Dissecting *cis*-regulatory effect on alternative polyadenylation using hybrid mice

A Dissertation
Submitted in Partial Fulfilment of the Requirements for
the Degree of
Doctor rerum naturalium (Dr. rer. nat.)
to the Department of Biology, Chemistry and Pharmacy
of the Freie Universität Berlin

by
Meisheng Xiao
肖梅生

Berlin, 2016

First supervisor: Prof. Dr. Wei Chen

Second supervisor: Prof. Dr. Markus Wahl

Date of defense: October 21, 2016

Declaration

I hereby declare that this thesis is my own original work, which was conceived and performed in Chen lab in Berlin Institute for Medical Systems Biology at Max Delbrück Center for Molecular Medicine in the Helmholtz Association under the supervision of Prof. Dr. Wei Chen. It has not been submitted to any other places for application of any degree. Contributions from others have been clearly mentioned in the preface and acknowledged in the text.

Xiao, Meisheng
Berlin, 2016

Preface

All the results I presented here are obtained by collaboration with Bin Zhang in Chen lab. Prof. Dr. Wei Chen and I conceived and designed all the experiments. I performed all the experiments, analyzed the data together with Bin Zhang. I interpreted the results together with Prof. Dr. Wei Chen and Bin Zhang. Yisheng Li contributed to the validation experiments. Wei Sun provides the fibroblast cell lines. Qingsong Gao and Xi Wang provide the RNA decay and translation efficiency data. Madlen Sohn, Claudia Quedenau, Mirjam Feldkamp, and Claudia Langnick contributed to the next generation sequencing and PacBio sequencing.

Xiao, Meisheng
Berlin, 2016

Acknowledgements

Time flies! Four years ago, I came to Berlin, joined MDC and started my scientific career abroad. During the passed time, I benefited a lot from both studying and living in Berlin. To those people who helped me, I would like to express my gratitude to him or her from the bottom of my heart!

Firstly, I would like to express my sincere gratitude to my mentor and supervisor Prof. Dr. Wei Chen for providing me such a great opportunity to do the exciting projects in his lab, and for his brilliant tutoring, encouragement, and motivation during the last four years. Thanks him for guiding me into the field of system biology. I would also like to thank him for the critical comments and proofreading of my thesis.

I would also like to show my greatest appreciation to my university supervisor Prof. Dr. Markus Wahl for offering me the opportunity to be one of the members in the Free University Berlin. Thanks him for helping me with all the process in the university and his critically reviewing my thesis.

Then, I would like to thank Bin Zhang for analyzing the sequencing data together with me. In addition, I want to than Yisheng Li for his assistance when I was doing the 3’ RACE and PAS motif validation experiments. Thanks Wei Sun for provide me the fibroblast cell lines. Thank Qingsong Gao and Xi Wang for sharing the decay and translation data with me. Moreover, I would like to thank Claudia Langnick, Mirjam Feldkamp, Madlen Sohn, and Claudia Quedenau for the sequencing and their excellent lab administration. I would also like to thank all my other labmates in Chen lab for the help and support all the time.

During the last four years in MDC, Sylvia Sibilak, Jennifer Stewart, Sabrina Deter, Annette Schledz, and Michaela Herzig really helped me a lot and I would like to thank them particularly.

Thanks my friend Dr. Thomas Rathjen for all his kind help when I was in Poy lab.

I also would like to thanks China Scholarship Council (CSC) for supporting my study in Germany.

Last but not the least, I would like to thank my wife Run Wang especially for her full support all the time and my little daughter who always bring a lot of fun to our family. In addition, I also want to thank my parents and my sister for their supporting. Without your persistent support and encouragement in Germany and China, it would be impossible for me to finish my study abroad!

Xiao, Meisheng
Berlin, 2016

Contents

Declaration I

Preface II

Acknowledgements III

Contents V

Summary 1

Zusammenfassung 3

Chapter 1 Introduction of alternative polyadenylation 5

1.1 Overview of APA 5

1.2 Discovery of APA 5

1.3 Different types of APA 6

1.4 Cis-elements, core complex, and RNA binding proteins (RBPs) 7

1.5 Genome wide analysis of APA 10

1.5.1 EST data *10*

1.5.2 Microarray data *11*

1.5.3 Next generation sequencing (NGS) *12*

1.6 Functional roles of APA 16

1.6.1 Localization *16*

1.6.2 Stability *17*

1.6.3 Translation *17*

1.6.4 Scaffold *18*

1.7 Dynamic regulation of APA 18

1.7.1 Extracellular stimulation on APA regulation *18*

1.7.2 Tissue specific APA *19*

1.7.3 Proliferation, differentiation, and reprogramming *19*

1.7.4 Cancer *20*

1.8 Mechanisms of APA 21

1.8.1 Interplaying between transcription and APA *21*

1.8.2 Splicing and APA *23*

1.8.4 Epigenetics and APA *24*

1.9 APA and human diseases 25

1.10 Evolution of APA, *cis*-elements, and RBPs 26

Chapter 2 Pervasive *cis*-regulatory alternative polyadenylation in mouse 29

2.1 Aim of this study 29

2.2 Materials and Methods....30

2.2.1 Cell culture and RNA extraction 30

2.2.2 3' READS....30

2.2.3 3' mRNA sequencing....31

2.2.4 Sequencing read alignment....32

2.2.5 pAs cluster identification 32

2.2.6 Filtering of C57BL/6J and SPRET/EiJ pAs reference 33

2.2.7 pAs annotation 34

2.2.8 pAs usage quantification....34

2.2.9 Creating mock F1 hybrids and pAs cluster filtering....34

2.2.10 PacBio sequencing and data analysis....35

2.2.11 Sequence variant density analysis....37

2.2.12 Local RNA secondary structure 37

2.2.13 Motif analysis....37

2.2.14 Vector construction and pAs reporter assay....38

2.3 Results....40

2.3.1 Construction of the pAs reference 40

2.3.2 APA divergence between C57BL/6J and SPRET/EiJ....46

2.3.3 Predominant Cis-regulatory pAs divergence between C57BL/6J and SPRET/EiJ ... 48

2.3.4 Genetic variants flanking the cleavage site contribute to cis-divergent pAs....53

2.3.5 RNA secondary structure in upstream proximal region inhibits pAs usage....56

2.3.6 cis-elements associated with divergent pAs expression 58

2.3.7 Functional consequence of cis-divergent pAs 62

2.4 Discussion....64

Bibliography....69

Publications 83

Curriculum Vitae 85

Summary

The 3’ ends of most eukaryotic mRNAs are cleaved and polyadenylated at the last step of transcription. Recent studies revealed that more than 70% mammalian genes have multiple polyadenylation sites (pAs) leading to the generation of multiple mRNA isoforms with different coding region or 3’ untranslated region (3’ UTR) from the same gene locus and contributes to the complexity of transcriptome and proteome by regulating their stability, localization, translation, and function. Boosted by the large-scale analysis technologies, extensive and dynamic regulation of 3’ UTR by alternative polyadenylation (APA) has been observed in different tissues; different cellular conditions (proliferation, differentiation, and development); and response to stimuli. Although the exact underlying mechanisms of APA remains under investigation, it should be in general regulated via the interaction between *cis*-regulatory elements residing at the DNA/RNA and *trans*-factors including polyadenylation cleavage core protein complex as well as other accessory RNA binding proteins (RBP).

Change of APA pattern during evolution remains underexplored. Such changes could arise from the divergence in *cis*-regulatory elements and/or *trans*-acting RBPs. The divergences of the two factors with different extent of pleiotropic consequences undergo distinct evolutionary trajectories. Therefore, to better understand evolution in APA, it is important to distinguish the relative contribution of *cis*- and *trans*-effects. In this project, to comprehensively investigate the contribution of *cis*-elements and *trans*-factors in the process of APA in a mammalian system, we identified and quantified pAs usage difference between two parental strains (C57BL/6J and SPRET/EiJ) and between the two alleles in the F1 hybrids with 3’ read capturing and sequencing (3’ READS) and 3’ mRNA-Seq methods, respectively. In total, we identified 3747 parental divergent pAs across five types of APA, between the two parental mouse strains. By comparing the parental divergent pAs with those in F1 hybrids, we observed a predominant contribution of *cis*-regulatory effect on pAs usage, which is mediated by genetic variants between two species around the pAs. Further sequence feature analysis demonstrated that the unstable secondary structure and a novel hexamer UUUUUU in the upstream region of pAs could enhance and inhibit the pAs usage, respectively.

Zusammenfassung

Die 3'-Enden der meisten eukaryotischen mRNAs werden im letzten Schritt der Transkription geschnitten und polyadenyliert. Jüngere Studien haben gezeigt, dass mehr als 70% der Gene von Säugetieren mehrere Polyadenylierungsstellen (pAs) haben. Diese ermöglichen die Generierung mehrer mRNA-Isoformen mit unterschiedlichen kodierenden oder 3‘-untranslatierten Regionen (3'UTR) aus demselben Genlokus und tragen zur Komplexität des Transkriptoms und des Proteoms bei durch Regulation ihrer Stabilität, Lokalisierung, Translation und Funktion. Mittels Einsatz von „large scale"-Technologien konnte die umfassende und dynamische Regulation des 3'UTRs durch alternative Polyadenylierung (APA) in verschiednenen Geweben gezeigt werden, sowie in verschiedenen zellulären Kontexten (Proliferation, Differenzierung und Entwicklung) und als Antwort auf Stimulation. Obwohl der genaue APA-Mechanismus noch untersucht wird, sollte er generell durch die Interaktion von cis-regulatorischen Elementen innerhalb der DNA oder RNA mit trans-Faktoren inklusive des *polyadenylation cleavage core protein complex* sowie zusätzlicher RNA-bindender Proteine (RBPs) vermittelt werden.

Die Veränderungen von globalen APA-Mustern während der Evolution sind wenig untersucht. Solche Veränderungen können aus der Divergenz von cis-regulatorischen Elementen und/oder trans-agierenden RBPs entstehen. Die Divergenzen der zwei Faktoren mit unterschiedlichem Ausmass an pleiotropen Konsequenzen verlaufen in unterschiedlichen evolutionären Bahnen. Deswegen ist es wichtig, die relativen Beiträge von cis- und trans-Effekten zu unterscheiden, um die Evolution von APA besser zu verstehen. Um den Beitrag von cis-und trans-agierenden Faktoren im APA-Prozess in einem Säugetiersystem umfassend zu untersuchen, haben wir in diesem Projekt *pAs usage* zwischen zwei parentalen Mausstämmen (C57BL/6J and SPRET/EiJ) und zwischen zwei Allelen von Hybriden der F1-Generation mittels *3' read capturing and sequencing* (3' READS) and *3' mRNA sequencing* identifiziert und quantifiziert. Insgesamt haben wir 3850 parentale, divergente pAs innerhalb von 5 APA-Typen zwischen den beiden parentalen Mausstämmen identifiziert. Durch den Vergleich der parental-divergenten pAs mit denen der Hybride der F1-Generation konnten wir beobachten, dass vorrangig cis-regulatorische Elemente einen Effekt auf die *pAs usage* haben, welcher durch genetische Varianten um die pAs herum herbeigeführt wird. Weitere Analysen der Sequenzeigenschaften konnten demonstrieren, dass instabile Sekundärstruktur sowie ein neuartiges UUUUUU-hexamer in der der pAs vorangehenden Region die *pAs usage* verstärken bzw. inhibieren können.

Chapter 1 Introduction of alternative polyadenylation

1.1 Overview of APA

Transcriptional end processing is an essential step during the maturation of a vast majority of eukaryotic genes transcribed by RNA polymerase II except for some replication dependent histone mRNAs (Proudfoot, 2011; Shi, 2012; Tian and Manley, 2013). It involves endonucleolytic cleavage of the nascent mRNA followed by adding a long stretches of untemplated poly (A) tail at the end of each transcript (Shi, 2012; Elkon et al., 2013). The length of poly (A) tail in different species is flexible with the highest average number of poly (A) in human mRNA (around 250 nt) comparing with about 70 - 80 nt in yeast (Elkon et al., 2013). mRNA with short poly (A) tail in general was thought to be subjected to degradation and/or have reduced translation efficiency. More than 70% of the mammalian genes and around half of the genes in flies, worms, and zebrafish were reported to have multiple polyadenylation sites (Jan et al., 2011; Derti et al., 2012; Smibert et al., 2012; Ulitsky et al., 2012; Hoque et al., 2013). APA could give rise to isoforms with different 3’ end of a gene during eukaryotic gene expression, thus significantly increased the diversity of transcripome and proteome. Mutations that affect different pAs usage were discovered in a series of human diseases including thrombophilia, thalassemias, and metachromatic leukodystrophy (Danckwardt et al., 2008; Elkon et al., 2013).

1.2 Discovery of APA

The earliest discovery of mRNA containing a terminal poly (A) came from the experiment, in which polysome associated mRNAs extracted from mammalian cells, were digested with both RNase (cuts C and U) and T1 RNase (cut G) (Lim and Canellakis, 1970 Edmonds et al., 1971; Adesnik et al., 1972; Mendecki et al., 1972). The digestion resistant fraction was thus thought to be poly (A). Subsequently, Winters and Edmonds identified, purified, and characterized the function of the polymerase, which is responsible for the synthesis of poly (A) tail at the end of mRNA, from calf thymus in 1973 (Winters and Edmonds, 1973b; Winters and Edmonds, 1973a). Later, in 1976 Proudfoot identified for the first time a common sequence AAUAAA located in the upstream 20 - 30 nt region of 3’ end (Proudfoot, 1976). This short sequence was then predicted to be polyadenylation signal (PAS).

Later, the APA was observed and one of the classic examples is DHFR (dihydrofolate reductase), which generates four distinct mRNA isoforms with different length of 3’ UTR ranging from 750 to 1600 nt without affecting the protein coding region (Setzer et al., 1980). Another example of APA comes from the IgM (immunoglobulin heavy chain), which was reported by three groups in 1980, independently (Alt et al., 1980; Early et al., 1980; Rogers et al., 1980). IgM produces two distinct mRNA isoforms by APA with one encoding for membrane targeted protein and the other one as secreted protein. Upon activation of the B cells, the membrane-bound protein was switched to secreted form and this phenomenon was reproduced by overexpression of CstF-60 in a B cell line (Takagaki et al., 1996). In 1997, Edwalds-Gilbert summarized about 120 reported genes that were subjected to APA regulation (Edwalds-Gilbert et al., 1997). The first global analysis of the prevalence of APA were performed by Tian et al. by using EST data sets and they identified ~ 50% and ~ 30% of genes have APA in human and mouse, respectively (Tian et al., 2005). Boosted by the next generation sequencing technology, a series of 3’ end sequencing methods were developed to comprehensively and accurately identify polyadenylation site (pAs) in different species and physiological/biological conditions. The details will be introduced in the following sections.

1.3 Different types of APA

APA could be grouped by different standards and here I classify it into five types according to a review written by Elkon et al. in 2013 (Figure 1) (Elkon et al., 2013). The first and most frequent type is called tandem 3’ UTR APA (including the terminal APA), which involves PAS in the terminal exon and this will produce mRNA isoforms with different length of 3’ UTR without affecting the protein-coding region. The second type is alternative last exon APA, which is regulated by alternative splicing. This type will give rise to mRNA isoforms with difference in both 3’ UTR length and protein-coding region. For PAS located in the intron region, which is defined as intronic APA, it creates an isoform with the last exon extended into the downstream intron region. Both 3’ UTR and protein coding region were different between the mRNA isoforms generated by intronic APA. The last and most rare one is called exonic APA. This type of APA will produce C terminal truncated proteins.

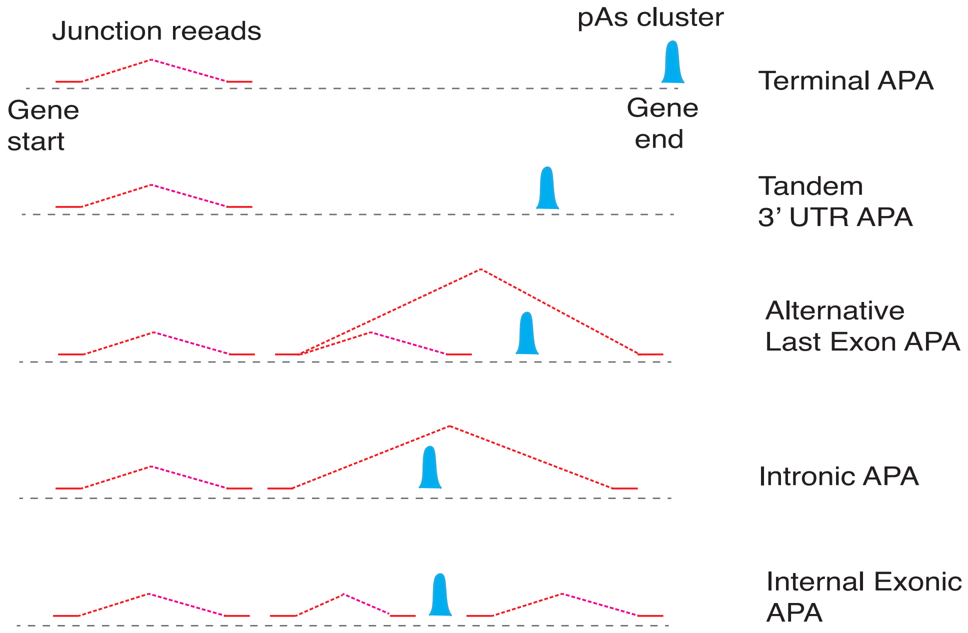

Figure 1.1 Five different types of APA.

1.4 Cis-elements, core complex, and RNA binding proteins (RBPs)

Earlier work on APA has identified a set of *cis*-elements and core complexes together with some auxiliary RBPs, which are all actively involved in the process of APA. Nucleotide composition around the cleavage site is generally AU rich (Figure 1.2). The 6 nt PAS motif AAUAAA predicted by Proudfoot in 1970s was confirmed by genome-wide analysis of the sequence features around the cleavage sites in both human and mouse using EST data sets (Proudfoot, 1976; Tian et al., 2005; Yan and Marr, 2005). Beside the canonical 6 nt motif AAUAAA, PAS can also adopt other close variants (Hu et al., 2005; Tian et al., 2005). The PAS is typically located in the upstream -30 to -15 nt region (Elkon et al., 2013). About 50% - 60% of the pAs in either human or mouse has the AAUAAA motif, which is considered as the strongest PAS, and approximately 15% to 20% of the pAs contain the AUUAAA (Tian et al., 2005). However, beside other rare PAS motifs, there are still 10% of the identified pAs seems do not harbor any known PAS. In addition to the upstream PAS located in region (-40 nt to 0), U- or GU-rich elements were also identified in the downstream (0 to 40 nt) region. Elements in these two regions were called core upstream elements (CUEs) and core downstream elements (CDEs), respectively. Apart from the region nearby cleavage site, elements in upstream of PAS and downstream of U- or GU-rich region were defined as auxiliary upstream elements (AUEs) and auxiliary downstream elements (ADEs) and they are also engaged in the regulation of pAs selection (Hu et al., 2005). Among the AUEs, UGUA motif plays an essential role in the APA when the canonical PAS AAUAAA is lacking. Furthermore, the secondary structure of the flanking region of pAs was also reported to be one of the effector on alternative polyadenylation (Bin Tian, 2008).

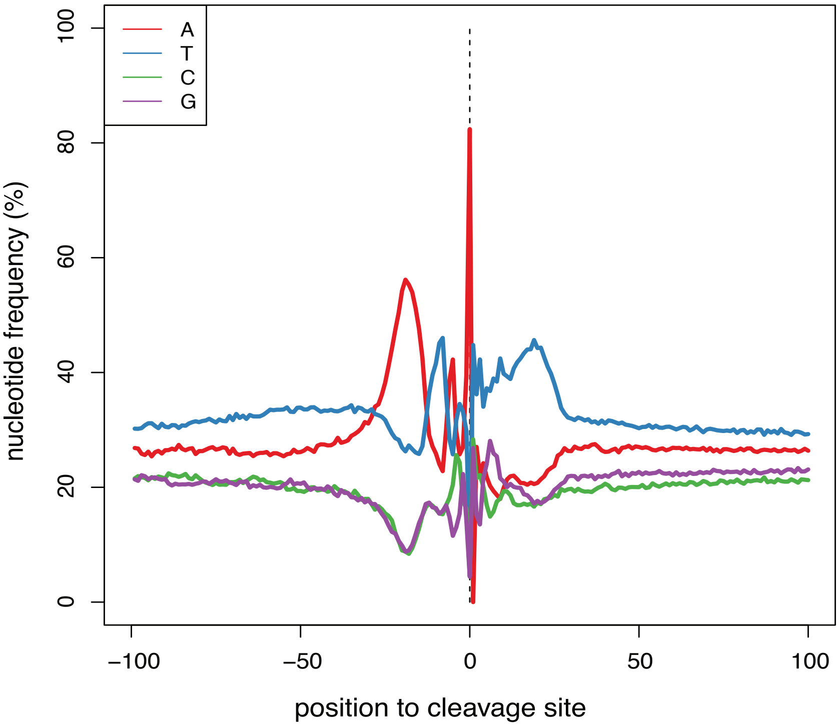

Figure 1.2 Nucleotide compositions around polyadenylation cleavage sites.

Approximately 16 core proteins were found to be actively involved in APA by using the method of affinity-purification to dissect the 3' end-processing complex of mammalian mRNAs (Millevoi et al., 2006; Martin and Keller, 2007; Mandel et al., 2008). Some of the peptides form different small complexes including: cleavage and polyadenylation specific factor (CPSF), cleavage stimulation factor (CstF) and cleavage factors I and II (CFIm and CFIIm) (Figure 3) (Elkon et al., 2013). The CPSF complex, which is responsible for recognition of AAUAAA, contains six proteins: CPSF160, Wdr33, CPSF 100, CPSF 30, CPSF 73, and Fip1 (Mandel et al., 2008; Shi et al., 2009). Previous studies suggested that the AAUAAA motif is targeted by CPSF 160, however, in 2014 two groups independently discovered that the CPSF 30 and Wdr33 are the direct binding protein instead of CPSF 160 (Keller et al., 1991; Chan et al., 2014; Schonemann et al., 2014). CPSF 73 has the endonuclease activity and is responsible for the cleavage of mRNA. The CstF complex, which contains three peptides, CstF-50, CstF-64, and CstF-77, exerts its function in selection of pAs (Elkon et al., 2013). Among them, CstF-64 binds to U- or GU-rich sequence downstream of the cleavage sites. CstF-64 is a rate-limiting protein in the formation of CstF complex and overexpression of it in human B cells could result in the switch of IgM protein from membrane bound to secret form (Schonemann et al., 2014). CFIm complex was composed of two CFIm 25 proteins, which can recognize the upstream UGUA motif and two other proteins (Coseno et al., 2008; Yang et al., 2011). A recent study reported that CFIm25 functions as a broad repressor of proximal pAs usage, when depleted, could increase the cell proliferation rate (Masamha et al., 2014). Additional proteins are also required for APA, including poly (A) polymerase, simplekin, and poly (A) binding proteins. Finally, since the transcription and 3' end processing of mRNA is coupled, RNA Po II is also actively engaged in the process of APA through the interaction between C terminal domain (CTD) and other APA related proteins (Figure 1.3) (Shi et al., 2009).

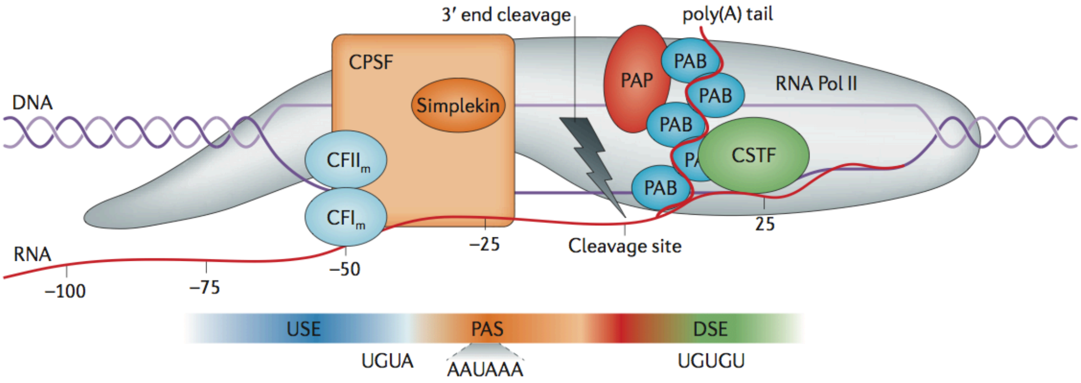

Figure 1.3 Core factor of the APA machinery.

Details of the machinery are depicted in the text. This figure is taken from (Elkon et al., 2013) with permission from Nautre Publishing Group (License Number: 3812481028450).

Apart from the core proteins involved in process of mRNA termination, a set of other RBPs has been reported to regulate the APA by direct and/or indirect contact with the APA machinery. To comprehensively examine the composition of 3' processing machinery and other proteins which are functional in modulating the pAs selection, Shi et al. purified the functional human pre-mRNA 3' processing complex and identified around 85 proteins are participating in regulation of pAs selection (Shi et al., 2009). Among the genes, there are transcription, splicing, and translation related proteins and this indicates that the 3’ end processing is a dynamic process and is coupled with different level’s regulation of gene expression. Takes U1 small nuclear ribonucleoprotein (snRNP) as an example, it participates in splice-site selection by base paring with 5’ splice site which is an essential step for the assembling of spliceosome complex. Physiological level of U1 snRNP would suppress the recognition of PAS in proximal intron region by APA complex and thus lead to longer mRNA isoform (Gunderson et al., 1998; Berg et al., 2012; Spraggon and Cartegni, 2013). Knockdown of U1 snRNP was demonstrated to result in the usage of proximal pAs located in intron region, implying that it plays a crucial role in preventing the premature termination of mRNA (Kaida et al., 2010). Another important RBP that can regulate the pAs usage is the ELAV, which is expressed in nervous system of Drosophila (Hilgers et al., 2012). This protein binds to the vicinity of proximal PAS and generates mRNA with extended 3’ UTR in neuron tissues by suppressing the proximal pAs usage (Oktaba et al., 2015). The human homologue of ELAV is HuR, which is also reported to prevent the usage of cleavage and polyadenylation in proximal region of 3’ UTR in neurons (Mansfield and Keene, 2012).

1.5 Genome wide analysis of APA

1.5.1 EST data

Through bioinformatics analysis of the EST data sets, Tian et al. identified ~54% and ~32% genes contain multiple pAs in human and mouse, respectively, for the first time (Tian et al., 2005). The characteristics of APA features between human and mouse are found to be extremely similar. The distance between the cleavage site and PAS ranging from 15 to 30 nt with the average of 20 nt in the upstream region. More than 90% of the pAs harbor canonical PAS or its close variants in both human and mouse (Table 1.1). Above 50% of human and mouse pAs have more than one cleavage site and this results in the heterogeneity in the cleavage sites. They also analyzed the nucleotide composition in the flanking region of cleavage site for different types of pAs, finding that the nearby region of cleavage sites is AU-rich.

Table 1.1 PAS frequencies in human and mouse (Tian et al., 2005)

PAS	Human (%)	Mouse (%)
AAUAAA	53.18	59.16
AUUAAA	16.11	14.9
UAUAAA	4.37	3.79
AGUAAA	3.72	3.28
AAGAAA	2.99	2.15
AAUAUA	2.13	1.71
AAUACA	2.03	1.65
CAUAAA	1.92	1.80
GAUAAA	1.75	1.16
AAUGAA	1.56	0.90
UUUAAA	1.2	1.08
ACUAAA	0.93	0.64
AAUAGA	0.60	0.36

1.5.2 Microarray data

Although EST data based analysis has greatly increased the number of pAs, its usage is cumbered by how the EST libraries were prepared and in which biological conditions. Microarray based gene expression analysis method was considered to be one of the alternative ways to quantify the pAs usage by integrating the pAs annotation derived from EST data sets. In brief, for each gene two probes are designed and one targets the common region of the two isoforms whereas the other probe targets the distal isoform specific region. The ratio of signal density between the isoform specific probe and common probe represents the relative usage of distal pAs and proximal pAs. By using this approach, Sandberg *et al.* firstly found a conserved increment of mRNA terminated in the proximal pAs upon the activation of T lymphocytes (Sandberg et al., 2008). Following the discovery of global shifted pattern of pAs usage in activated T cells, increased distal pAs usage was identified in neurons, resulting in mRNA with extended 3’ UTR (Miura et al., 2013). Furthermore, by using the microarray data of different conditions from databases, Ji *et al.* found that the extended 3’ UTR length in differentiated cells could be shortened by reprogramming those cells to iPS cells (Ji and Tian, 2009). There results indicate that the regulation of 3’ UTR length is related with different cell types, conditions, and status. Though microarray based approaches is powerful in characterizing the pAs usage pattern across different biological conditions, there are several limitations for this method. Firstly, the comparisons of pAs usage between different biological conditions is restricted by the probe designing. Secondly, it also depends on the previous defined pAs annotation. The third one is that it could only be used for comparing the gene with two pAs and genes contain more than two pAs would make it difficulty to quantify pAs usage. Finally, since it is based on probe hybridization and thus subject to issue of cross-hybridization.

1.5.3 Next generation sequencing (NGS)

1.5.3.1 RNA-Seq

High-throughput sequencing technology makes it possible to circumvent the limitations derived from using ESTs and microarray datasets. It could determine the expression level of different isoforms generated by alternative splicing and alternative polyadenylation. Wang *et al.* investigated the alternative mRNA isoform regulation in different human tissues systematically by sequencing 15 diverse human tissues and cell lines (Wang et al., 2008). They identified in total 5136 3’ UTR isoforms generated by alternative polyadenylation and around 75% of the events was differentially regulated in different tissues. In order to make better use of RNA-Seq data, a software named as Dapars was developed to infer the pAs usage dynamics directly with deep RNA-Seq data from different conditions (Xia et al., 2014). In the study, they analyzed 358 tumor samples with Dapars and found that 1346 genes were modulate by APA and the vast majority of the affected genes switched to use the proximal pAs, which can avoid degradation potentially mediated by specific microRNA.

1.5.3.2 3’ end-Seq

RNA-Seq is a powerful method for genome wide analysis of gene expression including quantification of gene expression level, alternative splicing pattern, alternative polyadenylation in 3’ UTR, but it cannot identify novel pAs efficiently. To comprehensively determine pAs at the genome-wide level, several high-throughput sequencing based 3’ end-profiling methods were developed in recent years (Sun et al., 2012; Elkon et al., 2013). In general, these methods could be divided into three groups: oligo (dT) based reverse transcription (1-4); ligation based reverse transcription (5-6), and direct single molecule RNA sequencing (7).

(1) 3'-end RNA-Seq (Yoon and Brem, 2010)

In this protocol, the poly (A) containing RNA was enriched after fragmentation. Anchored oligo (dT) primers were used to synthesize the cDNA. Illuminar paired-end adapters were ligated to cDNA ends and then PCR was performed to amplify the library. The purified PCR products were used for Illuminar paired-end sequencing.

(2) PAS-Seq (Shepard et al., 2011)

Poly (A) RNA were enriched and fragmented and followed by reverse transcription using SMART RT system. Here the anchored oligo (dT) primers linked with Illuminar sequence was used for RT. After second strand cDNA synthesis, PCR was performed to amplify the library. A customized primer with oligo (dT) extension in the 3' end was used for single end sequencing on Illuminar sequencing platform.

(3) Poly (A)-Seq (Derti et al., 2012)

This method is similar to PAS-Seq, it uses Illuminar adaptor sequence linked with random primers for second strand cDNA synthesis instead of using template switch.

(4) A-Seq (Martin et al., 2012)

It is similar as small RNA seq protocol but uses a stem containing oligo (dT) for reverse transcription. Then the amplified PCR products were used for Illuminar sequencing.

(5) 3-P Seq (Jan et al., 2011)

This method begins with a splint-ligation, which prefers to bind the end of poly (A) containing RNA. After partial digestion by RNase T1, poly (A) containing 3' end of mRNA were reverse transcribed by customized primers and dTTP only. Then the DNA-RNA hybrids were partial digested by RNase H and the remaining short poly (A) containing RNA were followed by adapter ligation, reverse transcription, PCR, and Illuminar sequencing (Figure 4).

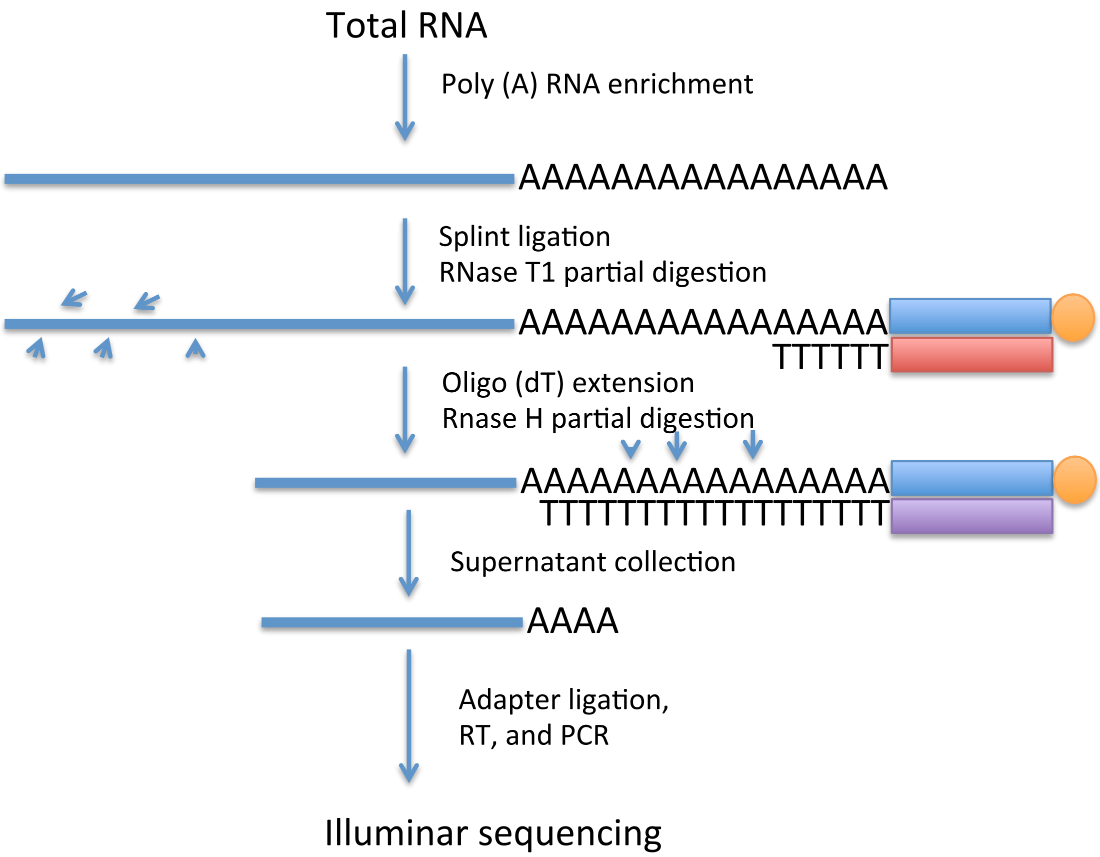

Figure 1. 4 Experimental procedure of 3P-Seq. Poly (A) containing RNAs are enriched with oligo (dT) beads and followed by splint ligation. Then 5' part of the transcripts are digestion with RNase T1 partially. RNase H digestion is performed to release the transcripts with short length of poly (A). The purified products are then used for adapter ligation, RT, PCR, and sequencing. Reads with non-genomic T will be used for pAs calling. Blue and orange stand for splint; purple represents RT primer; yellow is streptavidin bead.

(6) 3' READS (Hoque et al., 2013)

In this method, two rounds of poly (A) containing RNA enrichment were employed with commercial oligo (dT) beads and customized chimeric rU_5T_{45} conjugated streptavidin My One beads, respectively. The long Poly (A) tail was shortened to contain around 5 As on average for each transcript by using RNase H and followed by adapter ligation, RT, and PCR amplification. The purified PCR products were sequenced on Illuminar sequencing platform. After sequencing, only reads with at least 2 non-genomic T were considered as pAs supporting (PASS) reads and only PASS reads were used for pAs calling (Figure 5).

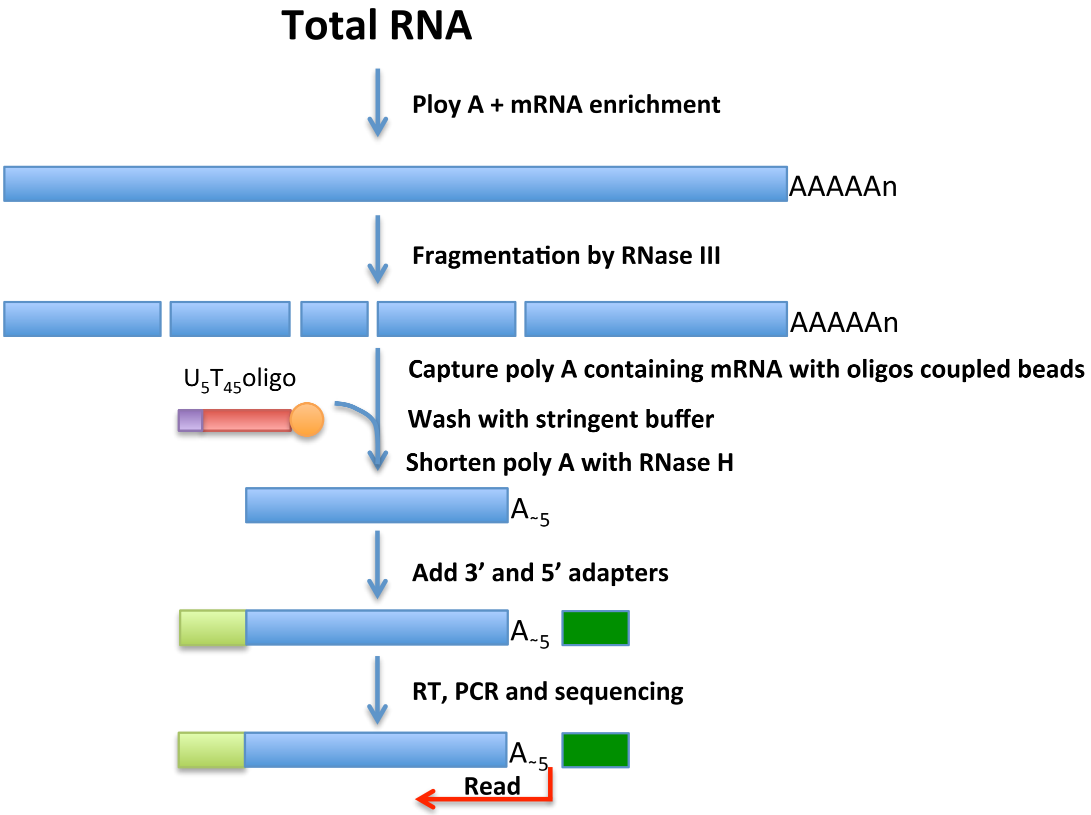
Total RNA
Ploy A + mRNA enrichment
AAAAAn
Fragmentation by RNase III
AAAAAn
U5T45oligo
Capture poly A containing mRNA with oligos coupled beads
Wash with stringent buffer
Shorten poly A with RNase H
A~5
Add 3' and 5' adapters
A~5
RT, PCR and sequencing
A~5
Read

Figure 1.5 Schematic map of 3' READS method. In brief, 30 μg total RNA are used as starting material for poly (A) containing RNA enrichment and then on bead digestion is performed to remove away the 5' part of each poly (A) containing RNA. Chimeric rU_5T_{45} oligos coupled streptavidin My One beads are used to further enrich the poly (A) containing RNA and followed by RNase H digestion to remove the long stretches of poly (A) with about 5 As in the end of each fragment. After 3' and 5' adapter are added sequentially, RT is performed with specific primers and Illuminar adapter sequence compatible PCR primers are used to amplify the library. The library is sequenced from the 3' end of each transcript. Reads with at least 2 non-genomic T will be used to pAs supporting reads and infer the pAs.

(7) DRS (Ozsolak et al., 2010)

This method is based on single molecule sequencing. Poly (A) containing RNA are captured by oligo (dT) coated flow cell and the sequencing begins from the 3' end without selection, digestion, ligation, and PCR reaction.

For method 1-4, oligo (dT) primers were used for RT and this will cause internal priming at genomic A-rich region, resulting in false positives of identified pAs. To rule out the potential internal priming problem, different bioinformatic approaches were developed to remove the reads potentially derived from internal priming, but these filtering cannot guarantee the accuracy and efficiency. In addition, some genome regions around true pAs are A-rich and would be filtered out, therefore resulting in false negatives. To avoid the internal priming issue, both 3-P and 3' READS method, did not use oligo (dT) for RT, and only consider the sequencing reads with non-genomic T as potential poly-A derived reads, which were used for identifying pAs. These two methods greatly increased the accuracy and efficiency of pAs identification although the methods are technically much more complicated and cumbersome. For the single molecule direct RNA sequencing, though oligo (dT) is also utilized during sequencing procedure, the authors found no or few pAs were generated by internal priming (Ozsolak et al., 2010).

1.6 Functional roles of APA

Based on the recent global analysis using next generation sequencing technologies, alternative polyadenylation was found to be widespread and more than half of the genes have multiple pAs in eukaryotic cells. APA could affect gene expression post-transcriptionally or co-transcriptionally by several mechanisms.

1.6.1 Localization

APA isoforms with different length in 3' UTR could modulate the localization of mRNA if the *cis*-elements in the differential 3' UTR could be targeted by specific RBP, which can further interact with motor protein (Ephrussi et al., 1991; Bertrand et al., 1998). A number of mRNAs were transported to specific subcellular compartment by RBP and elements in the 3' UTR (An et al., 2008; Loya et al., 2008; Vuppalanchi et al., 2010). This phenomenon is much more prevalent in polarized cells (Jung et al., 2012). BDNF (brain derived neurotropic factor), which plays an important role in synaptogenesis and synaptic plasticity, is a typical example. This gene could generate two different 3' UTR isoforms, the longer isoform localizes to dendrites, where it regulates the memory formation, while the short version is retained in soma region (An et al., 2008). The localization of mRNA to the dendrites is impaired when the long 3' UTR isoform is truncated in a mutant mouse. Even modest regulation of BDNF isoforms could have severe impact on behavior, memory, and emotion (An et al., 2008). The other example is IMPA1 (myo-inositol monophosphatase-1), which can also produce two 3' UTR isoforms with the shorter one localizes to soma and the longer one to axon (Andreassi et al., 2010). In that study, they found that the axon localized IMPA1 mRNA could be translated locally in response to the nerve growth factor (NGF) stimulation and impaired synthesis of IMPA1 in axons resulted in axonal degeneration. These results highlight the important role of mRNA localization in establishing cellular polarity and specific functions.

1.6.2 Stability

If a RBP, which could recruit/avoid deadenylation and/or decapping factors, binds to 3' UTR, the half-life or stability of the affected mRNA would be affected. *Cis*-elements controlling the stability of mRNA including: AU rich elements and microRNA binding sites. AU rich elements recognized by AU-rich RBPs and 3' UTR region targeted by microRNAs leads to increased degradation rate. The reduced expression level of mRNA will result in lower protein abundance of the affected gene. Takes the oncogene IGF2BP1 for an example. It expresses both short and long 3' UTR isoforms. The short isoform generates sufficient amount of proteins for the transformation of fibroblast in mice, while the longer isoform, which harbors repressive elements gives rise to lower amount of proteins, lost the transformation ability (Mayr and Bartel, 2009). This result suggests that short 3' UTR would escape the regulation by microRNA or AU rich elements, and thus produce more proteins and results in increased transformation ability. However, in a large-scale analysis of correlation between alternative 3' UTR usage and RNA stability, only week correlation was observed between short 3' UTR and increased mRNA stability at the genome-wide level (Spies et al., 2013; Masamha et al., 2014). In line with this, a recent study reported that the number of stabilizing and destabilizing *cis*-elements in 3' UTR is comparable (Oikonomou et al., 2014).

1.6.3 Translation

3' UTR could affect the translation efficiency through interaction between RBPs and *cis*-elements. Including or excluding the potential *cis*-elements in 3' UTR region by APA could regulate translation. Again, takes the BDNF gene for an example, it is reported that the short version of 3' UTR isoform is translated in the steady state of hippocampal neurons, however once the neuron is activated, the translation will switch to isoforms with long 3' UTR (An et al., 2008). Pinto *et al.* reported that polo gene, a cell-cycle gene, generates two isoforms by APA (Pinto et al., 2011). The longer 3' UTR isoform was found to have higher translation efficiency than the short one. Intriguingly, mutation of the distal PAS of polo gene, which would have reduced expression of long 3' UTR isoform, the flies died at pupa stage caused by proliferation defects of the abdomen precursor cells, whereas, disruption of the proximal PAS had no effect. This may suggest that the longer 3’ UTR isoform with higher translation efficiency is essential for the normal development of precursor cells in abdomen of flies.

1.6.4 Scaffold

A recent study focused on the CD47 gene, which generates two 3’ UTR isoforms but encoding the same protein, identified a new function of 3’ UTR on protein localization, which was independent of the mRNA localization (Berkovits and Mayr, 2015). The protein translated from the short isoform (CD47-SU) was predominantly staying in the endoplasmic reticulum where it is synthesized, however, the protein generated from the long one (CD47-LU) was mostly localized to membrane (Berkovits and Mayr, 2015). In the paper, the author demonstrated that the 3’ UTR region of the longer isoform was targeted by RNA binding protein HuR (also known as ELAV in fly), which resulted in the recruitment of SET to the translation site. Then the SET was transferred to the newly synthesized region of CD47 protein and created an interaction between CD47 and SET. SET was further recognized by RAC1 and the activated form of RAC1 would translocate the CD47 to membrane. In contrast, for the short 3’ UTR isoform, since there is no binding site for HuR and thus SET is not recruited. This will make the translated protein from short 3’ UTR mRNA retain in endoplasmic reticulum. The mechanism of 3’ UTR dependent protein localization is potentially a widespread phenomenon, since HuR has thousands of binding sites in different mRNAs. Through validation of other three genes CD44, ITGA1, and TNFRSF13C, which all have the binding site by HuR in the longer 3’ UTR region, they found that the membrane localized protein level was higher comparing with the corresponding mRNA with shorter 3’ UTR. This indicates that the new identified function of 3’ UTR affecting the localization of the translated protein plays an essential role in post-transcriptional regulation.

1.7 Dynamic regulation of APA

1.7.1 Extracellular stimulation on APA regulation

Previous work on APA discovered that the activation of T cells could regulate the pAs usage of different individual genes (Peattie et al., 1994; Edwalds-Gilbert et al., 1997; Chuvpilo et al., 1999). Sandberg *et al.* performed a global analysis of the APA regulation during T cell activation for the first time using microarray data and found that upon T cell activation, the pAs usage pattern was shifted from distal to proximal and this leaded to the increased expression level of 3' UTR isoforms terminating in the upstream pAs (Sandberg et al., 2008). In contrast, a later study on the same cell line by NGS did not identify significant correlation between 3' UTR length and expression level of 3' UTR isoforms, though the shifted pattern of pAs usage was recapitulated (Gruber et al., 2014). In another study, Flavell *et al.* found that a number of genes have increased expression level of 3' UTR isoform using the proximal pAs when neuronal cells were activated (Flavell et al., 2008). In the same study, the data from several other stimulations including EGF (epidermal growth factor), interleukin-2, and anti-IgM on different cell types were analyzed and identified the distinct pattern of pAs uage change for different stimulations. These results suggest that APA can be modulated in a pathway specific manner in response to different kinds of extracellular induction.

1.7.2 Tissue specific APA

Global analysis of APA isoforms in different tissues of both human and mouse by using EST data sets has found biased expression of APA isoforms (Beaudoing and Gautheret, 2001; Zhang et al., 2005). Genes expressed in neurons or brain tissues more preferred to use distal pAs, which generates long 3' UTR containing mRNA, while the pAs in other tissues including placenta, ovaries, and blood showed the opposite pattern (Zhang et al., 2005). Peter *et al.* analyzed the tissue specific pattern of APA in fly with deep RNA-Seq data and found similar trend for 3' UTR lengthening in central nervous system (CNS) and shortening in the testis (Smibert et al., 2012). The pAs usage pattern is more similar among the same tissues of different species than different tissues from the same species (Derti et al., 2012). This pattern is even more obvious for brain and liver samples.

1.7.3 Proliferation, differentiation, and reprogramming

Apart form the finding of a general 3' UTR length shortening in T cells upon stimulation based on microarray data, the other crucial discovery of the same study was that the global shifted promoter-proximal pAs pattern was accompanied by the increased cell proliferation (Sandberg et al., 2008). And similar pattern was found when human B cells and monocytes were stimulated (Elkon et al., 2013). In line with these findings, a comprehensive analysis of a number of human tissues and cell lines showed a negative correlation between the 3’ UTR length and proliferation rate (Elkon et al., 2013). Thus according to these observations, it is speculated that the enhanced proximal pAs usage is associated with increased cell proliferation. The striking trend might be explained as the long 3’ UTR containing mRNA would rend it to be targeted by microRNA, whereas the short 3’ UTR isoform could thus avoid the degradation by microRNA. However, one recent study showed that the relationship between 3’ UTR length and cell proliferation could be much more complex than what were thought before (Fu et al., 2011).

Ji *et al.* examined the APA regulation during mouse embryonic development using SAGE (serial analysis of gene expression) and microarray data sets identified that embryonic development is associated with increased usage of distal pAs (Ji et al., 2009). Furthermore, in the study they also compared the pAs usage pattern between myoblasts of C2C12 and its differentiated form myotubes, and the results are consistent with that observed during embryonic development. The progressive lengthening of 3’ UTR along with development was further consolidated by additional global studies in different species including fly and zebrafish (Shepard et al., 2011; Hilgers et al., 2012; Smibert et al., 2012; Ulitsky et al., 2012).

As the opposite direction of differentiation, reprogramming could be used to generate iPS cells (iduced pluripotent cells) from differentiated cells. Ji *et al.* examined APA dynamics with different microarray data sets and revealed that the 3’ UTR length was shortened during the processes of reprogramming of somatic cells to iPS cells with the exception of spermatogonial cells, which were accompanied by increased length of 3’ UTR (Ji and Tian, 2009). The special case maybe explained by the fact that spermatogonial cells are more proliferative than ES cells. The finding indicates that reprogramming of 3’ UTR length is an integral part of the process of iPS cell generation although the underline mechanism is still unknown.

1.7.4 Cancer

One of the hallmarks of cancer is the uncontrolled cell proliferation. Consistent with the above context, comparing with the non-transformed cell lines, the 3’ UTR length of many genes in cancer cell lines were found to be substantially shortened, which is mediated by APA (Mayr and Bartel, 2009). The functional consequence of short 3’ UTR is the increased mRNA stability, which is in part regulated by the depleted microRNA binding sites, and thus results in the augmented protein production. Following this discovery, the pattern of biased length of 3’ UTR was further observed in other cancers including lymphoma, colorectal cancer, breast cancer, and lung cancer (Singh et al., 2009; Lembo et al., 2012; Morris et al., 2012). Interestingly, the changes of APA profile could also be used to separate distinct subtypes of cancers, which represent different stages during caner development (Singh et al., 2009). Both breast and lung cancer cells expressing mRNA with shorter 3’ UTR presented to be much more aggressive (Lembo et al., 2012). This makes the dynamic patterns of APA to be a potential diagnostic marker for some cancers. A recent study on 358 tumor/normal pairs of samples from seven tumor types have demonstrated that more than 90% of the APA regulated genes have shortened 3’ UTR, which can avoid repression mediated by microRNA (Xia et al., 2014). However, in another study comparing the tandem 3’ UTR usage among three cell lines (MCF7, MCF10A, and MB231), Fu *et al.* identified that MCF7 had shortened 3’ UTR, while MB231 had lengthened ones (Fu et al., 2011). These controversial results suggested the complexity of APA regulation in cancer cells was far more than our previous thoughts.

1.8 Mechanisms of APA

1.8.1 Interplaying between transcription and APA

Recent studies begin to reveal that the regulation of gene expression is coordinated through a complex and extensively coupled network in which different functional parts or machineries are interacted to form a huge platform. This can maximize the efficiency and specificity of each single regulatory step (Maniatis and Reed, 2002). The coupling between transcription and 3’ end processing is mediated in a number of ways such as: transcription initiation, transcription activity, transcription rate, RNA polymerase II modification, splicing factors linked to transcription machinery (Shi et al., 2009; Ji et al., 2011; Hilgers, 2015).

PC4, a human transcriptional co-activator, interact with CstF-64, which is a core component of the alternative polyadenylation machinery. The homologs of the two proteins in yeast also connect with each other and this interaction is required for termination of transcription. The observations raised a model to explain the dynamic processes of transcription and termination. In the model, PC4 first involves in the assembling of transcription complex and then during transcription elongation stage it bind to CstF-64 and inhibit the premature transcription termination. Once the transcription machinery recognized the 3' end processing signal, PC4 would dissociated from the complex and allow the CstF-64 to exert its function in cleavage and polyadenylation (Calvo and Manley, 2001). Transcription elongation complex PAF1C, which is required for efficient 3' end processing, provides another example. Knocking down the CDC73 subunit of the PAF1C complex leads to decreased loading of CPSF and CstF in the PAF1C target genes locus, which results in augmented distal pAs usage (Rozenblatt-Rosen et al., 2009). Furthermore, Ji *et al.* observed that highly expressed genes generated relatively more mRNA with short 3' UTR while lowly expressed genes produced more mRNA with longer 3' UTR through analyzing human and mouse transcriptomes. Additional reporter assay demonstrated that the pAs selection is modulated by transcription activity (Ji et al., 2011). A recent study in flies identified that gene specific promoter sequence could decide the 3' UTR length. This process involves the ELAV (embryonic lethal abnormal vision) protein whose homolog in human is HuR protein and it could bind to promoter, intron, and 3' UTR region of the regulated genes (Hilgers et al., 2012; Oktaba et al., 2015). The binding of ELAV in the promoter region mediates the neuron specific 3' UTR extension in flies by suppressing the proximal pAs usage. Replacing the promoter sequence with the native promoter sequence of another 3' UTR extendable gene showed the same phenomenon, while exchange with a strong synthetic promoter led to failed 3' UTR extension mediated by ELAV (Oktaba et al., 2015). This discovery suggests that the ELAV mediated 3' UTR extension is promoter sequence dependent and there is a potential coupling between alternative polyadenylation and transcription initiation.

During transcription, generally proximal pAs is transcribed first and thus rend it to the 3' end processing machinery earlier. This makes it more likely to be used than the distal one. So in theory, the pAs selection was decided based on both the distance between two adjacent pAs and the transcription rate. In line with this concept, Pinto *et al.* used a D. melanogaster strain as a model and found that a mutant RNA Po II with lower elongation rate resulted in increased usage of proximal pAs, while the wide type RNA Po II used less proximal pAs (Pinto et al., 2011). Similar mechanism was found to regulate alternative splicing previously (de la Mata et al., 2003). Although the underling mechanisms of transcription rate on pAs selection are still elusive, epigenetic factor may play a role with accumulating evidence (Brown et al., 2012).

1.8.2 Splicing and APA

Accumulating evidence indicates that splicing and polyadenylation is coupled during transcription (Proudfoot et al., 2002; Millevoi et al., 2006; Tian et al., 2007; Shi et al., 2009). Among the different types of alternative polyadenylation, the generation of both alternative terminal exon and intronic pAs are regulated by alternative splicing together with 3' end processing. Tian *et al.* performed a global analysis of human EST data sets of all human genes and identified that around 20% of them have at least one intronic pAs. The week 5' SS and long intron size leads to increased usage of intronic pAs (Tian et al., 2007). This discovery suggested that there is a dynamic competition between 5' splice site (5' SS) and its downstream APA. The coupling between alternative splicing and polyadenylation is further consolidated by examining the deep RNA-Seq data of a diverse human tissues and cell lines (Wang et al., 2008). In the study, it was observed that the pattern of alternative splicing and APA among different tissues was highly correlated, indicating coordinated regulation of the two processes. Subsequent sequence analysis of the shared regulatory motifs between alternative splicing and APA showed extreme conservation, suggesting a common and essential role of the corresponding binding proteins in regulating both alternative splicing and polyadenylation regulation.

Several splicing related factors were reported to affect the 3' end processing (Millevoi and Vagner, 2010). On examples is U2AF 65 (U2 snRNP Auxiliary Factor 65), which plays an important role in splicing and engages in coupling of splicing and APA. U2AF 65 first binds to pyrimidine tract in 3' SS of the last intron and then recruits CFIm complex to its downstream pAs, thereby stimulates the cleavage and polyadenylation. At the same time, the U2AF 65 is tethered by poly (A) polymerase (PAP) and activates the splicing reaction in turn (Millevoi et al., 2006). Similarly, SF3b, which is a subunit of U2 snRNP, directly interact with CPSF complex and the interaction was found to be required for both efficient splicing and alternative polyadenylation. Either depletion of CPSF or mutation of U2 snRNP leads to impaired coupling of splicing and 3' end processing (Kyburz et al., 2006).

Another example is the U1 snRNP, which is a crucial component of splicesome and functions in defining the 5' SS by RNA-RNA base paring through the U1 snRNA. The abundance of U1 snRNP was estimated to be ~ 10^6 molecule per cell and that is much more than other U snRNPs. By knocking down the U1 snRNP with antisense morpholino oligocleotide, Kaida *et al.* observed in addition to the inhibition of splicing, increased level of pre-mature cleavage and polyadenylation, which utilized the intronic cryptic pAs. However, the phenomenon was not observed when U2 snRNP was depleted with the same method. This demonstrated that the U1 snRNP protects pre-mRNA from premature cleavage and polyadenylation in upstream intron region independent of splicing (Kaida et al., 2010). In another study led by the same group, they found that moderate decrease of U1 snRNP resulted in insufficient inhibitory of splicing and the pattern of pre-mature cleavage and polyadenylation was shifted to proximal pAs in 3’ UTR region in a dose-dependent manner. Furthermore, the feature of shortened 3’ UTR in activated neurons could be recapitulated by reducing the U1 snRNP level while overexpression of U1 snRNP prevents the shortening of mRNA in activated neurons (Berg et al., 2012). A recent study of U1 snRNP and PAS on transcription directionality revealed that the U1 snRNP recognition sites and PAS were enriched and depleted in the sense direction, respectively, while the opposite pattern was discovered in the antisense direction. This suggested a functional role of U1 snRNP-PAS axis in determining the transcriptional direction and limits the pervasive transcription in the genome (Almada et al., 2013).

1.8.4 Epigenetics and APA

Accumulating evidence showed epigenetic modifications regulate alternative splicing (Colgan and Manley, 1997; Fox-Walsh and Fu, 2010; Luco et al., 2010; Luco et al., 2011; Zhou et al., 2014). Connection between alternative polyadenylation and histone positioning was first identified in yeast and a significant depletion of nucleosome was observed around the flanking region of the pAs (Mavrich et al., 2008; Shivaswamy et al., 2008). Similarly, a following study on human T cells also found strong nucleosome depletion around pAs together with enriched nucleosome downstream of the pAs. For genes with multiple pAs, the higher nucleosome density in the downstream of pAs was correlated with higher pAs usage (Spies et al., 2009). The nucleotide composition around the pAs was normally AT rich and this might reduce nucleosome affinity. On the other hand, this AT rich region could also be the target of other proteins that inhibit the assembling of nucleosome. However, current observation is just a correlation between the nucleosome density and APA, the cause-effect relationship is not determined yet.

Genomic imprinting was also found to be involved in regulating APA (Wood et al., 2008). Alternative polyadenylation in H13, a murine imprinted gene, was regulated in allelic specific manner with the maternal allele utilizing proximal pAs and the paternal allele using distal pAs. The proximal and distal pAs was separated by CpG island (Wood et al., 2008). The underlying mechanism of the CpG island regulating pAs selection is unclear and it is possible that the CpG island recruits inhibitory protein factors to prevent the usage of proximal pAs or CpG island itself inhibits the assembling of APA machinery on maternally derived allele. Similar as the effect of nucleosome position on pAs selection, the exact underling mechanisms of epigenetic markers regulating pAs usage are still elusive and much more work is needed.

1.9 APA and human diseases

Mutations affecting the PAS and/or other polyadenylation-regulating *cis*-elements have been reported to be responsible for a number of human diseases including Alpha-thalassemias (Higgs et al., 1983; Harteveld et al., 1994), Beta-thalassemias (Orkin et al., 1985; Jankovic et al., 1990), PEX syndrome (Bennett et al., 2001), Metachromatic leukodystrophies (Gieselmann et al., 1989), Fabry disease (Yasuda et al., 2003), Acetyltransferase 1 polymorphism in colorectal cancer (Bell et al., 1995), Insulin-like growth factor 1 deficiency (Bonapace et al., 2003), X-linked severe combined immunodeficiency (Bonapace et al., 2003), and some other diseases (Danckwardt et al., 2008). The mutations cause human diseases through either gain-of-function or loss-of-function, which results in alteration in gene expression. Take the PEX syndrome as an example, it was caused by a loss-of-function mutation in FOXP3 gene, which is a DNA binding protein, and malfunction of the gene leads to dysfunction of regulatory T cells and causes autoimmunity (Bennett et al., 2001). However, beside the general mutation affecting the protein binding ability, Bennett *et al.* identified another rare mutation in PAS of this gene and this mutation resulted in the increased usage of distal pAs in the 3’ UTR region, which is 5.1 Kbp away from the proximal pAs. The extended 3’ UTR region may contain microRNA binding site or RNA degradation related RBP binding sites, which leads to the decreased mRNA level. As another example, through comparing the gene expression level of Cyclin D1 between MCL (mantle cell lymphoma) patient with short 3’ UTR and long 3’ UTR, Wiestner *et al.* found that the shorter 3’ UTR isoform leads to worse survival. Subsequent sequencing of the patient samples with shorter 3’ UTR, they discovered a mutation in the PAS of the proximal pAs and this mutation could strengthen the PAS signal (Wiestner et al., 2007). This is consistent with previous findings that mRNA with shorter 3’ UTR is more stable and generates more proteins in different cancers (Mayr and Bartel, 2009).

One of the typical features of cancer is uncontrolled cell proliferation. Global analysis of APA regulation in different conditions or contexts found that the length of 3’ UTR is correlated with cell proliferation and differentiation in opposite direction. Mayr *et al.* found cancer cells prefer to use proximal pAs by comparing the cancer cells and non-transformed cell lines (Mayr and Bartel, 2009). They speculated that mRNA with shorter 3’ UTR have increased stability by avoiding the repressing from microRNA and finally leads to more amount of protein generated from the affected genes, although the finding is challenged by later genome wide analysis in 3T3 cell line (Spies et al., 2013). This study indicates that the regulation of APA may play an import role in the activation of oncogenes. In line with this finding, another study discovered a global pattern of increased usage of proximal pAs in breast cancer cell line MCF7 comparing with the immortalized but non-transformed cell line MCF10. However, the similar trend was not observed in another breast cancer cell line MB231 (Fu et al., 2011). This make the relationship between APA and cancer much more complicated than previous expectation. A more comprehensive APA analysis was performed in five different cancers and their corresponding normal tissues identified shorter 3’ UTR mRNA isoforms were more enriched in cancer tissues comparing with the matched normal tissues (Lin et al., 2012). Furthermore, assessing the pattern of APA alteration during colorectal cancer development revealed progressive changes in APA and this may be a potential biomarker for disease diagnosis in the future (Lin et al., 2012).

1.10 Evolution of APA, *cis*-elements, and RBPs

The number of protein coding genes and the length of coding region has not changed a lot from worms to humans during evolution, while the phenotype was surprisingly different (Lander et al., 2001; Hillier et al., 2005; Ezkurdia et al., 2014). In contrast, the number of APA in human has doubled comparing with that in worms and the length of poly (A) was also significantly increase during evolution (Mayr, 2015). These discoveries suggest that the APA plays a pivotal role in shaping the transcriptional and post-transcriptional regulation of gene expression during evolution.

Phylogenetic analysis of the sequence around the cleavage site of human, mouse and chicken genes shown that most of the *cis*-elements are similar (Tian 2008). Except for the downstream UG- and G-rich elements, the elements or motifs in the upstream region and downstream nearby region were conserved from Human to Yeast (Table 2) (Bin Tian 2008). Nematode, plant, and yeast contain the identical *cis*-elements including PAS and upstream UGUA motif, however, the downstream elements are not well conserved across different species. This may indicate that the sequence features in the downstream region especially the distal region were quit variable.

Table 1.2 The conservation of *cis*-elements in the flanking region of pAs.

	UA-rich	UGUA	U-rich	PAS	U-rich	UG-rich	G-rich
Human	+	+	+	+	+	+	+
Fish	+	+	+	+	+	+	−
Fly	+	+	+	+	+	+	−
Nematode	+	+	+	+	+	−	−
Plant	+	+	+	+	+	−	−
Yeast	+	+	+	+	+	−	−

Note: ‘+’ indicates presence of the specific elements; ‘−’ indicates the absence of the specific elements.

Though the proteins comprise the polyadenylation machinery are well conserved across different species in evolution, RBPs that are actively involved in the regulation of APA usage were found to be very divergent during evolution (Bin Tian, 2008).

Chapter 2 Pervasive *cis*-regulatory alternative polyadenylation in mouse

2.1 Aim of this study

Change of APA pattern during evolution remains underexplored. Such changes could arise from the divergence in *cis*-regulatory elements and/or *trans*-acting RBPs. The divergences of the two factors with different extent of pleiotropic consequences undergo distinct evolutionary trajectories. Therefore, to better understand evolution in APA, it is important to distinguish the relative contribution of *cis*- and *trans*-effects. Here, to comprehensively investigate contribution of *cis*-elements and *trans*-factors in the process of APA in mammalian system, we utilized the methods of 3' region extraction and deep sequencing (3' READS) and 3' mRNA-Seq to identify and quantify the pAs usage difference between two parental strains (C57BL/6J and SPRET/EiJ) and two alleles in the F1 hybrids, respectively. Since two alleles of F1 hybrids were exposed in the same *trans*-regulatory cellular environment, the identified pAs usage difference between two alleles should only reflect *cis*-regulatory divergence. The fraction of pAs usage difference between parental strains, which could not be explained by *cis*-regulatory divergence, is attributed to *trans*-regulatory pattern. The parental mouse strains used in this study diverged ~1.5 million years ago and this resulted in about 35.4 million single nucleotide polymorphisms (SNPs) and 4.5 million insertions and deletions (indels) when comparing two genomes (Keane et al., 2011). Such high genetic divergences between two mouse strains allowed us to accurately identify the origin of a large amount of sequencing reads and quantify the pAs usage between two alleles of F1 hybrids. In total, we identified 3747 parental divergent pAs, which are distributed in five types of APA, between the two mouse strains. By comparing them with F1 hybrids, we observed a predominant contribution of *cis*-regulatory effect on pAs usage. Our results showed that the formation of unstable secondary structure in the upstream region of pAs enhances the pAs usage and the existence of hexamer UUUUUU in the same region is able to confer inhibitory effect on APA.

2.2 Materials and Methods

2.2.1 Cell culture and RNA extraction

The fibroblast cell lines derived from two parental mouse strains and the F1 hybrids used here were the same as our previous study (Gao et al., 2015). Cells were cultured in RPMI 1640 Medium (Gibco, Life Technologies) with 10% FBS and 1% P/S. Total RNA from cultured cells were extracted using Trizol reagent according to the manufacture's protocol (Life Technologies). The integrity of total RNA was estimated by using Agilent Bioanalyzer with RNA Nano kit (Agilent Technologies) and RNA integrity number (RIN) above 9.0 were used for subsequent experiments.

2.2.2 3' READS

To construct a reference pAs annotation for the fibroblast cells of the two parental strains, the method of 3'READS previously developed by Hoque et al. was utilized with some minor modifications (Hoque et al., 2013). In brief, 30 μg total RNA was subjected to poly (A) selection with 200 μL Ambion oligo $(dT)_{25}$ beads (Ambion). Poly (A) containing RNA was then digested by 4 U RNase III (NEB) in the volume of 50 μL reaction system at 37 ^{o}C for 45 min. After discarding the supernatant and the RNA-bound beads were washed twice with binding buffer, the short poly (A) containing RNA were eluted by 100 μL Elution buffer. Following binding rU_5T_{45} oligos with biotin modification at 5' end (IDT) to MyOne streptavidin C1 beads (Life Technologies), the eluted RNA were incubated with the rU_5T_{45} oligos coated beads in binding buffer for 1 h at room temperature on a rotor with gentle shaking to avoid the precipitation of beads and increase the binding efficiency. After second round of capturing, the supernatant was discarded and the beads were washed with stringent washing buffer, followed by 1 U RNase H (NEB) digestion in the volume of 50 μL at 37 ^{o}C for 30 min. Then the supernatant was collected and RNA were purified by phenol-chloroform extraction and ethanol purification. The pellet was dissolved in 8 μL ddH_2O.

Six μL digested RNA were used for 3' adaptor ligation with the truncated T4 RNA ligase II (NEB), followed by 5' adaptor ligation with T4 RNA ligase I (NEB). The adaptor containing RNA were then reversed transcribed to cDNA using Superscript III Reverse Transcriptase (Life Technologies). The cDNAs were separated on 6% urea polyacrylamide gel and only the range between 100 nt and 600 nt were exercised for gel extraction. Then, the selected cDNA were amplified by indexed primers, which are compatible with Illumina sequencing. Finally, Ampure beads (Beckman Coulter) were used to purify the PCR products. The libraries were sequenced in single end 1×101 nt format on Illumina HiSeq 4000 platform. All the sequence of adaptors and primers are listed in Table 2.1.

Table 2.1 Primers for 3'READS library preparation and sequencing

Names	Primers (5'-> 3')
Chimeric_oligo_dT	Biotin-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTUUUUU
3' adapter	rApp/NNNNGATCGTCGGACTGTAGAACTCTGAAC/3ddC/
5' adapter (RNA)	CCUUGGCACCCGAGAAUUCCANNNN
RT primer	GTTCAGAGTTCT ACAGTCCGACGATC
PCR_forward	AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA
PCR_index_reverse_2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
PCR_index_reverse_7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
PCR_index_reverse_11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
PCR_index_reverse_12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA

Note: the underline indicates the indexes for multiplex sequencing.

2.2.3 3' mRNA sequencing

The usage of pAs was quantified by using the 3' mRNA-Seq Library Prep Kit (Lexogen). Briefly, 500 ng total RNA were used as starting material. Poly (A) containing RNA were reverse transcribed by anchored oligo (dT) primers, followed by second strand cDNA synthesis with random primers containing part of the Illumina adaptor sequence. PCR amplification was then performed to get the Illumina sequencing libraries. The libraries were sequenced in single end 1×101 nt format on

HiSeq 2000/2500 platform. The detailed information of adaptors and primers could be found on the website of Lexogen (https://www.lexogen.com).

2.2.4 Sequencing read alignment

The reference genome (mm10) and Ensembl gene annotation (GRCm38, release 74) of C57BL/6J were downloaded from Ensembl (http://www.ensembl.org), while the SPRET/EiJ reference genome and gene annotation were created as described in our previous study (Gao et al., 2015).

For the 3' READS sequencing reads, the leading T and tailed adaptor sequence were removed by using customized scripts and Cutadapt (http://cutadapt.readthedocs.io/en/stable/guide.html). Only the reads not shorter than 15 nt were aligned to reference genome using Tophat2 with the default parameters (Kim et al., 2013). Only the uniquely mapped reads with MAPQ >= 10 were retained for further analysis.

For 3' mRNA-Seq sequencing reads, after removing the adaptor sequence at 3' end, 12 nt from the 3' end corresponding to the random primer sequence used during the second strand cDNA synthesis were further trimmed off. After removing the first nucleotide in 5' end of the reads, only the reads not shorter than 15 nt were aligned to reference genome by Tophat2 with default parameters (Kim et al., 2013). For sequencing data from parental fibroblast, reads were aligned to the corresponding genome references, whereas for the F1 hybrid and mock F1 hybrid data, reads were aligned to both genomes and then assigned to the corresponding parental allele with smaller mapping edit distance. Reads with identical edit distance to both parental alleles were discarded and the remaining allele-specific reads were retained for further analysis.

2.2.5 pAs cluster identification

The uniquely mapped reads from 3' READS data were compared with the genome sequences to which they were aligned and only those with at least 2 non-genomic T (2T), which was defined as PASS reads, were used to call pAs clusters with the method described in the previous studies (Tian et al., 2005; Hoque et al., 2013). In brief, 3' end of the alignment represents the cleavage site and the cleavage sites located within 24 nt from each other were clustered together. If the cluster size was no more than 24 nt, it was considered as one pAs cluster and the position with the highest read coverage was defined as the representative position for the pAs. If a cluster was larger than 24 nt, the position with the highest read coverage will be defined as representative cleavage site and the remaining reads located at least 24 nt away from the position were re-clustered as described above. This process was repeated until all PASS reads within the cluster were determined.

We then determined the potential false positives of the pAs identified. First, we estimated the enrichment of PASS reads within the pAs using our previously published standard poly (A) RNA-Seq data (Gao et al., 2015) as background. In brief, for each pAs, we counted two numbers, i.e. 1) PASS reads within the cluster (R), 2) the ratio of the PASS read coverage within the identified pAs cluster to the RNA-Seq read coverage in the region of (-500, 50) flanking the representative pAs position (E). Second, we repeated the pAs calling using the reads without non-genomic T (0T) and then determined R and E value for each of these peusdo-pAs as described above. Then, for each of the real pAs, we determined a P value for each cluster as following: In a total of N clusters, given any cluster i (R_i, E_i), there would be n_i clusters with both higher R and E, its P was calculate as,

$$P_i = n_i/N$$

For each pAs, we calculated the P value based on both real (P_i (2T)) and peusdo-pAs dataset (P_i (0T)), representing positive (TP) + false positive (FP) and also positive (FP), respectively. To balance sample sizes of real pAs and peusdo-pAs dataset, we subsampled the peusdo-pAs dataset with the same sample size as the real one for 100 times. The average of P_i (0T) (E (P_i (0T))) was used to estimate Q value as below.

$$Q_i = FP/(FP + TP) = E\ (P_i\ (0T))/P_i\ (2T)$$

Finally, based on our replicate data, we performed IDR analysis to retain only the reproducible pAs (Li et al., 2011). Here, IDR analysis was based on the Q value determined in each of the two replicates. Only the pAs clusters with IDR value less than 0.05 were considered as putative pAs clusters.

2.2.6 Filtering of C57BL/6J and SPRET/EiJ pAs reference

Given the potentially incomplete/incorrect assembly of SPRET/EiJ genome reference sequences, some of the pAs identified in one strain might not be unambiguously mapped in the genome references of the other strain, for which it was infeasible to directly compare the usage between these two strains/alleles. Therefore we retained only the pAs, of which their flanking region (-124 nt to 24 nt) could be reciprocally aligned between the two strains based on LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

2.2.7 pAs annotation

According to the Ensembl gene annotation, we assigned each pAs cluster to protein-coding gene, lincRNA, other ncRNA, as well as the intergenic regions. For the pAs assigned to protein coding gene, we further classified the pAs into five categories based on Ensembl annotation as well as the RNA-Seq data including terminal pAs (annotated Ensembl gene end), tandem 3' UTR pAs, alternative last exon pAs, intronic pAs and internal exonic pAs.

2.2.8 pAs usage quantification

The 3' mRNA-Seq data were used to quantify pAs usage by counting the number of reads with the 5' end located within the reference pAs cluster. If there were more than one pAs within a protein-coding gene (reads counts of gene >= 20, average reads count of pAs in either strain >= 5 reads), we compared the strain or allelic difference of their usage with DEXSeq (Anders et al., 2012).

2.2.9 Creating mock F1 hybrids and pAs cluster filtering

In F1 hybrids, only the reads unambiguously assigned to specific genome were kept for the quantification of pAs usage. Thus the region with low density of sequence variants, therefore covered with fewer allelic specific reads and this could potentially increase the inconsistence of pAs usage calculated from parental strain and F1 hybrids. In order to reduce the potential errors, we created the a mock F1 hybrids by combining the reads from two parental mouse strains together and then performed the same analysis as for real F1 hybrids. We assumed the assignment of each reads to its corresponding strain is a Bernoulli trial, while for a pAs cluster with n reads in parental strain, the assigned read number in mock F1 hybrids followed a binomial distribution B (n, p). The p was estimated from the genome-wide average proportion of reads could be unambiguously assigned to the allele (p = Total_mock/Total_parental, where Total_parental and Total_parental represents the number of all reads from all reference pAs mapped in the original parental strain data and Total_mock represents those that could be unambiguously assigned to the allele). We filtered out the pAs that was not followed the binomial distribution (single tail binomial test, $P < 0.05$) or with Readsparental < Readsmock (Readsparental stands for read count assigned to parental strain before pooling and Readsmock represents read count assigned to parental strain after pooling) for either allele and the remaining ones were retained for allelic pAs usage quantification in F1 hybrids.

2.2.10 PacBio sequencing and data analysis

Total RNA was extracted from the F1 fibroblast cell line using Trizol reagent (Life Technologies). Anchored oligo (dT) primers linked with common sequence were used for reverse transcription of the first strand cDNA by superscript III (Life Technologies) following the manufacture's protocol. Gene specific primers, which do not contain variants between C57BL/6J and SPRET/EiJ genome, together with primers targeting the common sequence were used to amplify the different isoforms. Two µL cDNA were used as template in a volume of 50 µL PCR reaction system. PCR program was as follows: 5 min at 95 ^{o}C; followed by 5 cycles of 30 sec at 95 ^{o}C, 30 sec at 68 ^{o}C/58 ^{o}C, 30 sec at 72 ^{o}C and 30 cycles of 30 sec at 95 ^{o}C, 30 sec at 65 ^{o}C, 30 sec at 72 ^{o}C with 0.5 ^{o}C reduction of Tm value for each cycle and a final extension at 72 ^{o}C for 5 min; then hold at 4 ^{o}C. The PCR products were purified by Agencourt Ampure XP system (Beckman Coulter)/gel extraction (Qiagen) and then quantified by Qubit HS dsDNA reagent (Life Technologies). Finally equal amount of each samples were mixed and sequenced on PacBio RS SMART platform according to the manufacture's protocol. All the primers used for PacBio validation are listed in Table 2.2.

Table 2.2 Primers used for PacBio validation

Names	Primers (5'-> 3')
3'_RT	GCAGTGGTATCAACGCAGAGTAC(T)18 V N
3'_PCR_Com	GCAGTGGTATCAACGCAGAGTAC
zfp229_P	GAATGCTGTCCTTAGGCACCACTGC
zfp229_D	GTTACGATCTGCCATCAGGTATGGG
Rsad1_P	GGCTTTTGTTGACTTTTTGGACACC
Rsad1_D	TCAAGCCAGACAACCTCCTTCTATGG
Zfand1_P	GAGCAAGGAAAAAGCAAAGCCATG
Zfand1_D	CCTGGTAGACTAATGGCAGCCATTG
Txndc16_P	ACCAGTGTGCTTGACCTGGGTCTAG

Txndc16_D	AGATGGATCCTCCATCAAAGAATTGTC
Smim14_P	AGTTTGGATGGTATTGTTTTCATGAGC
Smim14_D	CCCATGCTTCCACTTAGCGTTTGTG
Efhc1_P	TGGAGAGCAACGCTTCCCAGTATTC
Efhc1_D	TTGGATGCACTGATAGACCAAATCC
Gatb_P	TCATTGTGAGATGTGTGGGGTCC
Gatb_D	TGTGTGTGTCCTATGAGACCCCATG
Tnfrsf23_P	TGCGAGTGCCAAATAGGTCTTTACTAC
Tnfrsf23_D	GATGCCCCAATGTCAGCAGGAAG
Rnf150_P	GTGGAAAGGCAGATTTTACCAAGTG
Rnf150_D	TGCCTACCCGACTCTTTTAAAGCAC
Pla2g12a_P	GAGAGCAGGCGAGAATGGAGGAC
Pla2g12a_D	GTGTTGTTTCTCTGAAGCCCACTAAG
Snp8_P	TTATCCCTTTTGGTCTCCCTTCCTTG
Snp8_D	TTGGGTTATTTTGTTGGTGGTGGTG
Acap2_P	GTGGCCCACACTTTACTCTTTAGATTC
Acap2_D	TTGCTGTGTTATCGAGATTTTTAGCAC
Wdr60_P	GGGAATATCAAGTGATGTCCAGAAAGC
Wdr60_D	TGATGGCTCAGCACCTCCTAATACC
Prcp_P	GCATGCCTGAATATTTCACAAACGAC
Prcp_D	TAGCATCAGGTGTGTGGGGACTTTC
Fam198b_P	AGTCTCAACTAATTCTCCCCCAAATCC
Fam198b_D	ACCCTGTCTCAAAAACAAATGAAAACC
Bcas2_P	AGATTTGAACTGGCAGCGAAAGAAC
Bcas2_D	CCAACAAGGAAAACATCCGCCAAG
Scrn3_P	TGTAAACATCCTAGCGGAGGTTAGCC
Scrn3_D	ACCAGTGGAAAGAATTTAGGGAAGTGG
Utp23_P	GGACAAGCCTTCTCCCAGAACAGTG
Utp23_D	GTCCTCTGGAGAACTGTCATGGTAGGC
Lsp1_PD	CAGCCCTGACCAAGAAATTGCTTC
Mrps14_P	AGAGACGGAAAATGGCTTATGAATATG
Mrps14_D	ATAATGTGGCTATTTGTCATTGAATTGTC

Note: P and D stand for proximal and distal pAs, respectively.

Sequencing reads from the PacBio RS SMRT chip were processed through PacBio's SMRT-Portal analysis suite to generate circular consensus sequences (CCSs). The CCSs were then mapped to pAs site region of both C57BL/6J and SPRET/EiJ using BLASTN with default parameters. Reads were assigned to a specific allele with better alignment score. The number of reads assigned to proximal and distal pAs of each allele was used to calculate the ratio of proximal to distal for the two alleles and then compared to that determined by 3' mNRA-Seq.

2.2.11 Sequence variant density analysis

We calculated the density of sequence variants between the genomes of the two mouse strains, including both SNPs and indels, in the regions flanking of the cleavage site for both *cis*-divergent and non-divergent control pAs. According to previous studies (Hu et al., 2005), the flanking region was separated into four windows: [-100, -40), [-40, 0), (0, 40], and (40,100], which contain different *cis*-elements or motifs. The variant density was calculated and compared for each window separately. To further determine the variant density in the flanking region ([-100, 100]) with higher spatial resolution, we calculated the density of sequence variants in a 8 nt sliding window with the step size of 1 nt. The density of variants was compared between *cis*-divergent pAs and controls in each window by Mann–Whitney U-test and the *P* value was adjusted by "Bonferroni" method.

2.2.12 Local RNA secondary structure

Local RNA secondary structure minimum free energy (MFE) was calculated using RNAfold from the ViennaRNA package version 2.1.9 with default parameters at a temperature of 37 °C (Lorenz et al, 2011). We compared the MFE of the four mRNA regions (as described above) flanking the cleavage site between the two alleles for the three pAs groups separately, i.e. pAs usage biased towards SPRETS/EiJ allele, pAs usage biased towards C57BL/6J allele and pAs without allelic bias (controls). Kolmogorow-Smirnow-Test was used to estimate the statistical significance of the difference between the three pAs groups.

2.2.13 Motif analysis

To investigate the influence of *cis*-element on pAs usage, we divided the *cis*-divergent pAs into two groups, (1) pAs with usage biased towards C57BL/6J allele, and (2) pAs with usage towards SPRETS/EiJ allele. For each of the *cis*-divergent pAs, we counted the hexmer occurrence in pAs flanking region from C57BL/6J and SPRETS/EiJ genome, respectively. The allelic difference in hexmer frequency was then calculated, summed up, and compared between the two groups. The hexmer with higher frequency both in C57BL/6J genome for group 1 and in SPRETS/EiJ for group 2 was considered as pAs usage enhancer. In contrast, the hexmer with lower frequency both in C57BL/6J genome for group 1 and in SPRETS/EiJ genome for group 2 was thought to be pAs usage repressor. For the motifs have potential significant impact on pAs usage, including AAUAAA, AUUAAA, and UUUUUU, we estimated their effect size by the change of pAs usage when they were mutated, respectively.

2.2.14 Vector construction and pAs reporter assay

To investigate the effect of allelic sequence difference in the flanking regions on allelic pAs usage, the flanking region of 10 pAs pairs from 7 genes were PCR amplified from the genomic DNA of both mouse strains, separately. Genomic DNA from the two strains was extracted using DNeasy Blood Tissue Kit according to the manufacture's guide (Qiagen). Then the target regions were amplified by using Phusion High-Fidelity DNA Polymerase (NEB) from genomic DNA with homology sequence conjugated gene specific primers. 10 ng genomic DNA was used as template in a volume of 50 µL PCR reaction system. PCR program was as follows: 5 min at 98 ^{o}C; followed by 30 cycles of 30 sec at 98 ^{o}C, 30 sec at 60 ^{o}C, 30 sec at 72 ^{o}C and a final extension at 72 ^{o}C for 5 min; then hold at 4 ^{o}C. After PCR amplification and purification with Ampure beads (Beckman Coulter), the fragments were inserted into a customized vector (pRIG) by using In-Fusion HD Cloning Kit (Clontech) following manufacture's protocol (Figure 2.18 B) (Ji et al., 2009). Constructs with correct inserts were validated by Sanger sequencing.

The 3T3 cell line were maintained in DMEM (Gibco) with 10% FBS (Gibco). For 24-well plate, 3×10^4 3T3 cells were seeded in each well one day before transfection. 500 ng plasmids were transfected to each well using 1.5 µL Lipofectamine 20000 reagent (Life Technologies) in three replicates according to the manufacture's protocol. After incubation for 24 hours, the cells were harvested and lysised with TRIzol reagent (Life Technology) followed by extracting the total RNA using Direct-zol RNA kit (Zymo Research). To remove the potential remaining DNA contamination, 1 µg total RNA were treated with TURBO DNase (Ambion) in 20 µL reaction system following the manufacture's protocol. Reverse transcription was performed to generate cDNA from the treated total RNA with oligo (dT) primers by Superscript III Reverse Transcriptase (Life Technologies) according to the manufacture's guide. To compare the cleavage and polyadenylation efficiency between the two alleles, quantitative real-time PCR (qRT-PCR) was carried out with primers targeting RFP and EGFP regions, respectively, using SYBR Green Masrermix I (Roche) on LightCycler 480 (Roche) according to manufacturer's procotocol. The relative pAs strength was estimated as the RFP mRNA level normalized by the EGFP mRNA level, and compared between the two alleles.

To assess the potential role of the UUUUUU motif on cleavage and polyadenylation, we chose one gene containing the *cis*-regulatory pAs events, which has the UUUUUU motif in the upstream region of pAs from one strain but not the other, for further reporter assay. We created two mutants for pAs by exchanging the UUUUUU in one allele and its counterpart in the other allele without affecting other variants. To construct the mutants, 20 ng total plasmid DNA were used as template for PCR with mutation site containing primer pairs by using the Site-Directed Mutation System (Life Technology) according to the manual. Mutants were again validated by Sanger sequencing and were used for further transfection. After transfection, total RNA were extracted and treated with TURBO DNase followed by RT and qPCR. The details of the procedures were the same as depicted above. All the primers used for cloning, mutagenesis, and qPCR are listed in Table 2.3.

Table 2.3 Primers used for minigene validations

Names	Priemra (5'-> 3')
Zfand1_F	GTCGACTGCAGAATTCAACACATAATTAAACACCAACATGG
Zfand1_R	CTCAAGCTTCGAATTTTAAAAAATGTTGACTCTTACTTGG
Frk_F	CTCAAGCTTCGAATTGTCTGAATGTAGTTATGAACTGTAGC
Frk_R	GTCGACTGCAGAATTTCCTTGTGAGCATCTAGATAGTCC
Prcp_F	CTCAAGCTTCGAATTGGGCCCATCATTTAGATCTCCG
Prcp_R	GTCGACTGCAGAATTACCCCTGTCACTCTCTTGTGTCTTG
Lpar2_F	CTCAAGCTTCGAATTATATGCATAGGACCACTCTCCTC
Lpar2_R	GTCGACTGCAGAATTATTCTTCTGACCTCAAGAGCATC
Rasd1_P_F	CTCAAGCTTCGAATTAGACACCTGTGTGGTGCATAGATATTC
Rasd1_P_R	GTCGACTGCAGAATTACAGAGAGACCCCATCTTGAAAATCC

Rasd1_D_F	CTCAAGCTTCGAATTTGTTGTCTGTGTGTCTATGACACTGG
Rasd1_D_R	GTCGACTGCAGAATTCCTAGCACAATGAAACAAACCACAC
Zfp229_P_F	CTCAAGCTTCGAATTATGCCTCACTGCACAGTAGAAATTCC
Zfp229_P_R	GTCGACTGCAGAATTGTTCTCACCATCTTGTGATTGACAGG
Zfp229_D_F	CTCAAGCTTCGAATTTCATAAATTATGAGAATCATGTTACTC
Zfp229_D_R	GTCGACTGCAGAATTATCTAGTAATTTTAGACACACCTGTTC
Txndc16_P_F	CTCAAGCTTCGAATTTTGCATAATTAGCAACCTTGTAGTAGC
Txndc16_P_R	GTCGACTGCAGAATTCCAATGCCCTCTTCTGTTCTCTG
Txndc16_D_F	CTCAAGCTTCGAATTCTAGGAACCAAACTGGGGTCCTC
Txndc16_D_R	GTCGACTGCAGAATTGATAAGGGCTCACTTAAGACTCCCT
Alg10b_F	CTCAAGCTTCGAATTTTCAAACTGTATTCAGATAAAATCATG
Alg10b_R	GTCGACTGCAGAATTATGTTTATGGGCTTCATGGATG
Alg10b_B2S_F	CCAGAATAAAAAGACAAATTTTTTGTTGAAGGACGGTTGT
Alg10b_B2S_R	ACAACCGTCCTTCAACAAAAAATTTGTCTTTTTATTCTGG
Alg10b_S2B_F	CCAGAATAAAAAGACAAATTGTTGAAGGAAGGTTGT
Alg10b_S2B_R	ACAACCTTCCTTCAACAATTTGTCTTTTTATTCTGG
PUF60_q_F	AAAACCCAGAGGAAACAAGGAACA
PUF60_q_F	GCAAGGTGGTGGCTGAAGTGTA
RFP_q_F	CTTCAGGGCCTTGTGGATCT
RFP_q_R	CTTCAGGGCCTTGTGGATCT
EGFP_q_F	GGGCACAAGCTGGAGTACAACT
EGFP_q_R	ATGTTGTGGCGGATCTTGAAG

Note: The underline indicates homologous recombination arm sequence for cloning; P and D indicate proximal and distal pAs, respectively; B2S and S2B indicate mutagenesis between the two alleles.

2.3 Results

2.3.1 Construction of the pAs reference

To build a reference for pAs usage quantification, we utilized the 3' READS method (Hoque et al., 2013), which can effectively identify the pAs and eliminate the potential false positives caused by internal priming, with minor modifications to generate the pAs clusters for both C57BL/6J and SPRET/EiJ with two replicates (Figure 2.1; Materials and Methods). On average there are 63 millions filtered reads for each replicates and around 80% of them could be uniquely mapped to mouse genome by Tophat2 and nearly 40% of them harbored at least 2 non-genomic T at the 5' end and were defined as PASS reads. Only the PASS reads were used for subsequent pAs identification. The raw pAs identified in two replicates of each mouse strain were highly correlated, indicating high reproducibility of the method (Figure 2.2 A and B).

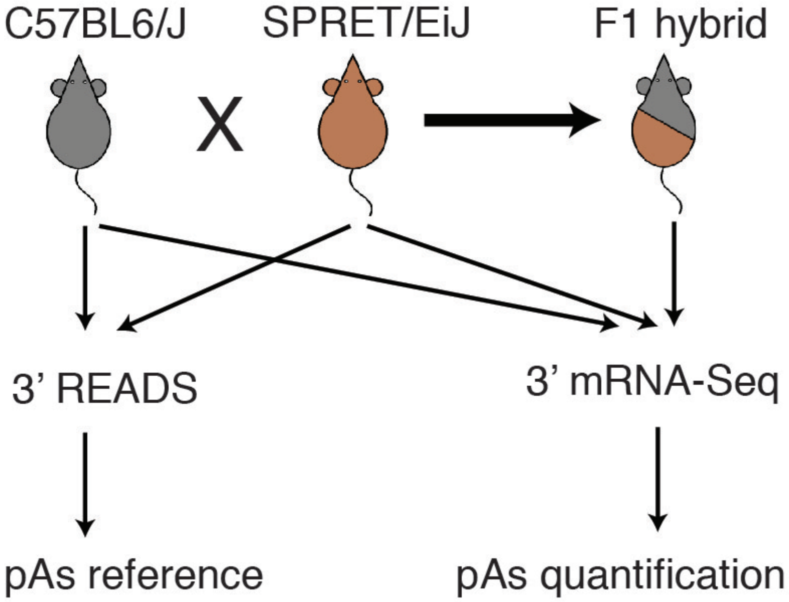

Figure 2.1 Study design. Fibroblast cell lines derived from C57BL/6J and SPRET/EiJ were used for creating the pAs reference with 3' READS method. Parental mouse strains and F1 hybrids were sequenced by 3' mRNA-Seq to quantify pAs usage.

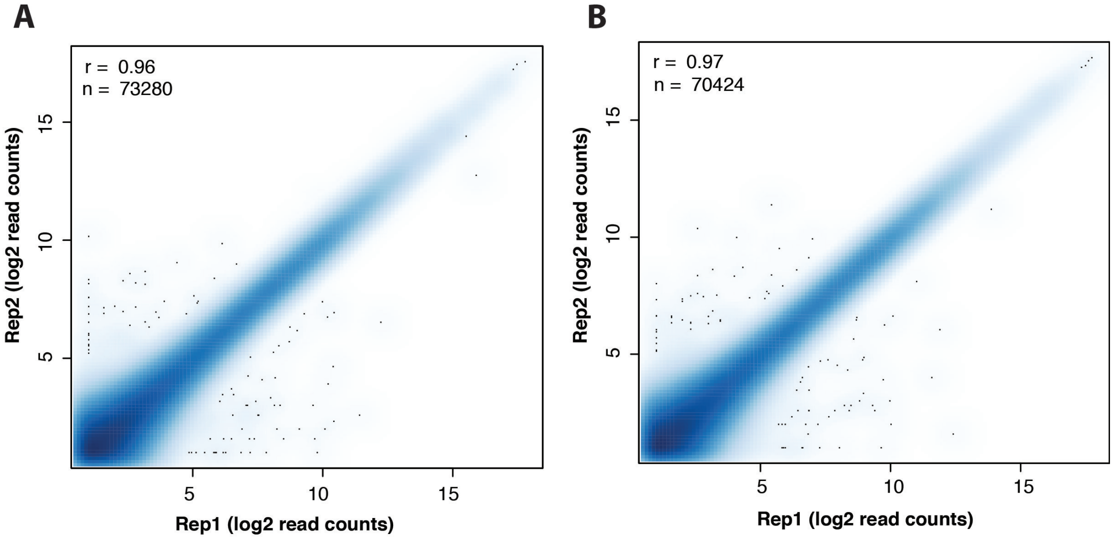

Figure 2.2 Reproducibility of 3' READS data. A and B, Reproducibility of pAs isoforms identified by 3' READS for C57BL/6J and SPRET/EiJ, respectively.

To further increase the accuracy of pAs identification, RNA-Seq data of the two strains from our previous study were used as sequencing background and 0 T reads (reads without non-genomic T) were used as control to filter the raw pAs clusters and this resulted in 38218 and 38836 highly confident pAs clusters for C57BL/6J and SPRET/EiJ, respectively (Materials and Methods). The nucleotide composition in the flanking region of the pAs cleavage sites of the two strains resembled that reported before (Figure 2.3 A and B) (Tian et al., 2005). Consistent with previous studies, above 50% of the filtered pAs from either C57BL/6J or SPRET/EiJ containing the canonical PAS motif AAUAAA in the upstream 100 nt region, however, the frequency reduced to 30% and background level when the 2 T reads and 0 T reads defined raw pAs clusters were examined, respectively (Figure 2.2 C and D (Tian et al., 2005).

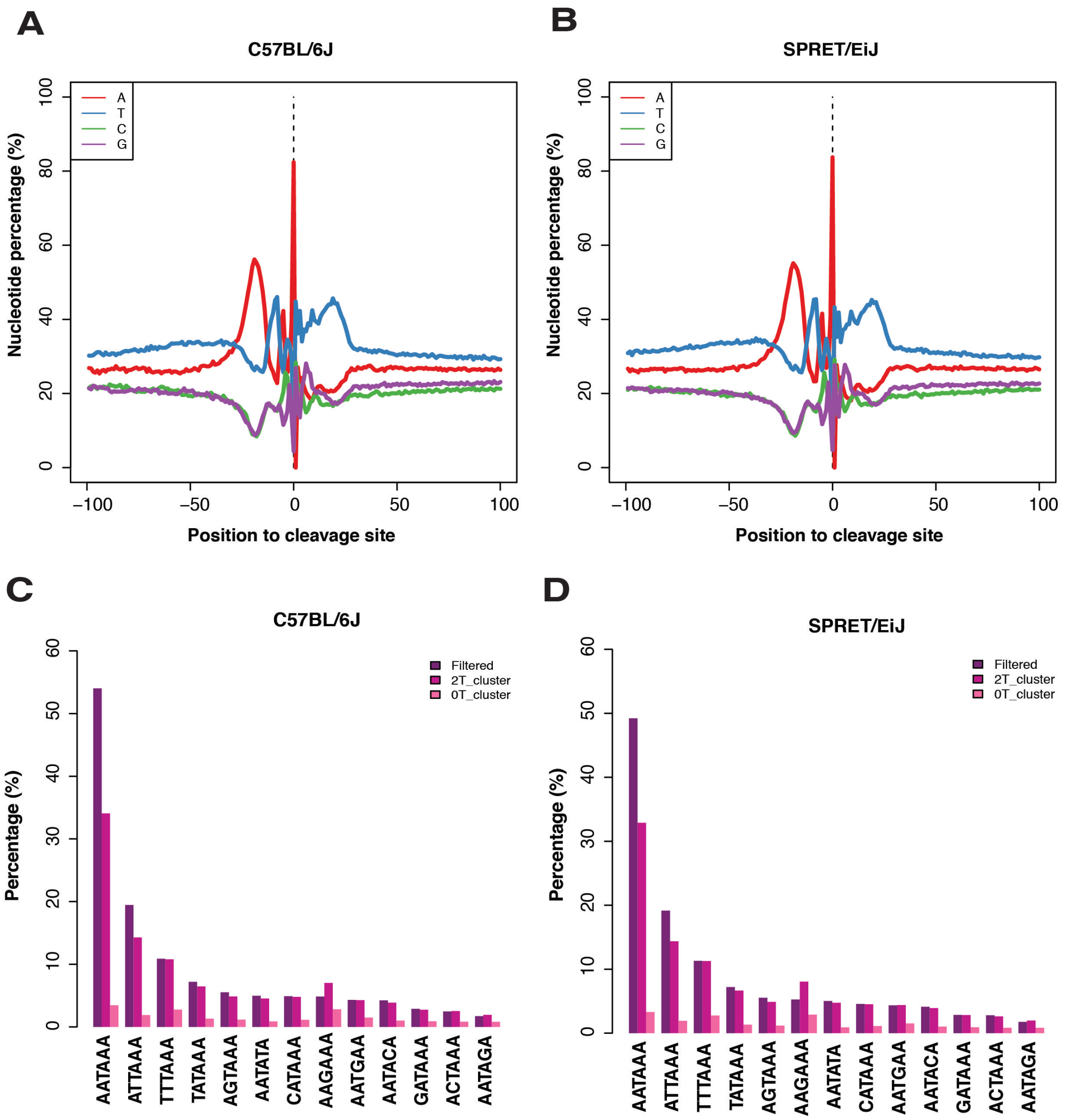

Figure 2.3 Features of the identified pAs (1). A and B, Nucleotide composition in the flanking region of pAs identified from C57BL/6J and SPRET/EiJ, respectively. The raw pAs were filtered and only pAs with IDR value less than 0.05 were analyzed; C and D, PAS motifs frequency of filtered pAs, raw pAs (2T), and pAs identified with non-PASS reads (0T) from both strains.

In order to make the comparison of pAs between two mouse strains much more reasonable and effective, we filtered out the pAs clusters identified in one strain, but could not be unambiguously mapped to the genome of the other strain, the remaining pAs from the two mouse strains were combined as a comprehensive reference pAs set containing 51446 pAs clusters for both strains (Figure 2.4 A; Materials and Methods). Comparing with the expression level of shared pAs isoform between two strains, the strain specific pAs isoforms have significantly decreased expression level (Figure 2.4 B). To estimate the accuracy of the pAs identified here, we also compared the combined pAs with Ensembl annotation and this gave rise to a sharp peak centered distribution of the pAs on the known pAs indicating the high quality of the identified pAs in this study (Figure 2.4 C). As shown in Figure 2.4 D, more than 60% of the genes have at least two pAs (Materials and Methods). Based on the 3' READS defined pAs and RNA-Seq data, we classified the pAs into five groups including Terminal pAs, Tandem 3' UTR, Alternative Last Exon, Intronic pAs, and Internal Exonic pAs (Figure 2.5 A; Materials and Methods). Nearly 80% of the pAs are located in the terminal exon (tandem 3' UTR and terminal pAs) while the remaining pAs were assigned to upstream exon, intron, and alternative last exon (Figure 2.5 B).

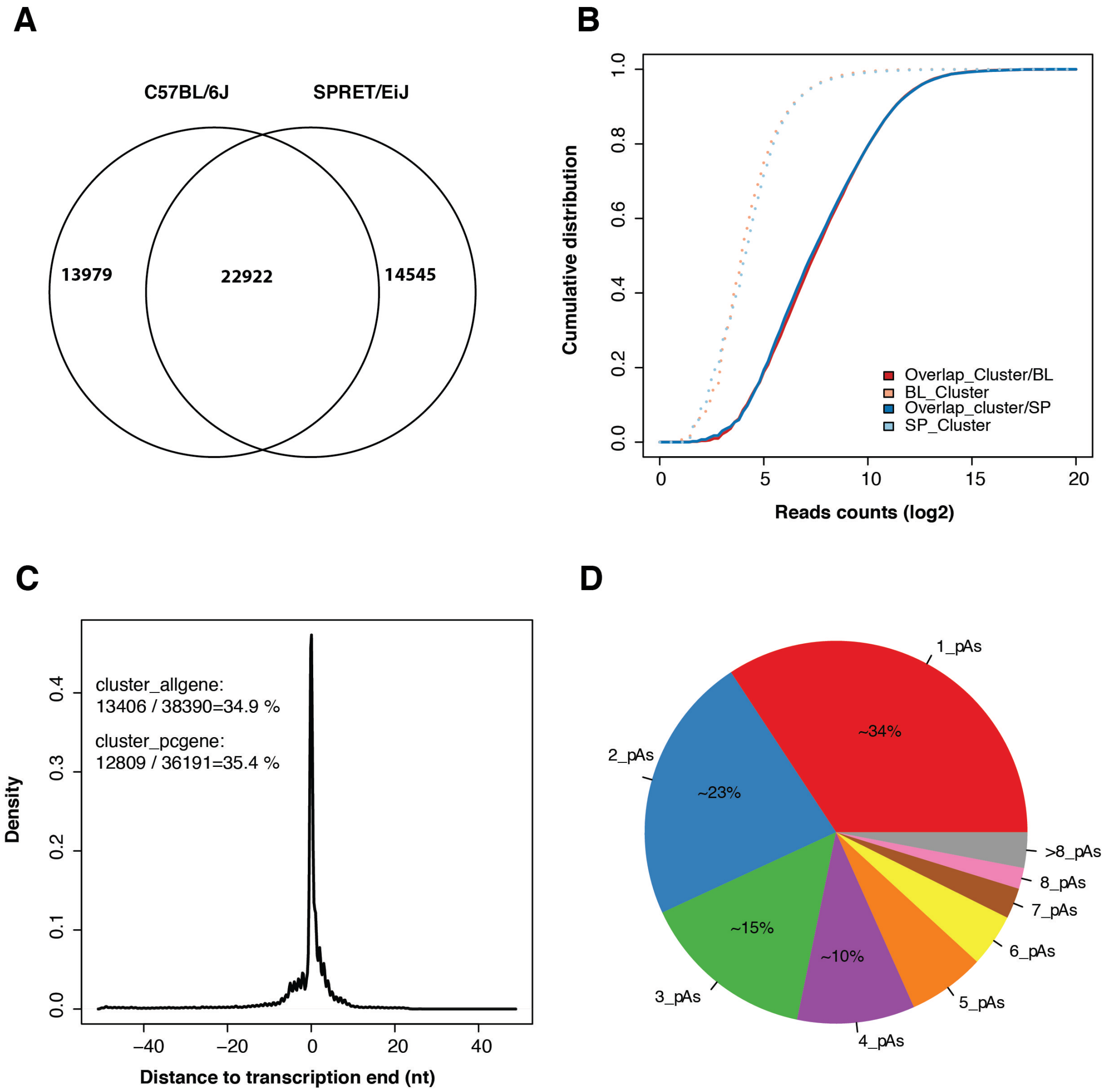

Figure 2.4 Features of the identified pAs (2). A, Venn diagram shows the shared and species specific pAs number; B, CDF shows the comparison of expression level of shared and allelic specific pAs; C, Comparison of the identified pAs with Ensembl annotated transcript end. 13406 (35%) pAs were located within 50 nt away from the transcript end in Ensembl. X-axis shows the distance between pAs

identified in this study and the nearest known pAs in Ensembl annotation. Y-axis represents the density;

D, The pie chart shows the distribution of genes with different number of pAs.

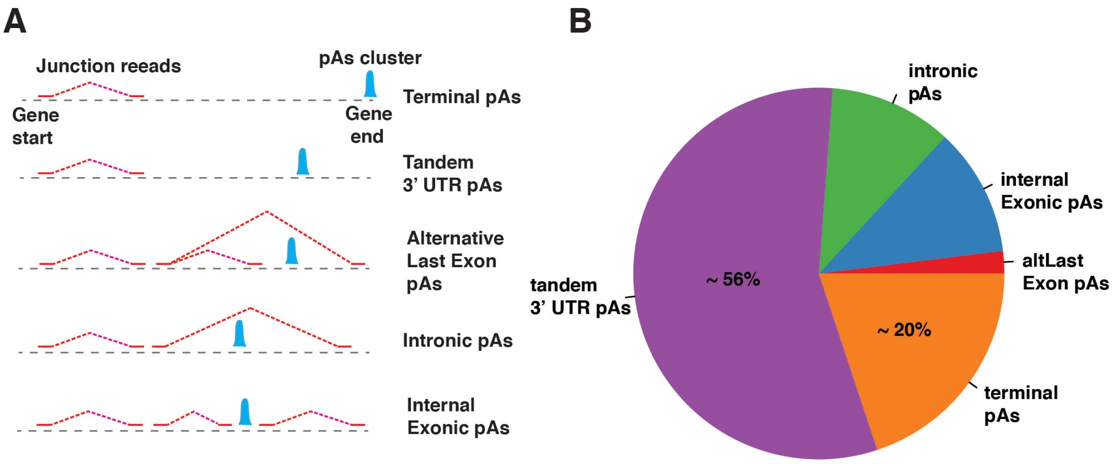

Figure 2.5 Classification of APA types. A, APA types defined based on the position of the identified pAs and junction reads of deep RNA-Seq data from the same cell line; Details are described in the Materials and Methods section. B, For all the pAs located in protein-coding region, we calculated the fraction of pAs numbers in each type of APA.

2.3.2 APA divergence between C57BL/6J and SPRET/EiJ

Though the 3' READS was highly effective in identifying pAs and could also be used for quantification of the APA isoform expression, the method was complicated requiring twice of bead enrichment for real poly (A) containing RNA, two steps of enzymatic digestions, which might affect the accuracy of the quantification. Therefore, a much simpler oligo (dT) priming and RNA-Seq based method (Materials and Methods), which was called 3' mRNA-Seq, was used to quantify the pAs usage in the following analysis. Given the oligo (dT) containing primers were used in this approach, a stretches of A or region of A-rich in the genome would be captured and sequenced, which leads to internal priming, in the subsequent analysis only 5' end of the sequencing reads mapped to the pAs clusters were retained for quantification of the usage of pAs identified by 3' READS.

To investigate the APA divergence between C57BL/6J and SPRET/EiJ, we sequenced the two mouse strain derived fibroblast cell lines with two replicates by 3' mRNA-Seq on Illumina HiSeq 2000/2500 platform (Figure 2.1; Materials and Methods) and obtained an average of 57.5 millions adaptor trimmed and filtered reads for each replicate. About 58.2 % could be uniquely mapped to either C57BL/6J or SPRET/EiJ genome by Tophat2 (Kim et al., 2013). The pAs usage between two replicates of both C57BL/6J and SPRET/EiJ was highly correlated (Figure 2.6 A and B) indicating good reproducibility of 3' mRNA-Seq method. Gene expression level calculated by counting all the reads located in all the defined pAs clusters of each gene was correlated well with that generated by mRNA-Seq implying the good quantitative characteristics of 3’ mRNA-Seq (Figure 2.6 C and D).

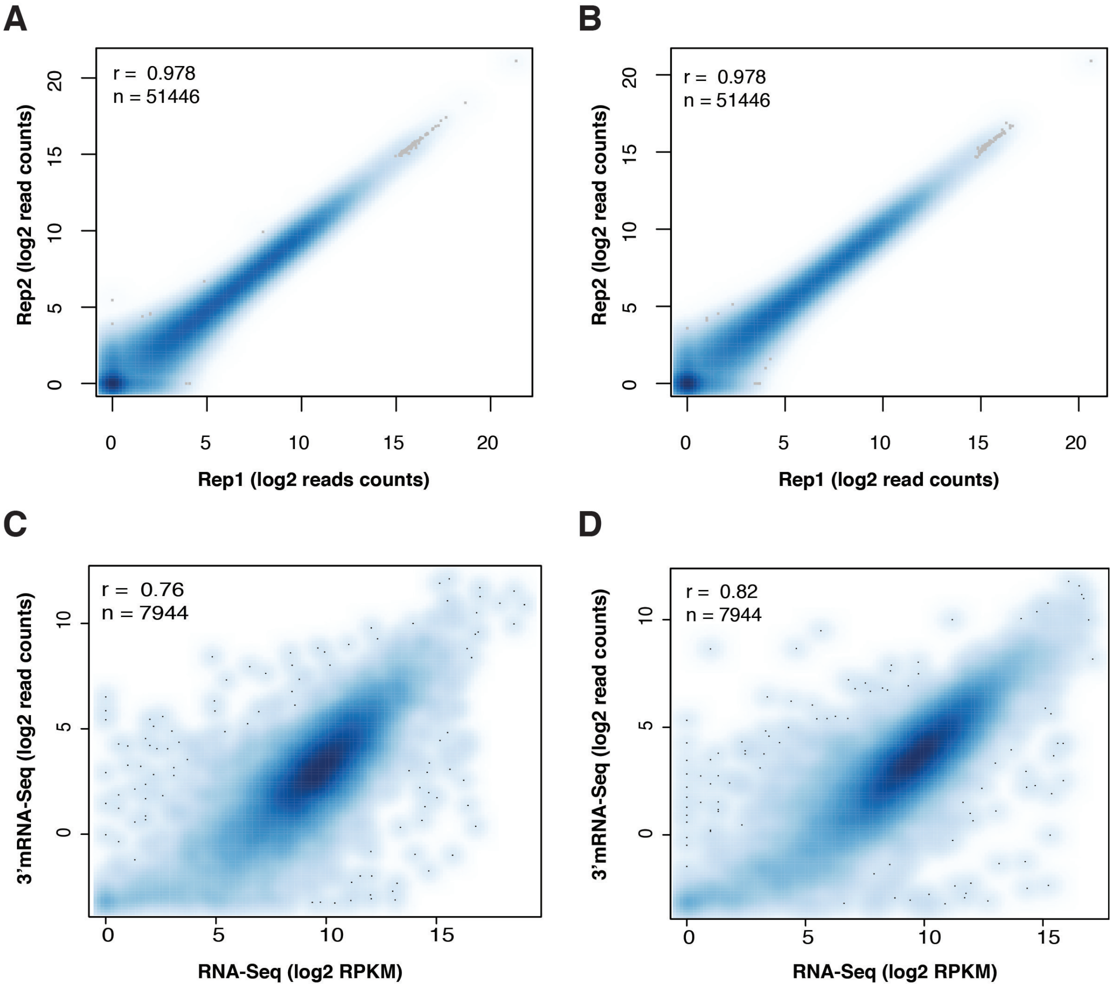

Figure 2.6 Reproducibility of 3’ mRNA-Seq data. A and B, pAs usage correlation between two replicates of two parental strains, respectively. The 3’ mRNA-Seq reads were mapped to the filtered pAs cluster created above; C and D, Comparison of gene expression level calculated by 3’ mRNA-Seq and RNA-Seq. For 3’ mRNA-Seq, reads within all defined pAs clusters were summed together and compared to that determined by mRNA-Seq.

To compare the pAs usage difference between two parental mouse strains, the uniquely mapped reads were mapped to the combined pAs reference created above. To ensure high accuracy, we considered only the multi-pAs containing genes which were expressed (>= 20 reads) in both fibroblasts and of which any pAs cluster was supported at least an average of 5 reads. Based on these criteria, we identified 3747 as differentially used pAs between two parental strains at a false discovery rate (FDR) of 5% by employing DEXSeq tool (Table 1; Materials and Methods) (Anders et al., 2012). These parental divergent pAs were spread across all five types of APA with the highest frequency in terminal APA (Table 1).

Table 2.1 Comparison of alternative polyadenylation between C57BL/6J and SPRET/EiJ.

APA types	Total expressed pAs	Divergent pAs (%)*
Alternative last exon	542	80 (14.8)
Internal exon	2196	240 (10.9)
Intronic	2363	270 (11.4)
Tandem 3'UTR	13874	2029 (14.6)
Terminal	5746	1128 (19.6)

*Indicates the proportion of divergent pAs in each type.

2.3.3 Predominant Cis-regulatory pAs divergence between C57BL/6J and SPRET/EiJ

Differential pAs usage between two mouse stains could arise from *cis*-regulatory elements and/or *trans*-acting factors. Given *cis*-regulatory element exerted its effect on the pAs selection in allele specific manner, while *trans*-acting factors regulated both two alleles of a diploid cell in the same way, we next utilized the F1 hybrids to distinguish *cis*- and *trans*-regulatory divergent pAs. In F1 hybrids, mRNAs derived from both parental alleles were subjected to the same *trans*-regulatory environment; thus the identified allelic specific pAs usage pattern suggested *cis*-regulatory divergence. The *trans*-acting divergence could then be inferred by comparing the interspecific divergence with allele specific divergence in F1 hybrids. Total RNA extracted from F1 hybrid fibroblast cell line were sequenced by 3' mRNA-Seq on Illumina HiSeq 2000/2500 platform for two replicates. Adaptor sequence trimmed and filtered reads were mapped to both C57BL/6J and SPRET/EiJ genome separately and uniquely mapped reads were assigned to either C57BL/6J or SPRET/EiJ according to the smaller mapping edit distance (Materials and Methods). On average, 47% of the F1 uniquely mapped reads could be assigned to C57BL/6J or SPRET/EiJ allele while the left reads are belonging to both alleles.

To avoid the potential errors in alignment and mis-assignment of F1 hybrid sequence reads to two parental genomes, we created a mock F1 hybrids by mix the reads from two parental strains and it was then analyzed in the same way as how we did with the real F1 hybrid data. The proportion of reads assigned to two parental strains was comparable with that of F1 hybrids. By comparing the reads number of each pAs between parental strain and the corresponding mock F1 hybrid derived allele, a total of 12809 pAs clusters were filtered out because of the biased assignment of reads and pAs could not be distinguished by genetic variants, and the remaining 16998 pAs were kept for further analysis (Figure 2.7 A-D; Materials and Methods).

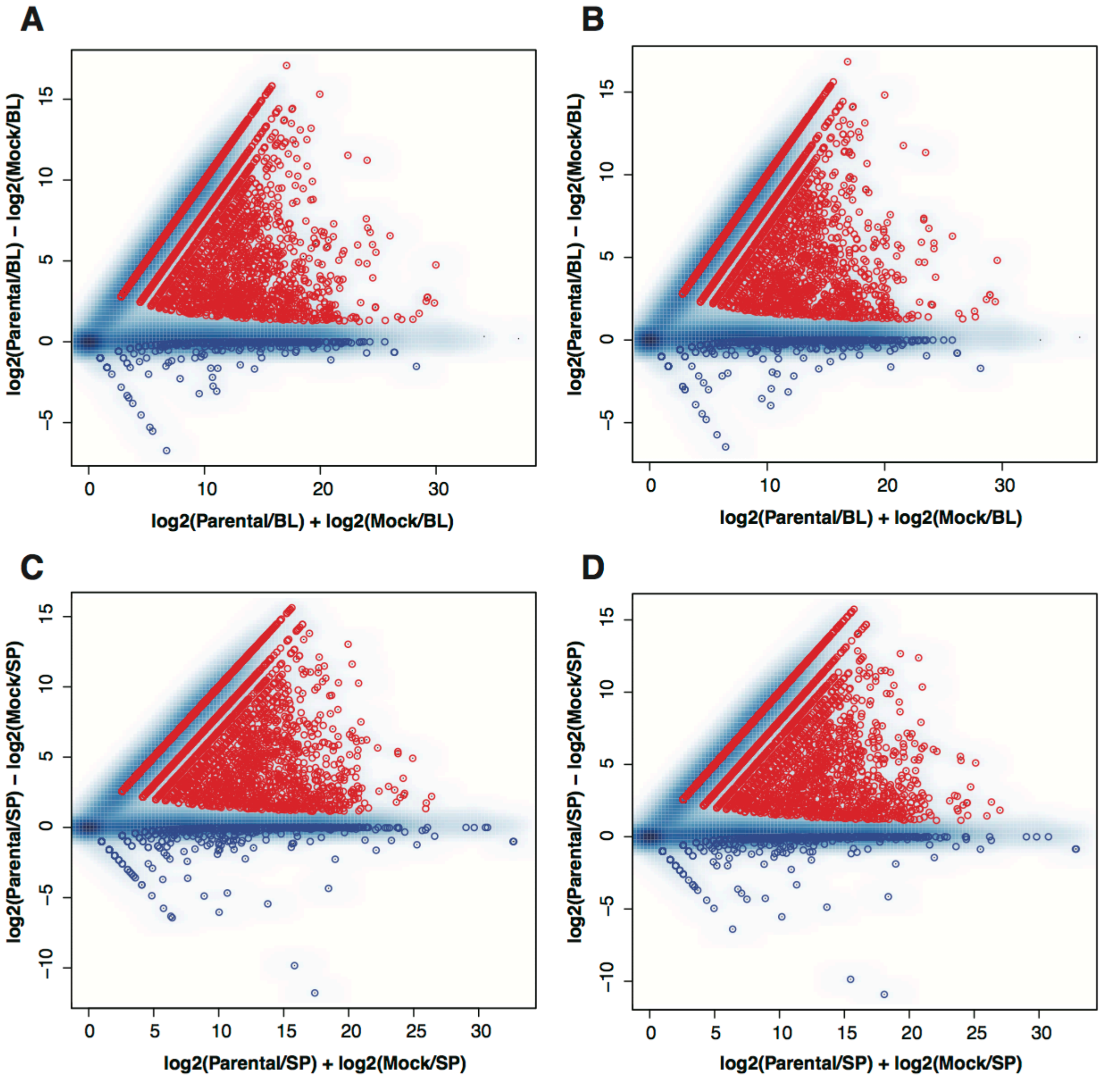

Figure 2.7 Filtering of pAs with mock F1 hybrids. A and B, Filtering of pAs by mock F1 hybrids for two replicates of C57BL/6J; C and D, Filtering of pAs by mock F1 hybrids for two replicates of SPRET/EiJ. Details of the filtering process are presented in the Materials and Methods section.

Of these, 11306 pAs were tested for the *cis*- and *trans*-regulatory divergence by comparing the allelic divergent pAs with that of parental strains with the same cut-off used for parental pAs divergent analysis. Among them, 8774 pAs showed neither evidence of pAs usage divergence between parental strains nor evidence of allelic specific pAs divergence. The other 2532 pAs were parental divergent events, including 1790 of the pAs exhibited significant *cis*-regulatory divergence, 456 pAs showed significant evidence of *trans*-regulatory divergence, and 86 pAs presented to be of both *cis*- and *trans*-regulatory divergence (Figure 2.8 A). Since APA had several different types, which might be regulated through different mechanisms, we calculated the frequency of both *cis*- and *trans*-regulated divergent pAs belonging to each type of APA. The divergent events spread across all the five types with more *cis*-regulatory pAs events over *trans*-regulatory divergent pAs in each type (Figure 2.8 B). Taken together, our data suggested extensive *cis*-regulatory effect on pAs divergence. Two typical examples of *cis*-regulatory divergent pAs were showed in Figure 2.8 C and D. In addition, when the divergent pAs events were divided into five groups based on different cut-offs of │ΔPPU│(delta percentage of pAs usage) values, the number of divergent pAs events caused by *cis*-regulatory elements always keeps dominant (Figure 2.9).

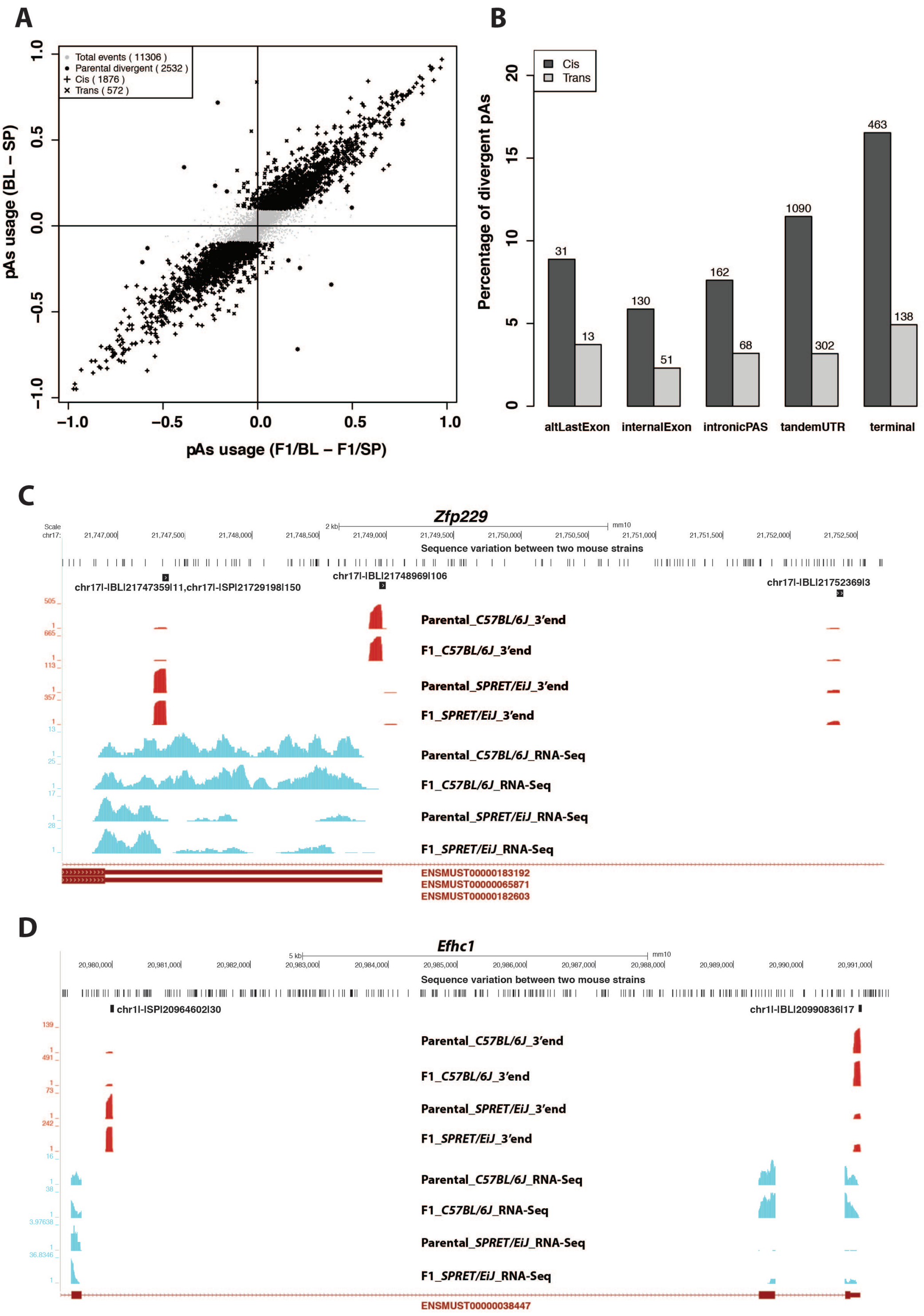

Figure 2.8 Dissection of the *cis*- and *trans*-regulatory pAs. A, Scatterplot comparing the parental pAs usage differences and allelic specific pAs usage differences; B, Number of *cis*- and *trans*-divergent pAs events across all five types of APA. Numbers of the pAs events were presented above the bars; C and D, Two representative examples of the *cis*-divergent pAs.

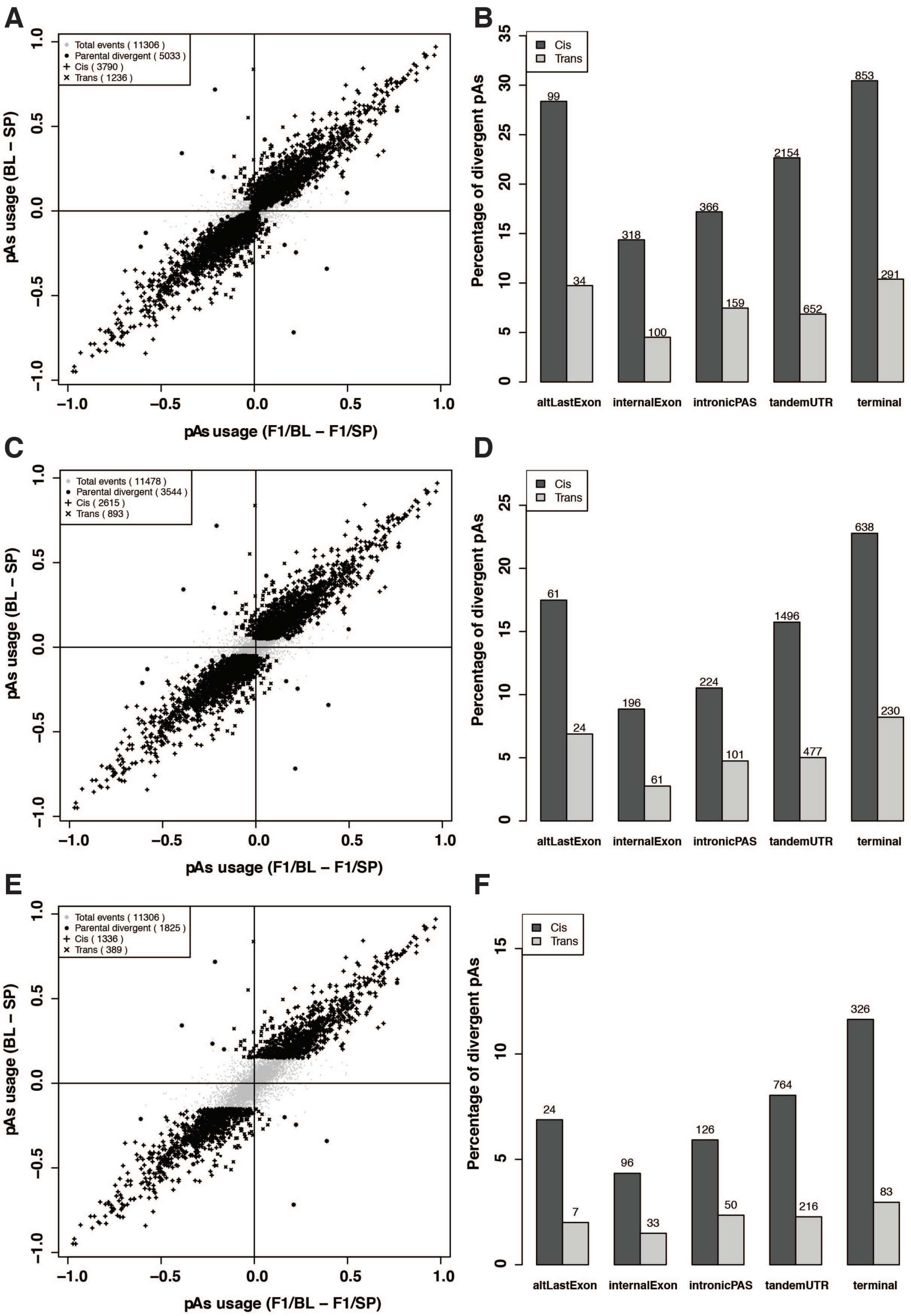

Figure 2.9 Predominant *cis*-regulatory pAs. A, C, and E, Scatterplot comparing the parental pAs differencies and allelic specific pAs differencies with different cut-offs of │ΔPPU│ (delta percentage of pAs usage) values (0, 0.05, 0.15); B, D, and F, Barplot showing the numbers of different types of APA.

To further assess the accuracy of the pAs usage quantification based on Illumina short sequencing reads between two alleles, we selected 20 genes for validation by using PacBio RS system. In brief, we performed 3' REACE for each pAs event with anchored oligo dT primers linked with common sequence and isoform specific primers without variants between two alleles in F1 hybrids. Around 400 nt region containing multiple variants at the end of each isoform was amplified and purified for deep-sequencing. The longer sequencing reads with more variants could improve the accuracy of assigning PacBio CCSs to different alleles and the read counts for each isoform were used to calculate the ratio of pAs usage between two alleles. As shown in Figure 2.10 the differences of pAs usage between the two alleles estimated in this way were highly correlated with those determined by 3' mRNA-Seq.

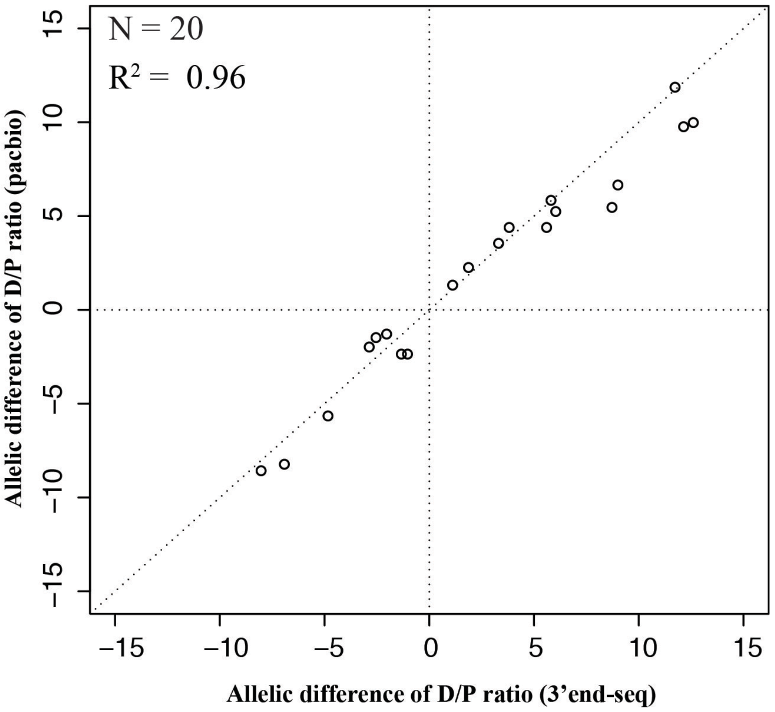

Figure 2.10 Allelic divergent pAs validation by PacBio sequencing. A, Scatterplot comparing the ratio of pAs usage between two alleles and that estimated by 3' mRNA-Seq data.

2.3.4 Genetic variants flanking the cleavage site contribute to cis-divergent pAs

Generally, the upstream and downstream 100 nt region of the cleavage site were considered as target for APA complex and auxiliary RBPs. In contrast to *trans*-regulatory divergence, *cis*-regulatroy divergence should be mainly caused by genetic variants in the flanking region of the affected pAs events, thus in order to investigate this, we calculated the frequency of both SNPs and indels in this region (-100 nt, 100 nt) from *cis*-regulatory divergent pAs and non-divergent controls. As expected, comparing with non-divergent controls, the flaking regions of *cis*-regulatory divergent pAs have significantly increased frequency of genetic variants (Figure 2.11 A). This observation was even more significant while only the upstream regions of the cleavage sits were examined (Figure 2.11 B and C).

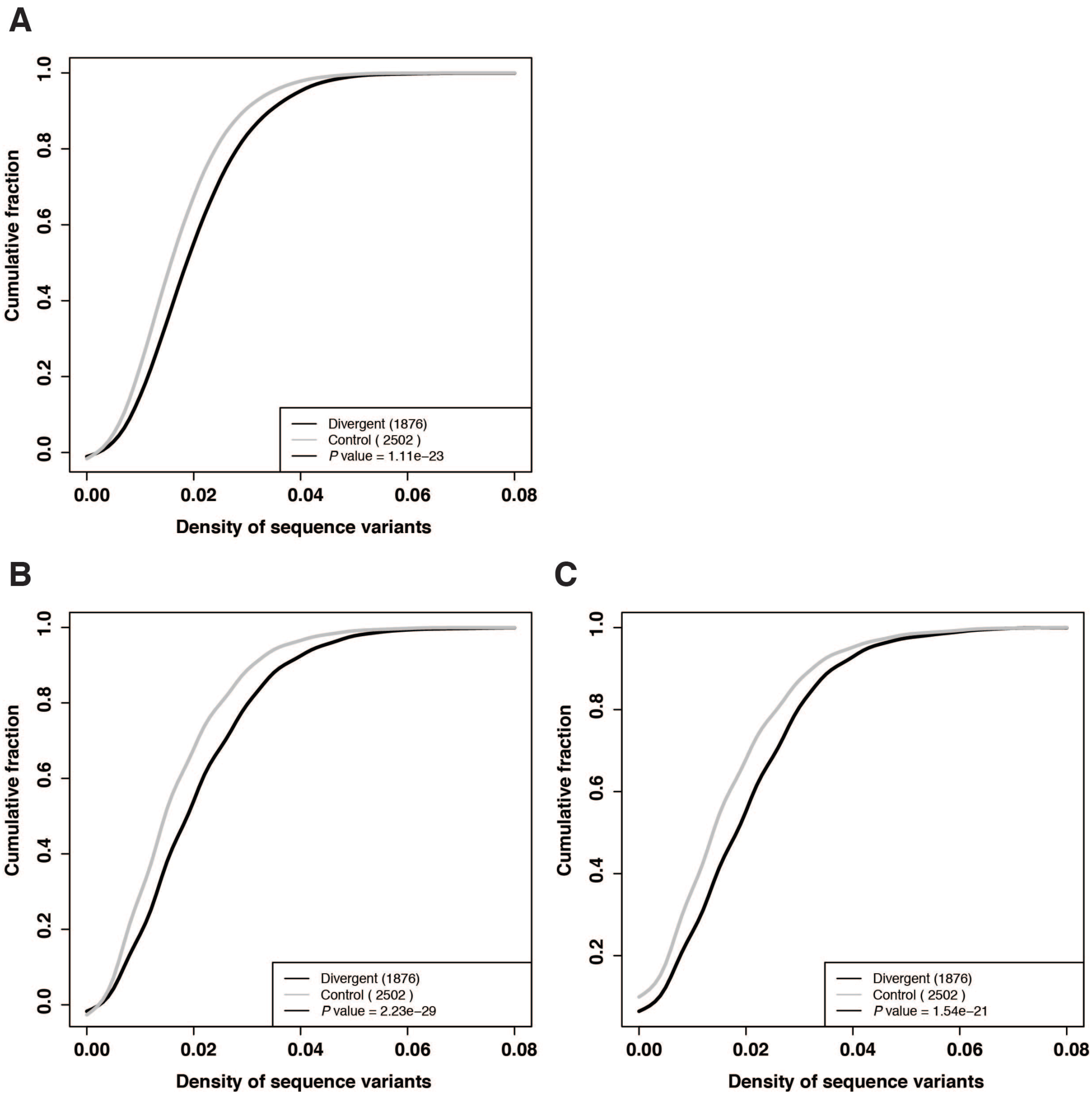

Figure 2.11 Genetic variants density around the pAs. A - C, CDF plots showing the genetic variant density in the flanking region, upstream region, and downstream region of pAs, respectively. Gray line indicates the non-divergent controls. Black line indicates the *cis*-divergent pAs events.

To further explore the distribution of genetic variants in the flanking region, we separated the region into four sub-groups: auxiliary upstream elements (AUE) [-100, -40), core upstream elements (CUE) [-40, 0), core downstream elements (CDE)

[0, 40), auxiliary downstream elements (ADE) [40, 100) according to previous study, and calculated the variants frequency. Comparing to the non-divergent controls, *cis*-divergent pAs showed higher variant density in both CUE and CDE region around the cleavage sites obviously (Figure 2.12 A). To better dissect the variant density along the flaking region with higher spatial resolution, we examined the distribution of variant frequency in the flanking region of cleavage site by slide window (Materials and Methods). As shown in Figure 2.12 B, the pAs with *cis*-divergence exhibited significant variant enrichment in the two proximal regions. Interestingly, the difference of variant density between divergent pAs and control pAs is most prominent in the narrow window (-30 nt to -15 nt) known to contain the PAS motif, i.e. AAUAAA and its variants. These results indicate that the raised variant density contributes to the regulation of pAs divergence.

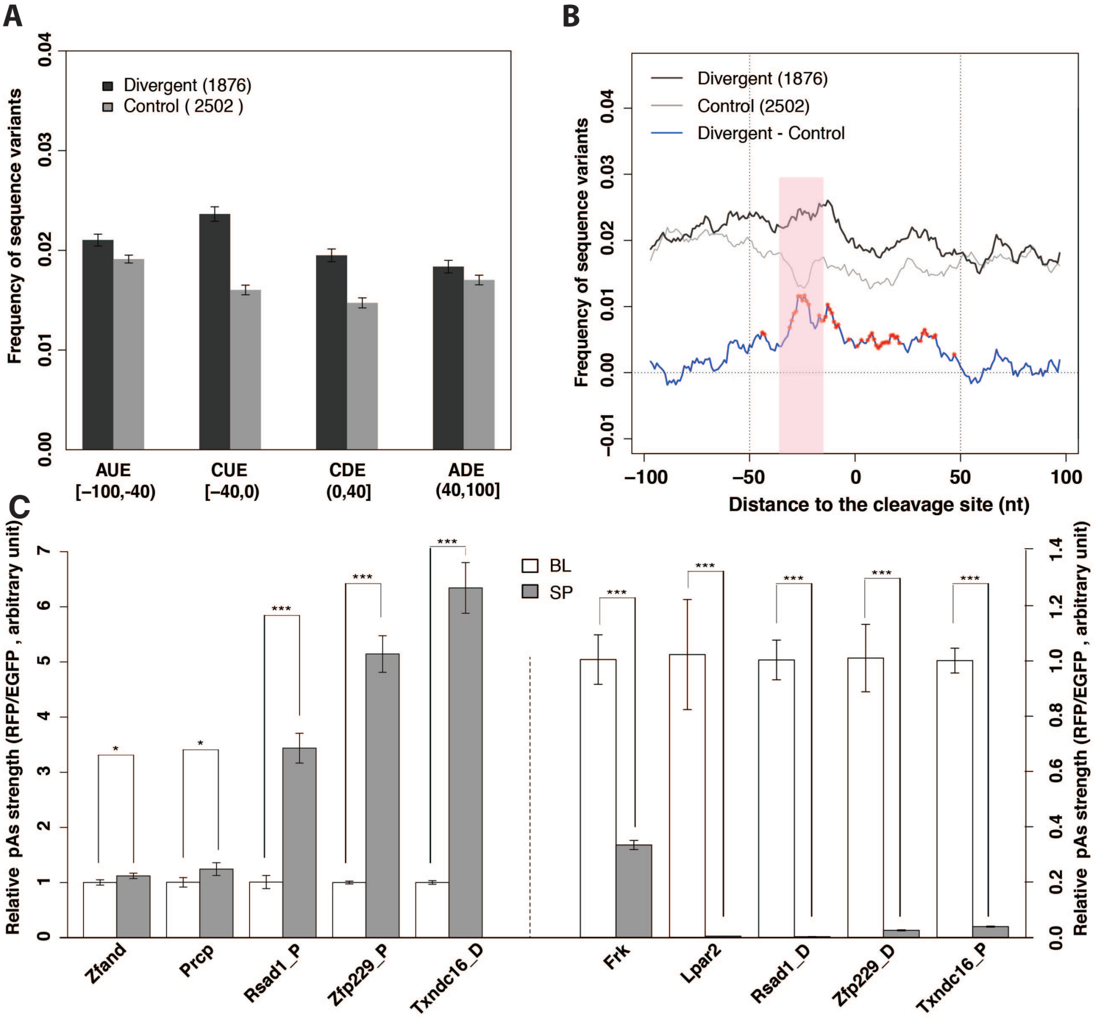

Figure 2.12 Dissecting the genetic variant around pAs. A, Barplot showing the variant density in four sub-regions around the cleavage sites; B, Analysis of genetic variant density in a 8 nt slide

window with step size of 1 nt around the cleavage site region. Gray and black line represents the genetic variant frequency of controls and *cis*-divergent pAs events, respectively. Red line shows the difference of variant frequency between *cis*-divergent and control pAs events. The pink sub window indicates the PAS region; C, Validation for the effect of genetic variants on pAs strength by reporter assay. The left and right panel shows the pAs strength biased to SPRET/EiJ and C57BL/6J allele, respectively. The biased pAs strength between two alleles is consistent with the 3' mRNA-Seq based quantification. X-axis shows different pairs of allele and the Y-axis indicates the relative strength of pAs. The higher ratio of RFP/EGFP means higher pAs strength. Details were written in the Materials and Methods.

To further examine whether the sequence difference in the flanking regions are able to drive the observed pAs divergence, we used a reporter assay to compare usage of the pAs flanked by the genomic sequences derived from the two alleles (Materials and Methods). As shown in Figure 2.12 C, 10 pAs derived from two allele pairs that were tested showed significant differential pAs usage biased towards the same alleles as observed in our global analysis. The results demonstrate that sequence variants in the flanking region in deed confer significant contribution to pAs usage.

2.3.5 RNA secondary structure in upstream proximal region inhibits pAs usage

RNA secondary structure is known to influence almost every step in the gene expression process, such as splicing, miRNA targeting and translation. However, its effect on pAs regulation is still not clear. To explore whether and how RNA secondary structure could regulate pAs usage, we compared the minimum free energy (MFE) of mRNA segments in the four windows flanking the cleavage sites between the two alleles (Materials and Methods), and correlated such difference to the observed allelic divergence in pAs usage. Interestingly, in the upstream proximal region (containing CUEs), the alleles with less stable local secondary structure were more likely to have higher pAs usage (Figure 2.13 A-D), indicating the formation of stable secondary structure in upstream proximal region of cleavage site could inhibit the definition of pAs, likely via making the PAS motif less accessible to CPSF complex. To further identify the affected region in more details, we also performed slide window analysis for MFE centered on the cleavage site of two alleles with different window sizes and got consistent results (Figure 2.14 A-C). These observations may suggest that apart from known PAS motifs and other auxiliary *cis*-elements, the formation of unstable secondary structure in close upstream region of cleavage site is another important modulator for the efficient cleavage and polyadenylation.

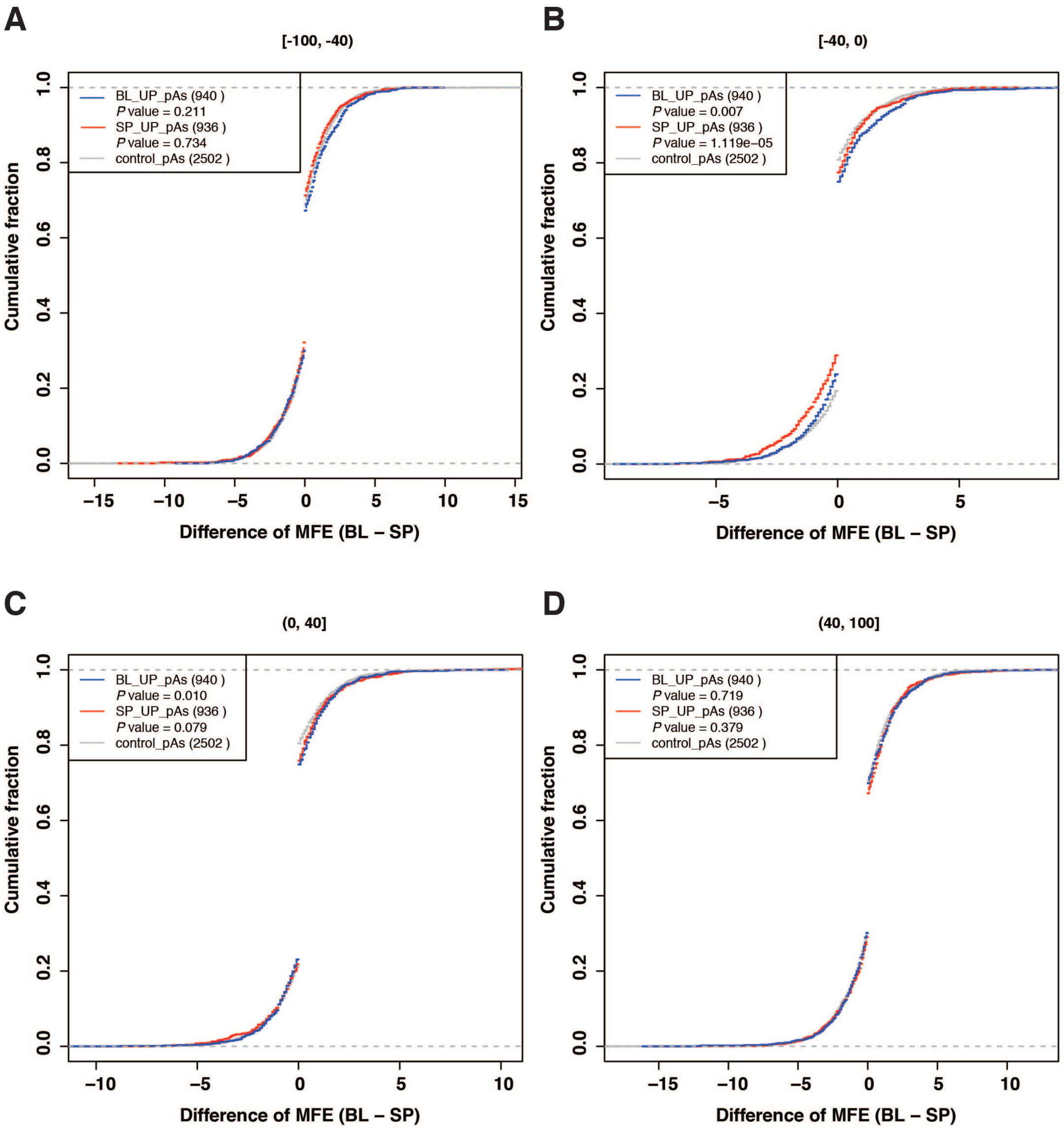

Figure 2.13 Effect of secondary structure on pAs usage. MFE differences of divergent pAs from C57BL/6J and SPRET/EiJ allele. All pAs are divided into three groups including C57BL/6J pAs usage up-regulated, SPRET/EiJ pAs usage up-regulated, and non-divergent controls. Difference > 0 indicates C57BL/6J allele has higher MFE and is less stable than SPRET/EiJ, and < 0 means SPRET/EiJ allele has higher MFE and is less stable than C57BL/6J. A and B, MFE in the upstream region of cleavage site are calculated; C and D, MFE in the downstream region of cleavage site are calculated.

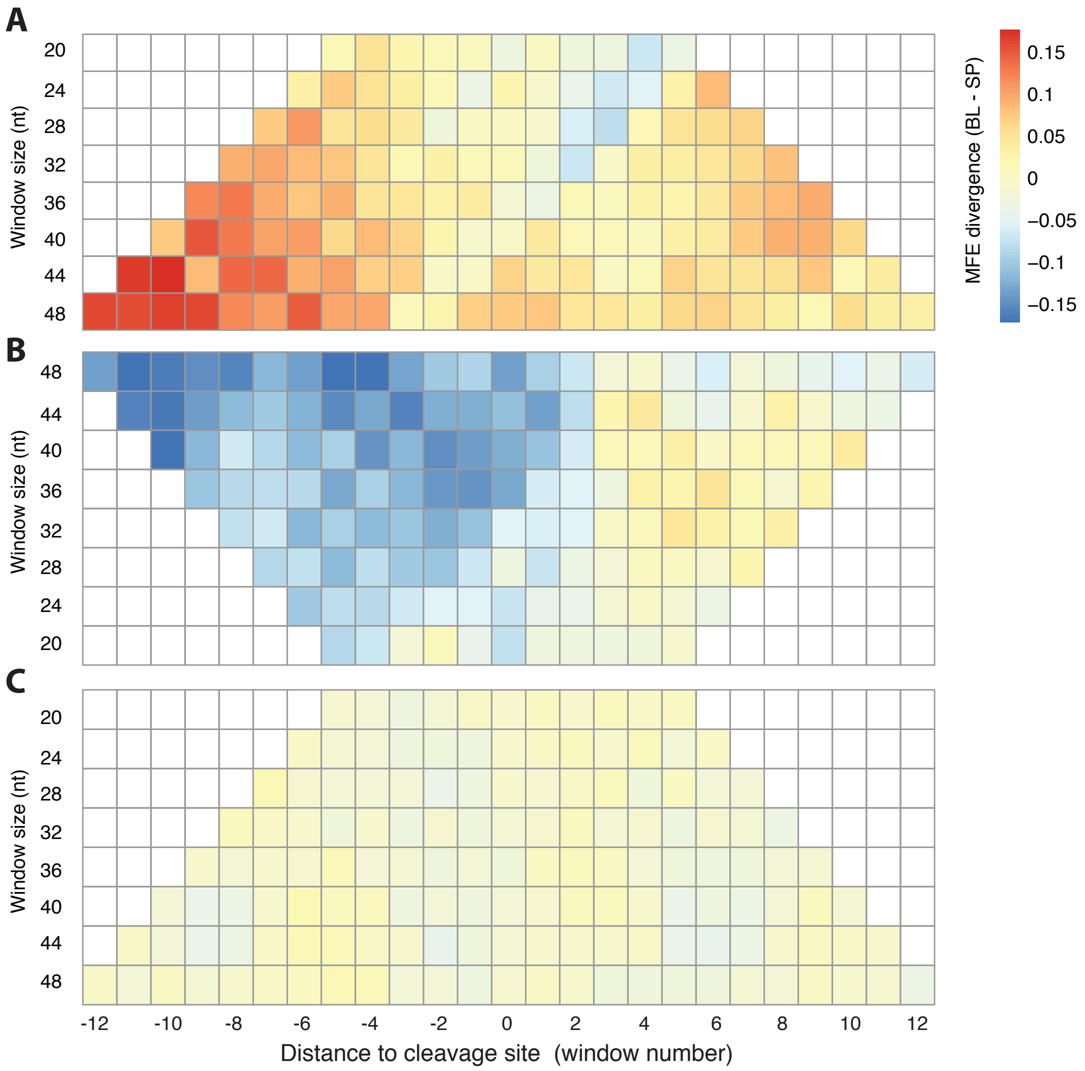

Figure 2.14 Slide window analysis of secondary structure on pAs usage. Heatmap showing the MFE differences between pAs from two alleles surrounding the cleavage site by slide window of different size by step size of 2 nt. X-axis indicates the window number with minus and plus standing for upstream and downstream of cleavage site, respectively. Y-axis indicates the different window size. A, pAs usage biased to C57BL/6J allele; B, pAs usage biased to SPRET/EiJ allele; C, non-divergent pAs controls. Details are written in the Materials and Methods section.

2.3.6 cis-elements associated with divergent pAs expression

To further identify the potential regulatory sequence elements affecting the pAs usage, we correlated the allelic difference in the frequency of all hexamers in the flanking region of the cleavage sites to the observed allelic pAs usage divergence. As shown in Figure 2.15 A and B, in the upstream region, three hexamers displayed obvious correlation: the frequency of AAUAAA and AUUAAA were positively correlated with increased pAs usage, whereas that of UUUUUU was negatively correlated. To exclude the potential bias introduced by regions with multiple Us or As and thus leads to overestimated contribution to pAs usage, we first extracted the pAs with one specific hexamer in the upstream region in either allele, then calculated the mutation frequencies of the specific hexamer in the upstream region of pAs whose usage down-regulated, up-regulated, and unchanged between two alleles. We found a higher frequency of mutation rate in the pAs whose usage down-regulated for both AAUAAA and its close variant AUUAAA (Figure 2.16 A and B). However, for UUUUUU, it turned out to be an opposite trend, which has significantly decreased mutation frequency in the down-regulated group (Figure 2.16 C). In addition, to substantiate this finding and assess whether the presence/absence of these hexamers could predict the allelic pAs divergence, we separated the pAs into two groups, one containing pAs with the presence of these hexamer in both alleles, and the other having pAs with the hexamers in only one allele, and compared the differences of allelic pAs usage between the two groups. As expected, the presence of the canonical pAs motif AAUAAA and the most frequently used variant AUUAA, significantly enhanced pAs usage (Figure 2.16 D and E). In contrary, the presence of UUUUUU inhibited the use of downstream pAs, although with much subtler effect (Figure 2.16 F). Interestingly, not only a stretches of six Us, but also a stretch of seven or eight Us could confer similar inhibitory effect (Figure 2.17). Moreover, the more severe the disruption of the poly U stretches, the more substantial the inhibitory effect, further suggesting the inhibitory role of poly(U) tract on pAs usage in the upstream region of the cleavage sites.

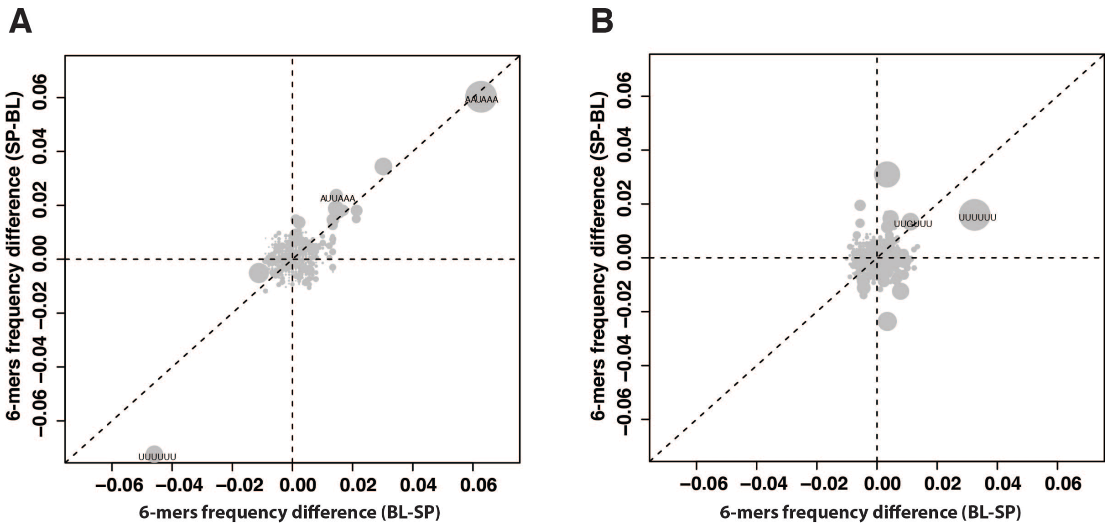

Figure 2.15 K-mer analysis. Divergent pAs were divided into two groups, which biased to C57BL/6J (BL) and SPRET/EiJ (SP) allele, respectively. The scatterplot showing the correlation of hexmer frequency differences of two alleles between two groups. A, Hexmer frequencies calculated in the upstream region of pAs; B, Hexmer frequencies calculated in the downstram region of pAs.

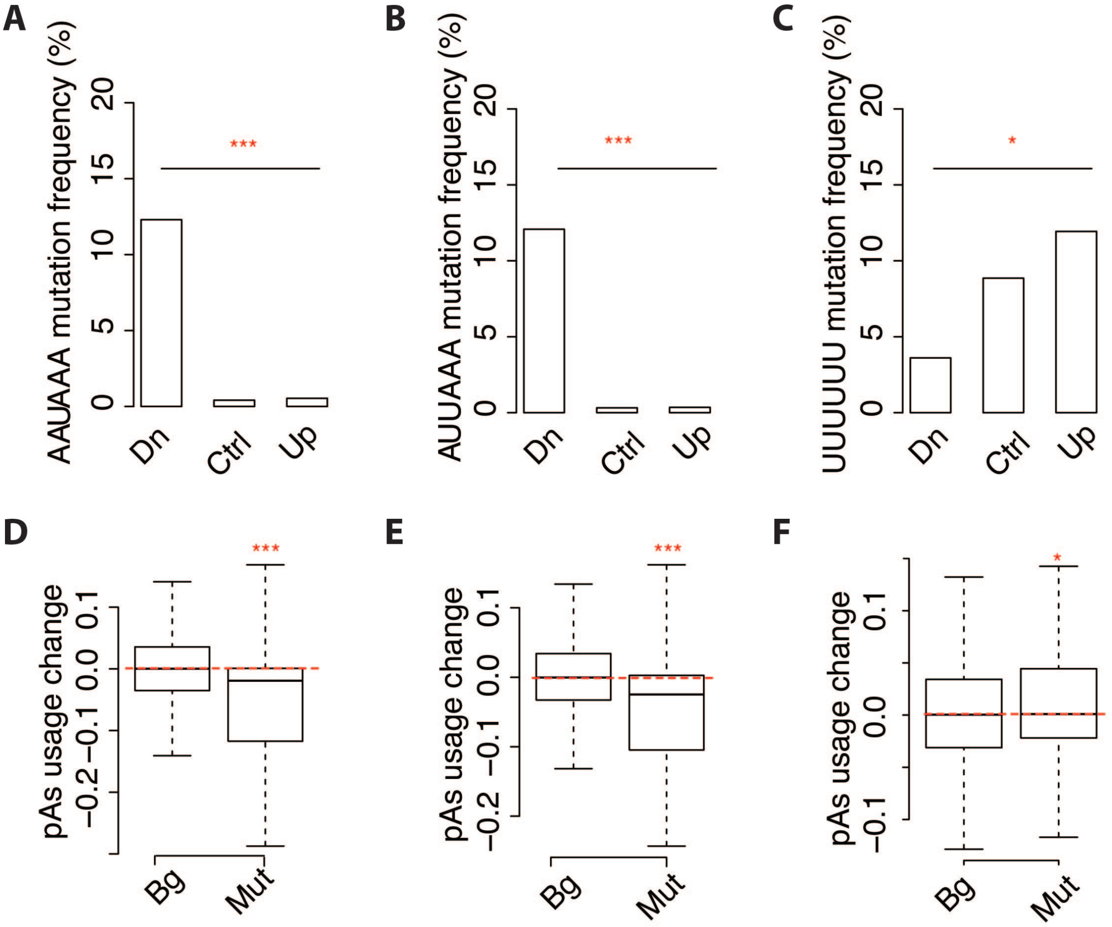

Figure 2.16 Effective size of PAS motifs and UUUUUU element. For all the pAs, either of the two alleles contains AAUAAA, AUUAAA, or UUUUUU in the upstream region of cleavage site were extracted and separated into three groups. For each of the three groups, pAs were further separated into

three subgroups according to whether there was biased pAs usage between two alleles. A, B, and C, showing the mutation frequency of AAUAAA, AUUAAA, or UUUUUU element in each subgroup, respectively; Total pAs were separated into two groups including both allele have the same PAS motifs or UUUUUU element and only one of the allele have the intact motifs or UUUUUU element. D, E, and F, showing the pAs usage differences of two alleles between the two groups.

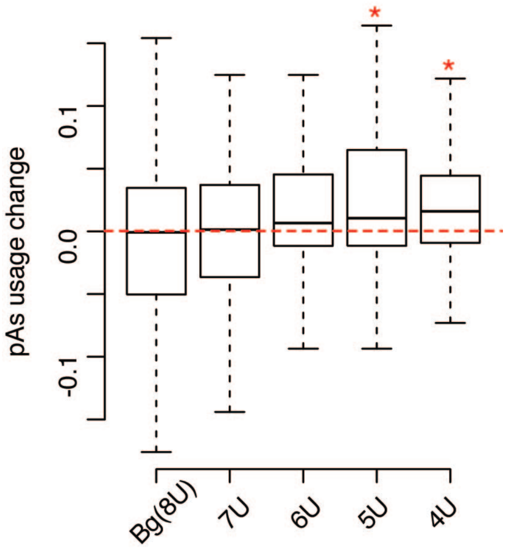

Figure 2.17 Effect of poly(U) tract on pAs usage. All pAs were divided into five groups according to the different number of poly Us in the upstream region of pAs between two alleles. X-axis indicates shared number of poly(U) between two alleles. Y-axis indicates the pAs usage differences between two alleles.

To experimentally validate the repressive role of UUUUUU element on pAs usage, we chose a divergent pAs from *Alg10b*, with UUUUUU in only one allele but not the other one. According to our 3' mRNA-Seq data, the pAs usage is significantly biased to the allele lacking the intact UUUUUU (Figure 2.18 A). Then by using the reporter assay, we first compared the pAs strength between two alleles (Materials and Methods). Consistent with our large-scale analysis, there is significant differential pAs usage biased against the allele containing intact UUUUUU element (Figure 2.18 B). Apart from the UUUUUU element, there are still some other variants between the two alleles, to determine directly the net effect of poly(U) tract on APA, we generated two additional constructs for each of the pAs. The only difference to their original vectors was the status of UUUUUU element (Materials and Methods). B (C57BL/6J with UU) had higher pAs strength than S (SPRET/EiJ with intact UUUUUU) as expected (Figure 2.18 B). The pAs strength decreased significantly by creating the intact UUUUUU element in B through introducing UUUU, in contrast, the pAs strength increased significantly upon removing UUUU in S (UUUUUU to UU) (Figure 2.18 B and C). These results demonstrate that intact UUUUUU element in the upstream region of cleavage site inhibit pAs usage.

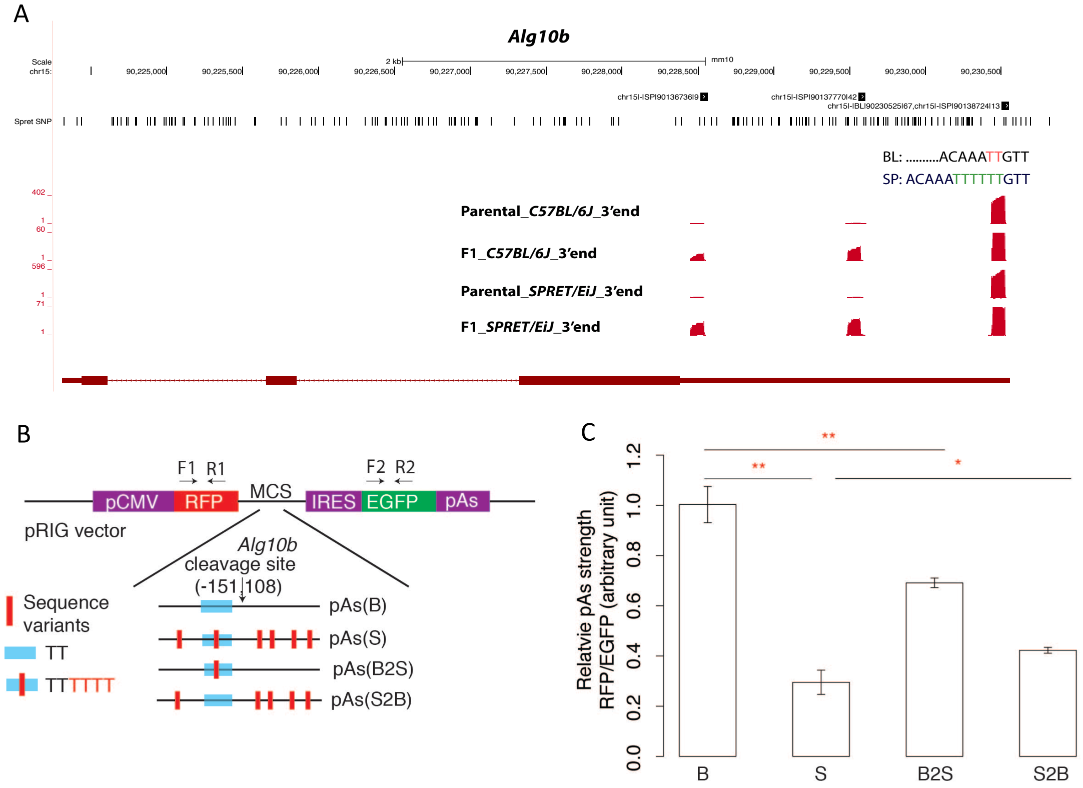

Figure 2.18 Validation of the effect of UUUUUU on APA. A, Example of destruction of intact UUUUUU element in the upstream region of pAs in *Alg10b* contributes to increased pAs usage; B, pRIG vector schematic map. The arrows stand for primers to amplify the REP and EGFP transcripts. The ratio of RFP/EGFP could be used for estimate the strength of different pAs with the higher strength standing for stronger pAs; C, Comparison of the strength of pAs in *Alg10b* between two alleles and their corresponding mutants. B and S stand for inserts from C57BL/6J and SPRET/EiJ, respectively. B2S indicates that UU in C57BL/6J is mutated to UUUUUU and for S2B is just the opposite mutation. The barplot shows the relative pAs strength for constructs with different inserts. Student's t-test is used to determine the difference of pAs strength between pAs derived from two alleles and *P* value less than 0.05 is thought to be statistically significant. Details on the cloning and quantification of pAs strength are written in the Materials and Methods section.

2.3.7 Functional consequence of cis-divergent pAs

Divergent pAs give rise to different mRNA isoforms with different length of 3' UTR and/or different coding region, which could regulate the metabolism of the affected transcripts. Shorter 3’ UTR containing mRNA was reported to be more stable and produce more amount of protein than the longer one from the same gene, though it was not generalized to other genes in 3T3 cell line in a later study (Mayr and Bartel, 2009; Spies et al., 2013). Among the divergent pAs, the vast majority was located in the last exon, only affecting the length of 3’ UTR. To explore the functional effect of the *cis*-divergent pAs on mRNA metabolism, we first compared the divergence of gene expression level between genes with *cis*-divergent pAs and that of non-divergent controls based on RNA-Seq data (Wei Sun, Qingsong Gao and *et al.*, unpublished data). As shown in Figure 2.19 A, genes contain divergent pAs have increased divergence of expression level between two alleles comparing with controls. As gene expression level is controlled by both transcription and degradation, to determine whether it is resulted from increased gene expression or altered mRNA half-life, we further compared gene expression level between genes with divergent pAs and controls utilizing the newly synthesized mRNA data and found no significant difference (Figure 2.19 B). So based on this finding, we expected that the divergent pAs can act on mRNA stability or half-life. Therefore we also compared the mRNA stability divergence between divergent pAs and controls, observing that genes contain divergent pAs have significantly incremented divergence of mRNA stability (Figure 2.19 C; Wei Sun, Qingsong Gao and *et al.*, unpublished data). Beyond mRNA level analysis, to examine the effect of alternative pAs on translation, we further compared the translation efficiency between genes with divergent pAs and controls by using our previous published data of translation efficiency in F1 hybrids (Hou et al., 2015). However, in contrast to mRNA stability, divergent pAs did not show regulatory effect on translation efficiency (Figure 2.19 D). Taken together, our data demonstrated that divergent pAs contribute to mRNA stability slightly but its impact on translation is limited.

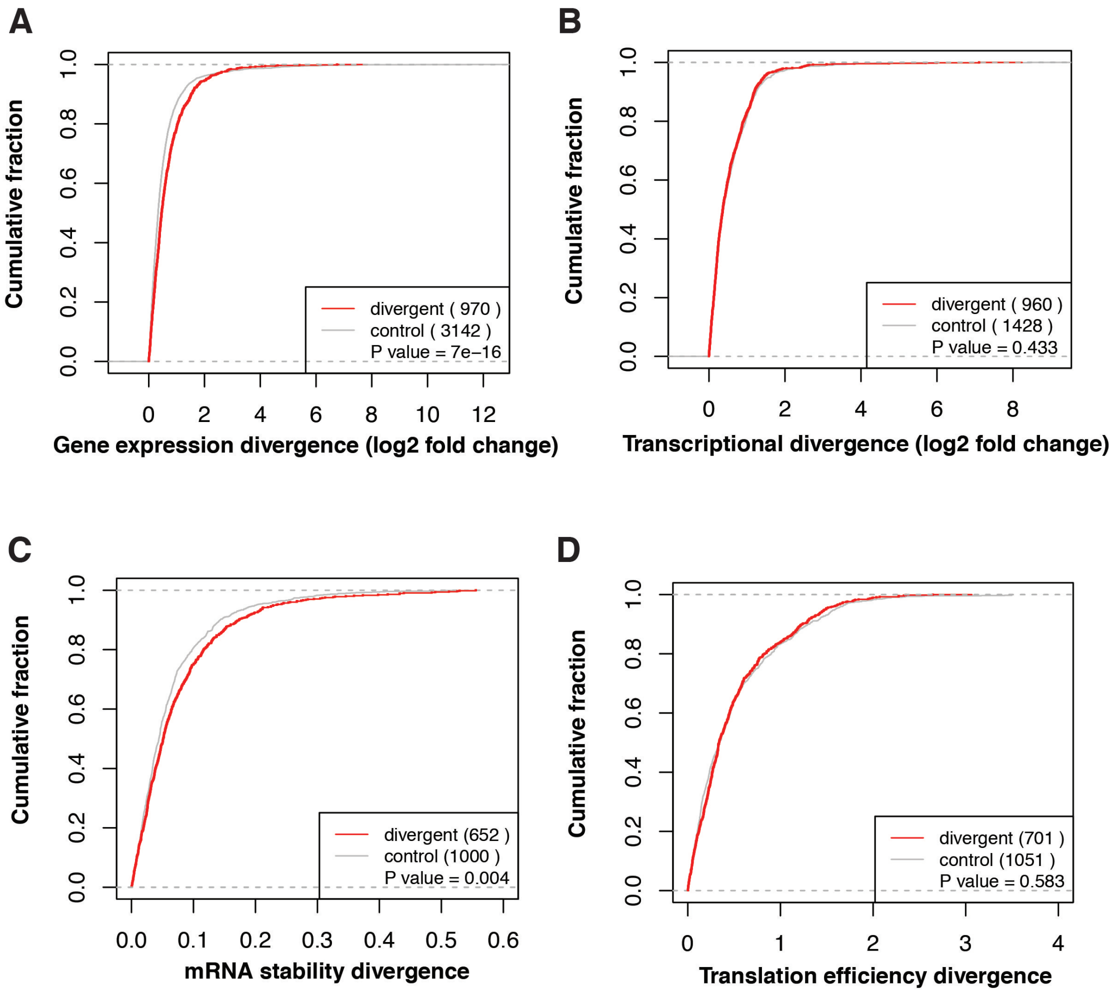

Figure 2.19 Functional consequence of divergent pAs on gene expression. Genes were divided into two groups according to whether there is divergent pAs or not. A, Cumulative fraction of total gene expression level divergence of genes with divergent pAs and those without (controls); B, Cumulative fraction of newly synthesized gene expression level of genes with divergent pAs and those without (controls); C, Cumulative fraction of mRNA stability divergence of genes with divergent pAs and those without (controls); D, Cumulative fraction of translation efficiency divergence of genes with divergent pAs and those without (controls).

2.4 Discussion

Comparing with the huge phenotypic divergences among different species, only small differences of gene number were identified by sequencing of different animal genomes. This led to the assumption that transcriptional and translational regulation paly a key role in shaping the phenotypic diversity during evolution. APA, which is regulated by both *cis*-elements and *trans*-factors, can affect transcriptome and proteome through generating mRNA isoforms that differ in coding or 3’ UTR regions and thus potentially regulate stability, translation, localization, and function of target mRNA (An et al., 2008; Loya et al., 2008; Mayr and Bartel, 2009; Elkon et al., 2013;

Tian and Manley, 2013; Berkovits and Mayr, 2015). To genome-widely investigate the contribution and underling mechanisms of *cis*- and *trans*-regulatory effect on APA divergence, we identified the pAs and quantified their usage in the fibroblast derived from two mouse strains and the F1 hybrids by 3’ READS and 3’ mRNA-Seq, respectively. Our data revealed that among the 2532 divergent pAs between two parental strains, 1790 (70.7%) showed clearly predominant contribution of *cis*-regulatory effect across different types of APA. Genetic variants and secondary structures in the close flanking region of the cleavage site contribute a lot to the observed extensive *cis*-regulatory divergent pAs. Furthermore, a poly(U) element upstream of the pAs was identified to be a potential repressor in APA.

Earlier studies on phenotypic diversity observed extensive *cis*-regulatory effect on gene expression (Goncalves et al., 2012), alternative splicing (McManus et al., 2014; Gao et al., 2015), and translation efficiency (Hou et al., 2015) during evolution through different model systems and methods. Consistently, our study also found pervasive and predominant contribution of *cis*-regulatory pAs across different types of APA by using F1 hybrid mouse model.

Comparing to the non-divergent controls, *cis*-divergent pAs have significantly higher variant density in the flanking region of cleavage site, suggesting that *cis*-regulatory pAs is actively affected by the genetic variants nearby (Figure 2.11). Furthermore, we also observed that the difference of sequence variant frequency between divergent pAs and controls is much more significant in the close vicinity of the pAs and with the highest difference in the upstream region (Figure 2.12 A and B), indicating the great importance of these two regions in APA. PAS and its variants located in the upstream -30 nt to -15 nt region of cleavage site and targeted by the CPSF complex could be identified in more than 90% pAs and intact motif is critical for effective APA (Tian et al., 2005; Proudfoot, 2011; Elkon et al., 2013). Interestingly, comparing to the non-divergent controls, *cis*-divergent pAs have significantly increased variant density in this local region (Figure 2.12 B). This observation implies that variants altering the integrity of PAS contribute to the pAs strength and thus results in the divergence of pAs during evolution. Apart from the PAS motif, recently by using NGS based analyses of RNA structure, Yiliang *et al.* found that the upstream -15 to -2 nt region of pAs is more structured than the average of the 100 nt region around the cleavage site in Arabidopsis (Ding et al., 2014). Nevertheless, how is the pAs strength affected and regulated by RNA secondary structure in the vicinity of pAs is unclear. Here our F1 hybrid mouse system provides an ideal model for comparing the pAs strength between two alleles, which contain multiple genetic variants and could potentially alter the local RNA structure. Indeed, comparing to non-divergent controls, pAs with increased strength or usage prefer to have less structure in upstream region of cleavage site significantly (Figure 2.13 and 2.14). The less structured region may increase the accessibility of pre-mRNA to APA machinery and/or auxiliary factors and thus enhance the cleavage efficiency. Further analysis with other datasets and possible functional experiments are needed to confirm it and further pinpoint the molecular mechanisms.

For *cis*-divergent pAs, we performed hexamer analyses and revealed that canonical PAS motif and its closest variant are positively correlated with increased pAs usage (Figure 2.15). Out of our expectation is that besides the canonical PAS, we also identified a UUUUUU element, which exerts its function as a potential pAs repressor. Genome-wide analysis of pAs strength of two alleles in F1 hybrids found that destroying the intact poly(U) element could slightly increase pAs usage and the trend could be even more obvious when multiple positions were mutated (Figure 2.17). Consistent with this, our cloning and mutagenesis results demonstrated that exchange affected UUUUUU element between the two alleles enhance or decrease the pAs strength. Consistent with our finding, this inhibitory role of poly(U) in the upstream region was also reported in a very recently paper(Gruber et al., 2016). In that study, the authors found that HNRNPC binds to poly(U) tract in close proximity of 3' end processing sites and this binding could inhibit cleavage and polyadenylation. However, the exact underlying mechanism of HNRNPC regulation and whether other RBPs are also involved in the process is still waiting for further studies.

By interrogating the RNA-Seq data generated from the same F1 hybrid fibroblast cell line, our results showed that the expression level divergence of newly synthesized mRNA of the divergent pAs containing genes was similar with that of non-divergent controls. This indicates that the transcriptional activity has limited influence on the divergence of pAs usage from the two alleles. This may also rule out the possibility that the identified divergent pAs is resulted from differential transcriptional activity between the two alleles. When the total mRNA expression level was compared, genes with divergent pAs have significantly increased divergence of mRNA expression, suggesting that different mRNA isoforms generated by APA could have variable degradation rate and ultimately lead to enlarged divergence of mRNA stability. The speculation was further confirmed by measuring the divergence of mRNA stability between genes with divergent pAs and the non-divergent controls. It has been reported that alteration of 3' UTR length could include or exclude functional motifs that regulate mRNA stability for a handful of genes in different cancer cell lines (Mayr and Bartel, 2009; Lin et al., 2012). However, when we only consider the divergent pAs within last exon, which generate isoforms with different length of 3' UTR, we observed no significant correlation between allelic mRNA stability divergence and allelic 3' UTR length difference. Consistent with our finding, two previous large-scale analyses found only subtle or limited correlation between 3' UTR length and mRNA stability (Spies et al., 2013; Gruber et al., 2014). The observation was further echoed by another study which showed that some shorter isoforms having higher and others having lower mRNA stability comparing to the corresponding longer isoforms, though transcripts stability could be regulated by APA (Gupta et al., 2014). Beyond mRNA level, APA was reported to regulate translation efficiency as well (Yu et al., 2006; Mayr and Bartel, 2009; de Klerk and t Hoen, 2015). However, we did not find obviously different divergence of translation efficiency between genes with divergent pAs and their corresponding non-divergent controls. In accordance with our finding, two previous genome-wide studies also failed to detect significant correlation between isoforms with different 3' UTR, indicating that 3' UTR might have limited influence on translation (Spies et al., 2013; Gruber et al., 2014). There might be two possible explanations. First, the translation efficiency can be regulated through different mechanisms and contribution from each way is limited or context dependent. Second, the regulatory role of 3' UTR on translation may be condition dependent.

In summary, our results revealed extensive *cis*-regulatory effect on APA during evolution and genetic variants in the flanking region contributed a lot to the divergent pAs. We also found that less stable local RNA secondary structure in the close upstream region of cleavage site is responsible for the increased pAs strength. Moreover, our data also showed that UUUUUU element in the close upstream region of pAs exerts as a potential pAs repressor.

Bibliography

Adesnik, M., Salditt, M., Thomas, W. and Darnell, J.E., 1972. Evidence that all messenger RNA molecules (except histone messenger RNA) contain Poly (A) sequences and that the Poly(A) has a nuclear function. J Mol Biol 71, 21-30.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B. and Sharp, P.A., 2013. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature 499, 360-3.

Alt, F.W., Bothwell, A.L., Knapp, M., Siden, E., Mather, E., Koshland, M. and Baltimore, D., 1980. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. Cell 20, 293-301.

An, J.J., Gharami, K., Liao, G.Y., Woo, N.H., Lau, A.G., Vanevski, F., Torre, E.R., Jones, K.R., Feng, Y., Lu, B. and Xu, B., 2008. Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. Cell 134, 175-87.

Anders, S., Reyes, A. and Huber, W., 2012. Detecting differential usage of exons from RNA-seq data. Genome Res 22, 2008-17.

Andreassi, C., Zimmermann, C., Mitter, R., Fusco, S., De Vita, S., Saiardi, A. and Riccio, A., 2010. An NGF-responsive element targets myo-inositol monophosphatase-1 mRNA to sympathetic neuron axons. Nat Neurosci 13, 291-301.

Baltz, A.G., Munschauer, M., Schwanhausser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., Wyler, E., Bonneau, R., Selbach, M., Dieterich, C. and Landthaler, M., 2012. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol Cell 46, 674-90.

Bandziulis, R.J., Swanson, M.S. and Dreyfuss, G., 1989. RNA-binding proteins as developmental regulators. Genes Dev 3, 431-7.

Beaudoing, E. and Gautheret, D., 2001. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. Genome Res 11, 1520-6.

Bell, D.A., Stephens, E.A., Castranio, T., Umbach, D.M., Watson, M., Deakin, M., Elder, J., Hendrickse, C., Duncan, H. and Strange, R.C., 1995. Polyadenylation polymorphism in the acetyltransferase 1 gene (NAT1) increases risk of colorectal cancer. Cancer Res 55, 3537-42.

Bennett, C.L., Brunkow, M.E., Ramsdell, F., O'Briant, K.C., Zhu, Q., Fuleihan, R.L., Shigeoka, A.O., Ochs, H.D. and Chance, P.F., 2001. A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA-->AAUGAA) leads to the IPEX syndrome. Immunogenetics 53, 435-9.

Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L. and Dreyfuss, G., 2012. U1 snRNP determines mRNA length and regulates isoform expression. Cell 150, 53-64.

Berkovits, B.D. and Mayr, C., 2015. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. Nature 522, 363-7.

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S.M., Singer, R.H. and Long, R.M., 1998. Localization of ASH1 mRNA particles in living yeast. Mol Cell 2, 437-45.

Bonapace, G., Concolino, D., Formicola, S. and Strisciuglio, P., 2003. A novel mutation in a patient with insulin-like growth factor 1 (IGF1) deficiency. J Med Genet 40, 913-7.

Brown, S.J., Stoilov, P. and Xing, Y., 2012. Chromatin and epigenetic regulation of pre-mRNA processing. Hum Mol Genet 21, R90-6.

Calvo, O. and Manley, J.L., 2001. Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination. Mol Cell 7, 1013-23.

Chan, S.L., Huppertz, I., Yao, C., Weng, L., Moresco, J.J., Yates, J.R., 3rd, Ule, J., Manley, J.L. and Shi, Y., 2014. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing. Genes Dev.

Chuvpilo, S., Zimmer, M., Kerstan, A., Glockner, J., Avots, A., Escher, C., Fischer, C., Inashkina, I., Jankevics, E., Berberich-Siebelt, F., Schmitt, E. and Serfling, E., 1999. Alternative polyadenylation events contribute to the induction of NF-ATc in effector T cells. Immunity 10, 261-9.

Colgan, D.F. and Manley, J.L., 1997. Mechanism and regulation of mRNA polyadenylation. Genes Dev 11, 2755-66.

Coseno, M., Martin, G., Berger, C., Gilmartin, G., Keller, W. and Doublie, S., 2008. Crystal structure of the 25 kDa subunit of human cleavage factor Im. Nucleic Acids Res 36, 3474-83.

Cowley, M., Wood, A.J., Bohm, S., Schulz, R. and Oakey, R.J., 2012. Epigenetic control of alternative mRNA processing at the imprinted Herc3/Nap1l5 locus. Nucleic Acids Res 40, 8917-26.

Danckwardt, S., Hentze, M.W. and Kulozik, A.E., 2008. 3' end mRNA processing: molecular mechanisms and implications for health and disease. EMBO J 27, 482-98.

de Klerk, E. and t Hoen, P.A., 2015. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends Genet 31, 128-39.

de la Mata, M., Alonso, C.R., Kadener, S., Fededa, J.P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D. and Kornblihtt, A.R., 2003. A slow RNA polymerase II affects alternative splicing in vivo. Mol Cell 12, 525-32.

Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M. and Babak, T., 2012. A quantitative atlas of polyadenylation in five mammals. Genome Res 22, 1173-83.

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C. and Assmann, S.M., 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505, 696-700.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R. and Hood, L., 1980. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. Cell 20, 313-9.

Edmonds, M., Vaughan, M.H., Jr. and Nakazato, H., 1971. Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship. Proc Natl Acad Sci U S A 68, 1336-40.

Edwalds-Gilbert, G., Veraldi, K.L. and Milcarek, C., 1997. Alternative poly(A) site selection in complex transcription units: means to an end? Nucleic Acids Res 25, 2547-61.

Elkon, R., Drost, J., van Haaften, G., Jenal, M., Schrier, M., Oude Vrielink, J.A. and Agami, R., 2012. E2F mediates enhanced alternative polyadenylation in proliferation. Genome Biol 13, R59.

Elkon, R., Ugalde, A.P. and Agami, R., 2013. Alternative cleavage and polyadenylation: extent, regulation and function. Nat Rev Genet 14, 496-506.

Ephrussi, A., Dickinson, L.K. and Lehmann, R., 1991. Oskar organizes the germ plasm and directs localization of the posterior determinant nanos. Cell 66, 37-50.

Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A. and Tress, M.L., 2014. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet 23, 5866-78.

Flavell, S.W., Kim, T.K., Gray, J.M., Harmin, D.A., Hemberg, M., Hong, E.J., Markenscoff-Papadimitriou, E., Bear, D.M. and Greenberg, M.E., 2008. Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. Neuron 60, 1022-38.

Fox-Walsh, K. and Fu, X.D., 2010. Chromatin: the final frontier in splicing regulation? Dev Cell 18, 336-8.

Fu, Y., Sun, Y., Li, Y., Li, J., Rao, X., Chen, C. and Xu, A., 2011. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res 21, 741-7.

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

Gieselmann, V., Polten, A., Kreysing, J. and von Figura, K., 1989. Arylsulfatase A pseudodeficiency: loss of a polyadenylylation signal and N-glycosylation site. Proc Natl Acad Sci U S A 86, 9436-40.

Goncalves, A., Leigh-Brown, S., Thybert, D., Stefflova, K., Turro, E., Flicek, P., Brazma, A., Odom, D.T. and Marioni, J.C., 2012. Extensive compensatory *cis*-trans regulation in the evolution of mouse gene expression. Genome Res 22, 2376-84.

Gruber, A.J., Schmidt, R., Gruber, A.R., Martin, G., Ghosh, S., Belmadani, M., Keller, W. and Zavolan, M., 2016. A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. Genome Res.

Gruber, A.R., Martin, G., Muller, P., Schmidt, A., Gruber, A.J., Gumienny, R., Mittal, N., Jayachandran, R., Pieters, J., Keller, W., van Nimwegen, E. and Zavolan, M., 2014. Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. Nat Commun 5, 5465.

Gunderson, S.I., Polycarpou-Schwarz, M. and Mattaj, I.W., 1998. U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. Mol Cell 1, 255-64.

Gupta, I., Clauder-Munster, S., Klaus, B., Jarvelin, A.I., Aiyar, R.S., Benes, V., Wilkening, S., Huber, W., Pelechano, V. and Steinmetz, L.M., 2014. Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. Mol Syst Biol 10, 719.

Harteveld, C.L., Losekoot, M., Haak, H., Heister, G.A., Giordano, P.C. and Bernini, L.F., 1994. A novel polyadenylation signal mutation in the alpha 2-globin gene causing alpha thalassaemia. Br J Haematol 87, 139-43.

Higgs, D.R., Goodbourn, S.E., Lamb, J., Clegg, J.B., Weatherall, D.J. and Proudfoot, N.J., 1983. Alpha-thalassaemia caused by a polyadenylation signal mutation. Nature 306, 398-400.

Hilgers, V., 2015. Alternative polyadenylation coupled to transcription initiation: Insights from ELAV-mediated 3' UTR extension. RNA Biol 12, 918-21.

Hilgers, V., Lemke, S.B. and Levine, M., 2012. ELAV mediates 3' UTR extension in the Drosophila nervous system. Genes Dev 26, 2259-64.

Hillier, L.W., Coulson, A., Murray, J.I., Bao, Z., Sulston, J.E. and Waterston, R.H., 2005. Genomics in C. elegans: so many genes, such a little worm. Genome Res 15, 1651-60.

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J.Y., Yehia, G. and Tian, B., 2013. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. Nat Methods 10, 133-9.

Hou, J., Wang, X., McShane, E., Zauber, H., Sun, W., Selbach, M. and Chen, W., 2015. Extensive allele-specific translational regulation in hybrid mice. Mol Syst Biol 11, 825.

Hu, J., Lutz, C.S., Wilusz, J. and Tian, B., 2005. Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation. RNA 11, 1485-93.

Huang, H., Chen, J., Liu, H. and Sun, X., 2013. The nucleosome regulates the usage of polyadenylation sites in the human genome. BMC Genomics 14, 912.

Jan, C.H., Friedman, R.C., Ruby, J.G. and Bartel, D.P., 2011. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature 469, 97-101.

Jankovic, L., Efremov, G.D., Petkov, G., Kattamis, C., George, E., Yang, K.G., Stoming, T.A. and Huisman, T.H., 1990. Two novel polyadenylation mutations leading to beta(+)-thalassemia. Br J Haematol 75, 122-6.

Ji, Z., Lee, J.Y., Pan, Z., Jiang, B. and Tian, B., 2009. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci U S A 106, 7028-33.

Ji, Z., Luo, W., Li, W., Hoque, M., Pan, Z., Zhao, Y. and Tian, B., 2011. Transcriptional activity regulates alternative cleavage and polyadenylation. Mol Syst Biol 7, 534.

Ji, Z. and Tian, B., 2009. Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. PLoS One 4, e8419.

Jiang, C. and Pugh, B.F., 2009. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 10, 161-72.

Jung, H., Yoon, B.C. and Holt, C.E., 2012. Axonal mRNA localization and local protein synthesis in nervous system assembly, maintenance and repair. Nat Rev Neurosci 13, 308-24.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L. and Dreyfuss, G., 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature 468, 664-8.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N.A., Eskin, E., Nellaker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T.G., Oliver, P.L., McIntyre, R.E., Bhomra, A., Nicod, J., Gan, X., Yuan, W., van der Weyden, L., Steward, C.A., Bala, S., Stalker, J., Mott, R., Durbin, R., Jackson, I.J., Czechanski, A., Guerra-Assuncao, J.A., Donahue, L.R., Reinholdt, L.G., Payseur, B.A., Ponting, C.P., Birney, E., Flint, J. and Adams, D.J., 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477, 289-94.

Keller, W., Bienroth, S., Lang, K.M. and Christofori, G., 1991. Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA. EMBO J 10, 4241-9.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36.

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

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Lee, C.Y. and Chen, L., 2013. Alternative polyadenylation sites reveal distinct chromatin accessibility and histone modification in human cell lines. Bioinformatics 29, 1713-7.

Legendre, M., Ritchie, W., Lopez, F. and Gautheret, D., 2006. Differential repression of alternative transcripts: a screen for miRNA targets. PLoS Comput Biol 2, e43.

Lembo, A., Di Cunto, F. and Provero, P., 2012. Shortening of 3'UTRs correlates with poor prognosis in breast and lung cancer. PLoS One 7, e31129.

Li, Q.H., Brown, J.B., Huang, H.Y. and Bickel, P.J., 2011. Measuring Reproducibility of High-Throughput Experiments. Annals of Applied Statistics 5, 1752-1779.

Lianoglou, S., Garg, V., Yang, J.L., Leslie, C.S. and Mayr, C., 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev 27, 2380-96.

Lim, L. and Canellakis, E.S., 1970. Adenine-rich polymer associated with rabbit reticulocyte messenger RNA. Nature 227, 710-2.

Lin, Y., Li, Z., Ozsolak, F., Kim, S.W., Arango-Argoty, G., Liu, T.T., Tenenbaum, S.A., Bailey, T., Monaghan, A.P., Milos, P.M. and John, B., 2012. An in-depth map of polyadenylation sites in cancer. Nucleic Acids Res 40, 8460-71.

Loya, A., Pnueli, L., Yosefzon, Y., Wexler, Y., Ziv-Ukelson, M. and Arava, Y., 2008. The 3'-UTR mediates the cellular localization of an mRNA encoding a short plasma membrane protein. RNA 14, 1352-65.

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R. and Misteli, T., 2011. Epigenetics in alternative pre-mRNA splicing. Cell 144, 16-26.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M. and Misteli, T., 2010. Regulation of alternative splicing by histone modifications. Science 327, 996-1000.

Lunde, B.M., Moore, C. and Varani, G., 2007. RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol 8, 479-90.

Majewski, J. and Pastinen, T., 2011. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet 27, 72-9.

Mandel, C.R., Bai, Y. and Tong, L., 2008. Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci 65, 1099-122.

Mangone, M., Manoharan, A.P., Thierry-Mieg, D., Thierry-Mieg, J., Han, T., Mackowiak, S.D., Mis, E., Zegar, C., Gutwein, M.R., Khivansara, V., Attie, O., Chen, K., Salehi-Ashtiani, K., Vidal, M., Harkins, T.T., Bouffard, P., Suzuki, Y., Sugano, S., Kohara, Y., Rajewsky, N., Piano, F., Gunsalus, K.C. and Kim, J.K., 2010. The landscape of C. elegans 3'UTRs. Science 329, 432-5.

Maniatis, T. and Reed, R., 2002. An extensive network of coupling among gene expression machines. Nature 416, 499-506.

Mansfield, K.D. and Keene, J.D., 2012. Neuron-specific ELAV/Hu proteins suppress HuR mRNA during neuronal differentiation by alternative polyadenylation. Nucleic Acids Res 40, 2734-46.

Martin, G., Gruber, A.R., Keller, W. and Zavolan, M., 2012. 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, 753-63.

Martin, G. and Keller, W., 2007. RNA-specific ribonucleotidyl transferases. RNA 13, 1834-49.

Masamha, C.P., Xia, Z., Yang, J., Albrecht, T.R., Li, M., Shyu, A.B., Li, W. and Wagner, E.J., 2014. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. Nature 510, 412-6.

Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I. and Pugh, B.F., 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res 18, 1073-83.

Mayr, C., 2015. Evolution and Biological Roles of Alternative 3'UTRs. Trends Cell Biol.

Mayr, C. and Bartel, D.P., 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138, 673-84.

McManus, C.J., Coolon, J.D., Eipper-Mains, J., Wittkopp, P.J. and Graveley, B.R., 2014. Evolution of splicing regulatory networks in Drosophila. Genome Res 24, 786-96.

Mendecki, J., Lee, S.Y. and Brawerman, G., 1972. Characteristics of the polyadenylic acid segment associated with messenger ribonucleic acid in mouse sarcoma 180 ascites cells. Biochemistry 11, 792-8.

Millevoi, S., Loulergue, C., Dettwiler, S., Karaa, S.Z., Keller, W., Antoniou, M. and Vagner, S., 2006. An interaction between U2AF 65 and CF I(m) links the splicing and 3' end processing machineries. EMBO J 25, 4854-64.

Millevoi, S. and Vagner, S., 2010. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. Nucleic Acids Res 38, 2757-74.

Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J.O. and Lai, E.C., 2013. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. Genome Res 23, 812-25.

Morris, A.R., Bos, A., Diosdado, B., Rooijers, K., Elkon, R., Bolijn, A.S., Carvalho, B., Meijer, G.A. and Agami, R., 2012. Alternative cleavage and polyadenylation during colorectal cancer development. Clin Cancer Res 18, 5256-66.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621-8.

Oikonomou, P., Goodarzi, H. and Tavazoie, S., 2014. Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts. Cell Rep 7, 281-92.

Oktaba, K., Zhang, W., Lotz, T.S., Jun, D.J., Lemke, S.B., Ng, S.P., Esposito, E., Levine, M. and Hilgers, V., 2015. ELAV links paused Pol II to alternative polyadenylation in the Drosophila nervous system. Mol Cell 57, 341-8.

Orkin, S.H., Cheng, T.C., Antonarakis, S.E. and Kazazian, H.H., Jr., 1985. Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. EMBO J 4, 453-6.

Ozsolak, F., Kapranov, P., Foissac, S., Kim, S.W., Fishilevich, E., Monaghan, A.P., John, B. and Milos, P.M., 2010. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell 143, 1018-29.

Ozsolak, F. and Milos, P.M., 2011. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12, 87-98.

Peattie, D.A., Hsiao, K., Benasutti, M. and Lippke, J.A., 1994. Three distinct messenger RNAs can encode the human immunosuppressant-binding protein FKBP12. Gene 150, 251-7.

Pinto, P.A., Henriques, T., Freitas, M.O., Martins, T., Domingues, R.G., Wyrzykowska, P.S., Coelho, P.A., Carmo, A.M., Sunkel, C.E., Proudfoot, N.J. and Moreira, A., 2011. RNA polymerase II kinetics in polo polyadenylation signal selection. EMBO J 30, 2431-44.

Proudfoot, N.J., 1976. Sequence analysis of the 3' non-coding regions of rabbit alpha- and beta-globin messenger RNAs. J Mol Biol 107, 491-525.

Proudfoot, N.J., 2011. Ending the message: poly(A) signals then and now. Genes Dev 25, 1770-82.

Proudfoot, N.J., Furger, A. and Dye, M.J., 2002. Integrating mRNA processing with transcription. Cell 108, 501-12.

Rogers, J., Early, P., Carter, C., Calame, K., Bond, M., Hood, L. and Wall, R., 1980. Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. Cell 20, 303-12.

Rozenblatt-Rosen, O., Nagaike, T., Francis, J.M., Kaneko, S., Glatt, K.A., Hughes, C.M., LaFramboise, T., Manley, J.L. and Meyerson, M., 2009. The tumor suppressor Cdc73 functionally associates with CPSF and CstF 3' mRNA processing factors. Proc Natl Acad Sci U S A 106, 755-60.

Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A. and Burge, C.B., 2008. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science 320, 1643-7.

Schonemann, L., Kuhn, U., Martin, G., Schafer, P., Gruber, A.R., Keller, W., Zavolan, M. and Wahle, E., 2014. Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. Genes Dev 28, 2381-93.

Setzer, D.R., McGrogan, M., Nunberg, J.H. and Schimke, R.T., 1980. Size heterogeneity in the 3' end of dihydrofolate reductase messenger RNAs in mouse cells. Cell 22, 361-70.

Shepard, P.J., Choi, E.A., Lu, J., Flanagan, L.A., Hertel, K.J. and Shi, Y., 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA 17, 761-72.

Shi, Y., 2012. Alternative polyadenylation: new insights from global analyses. RNA 18, 2105-17.

Shi, Y., Di Giammartino, D.C., Taylor, D., Sarkeshik, A., Rice, W.J., Yates, J.R., 3rd, Frank, J. and Manley, J.L., 2009. Molecular architecture of the human pre-mRNA 3' processing complex. Mol Cell 33, 365-76.

Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M. and Iyer, V.R., 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS Biol 6, e65.

Singh, P., Alley, T.L., Wright, S.M., Kamdar, S., Schott, W., Wilpan, R.Y., Mills, K.D. and Graber, J.H., 2009. Global changes in processing of mRNA 3' untranslated regions characterize clinically distinct cancer subtypes. Cancer Res 69, 9422-30.

Smibert, P., Miura, P., Westholm, J.O., Shenker, S., May, G., Duff, M.O., Zhang, D., Eads, B.D., Carlson, J., Brown, J.B., Eisman, R.C., Andrews, J., Kaufman, T., Cherbas, P., Celniker, S.E., Graveley, B.R. and Lai, E.C., 2012. Global patterns of tissue-specific alternative polyadenylation in Drosophila. Cell Rep 1, 277-89.

Spies, N., Burge, C.B. and Bartel, D.P., 2013. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. Genome Res 23, 2078-90.

Spies, N., Nielsen, C.B., Padgett, R.A. and Burge, C.B., 2009. Biased chromatin signatures around polyadenylation sites and exons. Mol Cell 36, 245-54.

Spraggon, L. and Cartegni, L., 2013. U1 snRNP-Dependent Suppression of Polyadenylation: Physiological Role and Therapeutic Opportunities in Cancer. Int J Cell Biol 2013, 846510.

Sun, Y., Fu, Y., Li, Y. and Xu, A., 2012. Genome-wide alternative polyadenylation in animals: insights from high-throughput technologies. J Mol Cell Biol 4, 352-61.

Takagaki, Y., Seipelt, R.L., Peterson, M.L. and Manley, J.L., 1996. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. Cell 87, 941-52.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R., Deaton, A., Andrews, R., James, K.D., Turner, D.J., Illingworth, R. and Bird, A., 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature 464, 1082-6.

Tian, B., Hu, J., Zhang, H. and Lutz, C.S., 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res 33, 201-12.

Tian, B. and Manley, J.L., 2013. Alternative cleavage and polyadenylation: the long and short of it. Trends Biochem Sci 38, 312-20.

Tian, B., Pan, Z. and Lee, J.Y., 2007. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. Genome Res 17, 156-65.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., Ibrahim, M., Omar, S.A., Lema, G., Nyambo, T.B., Ghori, J., Bumpstead, S., Pritchard, J.K., Wray, G.A. and Deloukas, P., 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39, 31-40.

Ulitsky, I., Shkumatava, A., Jan, C.H., Subtelny, A.O., Koppstein, D., Bell, G.W., Sive, H. and Bartel, D.P., 2012. Extensive alternative polyadenylation during zebrafish development. Genome Res 22, 2054-66.

Vuppalanchi, D., Coleman, J., Yoo, S., Merianda, T.T., Yadhati, A.G., Hossain, J., Blesch, A., Willis, D.E. and Twiss, J.L., 2010. Conserved 3'-untranslated region sequences direct subcellular localization of chaperone protein mRNAs in neurons. J Biol Chem 285, 18025-38.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B., 2008. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-6.

Wiestner, A., Tehrani, M., Chiorazzi, M., Wright, G., Gibellini, F., Nakayama, K., Liu, H., Rosenwald, A., Muller-Hermelink, H.K., Ott, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Vose, J., Armitage, J.O., Gascoyne, R.D., Connors, J.M., Campo, E., Montserrat, E., Bosch, F., Smeland, E.B., Kvaloy, S., Holte, H., Delabie, J., Fisher, R.I., Grogan, T.M., Miller, T.P., Wilson, W.H., Jaffe, E.S. and Staudt, L.M., 2007. Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. Blood 109, 4599-606.

Winters, M.A. and Edmonds, M., 1973a. A poly(A) polymerase from calf thymus. Characterization of the reaction product and the primer requirement. J Biol Chem 248, 4763-8.

Winters, M.A. and Edmonds, M., 1973b. A poly(A) polymerase from calf thymus. Purification and properities of the enzyme. J Biol Chem 248, 4756-62.

Wittkopp, P.J., Stewart, E.E., Arnold, L.L., Neidert, A.H., Haerum, B.K., Thompson, E.M., Akhras, S., Smith-Winberry, G. and Shefner, L., 2009. Intraspecific polymorphism to interspecific divergence: genetics of pigmentation in Drosophila. Science 326, 540-4.

Wood, A.J., Schulz, R., Woodfine, K., Koltowska, K., Beechey, C.V., Peters, J., Bourc'his, D. and Oakey, R.J., 2008. Regulation of alternative polyadenylation by genomic imprinting. Genes Dev 22, 1141-6.

Xia, Z., Donehower, L.A., Cooper, T.A., Neilson, J.R., Wheeler, D.A., Wagner, E.J. and Li, W., 2014. Dynamic analyses of alternative polyadenylation from

RNA-seq reveal a 3'-UTR landscape across seven tumour types. Nat Commun 5, 5274.

Yan, J. and Marr, T.G., 2005. Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. Genome Res 15, 369-75.

Yang, Q., Coseno, M., Gilmartin, G.M. and Doublie, S., 2011. Crystal structure of a human cleavage factor CFI(m)25/CFI(m)68/RNA complex provides an insight into poly(A) site recognition and RNA looping. Structure 19, 368-77.

Yasuda, M., Shabbeer, J., Osawa, M. and Desnick, R.J., 2003. Fabry disease: novel alpha-galactosidase A 3'-terminal mutations result in multiple transcripts due to aberrant 3'-end formation. Am J Hum Genet 73, 162-73.

Yoon, O.K. and Brem, R.B., 2010. Noncanonical transcript forms in yeast and their regulation during environmental stress. RNA 16, 1256-67.

Yu, M., Sha, H., Gao, Y., Zeng, H., Zhu, M. and Gao, X., 2006. Alternative 3' UTR polyadenylation of Bzw1 transcripts display differential translation efficiency and tissue-specific expression. Biochem Biophys Res Commun 345, 479-85.

Zhang, H., Lee, J.Y. and Tian, B., 2005. Biased alternative polyadenylation in human tissues. Genome Biol 6, R100.

Zhou, H.L., Luo, G., Wise, J.A. and Lou, H., 2014. Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. Nucleic Acids Res 42, 701-13.

Publications

1. **Xiao MS**, Zhang Bin, Li YS, Wei Sun, Chen Wei. 2016. Global analysis of regulatory divergence in the evolution of mouse alternative polyadenylation. (Under Review)
2. von Bernuth H, Ravindran E, Du H, Frohler S, Strehl K, Kramer N, Issa-Jahns L, Amulic B, Ninnemann O, **Xiao MS** et al. 2014. Combined immunodeficiency develops with age in Immunodeficiency-centromeric instability-facial anomalies syndrome 2 (ICF2). *Orphanet journal of rare diseases* **9**: 116.
3. Zhang W, **Xiao MS**, Ji S, Tang J, Xu L, Li X, Li M, Wang HZ, Jiang HY, Zhang DF et al. 2014. Promoter variant rs2301228 on the neural cell adhesion molecule 1 gene confers risk of schizophrenia in Han Chinese. *Schizophrenia research* **160**(1-3): 88-96.
4. Li WL, **Xiao MS**, Zhang DF, Yu D, Yang RX, Li XY, Yao YG. 2014. Mutation and expression analysis of the IDH1, IDH2, DNMT3A, and MYD88 genes in colorectal cancer. *Gene* **546**(2): 263-270.
5. **Xiao MS**, Chang L, Li WL, Du YS, Pan Y, Zhang DF, Wen Y, Luo J, Li XY, Yao YG. 2013. Genetic polymorphisms of the CASP8 gene promoter may not be associated with colorectal cancer in Han Chinese from southwest China. *PloS one* **8**(7): e67577.
6. **Xiao MS**, Zhang DF, Zeng Y, Cheng YF, Yao YG. 2011. Polymorphisms in the promoter region of the CASP8 gene are not associated with non-Hodgkin's lymphoma in Chinese patients. *Annals of hematology* **90**(10): 1137-1144.

Curriculum Vitae