PASs (Fig. 6G). This is consistent with previous findings in Drosophila[89] showing that faster elongation rates might lead to increased usage of distal PASs.

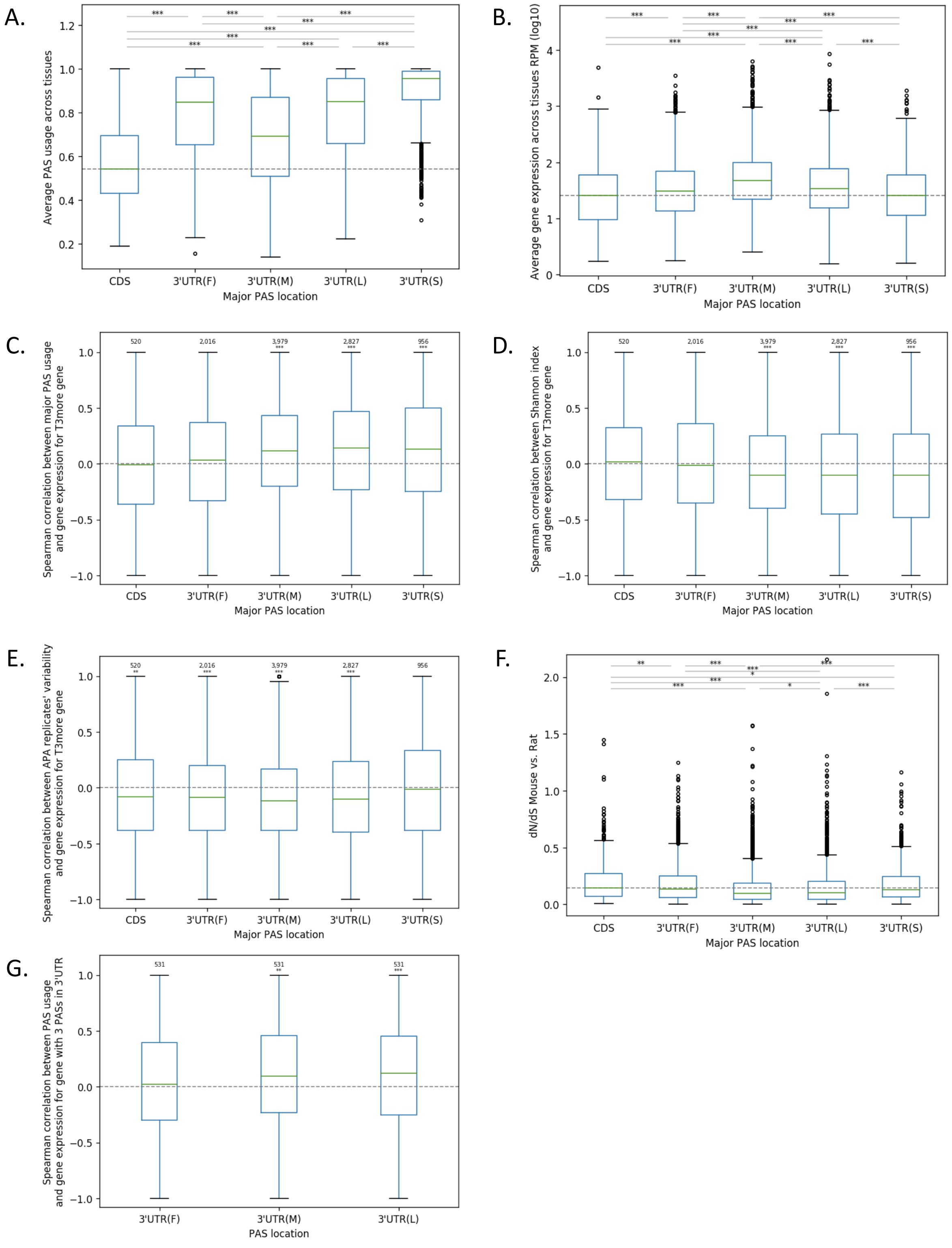


Figure 6. PAS location influences the APA regulation. A) Major PAS usage comparison among genes with different major PAS locations. CDS: upstream of stop codon, 3'UTR(F): first PAS in 3'UTR, 3'UTR(M): middle PAS in 3'UTR, 3'UTR(L): last PAS in 3'UTR, 3'UTR(S): single PAS in 3'UTR. B) Average gene expression level comparison among genes with different major PAS locations. C) Spearman's correlation coefficient between major PAS usage and gene expression level across tissues for each T3more gene (gene expressed in at least three tissues) with different major PAS location. D) Spearman's correlation coefficient between gene's APA diversity (Shannon index) and gene expression level across tissues for each T3more gene with different major PAS location. E) Spearman's correlation coefficient between

gene's APA variability in replicates and gene expression level across tissues for each T3more gene with different major PAS location. F) Comparison of dN/dS mouse vs. rat among genes with different location of major PASs. G) Spearman's correlation coefficient between PAS usage and gene expression level across tissues for genes with exact 3 PAS in 3'UTR and 3'UTR(M) major PAS. Binomial test was used to determine whether positive or negative correlated genes are significantly more than the random expectation of 50% in C-E) and G) (*** indicates p<0.001; ** p<0.01; * p<0.05). Mann-Whitney U test was used to determine the statistical significance in A), B) and F) (*** indicates p<0.001; ** p<0.01; * p<0.05).

Genes with more than one functional PAS

In contrast to error hypothesis that only one PAS is desired in each gene[99], we found that single-PAS gene has less conserved PAS than major PAS from multi-PAS gene (Fig 1C). By dividing major PASs into several groups according to their PAS usage in gene, we found that major PASs with different range of PAS usage all have higher PhastCons score in 50 nt upstream regions than single PAS, but their conservations are independent of their average PAS usage across tissues (Fig. 7A). It indicates that if purifying selection exists on cis-regulatory element for major PAS, the effects of selection does not depend on the diversity of APA in a gene. Then, we wonder whether there is competition between two PASs in one gene and checked the conservation scores of the rank 2 PAS (the PAS with second largest average PAS usage in gene). Interestingly, rank 2 PASs with high usage exhibit higher average PhastCons scores in the 50 nt upstream regions than the PAS with lower PAS usage (Fig. 7B). As expected, the PhastCons score difference between rank 1 and rank 2 PAS positively correlated with the PAS usage difference between rank 1 and rank 2 PAS (Fig. 7C), indicating that competition exist between the rank 1 and rank 2 PASs in those genes and that highly conserved rank 2 can intensify the competition for usage of rank 1 PAS. Interestingly, this competition affects only rank 2 PAS due to the fact that rank 1 PAS' PhastCons scores is independent of the PAS usage difference between rank 1 and rank 2 (Fig. 7D).

For the rank 2 PAS with similar PAS usage as the rank 1 PAS in same genes, we further checked whether the rank 2 PAS is also of functional importance. Because the 3'UTR between rank 1 PAS to rank 2 PAS may contain cis-elements involved in post-transcriptional regulation of rank 2 PAS (e.g., RBP or microRNA binding sites), if rank 2 PAS located downstream of rank 1 PAS (two sites should be both on 3'UTR). We compared these regions for genes with various range of usage difference between rank 1 and rank 2 PAS and found that the conservation of this region decreases with the magnitude of usage difference (Fig. 8A, see method). As we would expect the regions between functionally redundant PASs to be less conserved, this result suggests that the rank 2 PAS are with functional difference to rank 1 PAS in the genes where average PAS usages of rank 1 and rank 2 PAS are similar, compared to genes with the 'weakly' used rank 2 PAS. Further evidence that the high density of conserved microRNA target sites exists in these regions of genes with small usage differences between rank 1 and rank 2 PAS supports the functional importance of rank 2 PAS (Fig. 8B).

To check whether upstream regions of rank 1 and rank 2 PAS could both have high density microRNA binding sites in genes with small difference between rank 1 and rank 2 PAS usage, we scanned conserved microRNA binding sites in PAS' 300nt upstream region (limited in 3'UTR, see Method) in genes with rank 2 located downstream of rank 1 PAS. We found that genes with PAS usage difference less than 20% between rank 1 and rank 2 PAS have similar high microRNA targeting sites enrichment near the ends of both rank 1 and rank 2 PAS whereas microRNA targeting density of rank 2 PAS are quite lower in genes with larger PAS usage difference (>=40%) (Fig. 8C-D,G). Because the microRNA target site are preferentially enriched immediately upstream to PAS[34], it is also possible for us to test microRNA density for genes with rank 2 PAS located upstream of rank 1 PAS. Although the rank 2 PAS' microRNA density is not as high as rank 1 PAS' in different gene groups (Fig. 8E), genes with large difference between rank 1 and rank 2 PAS usage exhibits a significant lower microRNA density in rank 2 PASs than in rank 2 PASs of genes with less difference (Mann-Whitney U test, p-value: 0.03, Fig. 8F,H).

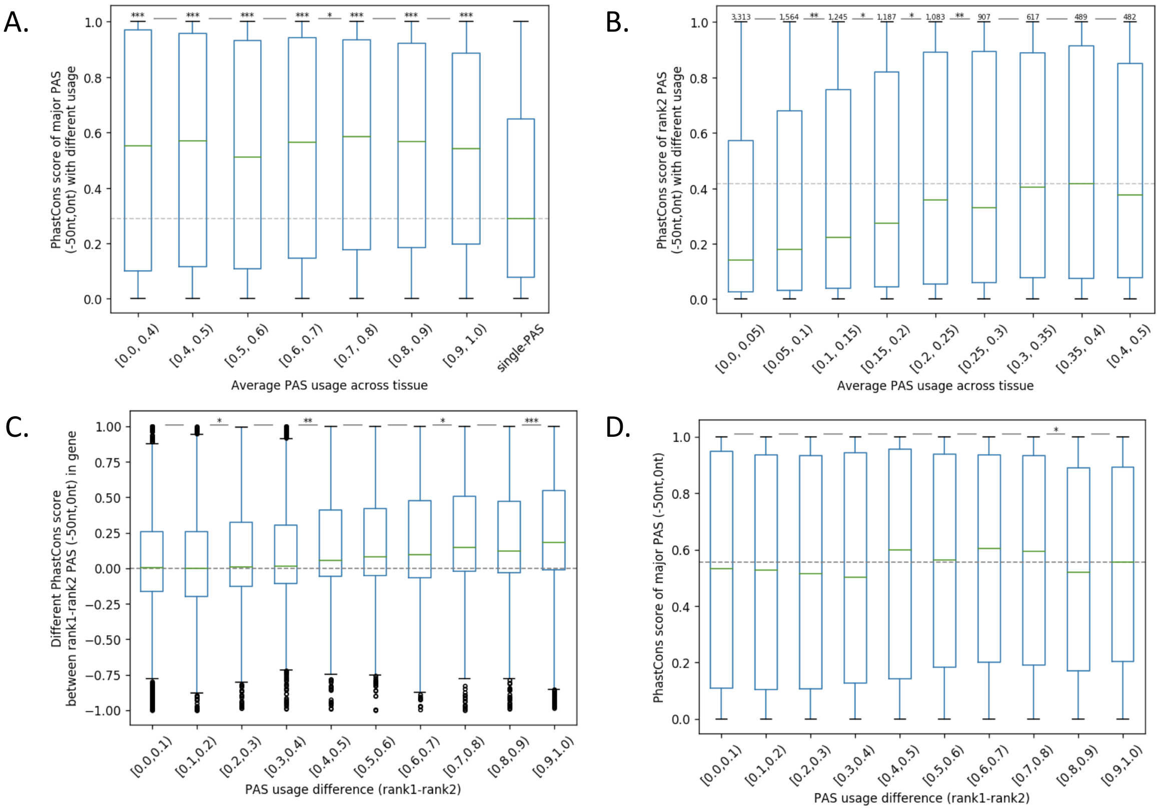


Figure 7. PhastCons score comparison between rank 1 and rank 2 PAS upstream flanking region (-50nt, 0nt). A) PhastCons score comparison among major PASs in genes with different range of major PAS usage and PASs of single-PAS genes. B) PhastCons score of rank 2 PAS comparison among genes with different range of rank 2 PAS usage. C) PhastCons score difference between rank 1 (major) PAS and rank 2 PAS in genes with different magnitude of usage difference between rank 1 and rank 2 PAS. D) PhastCons score of major PAS comparison among genes with different magnitude of usage difference between rank 1 and rank2 PAS. Mann-Whitney U test was used to determine the statistical significance, all major PAS PhastCons score groups are compared to single PAS in A. (*** indicates p<0.001; ** p<0.01; * p<0.05)

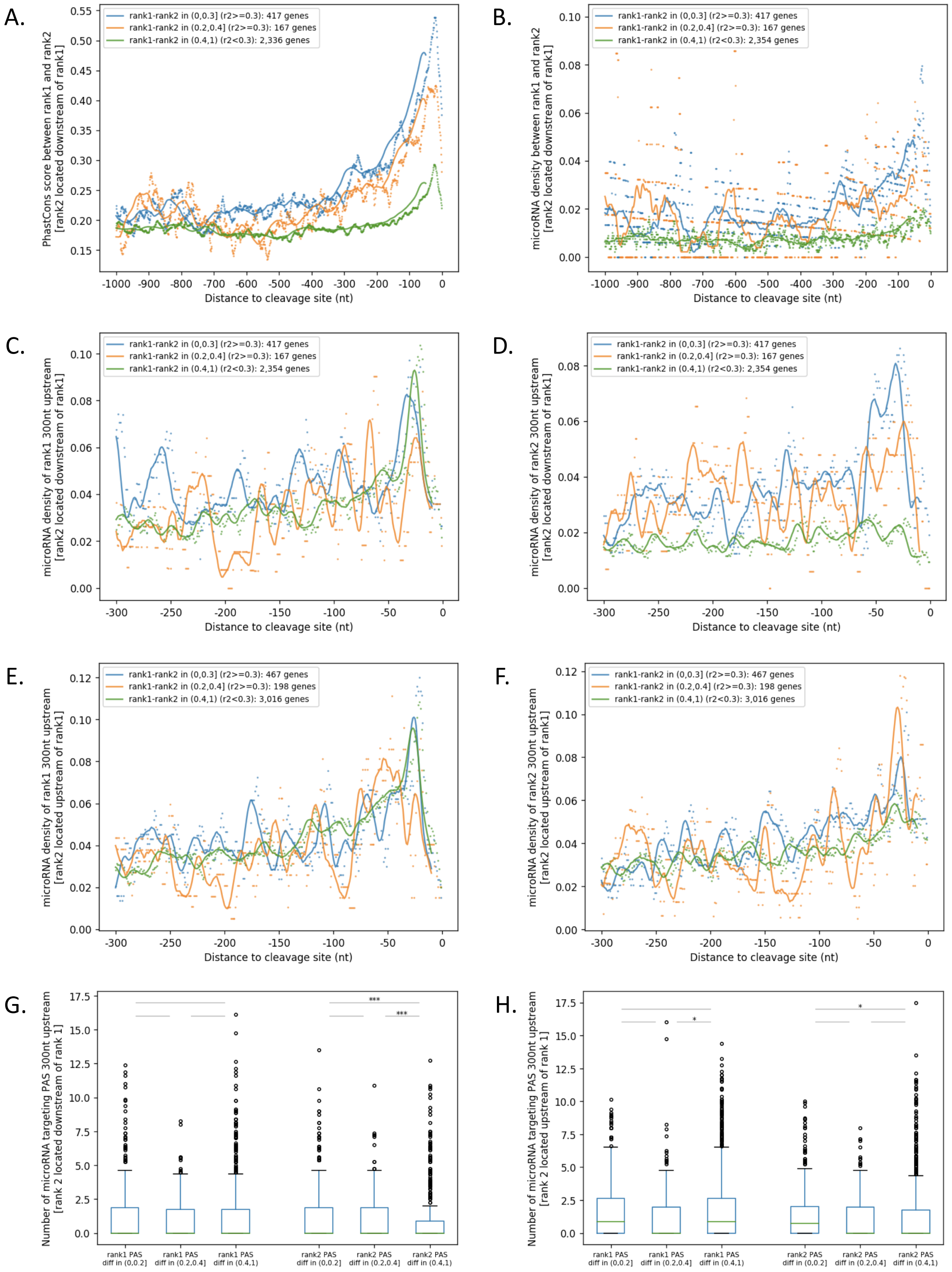


Figure 8. Rank 2 PAS is functional important in genes with small usage difference between rank 1 (major) and rank 2 PAS. A) PhastCons score of genes' 3'UTR between rank 1 to rank 2 PAS for genes with different magnitude of usage difference (rank 1 - rank 2). B) microRNA density in gene's 3'UTR between rank 1 to rank 2 PAS for genes with different magnitude of usage difference. C) microRNA density of rank 1 PAS upstream region (-300nt, 0nt) in genes where rank 2 PAS is located downstream of rank 1 PAS. D) microRNA density of rank 2 PAS upstream region (-300nt, 0nt) in genes

where rank 2 PAS is located downstream of rank 1 PAS. E) microRNA density of rank 1 PAS upstream region (-300nt, 0nt) in genes where rank 2 PAS is located upstream of rank 1 PAS. F) microRNA density of rank 2 PAS upstream region (-300nt, 0nt) in genes where rank 2 PAS is located upstream of rank 1 PAS. G) Number of microRNA targeting at PAS' upstream region (-300nt, 0nt) comparison in genes where rank 2 PAS is located downstream of rank 1 PAS. H) Number of microRNA targeting at PAS' upstream region (-300nt, 0nt) comparison in genes where rank 2 PAS is located upstream of rank 1 PAS. Mann-Whitney U test was used to determine the statistical significance (*** indicates $p<0.001$; ** $p<0.01$; * $p<0.05$)

Tissue-dependent functional PASs

On another aspect, because large number of genes with large maximum switch score (>=10%) do not exhibit a strong negative correlation with average gene expression (6,638 out of 11,037 in Fig. 5B), we checked whether for genes with significantly different APA patterns between tissues might reflect different PAS functional regulation rather than molecular error. Firstly, we applied DEXSeq to identify genes where at least one PAS shows switch-like pattern between any pairwise tissues (PAS usage difference larger than 50% and Bonferroni-adjusted P-value < 0.05, see Methods) and found 328 to 555 genes in each tissue and in total 831 genes across all tissues (Table 20). For these APA switch genes, their average mRNA expression levels are similar compared to other genes, but much lower in many individual tissues except in cortex and cerebellum (Fig. 9A), indicates that the APA variations between tissue pair might have strong tissue bias and might function importantly in brain. Next, we analyzed the correlation between APA variability and maximum switch score for these genes in individual tissue. As expected, no significant positive correlation is found (Table 21). Additionally, in all tissue pairs, most of them have negative correlation between switch score and APA variability of replicates (Table 22), in contrast to stronger positive correlations for all genes (Table 13-14), indicating that these genes' PASs are under strong tissue-specific *trans*-regulation. Interestingly, higher PhastCons score of major PASs region (-50nt, 0nt) are found in genes with differential APA patterns compared to normal genes (Fig. 9B), suggesting these genes' APA patterns might be well conserved during evolution.

Table 20. Number of genes with pairwise APA difference between tissues.

APA different genes	ESC	Cere-bellum	Cortex	Heart	Kidney	Liver	Lung	Muscle	Spleen	Total
ESC	-	257 (3.2%)*	255 (3.2%)	122 (1.6%)	117 (1.5%)	97 (1.3%)	117 (1.5%)	270 (3.9%)	116 (1.5%)	555 (6.7%)
Cerebellum	-	-	3 (0.0%)	55 (0.6%)	47 (0.5%)	81 (1.0%)	53 (0.6%)	210 (2.7%)	71 (0.8%)	525 (5.3%)
Cortex	-	-	-	60 (0.7%)	67 (0.8%)	94 (1.2%)	75 (0.8%)	170 (2.2%)	93 (1.1%)	488 (4.9%)
Heart	-	-	-	-	28 (0.3%)	31 (0.4%)	24 (0.3%)	126 (1.6%)	35 (0.4%)	336 (3.7%)
Kidney	-	-	-	-	-	13 (0.2%)	7 (0.1%)	134 (1.7%)	21 (0.2%)	328 (3.4%)
Liver	-	-	-	-	-	-	13 (0.2%)	145 (2.0%)	27 (0.3%)	330 (3.9%)
Lung	-	-	-	-	-	-	-	129 (1.6%)	8 (0.1%)	327 (3.3%)
Muscle	-	-	-	-	-	-	-	-	177 (2.3%)	515 (6.4%)
Spleen	-	-	-	-	-	-	-	-	-	393 (4.2%)

* gene ratio is calculated by # of genes with switch like APA patterns divided by total # of genes co-expressed in tissue pairs.

Table 21. Negative correlations between APA variability and switch score for genes with significantly APA difference.

Spearman-rho\ p-value	ESC	Cere-bellum	Cortex	Heart	Kidney	Liver	Lung	Muscle	Spleen
ESC	-	-0.21	-0.25	-0.19	-0.10	-0.26	-0.11	0.17	-0.16
Cerebellum	2.7e-03	-	-0.50	-0.17	-0.35	-0.38	-0.31	-0.22	0.04
Cortex	4.0e-04	0.67	-	-0.20	-0.27	-0.38	-0.28	-0.24	-0.23
Heart	0.08	0.25	0.16	-	0.03	-0.54	0.14	-0.21	-0.05
Kidney	0.35	0.03	0.05	0.90	-	-0.74	0.10	-0.31	-0.27
Liver	0.03	1.5e-03	5.5e-04	4.5e-03	0.04	-	-0.13	-0.27	-0.24
Lung	0.33	0.03	0.02	0.56	0.87	0.73	-	-0.10	-0.26
Muscle	0.02	2.1e-03	3.0e-03	0.02	8.4e-04	1.5e-03	0.29	-	-0.26
Spleen	0.17	0.77	0.05	0.81	0.33	0.40	0.53	1.2e-03	-

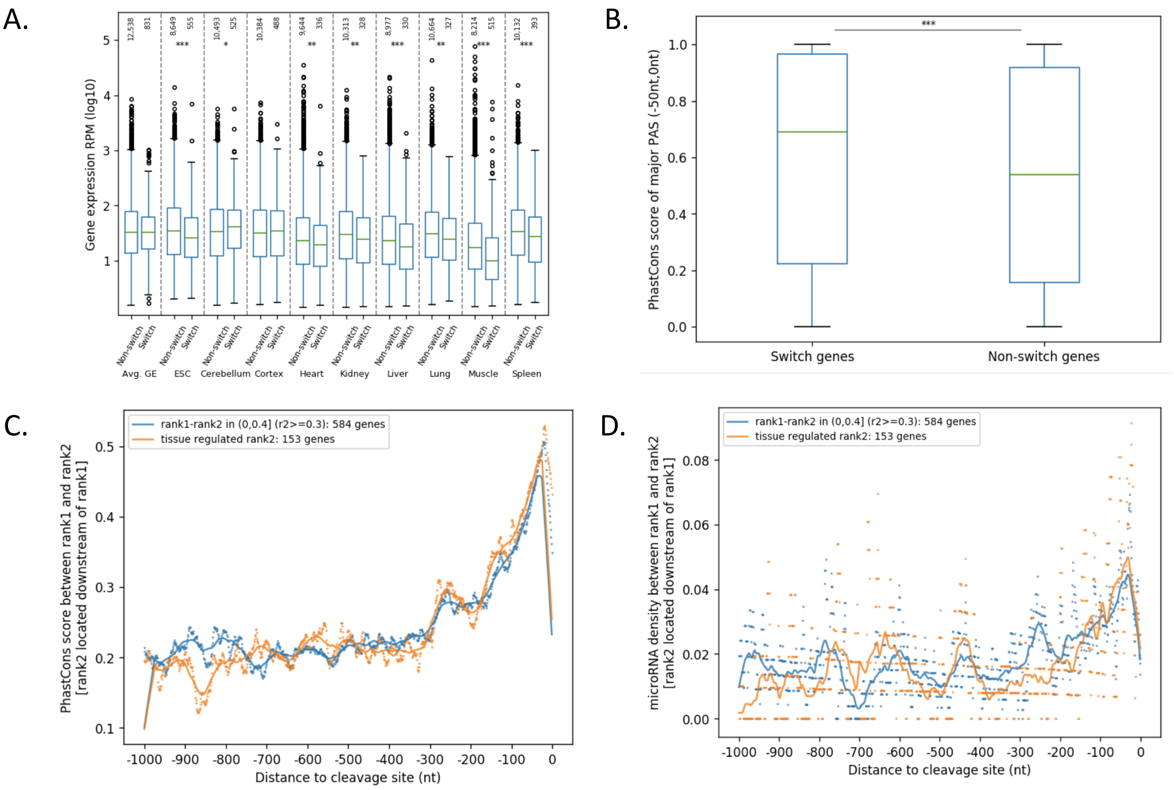
Table 22. Weak or negative correlation between maximum switch score and APA diversity for significant APA differential genes in each tissue.

APA diversity and maximum switch score	spearman-rho	p-value
ESC	-0.05	0.23
Cerebellum	-0.04	0.36
Cortex	-0.03	0.51
Heart	0.05	0.39
Kidney	-0.08	0.17
Liver	-0.16	4.31E-03
Lung	0.02	0.69
Muscle	-0.24	3.04E-08
Spleen	0.03	0.53

Among these APA switch genes, there are 377 genes with differential rank 1 or rank 2 PAS and with both rank 1 and rank 2 PAS located in genes' 3'UTR. In these 377 genes, 153 genes have their rank 2 PASs located downstream of rank 1 PASs, whose 3'UTR between rank1 and rank 2 PASs exhibit similar high PhastCons score and slightly higher microRNA binding density comparing to other genes with functional rank 2 PASs (rank1-rank2 <=0.4 as mentioned in above section, Fig. 9C-D). Whereas, only 67 (43.8%) of the 153 genes have PAS usage difference between rank 1 and rank 2 less than 40%, indicating that besides genes with small difference between rank 1 and rank 2 PASs usage genes with tissue-specifically regulated APA also favor functional rank 2 PAS.

Additionally, in 831 genes with differential PAS between tissues, there are 355 genes where their PASs are located upstream of genes' annotated last stop codon. Genes such as Gls (ENSMUSG00000026103), Klc1 (ENSMUSG00000021288) and Sept8 (ENSMUSG00000018398) have clear alternative last exons switched between brain and other tissues (Fig. 7E). In other 476 genes with significant differential PASs located in 3'UTR between tissues, there are 274 genes with differential rank 1 PASs, 187 genes with differential rank 2 PASs and 15 genes with differential minor PASs in lower ranks. Genes like Elavl1 (ENSMUSG00000040028), Mkln1

(ENSMUSG00000025609) and Mrpl35 (ENSMUSG00000052962) with differential rank 1 (major) PAS can reduce the proximal major PAS usage to increase all distal PAS usage in brain, which increases gene's APA diversities but with similar the expression level across tissues (Shannon index: 0.39 in other tissues vs. 1.64 in brain for Elavl1, 1.15 vs. 1.89 for Mkln1, 1.46 vs. 1.97 for Mrpl35). Whereas genes like Vezt (ENSMUSG00000036099) and Klhl36 (ENSMUSG00000031828) with differential minor PAS can specifically increase usage of the rank 1 PAS in brain without increase the APA diversity in tissues (Shannon index: 1.69 in other tissues vs. 1.57 in brain for Vezt and 1.26 vs, 1.02 for Klhl36).



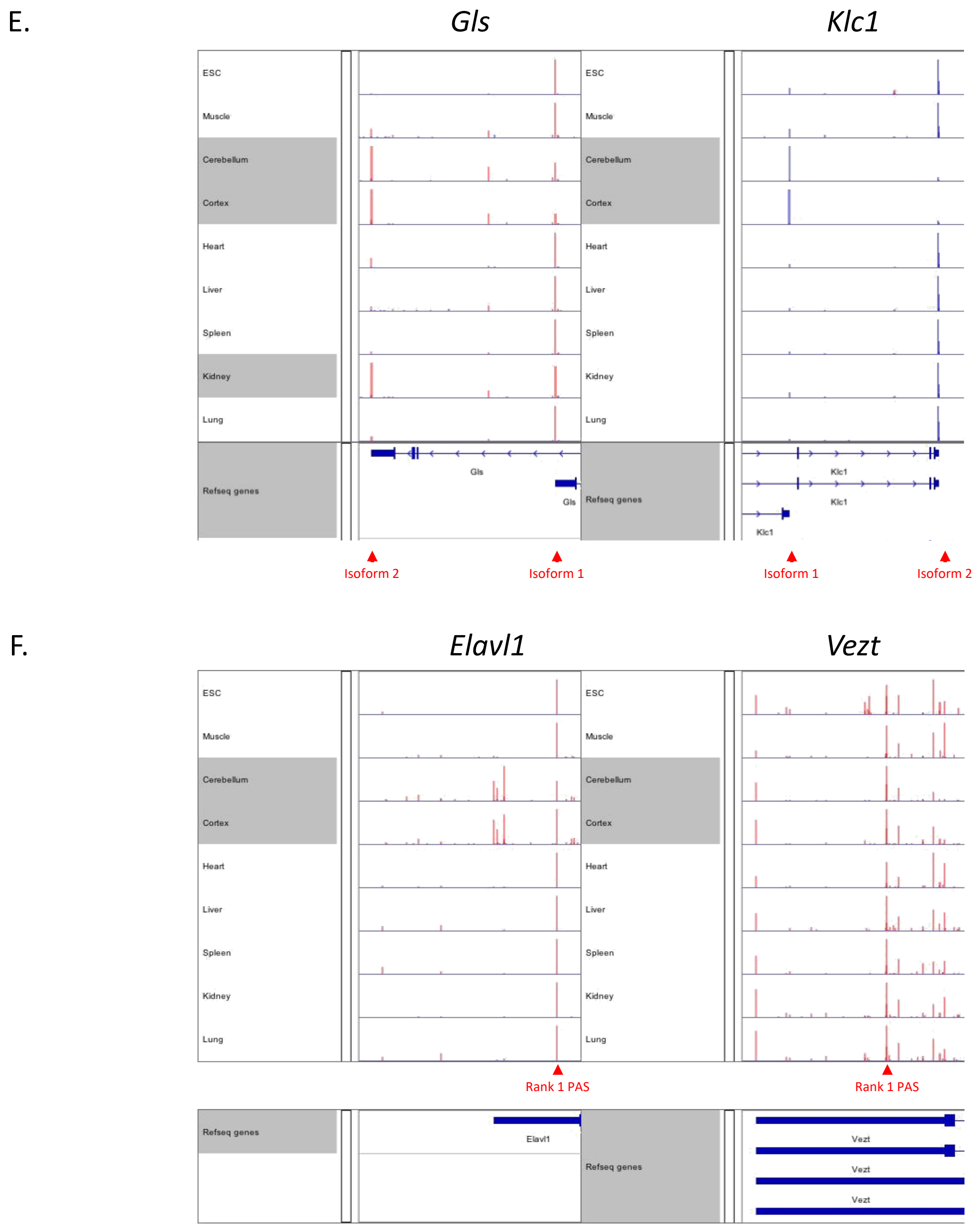


Figure 9. Tissue specific regulation in genes with switch-like APA pattern. A) Gene expression level comparison between switch genes and non-switch genes in average and in each tissue separately. B) PhastCons score of major PAS comparison between switch genes and non-switch genes. C) PhastCons score of gene's 3'UTR from rank 1 to rank 2 PAS for genes with functional rank 2 PAS and genes with switch-like pattern. D) microRNA density of gene's 3'UTR from rank 1 to rank 2 PAS for genes with functional rank 2 PAS and genes with switch-like pattern. Only genes where rank 2 PAS is located downstream of rank 1 PAS are selected in comparisons. E) The 3′-seq tracks of different tissues showing PAS usage switch between alternative last exons. F) The 3′-seq tracks of different tissues showing tissue regulated rank 1 PAS usage change in Elav1 and Vezt. Mann-Whitney U test was used to determine the statistical significance (*** indicates p<0.001; ** p<0.01; * p<0.05)

Allelic divergence

The F1 mouse model with unambiguous assigned allelic reads enables us to investigate allele-specific differences in PAS usage, reflecting cis-regulatory divergence between the two mouse strains and their interaction with trans-regulatory differences between tissues. If APA is mainly noise in regulation, according to the neutral hypothesis, the APA patterns of genes with lower expression levels are expected to diverge faster between species than highly expressed genes. As predicted, we found that in each tissue multi-PAS genes with major PAS diverged between C57BL/6J and SPRET/EiJ had lower expression levels than those with conserved APA patterns between alleles (Fig. 10A). Similarly, divergent genes also had higher dN/dS ratios than non-divergent genes (Fig. 10B). The magnitude of major PAS usage differences between C57BL/6J and SPRET/EiJ is negatively correlated with gene expression levels in all tissues (Table 23, see Methods), further supporting that genes with divergent APA patterns between the two mouse strains have been under relaxed selective constraints.

Table 23. Spearman correlation between magnitude of allelic major PAS divergence and gene's expression (GE) and dN/dS (Mouse vs. Rat).

Correlation	Spearman-rho		p-value	
	GE	dN/dS	GE	dN/dS
ESC	-0.18	0.08	<2.2e-16	6.6e-06
Cerebellum	-0.10	0.05	4.2e-13	1.4e-04
Cortex	-0.10	0.05	7.7e-13	2.6e-04
Heart	-0.11	0.05	5.9e-14	1.3e-03
Kidney	-0.08	0.03	2.2e-09	0.02
Liver	-0.10	0.02	4.5e-10	0.25
Lung	-0.16	0.04	<2.2e-16	8.2e-03
Muscle	-0.14	0.03	<2.2e-16	0.16
Spleen	-0.06	0.04	7.8e-05	0.01

The larger magnitude of APA differences between the two strains observed for genes with lower expression levels could be explained by two different, but not mutually exclusive scenarios: 1. Genes under reduced purifying selection accumulate cis-regulatory mutations faster than more conserved genes. 2. Genes under reduced selective constraints frequently have lower levels of major PAS usage than conserved. Therefore, they exhibit a scaling effect based on the kinetics of competition between PASs similar to that observed in alternative splicing, so that a single *cis*-regulatory mutation in such a gene with low major PAS usage (higher APA diversity) leads to a greater change than in genes with high major PAS usage.

To examine whether one or both of these scenarios can explain our findings, we divided all genes into three classes: non-divergent genes (no significant allele-specific PAS usage difference in any tissue), some-divergent genes (allele-specific PAS usage differences in some tissues, but not all) and all-divergent genes (allele-specific PAS usage differences in all tissues). To estimate whether genes under reduced selective constraints might accumulate cis-regulatory mutations at a faster rate, we then compared the SNP densities in the flanking regions of major and minor PAS sites between the three groups. We found, as expected, that all-divergent genes have the highest SNP densities and non-divergent genes the lowest, especially in the core region of PAS (-50nt, 0nt) (Fig. 10C). To examine the presence of a possible scaling effect, we asked whether in genes with divergent APA the major PASs with lower usages had larger differences between the two strains than those with higher usages (choosing the C57BL/6J allele arbitrarily as the starting value). This was indeed the case, also when correcting for sampling sequencing error, indicating a scaling effect (Fig. 10D).

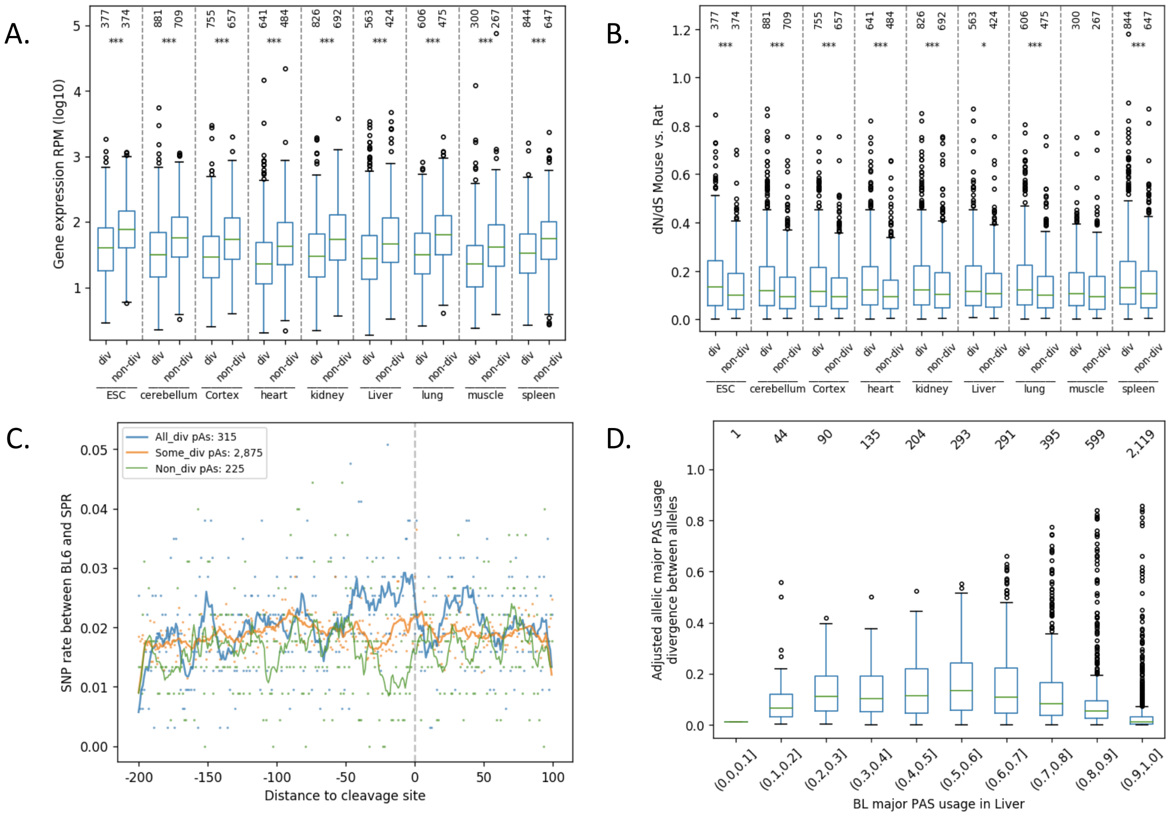


Figure 10. Allelic PAS divergences are under relaxed selective constraint. A) Gene expression comparison between genes with significant divergent major PAS between alleles and genes with non-divergent major PAS in each tissue. B) dN/dS

(mouse vs. rat) comparison between genes with divergent major PAS and genes with non-divergent major PAS in each tissue. C) SNP density in PAS for non-divergent, some-divergent and all-divergent genes in PAS flanking region (-200nt to 100nt). D) Scaling effect in liver for adjusted PAS usage divergence between two alleles with different range of C57BL/6J allele PAS usage as the starting value. Mann-Whitney U test was used to determine the statistical significance (*** indicates $p<0.001$; ** $p<0.01$; * $p<0.05$)

Evolutionary patterns of APA and mRNA expression levels

Different layers of gene regulation might be subject to different selection pressures and exhibit different evolutionary patterns due to differences in molecular mechanisms or in functional effects. Therefore, we compared allele-specific differences in APA in our F1 hybrid mouse model with those in mRNA levels. We found that in our data set samples clustered first by species based on APA patterns, but first by tissue based on mRNA levels (Fig. 11A-B). It appears therefore that tissue-dependent gene expression patterns are generally more conserved than tissue-dependent APA patterns. This conclusion is further supported by the fact that APA diversity across tissues (switch score) is more strongly correlated with variability between replicates (Fig. 3F, Table 11) than is the case for gene expression (Fig. 11E, Table 24, see Methods), indicating that noise plays a relatively larger role for tissue differences in APA compared to gene expression. Another factor contributing to these different patterns might be that allele-specific differences in APA are usually consistent across tissues (the same allele has higher major PAS usage in all tissues in which the gene is expressed). Only in 30 out of 3190 genes the direction of major PAS usage divergence is tissue-dependent, while there are 84 from 1347 cases in which different alleles have higher expression levels in different tissues (Fig. 11C-D, p-value < 2.2e-16, Fisher's extract test).

Table 24. Spearman correlation between adjusted gene expression diversity (adjusted squared coefficient) within each tissue and gene expression diversity across tissues.

Correlation of gene expression diversity within tissues and diversity across tissues	spearman-rho	p-value
ESC	0.01	0.26
Cerebellum	0.03	4.5e-03
Cortex	0.04	3.8e-06
Heart	0.08	1.1e-15
Kidney	0.11	<2.2e-16
Liver	0.07	8.4e-13
Lung	0.07	1.5e-14
Muscle	0.11	<2.2e-16
Spleen	0.20	<2.2e-16

In F1 hybrids, due to same *cis*-elements in alleles across tissues, the tissue differential PAS usage in tissue pairs for each allele can be introduced by APA noise or *trans*-factors regulation. If the PAS is under weak tissue specific regulation and with weak *cis*-elements regulation, the PAS usage can be randomly differential between tissues. Therefore, we checked differential major PAS usage between different tissue pairs for each allele. As shown in table 25, there are 224 (2%) to 1,103 (14%)* genes exhibiting similar allelic major PAS usage change between alleles pairwise tissues, whereas less than

1% genes differ oppositely in tissue pair (Table 26). Due to strong *trans*- regulation, almost all major PAS with differential PAS usage between pairwise tissue exhibit similar allelic PAS differential usage in pairwise tissue (Fig. 8F, Table 13), seldom or none of which exhibit opposite allelic PAS differential usage or only one allelic PAS differential usage (Table 14). As expected, major PASs of genes with opposite allelic PAS differential usage are less conserved (Fig. 8G). Compared to genes with non-differential major PAS usage in both alleles, majority genes (3,784 genes, 74.6%) are with high dN/dS which indicates relaxer selective constraints.

Table 25. Genes with similar allelic differential major PAS usage in pairwise tissues. The upper triangle indicates the number of genes with similar major PAS usage difference between two alleles in pairwise tissues and the percentage of these genes in total. The down triangle (gray) indicates the number of genes with tissue regulated APA pattern in tissue pair (right) and the number of genes showing similar major PAS usage difference between two alleles (left).

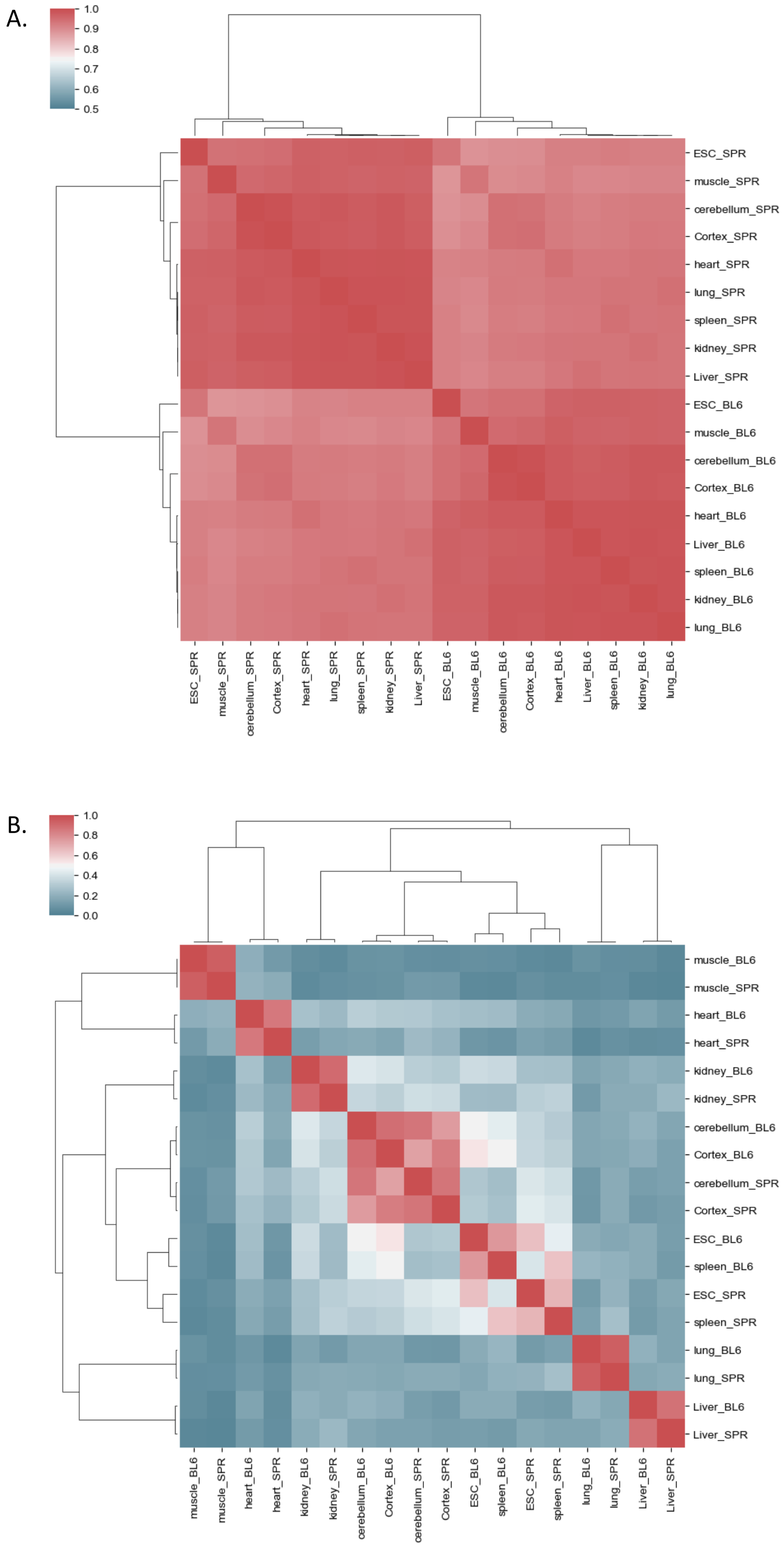
	ESC	Cerebellum	Cortex	Heart	Kidney	Liver	Lung	Muscle	Spleen
ESC	-	1,099 (14%)*	1,103 (14%)	761 (10%)	817 (10%)	642 (9%)	865 (11%)	904 (13%)	743 (9%)
Cerebellum	95/104	-	224 (2%)	565 (7%)	640 (7%)	640 (8%)	591 (7%)	919 (12%)	650 (7%)
Cortex	90/99	1/2	-	574 (7%)	626 (7%)	649 (8%)	626 (7%)	884 (11%)	700 (8%)
Heart	44/50	16/18	16/21	-	313 (4%)	335 (4%)	346 (4%)	685 (9%)	371 (4%)
Kidney	38/43	14/14	20/22	7/7	-	249 (3%)	342 (4%)	722 (9%)	340 (4%)
Liver	33/40	27/31	33/36	10/11	2/3	-	316 (4%)	666 (9%)	300 (4%)
Lung	45/48	20/21	30/31	3/6	3/4	4/5	-	782 (10%)	264 (3%)
Muscle	105/114	66/70	54/57	37/37	37/38	32/33	38/40	-	779 (10%)
Spleen	32/38	24/25	32/32	8/9	4/4	4/6	2/2	52/55	-

* gene ratio is calculated by # of genes with similar allelic major PAS usage change divided by total # of genes co-expressed in tissue pairs. Only genes with filtered allelic major PAS usage were taken into consideration, see Method in allelic PAS usage quantification.

Table 26. Genes with opposite allelic differential major PAS usage in pairwise tissues. The upper triangle indicates the number of genes with opposite major PAS usage change between two alleles in pairwise tissues and percentage of these genes in total. The down triangle (gray) indicates the number of genes with tissue regulated APA pattern in tissue pair (right) and the number of genes showing opposite allelic major PAS usage difference between two alleles (left).

	ESC	Cerebellum	Cortex	Heart	Kidney	Liver	Lung	Muscle	Spleen
ESC	-	94 (1.2%)*	77 (1.0%)	82 (1.1%)	86 (1.1%)	85 (1.2%)	89 (1.1%)	80 (1.1%)	73 (0.9%)
Cerebellum	0/104	-	45 (0.5%)	62 (0.7%)	44 (0.5%)	57 (0.7%)	80 (0.9%)	71 (0.9%)	66 (0.8%)
Cortex	0/99	0/2	-	56 (0.7%)	38 (0.4%)	62 (0.8%)	61 (0.7%)	70 (0.9%)	48 (0.6%)
Heart	3/50	0/18	0/21	-	63 (0.7%)	72 (0.9%)	83 (0.9%)	73 (0.9%)	71 (0.8%)
Kidney	0/43	0/14	0/22	0/7	-	67 (0.8%)	72 (0.8%)	60 (0.8%)	63 (0.7%)
Liver	1/40	0/31	0/36	0/11	0/3	-	87 (1.1%)	78 (1.1%)	60 (0.7%)
Lung	0/48	0/21	0/31	0/6	0/4	0/5	-	88 (1.1%)	78 (0.9%)
Muscle	1/114	0/70	0/57	0/37	0/38	0/33	0/40	-	69 (0.9%)
Spleen	1/38	0/25	0/32	0/9	0/4	1/6	0/2	0/55	-

* gene ratio is calculated by # of genes with opposite allelic major PAS usage change divided by total # of genes co-expressed in tissue pairs.


A.
1.0
0.9
0.8
0.7
0.6
0.5
ESC_SPR
muscle_SPR
cerebellum_SPR
Cortex_SPR
heart_SPR
lung_SPR
spleen_SPR
kidney_SPR
Liver_SPR
ESC_BL6
muscle_BL6
cerebellum_BL6
Cortex_BL6
heart_BL6
Liver_BL6
spleen_BL6
kidney_BL6
lung_BL6
B.
1.0
0.8
0.6
0.4
0.2
0.0
muscle_BL6
muscle_SPR
heart_BL6
heart_SPR
kidney_BL6
kidney_SPR
cerebellum_BL6
Cortex_BL6
cerebellum_SPR
Cortex_SPR
ESC_BL6
spleen_BL6
ESC_SPR
spleen_SPR
lung_BL6
lung_SPR
Liver_BL6
Liver_SPR

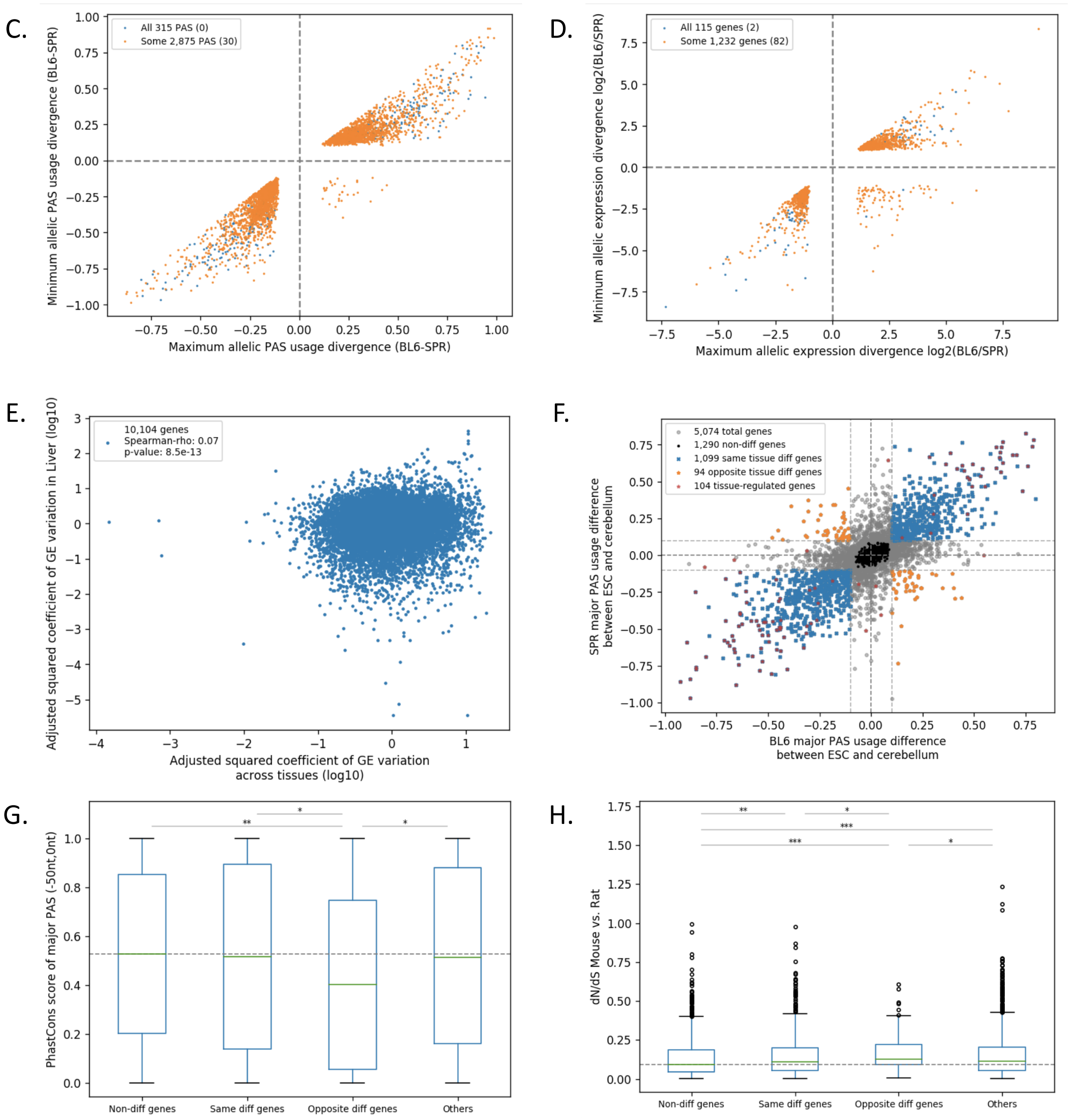


Figure 11. Allelic APA pattern across tissues. A) 2D clustering of PAS usage in 9 tissue for SPRET/EiJ and C57BL/6J allele. B) 2D clustering of gene expression in nine tissues for SPRET/EiJ and C57BL/6J allele. C) Allelic PAS usage divergence (C57BL/6J - SPRET/EiJ) across tissues. D) Allelic gene expression difference (C57BL/6J - SPRET/EiJ) across tissues. E) Correlation between adjusted squared coefficient of gene expression variants in liver and adjusted squared coefficient of gene expression variants across tissues. F) Gene's allelic major PAS differential usage between ESC and cerebellum. Each dot represents a gene's major PAS usage difference (major PAS usage in ESC-major PAS usage in cerebellum). X-axis is the usage difference in C57BL/6J allele. Y-axis is the usage difference in SPRET/EiJ allele. Grey dots are genes with at least one allelic difference less than 0.1. Blue dots are genes with similar PAS usage difference (>=0.1) in both alleles. Orange dots are genes with opposite allelic difference (>=0.1 in both alleles). Red dots are genes with significant differential APA between ESC and cerebellum. G-H) PhasstCons score of major PAS (-50nt, 0nt) and dN/dS (mouse vs. rat) comparison between genes in F). Genes are divided into "non-diff" genes (black genes in Figure F), 'same diff' genes (blue), 'opposite dff' genes (orange), and others (grey without overlap color).

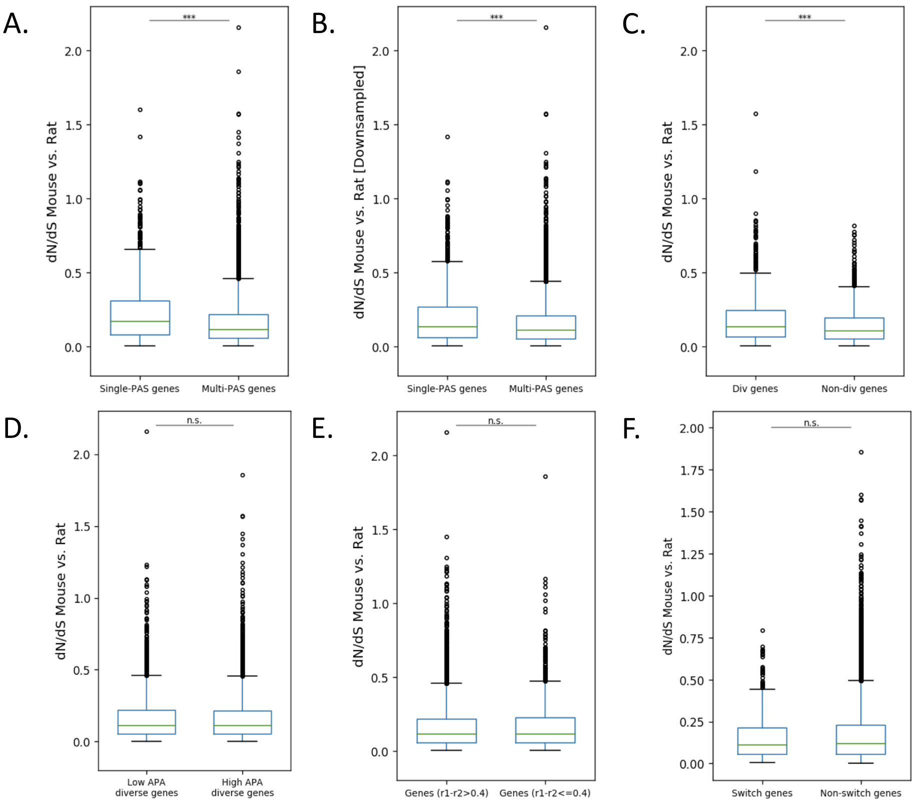


Figure 12. Summary of dN/dS comparison between different genes. A) Comparison between single-PAS genes and multi-PAS genes. B) Comparison between single-PAS genes and multi-PAS genes in downsampled dataset. C) Comparison between genes with divergent major PAS between two alleles and genes with non-divergent major PAS. D) Comparison betwen genes with low APA diversity in tissue (dominant PAS usage >=90%) and genes with high APA diversity (dominant PAS usage<90%). E) Comparison between genes with high difference between rank 1 and rank 2 PAS (>40%) and genes with low difference (<=40%). F) Comparison between genes with significant differential APA pattern across tissue and genes without differential APA pattern across tissue. Wilcoxon signed-rank test test was used to determine the statistical significance (*** indicates p<0.001; ** p<0.01; * p<0.05).

Discussion

APA is a common process in higher eukaryotes which has received increased attention in recent years with regards to its tissue-dependent regulation, function and evolution and its importance for multicellular complexity. However, cases where different functions of tissue-regulated APA isoforms have been clearly demonstrated are still rare. It has therefore been proposed that most observed APA is noise and that APA divergence between species is largely neutral or slightly deleterious. Here we addressed these issues in an F1 mouse model by investigating *trans*-regulatory differences between tissues and the *cis*-regulatory divergence between *Mus musculus* and *Mus spretus*.

Applying the neutral theory of molecular evolution to APA, we would expect that genes with higher APA diversity within a tissue, with larger APA differences between tissues or with larger divergence between alleles are under relaxed selective constraints compared to those genes with invariable APA patterns, which are expected to be under stronger purifying selection. This is indeed the case considering average gene expression levels as an indicator of selective constraint. However, using dN/dS ratios as an alternative measure, we only found that genes with divergent major PASs between two alleles are under relaxed selective constraint compared to non-divergent genes (Fig. 12A-C). In contrast, no differences in dN/dS ratios exist between multi-PAS genes with different degrees of APA diversity (Fig. 12D). Furthermore, single-PAS genes (i.e., genes with the lowest possible APA diversity) appear to be under relaxed selective constraints compared to multi-PAS genes. And although there are stronger negative correlations between gene expression and PAS usage for genes with dominant PAS usage above 0.9 within each tissue or major PAS usage above 0.9 across tissues, for genes with more diverse APA the negative correlation between gene expression and APA diversity becomes much weaker. Therefore, our results do only partially support the error hypothesis.

Several possible scenarios could explain why we find a discordance between different measures of selective constraint, where only gene expression levels correlate well with APA diversity: 1. Relaxed selective constraints on genes with lower average expression levels lead to weaker, more error-prone *cis*-regulation of APA, but not to faster coding sequence evolution; 2. Specific *trans*-factors increasing the cleavage and polyadenylation accuracy of their target genes are co-regulated with these in a tissue-dependent manner to precisely regulate tissue-specific APA pattern[80,113]; and this might result in increased polyadenylation error of genes in the tissues where they and their regulators are lowly expressed 3. General *trans*-factors are a limited resource and highly expressed genes compete more successfully for their binding, resulting in reduced polyadenylation accuracy of lowly expressed genes; 4. Transcriptional regulation is mechanistically coupled with polyadenylation in a way that leads to automatically increased major PAS usage with increased mRNA expression levels. These possibilities are not mutually exclusive. For example, PAS signals are recognized by CPSF in a concentration dependent way[114] and global APA changes can be introduced by different expression levels of *trans*-factors in 3' end complexes, such as U1 level for intronic PASs activation[111] and CFI for distal PASs[73]. Additionally, distal PAS usage can also be affected by the transcription elongation rate[89].

On the other hand, similar dN/dS ratios are found between genes with and without potentially functional rank 2 PASs and between switch genes and non-switch genes (Fig. 12E-F), indicating that genes with high APA diversity due to high rank 2 PAS usage and genes with high APA variability across tissues and potentially tissue-regulated APA patterns are under similar selective constraints as most other multi-PAS genes. Especially for genes with significant differential APA patterns between tissues, the fact that their APA diversities within a tissue are negatively correlated with their switch scores in pairwise tissue comparisons and that their major PAS regions are more highly conserved than those of other genes (Table 11) further supports the idea that there is a strong regulatory mechanism underlying APA variation for many genes. Additionally, these genes are highly conserved as shown in consistent tissue differential APA patterns between the C57BL/6J and SPRET/EiJ alleles in our F1 hybrid model (Fig. 11F).

Gene expression levels might correlate with a gene's major PAS usage not only because more highly expressed genes are under stronger purifying selection, but also because there might be a direct mechanistic or functional relationship between transcriptional regulation or mRNA levels and PAS choice. Major PASs are not randomly distributed, but are more likely to be located at the distal end of the UTR. And we found that PAS location plays an important role in PAS usage regulation. Although among genes with the major PAS located proximal to the stop codon in the 3'UTR (3'UTR(F)) there is a strong positive correlation between major PAS usage and gene expression levels within each tissue (Table 19), higher than among genes with the major PAS in 3'UTR(M) and 3'UTR(L), we found that the usage of the most distal PAS (3'UTR(L) PAS) for each gene was strongly positively correlated with gene expression level across tissues, even for minor distal PASs (Fig. 6C,D and G). These findings indicate that distal PAS usage automatically increases with higher mRNA levels and that major PASs located at the distal end of the 3’-UTR might be selectively favored, consistent with a previous study showing that 3'UTR(L) PASs are most conserved during evolution[115].

In contrast, genes with the major PAS located upstream of the stop codon have lower expression levels and their gene expression levels do not negatively correlate with APA diversity (Fig. 6B-D), indicating that usage of most of these PASs might be slightly deleterious. Additionally, although distal PAS usage increases with expression levels across tissues, we observed similar correlations between minor PAS usage for proximal and distal PAS localizations in 3'UTR within tissues when excluding minor PASs located upstream of the stop codon, contradictory to the previous conclusion that proximal minor PAS usage is generally more harmful than distal minor PAS usage[99]

In contrast to our expectations based on the neutral hypothesis, we found that the major PASs of multi-PAS genes are more conserved than those of single-PAS genes. If purifying selection was the only force during evolution balancing mutation and drift and therefore determining the degree of APA diversity as postulated by the neutral hypothesis [112], it would remove deleterious minor PASs and maintain the remaining single-PAS with greater functional importance and therefore with more conserved genomic sequences. Whereas, according to our data, there are 5,012 minor PAS with higher PhastCons score in their 50nt upstream regions than the median PhastCons score of the corresponding regions of single-PAS genes, indicating that many minor PASs might also be functional rather than deleterious as suggested in error hypothesis. This idea is further supported by the high density of microRNA targeting sites in the 3'UTR between the rank 1 PAS and rank 2 PAS in genes where the rank 2 PAS is located downstream. Another explanation is that competition between PASs shapes conservation patterns of their regulatory regions in a more complex way. Indeed, the conservation scores of rank 2 PASs correlate with their average usage, suggesting a competition between the rank 1 PAS and rank 2 PAS for the PAS usage in a gene.

Another unexpected finding is that between heart and skeletal muscle, unlike between the cerebellum and cortex, there are much more genes showing tissue-regulated APA (170 vs. 30 in the pairwise tissue comparisons). This is surprising as similar gene expression patterns are found both between heart and muscle and between cerebellum and cortex (Fig. 11B), but their APA patterns are clustered far away (Fig. 11A). Because many genes with differential major PAS usage between heart and muscle are consistent in two alleles, we think that genes with distinct APA patterns between muscle and heart might carry unique polyadenylation signatures important for driving tissue-specific features of the two organs.

Based on the results of this study, future research could address the influence of APA in mRNA translation efficiency or stability. Applying the F1 hybrid pan-tissue model with different *cis-* or *trans-* APA regulation, we can clearly identify genes with APA-regulated translation or stability differences between alleles. For example, as proposed in a recent study, the length of the poly(A) tail and translation efficiency are often coupled. A strong positive Spearman correlation between poly(A) tail length and mRNA translation efficiency (0.63~0.77) was found in early embryonic development[116], whereas shorter tails in somatic cells exhibit a length phasing pattern and tend to have high codon optimality and long half-lives[117]. The different PAS choices between alleles in a gene might differ the effects of PABP or microRNA binding to the gene's 3'UTR which leads to different poly(A) tail length in the mRNA's end to influence the mRNA's stability and translation efficiency.

In summary, general APA patterns of multi-PAS genes within and between tissues are consistent with the error hypothesis, with higher APA diversity being caused by noise resulting from weak *cis*-regulation. Besides the strength of *cis*-regulatory elements, PAS choice and APA diversity are also affected by PAS localization. However, not all minor PAS usage reflects molecular error, as there are many minor PASs with apparent functional importance, which are highly conserved and can compete with the major PASs. A small fraction of APA patterns is found to be regulated in a strongly tissue-dependent way. They are under intensive *trans*-regulation and conserved between the C57BL/6J and SPRET/EiJ alleles in our F1 hybrid mouse model. Additionally, some of the divergent functional APA might have shaped the unique phenotypes of the two species during evolution and contributed to their adaptation under natural selection in new environments.

Methods

RNA extraction from tissues and cultured cells

Tissues including cortex, cerebellum, heart, lung, liver, spleen, kidney and muscle were obtained from three adult F1 hybrid mice (C57BL/6J x SPRET/EiJ mouse strains). The mice were sacrificed by cervical dislocation, and tissues were dissected and stored in -80^{0}C before RNA extraction. The embryonic stem cells (ES cells) were cultured in Neuralbasal-DMEM F12 (Gibco) with N2B27 (Invitrogen), 2i (selleck), LIF (millipore). Total RNA was extracted from all tissues and cells using TRIzol reagent according to the manufacturer's protocol (Life Technologies). The integrity of purified total RNA was estimated by Agilent Bioanalyzer using RNA Nano kit (Agilent Technologies) before subsequent experiments. Total RNA with an RNA integrity number (RIN) above 9.0 was used for 3'mRNA library preparation.

3' mRNA library sequencing

QuantSeq 3' mRNA-Seq Library Prep Kit REV for Illumina (Lexogen) was applied to generate a library of sequences at the 3' end of RNA polyadenylation sites. In brief, 500 ng total RNA was taken as input. Polyadenylated RNA regions were reverse transcribed using an anchored oligo-dT primer, and second strand synthesis was initiated by random priming. PCR amplification was then performed to obtain an Illumina compatible sequencing library. All libraries were sequenced in paired-end 2 × 151 nt format on an Illumina HiSeq X Ten machine.

Alignment of sequencing reads

The C57BL/6J reference genome (mm10) was downloaded from Ensembl (http://www.ensembl.org). The SPRET/EiJ reference genome was created as described previously[106]. The reference set of PASs was obtained from PolyA_DB3 (http://exon.umdnj.edu/polya_db/), which provides precise polyadenylation positions of genes by 3'READS method[118]. We converted the PAS reference in mm9 to mm10 coordinates by g2gtools (version 0.1.29). To include imprecise cleavage reads near PAS we defined PAS clusters as a window of 48nt flanking the polyadenylation positions to quantified PAS usage reads. Genes with overlap PAS clusters and intergenic PASs from the database were removed in further analysis.

For the 3'mRNA-seq data, the 3' adaptor 5'-AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3' were first cut from forward reads, and 5'-AAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCGTCGTGTAGGG-3' from reverse reads by Cutadapt (version 1.18). The paired reads with each length longer than 15nt were mapped to mm10 and SPRET/EIJ genome separately by HISAT2 (version 2.0.1) with parameters --no-softclip --nodiscordant -x. Reads uniquely mapped to at least one of reference genomes were used. In these uniquely mapped reads, if a read could map to one genome with a shorter edit distance than to the other, it was assigned as unambiguously read in the specific allele with the shorter edit distance genome. If a unique mapped read had same edit distance to both parental genomes, it was assigned as common read. To unify the coordinates, reads assigned to SPRET/EIJ allele were converted to mm10 coordinates by g2gtools. The total unique reads from common reads, BL allele reads and converted SPRET/EIJ allele reads were used in hybrid system APA analysis.

PAS usage quantification

For PAS usage quantification in hybrid mice, PASs' cleavage position of genes are obtained from polyA_DB3 database[108]. A read with their 5'-end located in a window length of 48nt of flanking region of the PAS cleavage site (-24 nt to 24 nt) was counted as a PAS reads. Genes with a sum of PAS reads less than 20 were filtered out in each tissue separately. The PAS usage was calculated as its PAS reads divided by the sum of PAS reads in the gene. Genes with single identical PAS used in all tissues were viewed as single-PAS genes without alternative polyadenylation.

For allelic PAS usage, only reads unambiguously assigned to specific alleles were used. The allelic PAS usage was calculated as allele-specific PAS reads divided by the sum of allele-specific PAS reads in the gene. Known imprinted genes in mouse extracted from the Geneimprint database (http://www.gene-imprint.com/site/genes-by-species) and genes on the sex chromosomes or mitochondria were excluded from allelic PAS analyses. In order to avoid inaccurate calculation of allelic PAS usage due to sampling errors for low coverage PAS regions, PAS clusters with the sum of two allelic (C57BL/6J and SPRET/EIJ) PAS reads less than 5 were removed in each tissue. Genes with the sum of PAS reads in any allele less than 10 were removed. For the remaining PASs, we further estimated their relative PAS usages by using the combined unambiguously assigned allelic reads. If the difference between the relative PAS usage and total PAS usage estimated using the total reads was greater than 10%, we regarded the PAS cluster is with insufficient coverage of unambiguously assigned allelic PAS reads to calculate allelic PAS usage, and marked both allelic PAS usages (C57BL/6J and SPRET/EIJ) in the PAS cluster as missing value (NaN). Next, we compared the allelic PAS usage divergence between two alleles using DEXSeq as previous study[119]. Only PAS with Benjamini–Hochberg-adjusted P-value < 0.05 and delta percentage of pAs usage > 10% in all replicates is viewed as significantly divergent PAS.

In order to avoid sampling error in PAS identification caused by differences in sequencing depth between genes, we compared all results to those obtained when using downsampled data. In brief, we randomly picked the 20 reads from each tissue sample for all expressed genes, and use the selected reads to calculate the PAS usage as mentioned above. To validate the expression breadth difference between single-PAS genes and multi-PAS genes, a gene's sequencing reads are cumulated in all samples (normalized to same sequencing depth) and randomly picked 100 reads to tell if it is single-PAS or multi-PAS gene.

Gene expression quantification and adjusted variation calculation

Because PAS reads account for reads located in the end of transcripts directly, no gene length normalization was performed in gene expression's quantification. A gene' expression level was calculated as the sum of total PAS reads in the gene divided by sum of PAS reads from all genes in the sample and multiplied by 10^6 (referred as reads per million reads mapped, "RPM"). Allelic gene expression was calculated as the sum of unfiltered allele-specific PAS reads from the gene divided by sum of total PAS reads from all genes in the sample and multiplied by 10^6. Differential gene expression between alleles was analyzed with DESeq2[120]. Genes were considered as divergently expressed when the fold-change between alleles is greater than 2 in all replicates and the FDR is less than 0.05.

For coefficients of variation of gene expression between replicates and across tissue, we used following formula:

$$log10(CV_{gk}^2) = log10(CV_g^2) + \epsilon_{gk}$$

where ϵ_{gk} is independent distributed as a normal random variables for each gene g and sample k, and CV_g^2 is modeled as a smooth function of μ_g, the mean molecule count for gene g as described in calculating single cell transcriptional variation[121].

MicroRNA density scanning

To measure the density of microRNA targeting on mRNA's UTR region, conserved microRNA database from TargetScanMouse[122] (http://www.targetscan.org/mmu_72) was applied. Based on the usage difference of rank 1 and rank 2 PAS in a gene, tested genes were classified into three groups to compared the difference of microRNA density: 1) genes with usage difference between two sites large than 40%; 2) genes with usage difference in (20%, 40%]; 3) genes with usage difference equal or less than 20%. In the meantime, to ensure there exist a dominant PAS in gene (otherwise a gene could have highly diverse APA pattern), we further require genes in the last group have rank 1 PAS usage not less than 30%. For gene in these three groups, its 3' UTR from rank 1 PAS to downstream rank 2 PAS were selected and the density of microRNA targeting was calculated as average microRNA number at each position within all 1,000nt upstream the rank 2 PAS' cleavage site. If the distance from rank 1 to rank 2 is less than 1,000nt, the microRNA targeting numbers at positions larger than the actual distance were set as missing value (NAN). To calculate microRNA density in nearest end of rank 1 and rank 2 PAS upstream, 300nt upstream of each PAS' cleavage site was selected and the microRNA density were calculated as mentioned above. If the distance from gene's annotated last stop codon to PAS cleavage site is less than 300nt, the microRNA numbers at positions further than the actual PAS 3'UTR length to the cleavage site were set as missing value (NAN).

APA diversity in tissues, variability in replicates and APA variation between tissues

For APA diversity, we choose Shannon index defined as $-\sum_{i=1}^{S} p_i \ln p_i$ where p_i is the PAS usage in a gene. Shannon index measures the entropy of the alternative PAS usage. It can reflect both the number of different PASs and the evenness of the usage distribution across these sites.

To measure the PAS variability in tissue's replicates and PAS difference between tissue, we viewed each gene as vector containing its every PAS usage and used vector's maximum norm difference between repliactes as the index of APA replicates variability. In order to reduce potential error introduced by genes' different PAS number and reads depth, we further subtracted the vector's maximum norm difference by a mock difference. For this mock difference, we first created a mock data by pooling all reads in a gene from a tissue's three replicates and randomly separating the reads into three groups where gene's total reads number equal to original reads number in unshuffled data. The mock data was then used to calculated the mock difference in replicates. We repeated this calculation for mock difference 100 times and take the average maximum norm difference measured by these mock data as the subtracted mock difference to adjust the tissue replicates' variability.

For APA variation (switch score) between tissues, similar strategy was applied. The unadjusted APA variation is calculated as maximum norm difference between samples' PAS vector of gene's PAS usage. The mock data was created by combining a pair of tissues and randomly separating them into two tissue samples with same gene reads number as we did in APA variability mock. We define the gene's maximum switch score as the maximum adjusted APA variation in any pairwise tissues where the gene is expressed.

Estimation of sequence conservation, SNP densities and dN/dS ratios

PhastCons score from Glires Clade was used to estimate the sequence conservation. The PhastCons score data are obtained by PHAST (http://compgen.cshl.edu/phast/phastCons-tutorial.php)[123]. SNP density between C57BL/6J and SPRET/EIJ was calculated in chain file in our previous paper[106]. dN/dS ratios between the house mouse (*M. musculus*) and rat (*Rattus norvegicus*) were downloaded from the ENSEMBL database (ensembl.org), to estimate selective constraints on the amino acid sequences of proteins.

Bibliography

1. EDMONDS, M. & ABRAMS, R. Polynucleotide biosynthesis: formation of a sequence of adenylate units from adenosine triphosphate by an enzyme from thymus nuclei. *J. Biol. Chem.* **235**, 1142–1149 (1960).
2. Edmonds, M. & Caramela, M. G. The isolation and characterization of adenosine monophosphate-rich polynucleotides synthesized by Ehrlich ascites cells. *J. Biol. Chem.* **244**, 1314–1324 (1969).
3. Kates, J. & Beeson, J. Ribonucleic acid synthesis in vaccinia virus. II. Synthesis of polyriboadenylic acid. *J. Mol. Biol.* **50**, 19–33 (1970).
4. Darnell, J. E., Jelinek, W. R. & Molloy, G. R. Biogenesis of mRNA: genetic regulation in mammalian cells. *Science* **181**, 1215–1221 (1973).
5. Hammell, C. M. *et al.* Coupling of termination, 3' processing, and mRNA export. *Mol. Cell. Biol.* **22**, 6441–6457 (2002).
6. Brodsky, A. S. & Silver, P. A. Pre-mRNA processing factors are required for nuclear export. *RNA* **6**, 1737–1749 (2000).
7. Galy, V. *et al.* Nuclear retention of unspliced mRNAs in yeast is mediated by perinuclear Mlp1. *Cell* **116**, 63–73 (2004).
8. Hurt, E. *et al.* Mex67p mediates nuclear export of a variety of RNA polymerase II transcripts. *J. Biol. Chem.* **275**, 8361–8368 (2000).
9. Tseng, S. S. *et al.* Dbp5p, a cytosolic RNA helicase, is required for poly(A)+ RNA export. *EMBO J.* **17**, 2651–2662 (1998).
10. Stewart, M. Polyadenylation and nuclear export of mRNAs. *J. Biol. Chem.* **294**, 2977–2987 (2019).
11. Blobel, G. A protein of molecular weight 78,000 bound to the polyadenylate region of eukaryotic messenger RNAs. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 924–928 (1973).
12. Voeltz, G. K., Ongkasuwan, J., Standart, N. & Steitz, J. A. A novel embryonic poly(A) binding protein, ePAB, regulates mRNA deadenylation in Xenopus egg extracts. *Genes Dev.* **15**, 774–788 (2001).
13. Nicholson, A. L. & Pasquinelli, A. E. Tales of Detailed Poly(A) Tails. *Trends Cell Biol.* **29**, 191–200 (2019).
14. Eliseeva, I. A., Lyabin, D. N. & Ovchinnikov, L. P. Poly(A)-binding proteins: structure, domain organization, and activity regulation. *Biochemistry. (Mosc).* **78**, 1377–1391 (2013).
15. Malka, Y. *et al.* Post-transcriptional 3 -UTR cleavage of mRNA transcripts generates thousands of stable uncapped autonomous RNA fragments. *Nat. Commun.* **8**, 2029 (2017).
16. Richter, J. D. CPEB: a life in translation. *Trends Biochem. Sci.* **32**, 279–285 (2007).
17. Ortiz-Zapater, E. *et al.* Key contribution of CPEB4-mediated translational control to cancer

progression. *Nat. Med.* **18**, 83–90 (2011).

18. Calderone, V. *et al.* Sequential Functions of CPEB1 and CPEB4 Regulate Pathologic Expression of Vascular Endothelial Growth Factor and Angiogenesis in Chronic Liver Disease. *Gastroenterology* **150**, 982–97.e30 (2016).

19. Nevins, J. R. & Darnell, J. E. J. Steps in the processing of Ad2 mRNA: poly(A)+ nuclear sequences are conserved and poly(A) addition precedes splicing. *Cell* **15**, 1477–1493 (1978).

20. Fraser, N. & Ziff, E. RNA structures near poly(A) of adenovirus-2 late messenger RNAs. *J. Mol. Biol.* **124**, 27–31 (1978).

21. 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).

22. Hoque, M. *et al.* Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat. Methods* **10**, 133–139 (2013).

23. Ulitsky, I. *et al.* Extensive alternative polyadenylation during zebrafish development. *Genome Res.* **22**, 2054–2066 (2012).

24. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27**, 2380–2396 (2013).

25. Schaefke, B., Sun, W., Li, Y.-S., Fang, L. & Chen, W. The evolution of posttranscriptional regulation. *Wiley Interdiscip. Rev. RNA* e1485 (2018). doi:10.1002/wrna.1485

26. Filipowicz, W., Bhattacharyya, S. N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* **9**, 102–114 (2008).

27. To, K. K. W. *et al.* Escape from hsa-miR-519c enables drug-resistant cells to maintain high expression of ABCG2. *Mol. Cancer Ther.* **8**, 2959–2968 (2009).

28. Tranter, M. *et al.* Coordinated post-transcriptional regulation of Hsp70.3 gene expression by microRNA and alternative polyadenylation. *J. Biol. Chem.* **286**, 29828–29837 (2011).

29. 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).

30. Lackford, B. *et al.* Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO J.* **33**, 878–889 (2014).

31. Mayr, C. & Bartel, D. P. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138**, 673–684 (2009).

32. 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).

33. Ji, Z. *et al.* Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol. Syst. Biol.* **7**, 534 (2011).

34. Hoffman, Y. *et al.* 3'UTR Shortening Potentiates MicroRNA-Based Repression of Pro-differentiation Genes in Proliferating Human Cells. *PLoS Genet.* **12**, e1005879 (2016).

35. Kim, D., Kim, J. & Baek, D. Global and local competition between exogenously introduced

microRNAs and endogenously expressed microRNAs. *Mol. Cells* **37**, 412–417 (2014).

36. Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* **469**, 97–103 (2011).

37. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470**, 284–288 (2011).

38. Moraes, K. C. M., Wilusz, C. J. & Wilusz, J. CUG-BP binds to RNA substrates and recruits PARN deadenylase. *RNA* **12**, 1084–1091 (2006).

39. Gherzi, R. *et al.* A KH domain RNA binding protein, KSRP, promotes ARE-directed mRNA turnover by recruiting the degradation machinery. *Mol. Cell* **14**, 571–583 (2004).

40. Garneau, N. L., Wilusz, J. & Wilusz, C. J. The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.* **8**, 113–126 (2007).

41. Lal, A. *et al.* Concurrent versus individual binding of HuR and AUF1 to common labile target mRNAs. *EMBO J.* **23**, 3092–3102 (2004).

42. Ostareck, D. H. *et al.* mRNA silencing in erythroid differentiation: hnRNP K and hnRNP E1 regulate 15-lipoxygenase translation from the 3' end. *Cell* **89**, 597–606 (1997).

43. Goldstrohm, A. C., Hook, B. A., Seay, D. J. & Wickens, M. PUF proteins bind Pop2p to regulate messenger RNAs. *Nat. Struct. Mol. Biol.* **13**, 533–539 (2006).

44. Graham, R. R. *et al.* Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 6758–6763 (2007).

45. Mayr, C. Regulation by 3'-Untranslated Regions. *Annu. Rev. Genet.* **51**, 171–194 (2017).

46. Wang, E. T. *et al.* Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell* **150**, 710–724 (2012).

47. Spies, N., Burge, C. B. & Bartel, D. P. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* **23**, 2078–2090 (2013).

48. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

49. Neve, J. *et al.* Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation. *Genome Res.* **26**, 24–35 (2016).

50. Chen, L.-L. & Carmichael, G. G. Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. *Mol. Cell* **35**, 467–478 (2009).

51. Macdonald, P. M., Kerr, K., Smith, J. L. & Leask, A. RNA regulatory element BLE1 directs the early steps of bicoid mRNA localization. *Development* **118**, 1233–1243 (1993).

52. Lewis, R. A. *et al.* Conserved and clustered RNA recognition sequences are a critical feature of signals directing RNA localization in Xenopus oocytes. *Mech. Dev.* **121**, 101–109 (2004).

53. Chartrand, P., Meng, X. H., Singer, R. H. & Long, R. M. Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle in vivo. *Curr. Biol.* **9**, 333–336 (1999).

54. Ross, A. F., Oleynikov, Y., Kislauskis, E. H., Taneja, K. L. & Singer, R. H. Characterization of a beta-actin mRNA zipcode-binding protein. *Mol. Cell. Biol.* **17**, 2158–2165 (1997).

55. Blichenberg, A. *et al.* Identification of a *cis*-acting dendritic targeting element in the mRNA encoding the alpha subunit of Ca2+/calmodulin-dependent protein kinase II. *Eur. J. Neurosci.* **13**, 1881–1888 (2001).

56. Bashirullah, A., Cooperstock, R. L. & Lipshitz, H. D. Spatial and temporal control of RNA stability. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 7025–7028 (2001).

57. Martin, K. C. & Ephrussi, A. mRNA localization: gene expression in the spatial dimension. *Cell* **136**, 719–730 (2009).

58. An, J. J. *et al.* Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. *Cell* **134**, 175–187 (2008).

59. Ma, W. & Mayr, C. A Membraneless Organelle Associated with the Endoplasmic Reticulum Enables 3'UTR-Mediated Protein-Protein Interactions. *Cell* **175**, 1492-1506.e19 (2018).

60. Singh, I. *et al.* Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nat. Commun.* **9**, 1716 (2018).

61. Amara, S. G., Jonas, V., Rosenfeld, M. G., Ong, E. S. & Evans, R. M. Alternative RNA processing in calcitonin gene expression generates mRNAs encoding different polypeptide products. *Nature* **298**, 240–244 (1982).

62. Alt, F. W. *et al.* Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. *Cell* **20**, 293–301 (1980).

63. Pan, Z. *et al.* An intronic polyadenylation site in human and mouse CstF-77 genes suggests an evolutionarily conserved regulatory mechanism. *Gene* **366**, 325–334 (2006).

64. Higgs, D. R. *et al.* Alpha-thalassaemia caused by a polyadenylation signal mutation. *Nature* **306**, 398–400 (1983).

65. Wickens, M. & Stephenson, P. Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation. *Science* **226**, 1045–1051 (1984).

66. Montell, C., Fisher, E. F., Caruthers, M. H. & Berk, A. J. Inhibition of RNA cleavage but not polyadenylation by a point mutation in mRNA 3' consensus sequence AAUAAA. *Nature* **305**, 600–605 (1983).

67. Wang, R., Zheng, D., Yehia, G. & Tian, B. A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Res.* **28**, 1427–1441 (2018).

68. Kwok, C. K., Marsico, G., Sahakyan, A. B., Chambers, V. S. & Balasubramanian, S. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat. Methods* **13**, 841–844 (2016).

69. 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).

70. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18**, 18–30 (2016).

71. Gunderson, S. I., Polycarpou-Schwarz, M. & Mattaj, I. W. U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. *Mol. Cell* **1**, 255–264 (1998).

72. Devany, E. *et al.* Intronic cleavage and polyadenylation regulates gene expression during DNA damage response through U1 snRNA. *Cell Discov.* **2**, 16013 (2016).

73. Martin, G., Gruber, A. R., Keller, W. & Zavolan, M. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep.* **1**, 753–763 (2012).

74. Gruber, A. R., Martin, G., Keller, W. & Zavolan, M. Cleavage factor Im is a key regulator of 3' UTR length. *RNA Biol.* **9**, 1405–1412 (2012).

75. Zhu, Y. *et al.* Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. *Mol. Cell* **69**, 62-74.e4 (2018).

76. Thomas, P. E. *et al.* Genome-wide control of polyadenylation site choice by CPSF30 in Arabidopsis. *Plant Cell* **24**, 4376–4388 (2012).

77. Li, W. *et al.* Systematic profiling of poly(A)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet.* **11**, e1005166 (2015).

78. Kuhn, U. *et al.* Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. *J. Biol. Chem.* **284**, 22803–22814 (2009).

79. Chen, F., Chisholm, A. D. & Jin, Y. Tissue-specific regulation of alternative polyadenylation represses expression of a neuronal ankyrin isoform in C. elegans epidermal development. *Development* **144**, 698–707 (2017).

80. Smibert, P. *et al.* Global patterns of tissue-specific alternative polyadenylation in Drosophila. *Cell Rep.* **1**, 277–289 (2012).

81. Grozdanov, P. N., Li, J., Yu, P., Yan, W. & MacDonald, C. C. Cstf2t Regulates expression of histones and histone-like proteins in male germ cells. *Andrology* **6**, 605–615 (2018).

82. Harris, J. C. *et al.* The Cstf2t Polyadenylation Gene Plays a Sex-Specific Role in Learning Behaviors in Mice. *PLoS One* **11**, e0165976 (2016).

83. Takagaki, Y., Seipelt, R. L., Peterson, M. L. & Manley, J. L. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* **87**, 941–952 (1996).

84. Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469 (2008).

85. Hilgers, V., Lemke, S. B. & Levine, M. ELAV mediates 3' UTR extension in the Drosophila nervous system. *Genes Dev.* **26**, 2259–2264 (2012).

86. Zhu, H., Zhou, H.-L., Hasman, R. A. & Lou, H. Hu proteins regulate polyadenylation by blocking sites containing U-rich sequences. *J. Biol. Chem.* **282**, 2203–2210 (2007).

87. Yonaha, M. & Proudfoot, N. J. Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. *Mol. Cell* **3**, 593–600 (1999).

88. Cui, Y. & Denis, C. L. In vivo evidence that defects in the transcriptional elongation factors RPB2, TFIIS, and SPT5 enhance upstream poly(A) site utilization. *Mol. Cell. Biol.* **23**, 7887–7901 (2003).

89. Liu, X. *et al.* Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in Drosophila melanogaster. *RNA* **23**, 1807–1816 (2017).

90. Lutz, C. S. *et al.* Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro. *Genes Dev.* **10**, 325–337 (1996).

91. Kyburz, A., Friedlein, A., Langen, H. & Keller, W. Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3’ end processing and splicing. *Mol. Cell* **23**, 195–205 (2006).

92. Muller-McNicoll, M. *et al.* SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes Dev.* **30**, 553–566 (2016).

93. Hsieh, W.-P., Chu, T.-M., Wolfinger, R. D. & Gibson, G. Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* **165**, 747–757 (2003).

94. Shi, Y. *et al.* Molecular Architecture of the Human Pre-mRNA 3′ Processing Complex. *Mol. Cell* **33**, 365–376 (2009).

95. Tian, B. & Graber, J. H. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA* **3**, 385–396 (2012).

96. Lee, J. Y., Ji, Z. & Tian, B. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3’-end of genes. *Nucleic Acids Res.* **36**, 5581–5590 (2008).

97. Ara, T., Lopez, F., Ritchie, W., Benech, P. & Gautheret, D. Conservation of alternative polyadenylation patterns in mammalian genes. *BMC Genomics* **7**, 189 (2006).

98. Kainov, Y. A., Aushev, V. N., Naumenko, S. A., Tchevkina, E. M. & Bazykin, G. A. Complex Selection on Human Polyadenylation Signals Revealed by Polymorphism and Divergence Data. *Genome Biol. Evol.* **8**, 1971–1979 (2016).

99. Xu, C. & Zhang, J. Alternative Polyadenylation of Mammalian Transcripts Is Generally Deleterious, Not Adaptive. *Cell Syst.* **6**, 734-742.e4 (2018).

100. Schaefke, B. *et al.* Inheritance of gene expression level and selective constraints on *trans*- and *cis*-regulatory changes in yeast. *Mol. Biol. Evol.* **30**, 2121–2133 (2013).

101. Zhang, J. Neutral Theory and Phenotypic Evolution. *Mol. Biol. Evol.* **35**, 1327–1331 (2018).

102. Gruber, A. R. *et al.* Global 3’ UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nat. Commun.* **5**, 5465 (2014).

103. McManus, C. J., Coolon, J. D., Eipper-Mains, J., Wittkopp, P. J. & Graveley, B. R. Evolution of splicing regulatory networks in Drosophila. *Genome Res.* **24**, 786–796 (2014).

104. Wong, E. S. *et al.* Interplay of *cis* and *trans* mechanisms driving transcription factor binding and gene

expression evolution. *Nat. Commun.* **8**, 1092 (2017).

105. Xiao, M.-S. *et al.* Global analysis of regulatory divergence in the evolution of mouse alternative polyadenylation. *Mol. Syst. Biol.* **12**, 890 (2016).

106. Gao, Q., Sun, W., Ballegeer, M., Libert, C. & Chen, W. Predominant contribution of *cis*-regulatory divergence in the evolution of mouse alternative splicing. *Mol. Syst. Biol.* **11**, 816 (2015).

107. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).

108. Wang, R., Nambiar, R., Zheng, D. & Tian, B. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res.* **46**, D315–D319 (2018).

109. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27**, 2380–2396 (2013).

110. Mayr, C. Evolution and Biological Roles of Alternative 3'UTRs. *Trends Cell Biol.* **26**, 227–237 (2016).

111. Shi, Y. Alternative polyadenylation: new insights from global analyses. *RNA* **18**, 2105–17 (2012).

112. Xu, C. & Zhang, J. Alternative Polyadenylation of Mammalian Transcripts Is Generally Deleterious, Not Adaptive. *Cell Syst.* **6**, 734-742.e4 (2018).

113. Hilgers, V. Alternative polyadenylation coupled to transcription initiation: Insights from ELAV-mediated 3' UTR extension. *RNA Biol.* **12**, 918–921 (2015).

114. Schonemann, L. *et al.* Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes Dev.* **28**, 2381–2393 (2014).

115. Lee, J. Y., Ji, Z. & Tian, B. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. *Nucleic Acids Res.* **36**, 5581–5590 (2008).

116. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**, 66–71 (2014).

117. Lima, S. A. *et al.* Short poly(A) tails are a conserved feature of highly expressed genes. *Nat. Struct. &Amp; Mol. Biol.* **24**, 1057 (2017).

118. Zheng, D., Liu, X. & Tian, B. 3'READS+, a sensitive and accurate method for 3' end sequencing of polyadenylated RNA. *Rna* 1–9 (2016). doi:10.1261/rna.057075.116

119. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).

120. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

121. Kolodziejczyk, A. A. *et al.* Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell* **17**, 471–485 (2015).

122. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).

123. Siepel, A. & Haussler, D. Phylogenetic Hidden Markov Models. in *Statistical Methods in Molecular Evolution* 325–351 (Springer New York, 2005). doi:10.1007/0-387-27733-1_12

List of publications

Schaefke, B., Sun, W., **Li, Y.-S.**, Fang, L., & Chen, W. (2018). The evolution of posttranscriptional regulation. *Wiley Interdisciplinary Reviews. RNA*, e1485. https://doi.org/10.1002/wrna.1485

Xiao, M.-S., Zhang, B., **Li, Y.-S.**, Gao, Q., Sun, W., & Chen, W. (2016). Global analysis of regulatory divergence in the evolution of mouse alternative polyadenylation. *Molecular Systems Biology,* 12(12), 890. https://doi.org/10.15252/msb.20167375

Yang, J.-Y., Deng, X.-Y., **Li, Y.-S.,** Ma, X.-C., Feng, J.-X., Yu, B., … Gao, S. (2018). Structure of Schlafen13 reveals a new class of tRNA/rRNA- targeting RNase engaged in translational control. *Nature Communications,* 9(1), 1165. https://doi.org/10.1038/s41467-018-03544-x

Li, Y.-S., Qin, L.-X., Liu, J., Xia, W.-L., Li, J.-P., Shen, H.-L., & Gao, W.-Q. (2016). GIT1 enhances neurite outgrowth by stimulating microtubule assembly. *Neural Regeneration Research,* 11(3), 427–434. https://doi.org/10.4103/1673-5374.179054