Computational Study of Transcriptional Termination by Using Comprehensive High-Throughput Sequencing Data

Yufan Hui

A thesis presented for the degree of
Doctor of Philosophy

School of Computer Science
University of East Anglia
United Kingdom
May 2022

Computational Study of Transcriptional Termination by Using Comprehensive High-Throughput Sequencing Data

Yufan Hui

2021

Abstract

FCA is a plant-specific RNA-binding protein which regulates the expression of *Flowering Locus C (FLC)* through choice of poly(A) site in *COOLAIR*, the antisense transcript of *FLC*. While FCA is likely to be involved in the general transcription termination mechanism, little was known about the transcriptome-wide activity of FCA besides flowering control. With comprehensive high-throughput sequencing data, including enhanced Crosslinking and Immunoprecipitation (eCLIP)-seq, mRNA-seq, $3'$mRNA-seq and Chromatin-bond RNA (CB-RNA)-seq, we found the specific binding pattern of FCA to intronic regions, last exons and proximal downstream regions of transcriptional termination sites, with the binding of FCA in introns affecting co-transcriptional splicing efficiency of the introns and the binding near $3'$ end affecting the usage of poly(A) site of some genes. The binding downstream of PAS in *fpa-7* background was also found to increase in corresponding with the raised intergenic expression in *fpa-7*. With current results, we pictured the function of FCA as a $3'$ factor which promote the correct expression of genes through its binding to introns and $3'$ end of genes.

Computational Study of Transcriptional Termination by Using Comprehensive High-Throughput Sequencing Data

Yufan Hui

Manuscript in preparation

Arabidopsis FCA function as a sensor of aberrant transcription event at intron and 3$'$ end

I want to thank...

As a student of the Joint PhD Training Program offered by the University of East Anglia(UEA) and Southern University of Science and Technology (SUSTech), I want to thank my supervisors Taoyang and Alastair in UEA and Zhe(Jake) in SUSTech for their supervision. I want to thank all my labmates, both in UEA and SUSTech, especially Qiqi, Ling and Sixian for cDNA library construction of HTS data in FCA project. I would like to mention my dearest friend Deyue as well, for always been so supportive in research and in life. Also, I would never forget the time I spent in Bioinformatic Center of NJAU at the beginning of the second year of my PhD, where I started to learn basic bioinformatic skills for CLIP-seq analysis. I want to thank my friends in Room 241 of First Scientific Research Building in SUSTech as well, especially Ruiqing Fu and Bo Zhou from Jin lab for their help in many scientific issues. Thanks to all the people that helped me in life and in study during these four years of my postgraduate research.

Contents

List of Figures

List of Tables

1 Introduction

This thesis is composed of five more themed chapters besides the first introduction chapter: Chapter 2 deals with the biological background. In this thesis, we mainly focus on the transcriptional termination event through research of a plant specific RNA-binding protein FCA. As the background chapter, we would present an overview of the transcription events, from transcription initiation to termination, along with the RNA processing steps especially splicing and polyadenylation. Since the steps in termination are not isolated, the connection between termination and other steps would also be introduced. On the other hand, the transcription of mRNAs mainly depends on the RNA polymerase II (Pol II), and many RNA-binding proteins (RBPs) works closely with Pol II to promote transcription. There exist numerous RNA-binding proteins, working in a coordinated manner in the process of transcription. FCA is a typical RNA-binding protein (RBP) with two RNA recognition motifs (RRMs). FCA was detected as a suppressor of flowering, as well as a $3'$ processing factor, which affect the expression level of *Flowering Locus C (FLC)* through regulation of poly(A) site usage of *FLC*'s antisense transcript *COOLAIR* and affects flowering time. Previous research of FCA mainly focus on its impact on *FLC* expression level. Through the papers, we can also find close connection between FCA and other members of the autonomous pathway, especially FY and FPA. In the third section of this chapter, we would review the autonomous pathway to see how FCA and other members of the autonomous pathway affect the expression of *FLC*.

In Chapter 3, we would focus on next-generation sequencing (NGS) method. According to study of FCA before, the main concern was how it regulates the expression of *FLC* thus NGS technique was rarely applied to perform transcriptome-wide research on the function of FCA. As we are now interested in FCA function besides flowering control, it is necessary to introduce NGS into the project. Firstly, since FCA is an RNA-binding protein, we can apply the advanced technique of eCLIP-Seq data to define the binding site of FCA at single-nucleotide resolution. The position of binding may serve as a hint for possible function of FCA. But CLIP data alone is not enough to learn the transcriptome-wide impact aroused by FCA. Other sequencing data like mRNA-seq, $3'$mRNA-seq, CB-RNA-seq would also be applied to check the change in gene expression, alternative polyadenylation, splicing et al. By combination of data mentioned above, we hoped to figure out not only the position of FCA binding, but also what triggers FCA to bind the specific region of RNAs and what's the function of FCA. In this chapter, we would introduce the NGS method used in this research, including library construction and data analysis workflow. The methods are classified into three types according to the target RNAs: 1. mRNA-seq and $3'$mRNA-seq for mature messenger RNAs (mRNAs); 2. Chromatin-bond RNA-seq (CB-RNA-seq) and pNET-seq for RNA under construction; 3. CLIP-seq for RNA bond by protein or with modification. Followed by the result of analysis in Chapter 4 and Chapter 5.

In Chapter 4, we present the binding pattern of FCA gained from eCLIP-seq data. We've found a strong preference for FCA to bind intronic region and the $3'$ end including the downstream of the genes. Meanwhile, FCA binding might be affected by interaction with another member of autonomous pathway, FY. In the following chapter, more results of other sequencing data are shown to see why FCA specifically bind intronic and $3'$ end region and how the binding of FCA would affect transcription. We tested the alternative splicing along with the change of splicing efficiency in *fca-9* mutant in intronic regions and study the alternative polyadenylation near the $3'$. With all the results above, we would propose a model for possible function of FCA in transcription.

Chapter 6 is the last chapter for discussion of the result and future work. In this chapter, we discuss the shortcomings in the current method and provide a forecast for the future research. We would talk about the CLIP-seq method and the current shortcomings in detecting the strength of binding of RBPs. We are also interested in the potential application of long-read sequencing method in the research of FPA function to detect read-through transcript. At last, same unfinished part in FCA research would be listed as the potential start point for research in the future.

2 Biological Background

In this chapter, we present the necessary biological background to understand the motivation of this study. It starts with RNA metabolism to present how the RNA was produced and the events mRNAs experienced before their get decay. We will introduce in detail some models of splicing and transcriptional termination to help the readers to understand the importance of these events. Next, we are going to introduce a significant type of proteins called RNA-binding Proteins (RBPs), many of which play important roles in RNA transcription, processing, export and decay. In the end, we will focus on the autonomous flowering pathway, especially on our research target FCA, a typical RBP. We will discuss what other researchers have got in this field and what we can do to fill the gaps in current knowledge.

2.1 RNA Metabolism: Fate of RNAs from Synthesis to Degradation

RNA metabolism refers to the events during the life cycle of RNA molecules, including their synthesis, modification, processing, transportation, storage and degradation. A critical enzyme involved in transcription and processing of mRNA is RNA polymerase II (Pol II). Not only does it catalyze the transcription of DNA to synthesize precursor RNAs, but evidence also shows that it may also help to coordinate RNA transcription and RNA processing (Hsin and Manley, 2012). Therefore, we would describe the transcription and processing of pre-mRNAs from a Pol II-centered aspect.

2.1.1 Introduction to RNA Polymerase

RNAs are transcribed with the help of RNA Polymerase. But whether different types of RNAs are transcribed by the same kind of polymerase remained unknown for a long time. In 1969, researcher managed to purified and split the Polymerases from both sea urchin and mouse liver (Roeder and Rutter, 1969). They found out that there exist three distinct types of polymerases, and named them RNA Polymerase I/II/III (Roeder and Rutter, 1969). Both Pol I and Pol II are nuclear localized: Pol I is enriched in nucleolus to transcribe 45S rRNA (Goodfellow and Zomerdijk, 2013); Pol II located in nucleoplasm for transcription of mRNA and some snRNAs; Pol III is also nucleoplasm-located and is responsible for the synthesis of non-coding RNAs including tRNAs (Woychik and Hampsey, 2002), 5S rRNA and U6 RNA (Dieci et al., 2007). In Arabidopsis, there exist two additional plant-specific polymerase, Pol IV and Pol V, which are involved in siRNA-mediated gene silencing (Ream et al., 2009).

Pol II is by far the best studied of all these polymerases. It plays an important role in the synthesis of messenger RNAs, which would be later transcribed into proteins of various function for maintenance of physiological events within cells. Pol II contain 12 subunits, with a C-terminal domain (CTD) of the largest subunit Rpb1 is important for the maintenance of transcription. Pol II CTD consists of tandem heptapeptide repeats with a consensus sequence of tyrosine-serine-proline-threonine-serine-proline-serine $Y_1S_2P_3T_4$ $S_5P_6S_7$. The number of repeats differs among species, for example, there are 26 such repeats in yeast RNAP II, while 52 in the human counterpart (Chapman et al., 2008). The CTD forms an unstructured extension connected to the polymerase core by an 80-residue linker. The linker binds RNAP II subunit, Rpb7, which forms a subcomplex with Rpb4. Pol II with unphosphorylated CTD is recruited to a promoter and assembles with general transcription factors (GTFs) a pre-initiation complex (PIC). Ser5 phosphorylation (Ser5-P) are enriched at the promoter and decrease successively towards the $3'$ end of genes, while Ser2-P, Thr4-P and Tyr1-P levels increase during transcriptional elongation. One of the main function of

Ser5-P is to recruit the capping machinery. Ser2-P might be involved in the productive elongation phase of transcription, while Thr4-P is limited to the body region of genes and might contribute to transcriptional elongation/termination (Nojima et al., 2015). Pol II phosphorylation was also proved to incorporate Pol II into mediator condensates and regulate the switch between condensates of transcription and processing (Guo et al., 2019). The dynamic phosphorylation states of the Pol II CTD are important to the regulation of transcription(Figure 2.1).

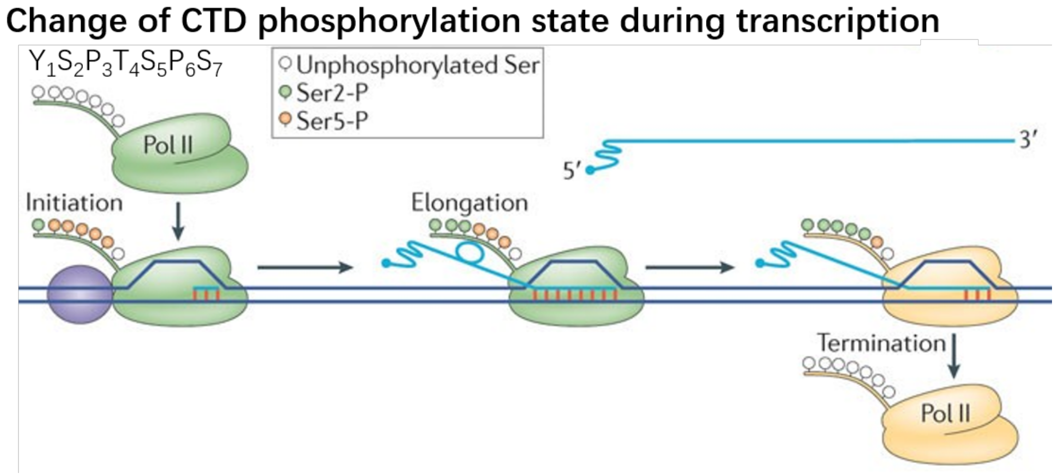


Figure 2.1: The phosphorylation state of Pol II changes dynamically during transcription. The pre-initiation complex recruitment was closely related with Ser5-phosphorylation, but Ser5P was removed after initiation, while Ser2 increased during transcription. This figure is from Kuehner et al. (2011).

2.1.2 RNA Metabolism: Transcription and Processing

2.1.2.1 RNA Transcription: Synthesis of Pre-mRNA

Initiation In the initiation phase, multiple events concerning Pol II are involved, including the recruitment of general transcription factors (GTFs), unwinding DNA templates, the initiation of RNA synthesis and Pol II promoter clearance. Although Pol II is responsible for catalyzation of DNA-dependent synthesis of mRNA, the complex alone is not sufficient to initialize the transcription in the absence of other factors. Thus, the assembly of the pre-initiation complex (PIC) of Pol II is a necessary step of locating a promoter to initialize transcription. In the case of DNA sequence that bearing a TATA box, a typical kind of promoter sequence, the PIC consists of GTFs including, TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH (Roeder, 1996). These factors help Pol II recruitment, leading to the association between Pol II and interact with the upstream promoter DNA. The synthesis then starts at the promoter site given the availability of Adenosine Triphosphate (ATP) and nucleoside triphosphates. Soon after that, it may enter a critical phase referred to as promoter clearance, in which the transcription complex tends to be physically and functionally unstable. After this step, Pol II pass through the promoter proximal region and loss the interactions with initiation factors, followed by the entrance to elongation stage known as transcriptional elongation (Luse, 2013)). It is caused by TFIIH phosphorylation of the Pol II carboxy-terminal domain (CTD) Ser2. Kin28, a subunit of TFIIH, is considered to be essential to this process (Wong et al., 2014). Clearance is complete when the Pol II elongation complex is fully stabilized. Some used to believe that transcription initiation happens at a single site for every gene, but in fact, transcription starting sites are heterogeneous for many genes. While it was widely accepted that the rate-limiting step during initiation is the recruitment of RNA polymerase, recent studies (Wade and Struhl, 2008) shows that the steps after Pol II recruitment could also be critical.

Elongation Production of mRNA depends critically on the rate of Pol II in transcriptional elongation. Elongation is the process of producing the RNA molecular from $5'$ end to $3'$ end, creating a corresponding RNA based on the template DNA strand. It starts after the completion of the promoter clearance. However, before entering the productive elongation phase, negative elongation factor (NELF) and 5,6-dichloro-1-β-D-ribofuranosylbenzimidazole (DRB) sensitivity-inducing factors (DSIF) mediate Pol II to pause in a promoter-proximal position (Yamaguchi et al., 2013). Interestingly, promoter-proximal pausing occurs at a point where it may serve to coordinate transcription elongation with pre-mRNA processing. RNA capping (Saunders et al., 2006) for example, happens soon after the initiation and could occur to help to increase the processivity of Pol II (Sims et al., 2004). While NELF and DSIF reduce the elongation rate, the positive transcription elongation factor (P-TEFb) helps Pol II to escape from the pausing state and enters into next state. P-TEFb triggers Pol II to release from a paused position and to recruit necessary factors for productive elongation such as transcription factor IIS (TFIIS) which aids in restarting arrested Pol II (Fish and Kane, 2002) and transcription factor IIF (TFIIF) which stimulate the elongation rate of Pol II in the absence of other factors (Price et al., 1989). The crucial role of P-TEFb also appears in the coupling of elongation and RNA processing (Zhou et al., 2012), especially in splicing.

Termination The last step of transcription is called transcriptional termination. When Pol II reaches the $3'$ end of the gene, elongation rate of Pol II would slow down, after which Pol II was then disassociated from the DNA template. This process is directly coupled to RNA processing. During or soon after transcriptional termination, RNA transcripts are released from the locus and be polyadenylated at its $3'$ end. A shocking fact about termination of transcription is that it occurs at all points in the transcription cycle, at the beginning, middle, and end of genes. The termination of transcription within gene body is known as pre-mature termination, while inefficient termination might result in the increase of transcription in intergenic regions so-called transcription read-through. Termination remained the least understood for an extended period. However, research (Kuehner et al., 2011) shows that the interconnection between transcription and other cellular processes tend to be strengthened as Pol II enters the last stage of RNA synthesis, adding to the importance of this step. Transcriptional termination is crucial for correct gene expression. The synthesis ends with the release of RNA, which will then be exported to the cytoplasmic after necessary processing and translated into proteins to fulfill its biological function. On the other side, the release of Pol II makes sure the availability of the subsequent round of RNA synthesis. Also, proper termination helps to restrict the extent of non-coding transcription and prevent the adjacent transcriptional units from interfering each other.

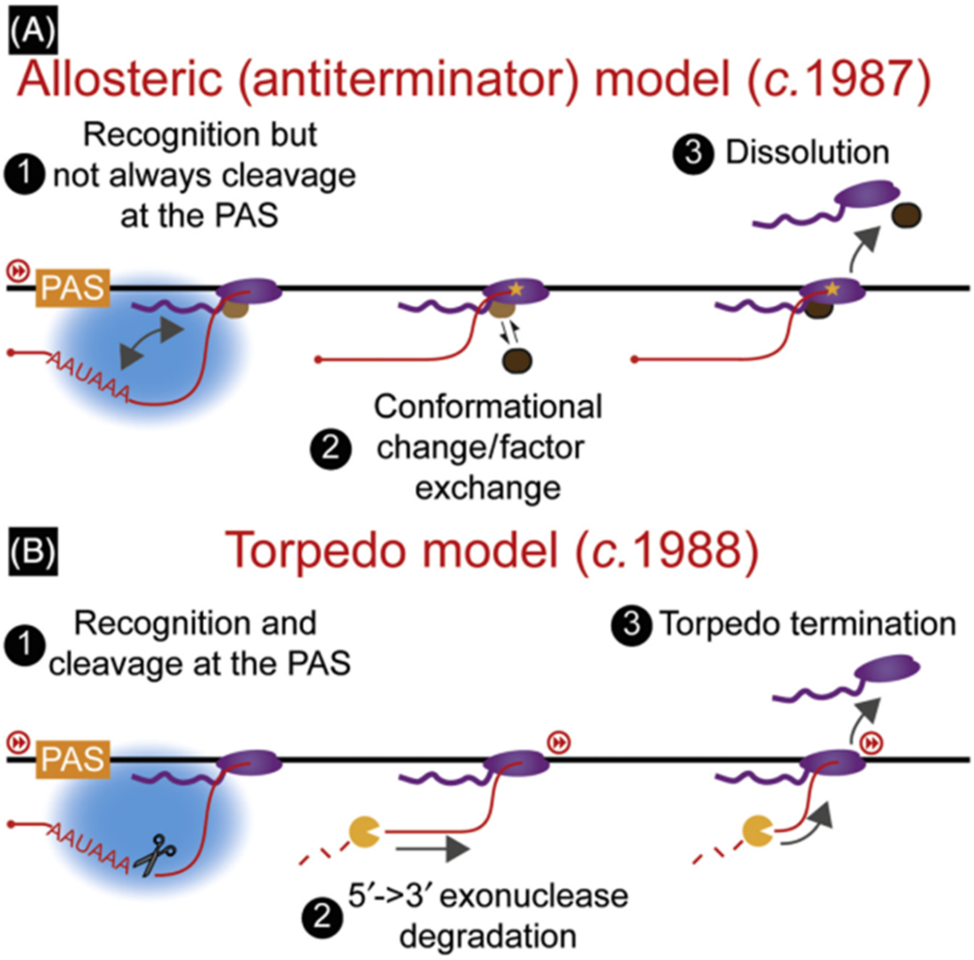

(A)
Allosteric (antiterminator) model (c.1987)
1 Recognition but not always cleavage at the PAS
3 Dissolution
PAS
AAUAAA
2 Conformational change/factor exchange
(B)
Torpedo model (c.1988)
1 Recognition and cleavage at the PAS
3 Torpedo termination
PAS
AAUAAA
2 5′->3′ exonuclease degradation

Figure 2.2: Two classical models of termination summarized by Eaton and West (2020), A. Allosteric Model proposes that termination was triggered by a conformational change or factors recruited to or dissociated from Pol II; B. torpedo model, in which the cleavage at PAS would generate a Pol-II-asscoiated RNA to be degraded by XRN2; Purple: Pol II complex with the CTD; XRN2, nuclear enzyme that degrades RNA in a 5′’→3′’ direction

Multiple models was introduced to describe the transcriptional termination process, the two most well-known models are allosteric model and torpedo model.

Shown in One of the well-known models is (Figure 2.2 A), also known as anti-terminator model. The model proposes that transcription of a PAS causes a conformational change within Pol II that promotes its termination and finally leads to the dissociation of anti-termination factors from the Pol II. Previous research has revealed that PAS-dependent termination (PADT) does not require poly(A) site cleavage (Zhang et al., 2015a). It was also detected that Pol II inhibitor *alpha*-amanitin which can inhibit the conformation change could also inhibit PADT. Meanwhile, the research on anti-termination factors is also important in understanding the mechanism of termination. SCAF4 and SCAF8 are two knowm anti-termination factors (Gregersen et al., 2019), both of which interact with Pol II and suppress pre-mature cleavage and termination (PCPA).

Another classical model for describing termination events is the torpedo model (Figure 2.2 B). When the Pol II reaches gene end, it first slows down over the terminator. The cleavage and polyadenylation (CPA) complex is recruited onto Pol II, helping to slow down the huge complex composed of Pol II (referred to as a 'juggernaut' in the paper by Proudfoot (2016)). The nascent RNA will often form an R-Loop with the DNA. During this period of time, mRNA is released from chromatin by CPA and eventually into cytoplasmic translation. After the release of the transcript, Pol II continues to transcribe. The transcription event would not stop until cleavage happens and the Pol-II-RNA of an exposed $5'$ end would be degraded by the $5'$-¿$3'$ exonuclease Xrn2 which would finally stop Pol II from transcription. This proposed mechanism is evocatively named the torpedo model where, in naval vernacular, Pol II is the battleship and Xrn2 the torpedo.

RNA Transport/Export After synthesis and processing, RNA will be released from the chromatin and transported out of the nucleus to the sites in the cell where it is in need by specific transport pathways. It is closely linked to the protein biosynthesis in the cell to which the transcription in the nuclear is one of the most significant controlling levels. In this phase, Pol II is not directly involved. During transcription, She2p is recruited to sites of active transcription binding to the elongating Pol II machinery and associates with nascent mRNAs. The recruitment depends on an interaction of She2p with transcription elongation factors Spt4 and Spt5 (Shen et al., 2010). Before nuclear export, the *ASH1* mRNA-She2p complex transits through the nucleolus. *ASH1* mRNA will pass through the nucleolus after transcription (Du et al., 2008).

RNA decay Journey of a message RNA ends in destruction of the transcript, known as RNA decay. The process of decay was at first viewed as an uncontrolled activity in the cell, but later confirmed to be highly regulated and of great importance. RNA decay determines the expression level of cellular RNA, limits the formation of aberrant transcripts through selective decay, removes byproduct of transcript including free introns. Most normal mRNAs are degraded through general default decay pathways. Initiated by de-adenylation, transcript with the Poly(A) tail removed will then be de-capped to carry on $5'$ to $3'$ degradation by Xrn1. In some cases, $3'$ to $5'$ degradation will happen independent of de-capping (John S.Jacobs Anderson, 1998). Interestingly, decay is promoter-influenced. The promoter swapping experiment of native upstream regulatory of RPL30 and ACT1 in yeasts resulted in huge change in the decay rate (Bregman et al., 2011), which also indicates that transcription and decay works in coordination. The nonsense-mediated decay (NMD) was the key to ensure accuracy of transcription (Kurosaki, 2019), so decay is also considered to be a quality control step of mRNA transcription.

2.1.2.2 Pre-mRNA Processing: from Pre-mRNA to mRNA

$5'$ Capping $5'$ capping is the very first step of mRNA processing, which happens at an early stage of transcription. Only Pol II transcripts are capped, mainly due to the existence of the unique CTD which recruits capping enzymes when Pol II CTD Ser2 was phosphorylated (McCracken et al., 1997). The reliability of the hypothesis was strengthened by the fact mRNAs produced by yeast Pol II mutants lacking an intact CTD are not efficiently capped (Cho et al., 1997). The process is largely influenced by the phosphorylation status of the CTD but finally completed with the actions by the capping enzyme.

Splicing Traditionally, it was believed that splicing happens after transcription, but more and more research proved that splicing happens co-transcriptionally at the majority of introns efficiently. With the cryo-electron microscope, structure of U1 snRNP-Pol II was revealed and a "Growing Intron Model" (Figure 2.3) was proposed based on the structure (Zhang et al., 2021). It is a model explaining the formation of the lariat RNA based.

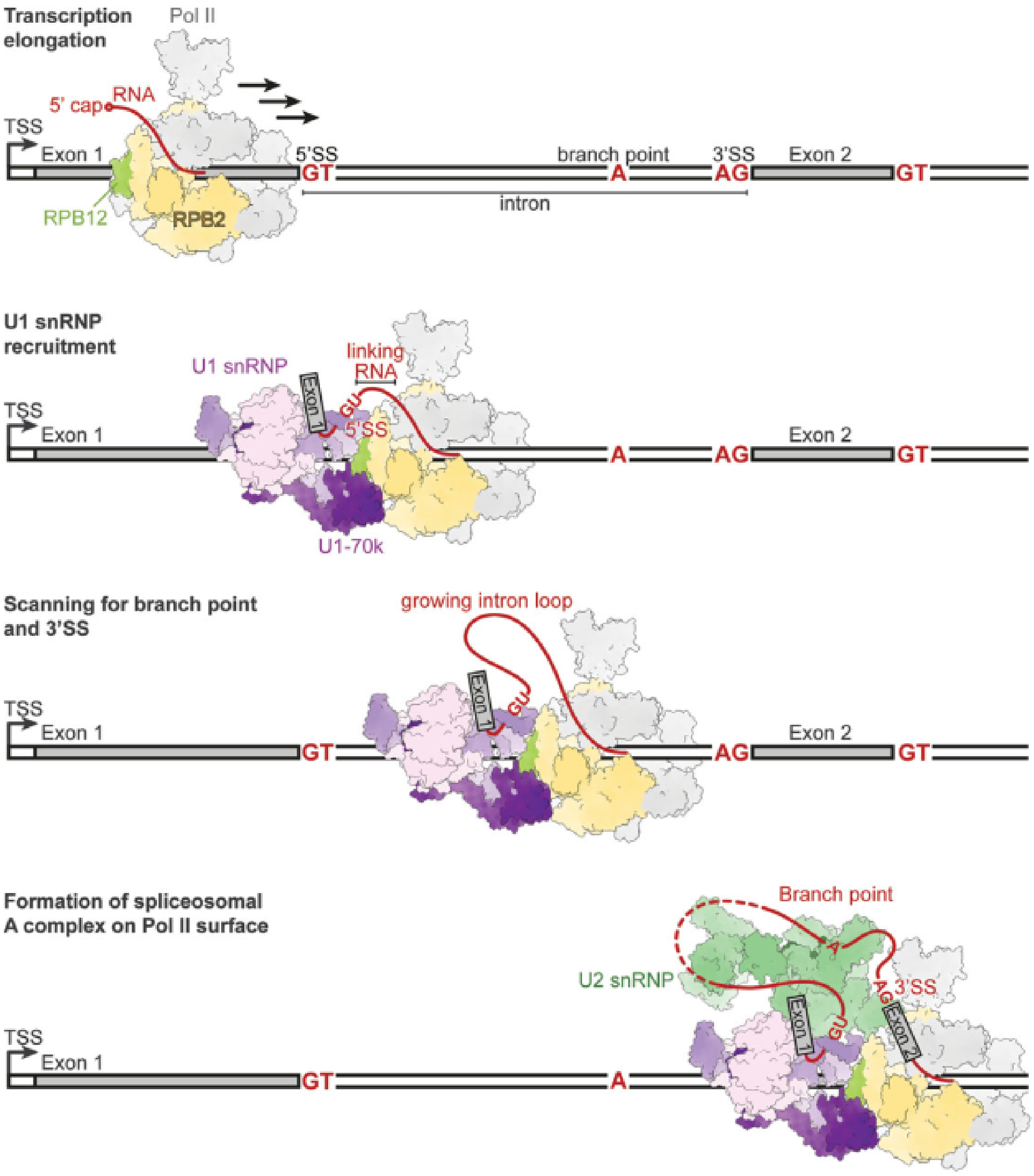


Figure 2.3: Growing intron loop model of co-transcriptional spliceosome assembly (Zhang et al., 2020). The intron is defined by the 5′ SS, branch point, and 3′ SS (conserved nucleotides in red). When the 5′ SS emerges in nascent pre-mRNA, U1 snRNP (purple) is recruited and directly binds Pol II subunits RPB2 (gold) and RPB12 (green). The 5′ SS is positioned near the existing site of transcription, so with transcription goes on a growing loop forms. When Pol II reaches the branch point, U2 snRNP would be recruited. The 5′ SS would finially attached with the branch point to form a lariat. Afterwords, the 5′ donor site would attack the 3′ acceptor site to release the intron.

Here shows a transcript under elongation process, of which the 5′ end has been capped. The huge complex of Pol II is moving forward fast on the DNA template to transcribe RNAs. The gene here contains an intron, defined by 5′ splicing site (5′ SS), branch point (BP) and 3′ splicing site (3′ SS). After 5′ SS transcription, the U1 snRNP is recruited. U1-snRNP is composed of U1-snRNA (also called the U1-RNA), the seven common core Sm proteins, and three U1-specific proteins (U1-70K, U1-A, and U1-C). The subunit U1-70K of U1 snRNP interact with RPB12 and RPB2 subunit of Pol II and the relative position between the two protein complexes remains unchanged. The 5′ SS is positioned near the existing site of RNA. One may imagine that Pol II is dragging U1 forward, while it continues to transcribe. With the extension of pre-mRNA, a growing intron loop forms between the 5′ SS and the RNA exit which are quite close spatially. When transcription reaches the branch point, U2 is recruited and the 5′ of intron connected to the branch point to form a lariat. Finally, the donor site will attack the acceptor site to finally liberate the intron. The U1 complex is conserved in mammalians but not in yeast, the introns of which are of smaller size to mammalians. It is highly suspicious that the splicing mechanism might be different between extremely long introns and short introns.

3′ Polyadenylation The poly(A) sequence at the 3′ end of the gene was detected in the 1970s (Lim and Canellakis, 1970; Edmonds et al., 1971). The long poly(A) tracts was not considered as DNA-templated and the idea was later confirmed by discovery of the poly(A) polymerase (Winters and Edmonds, 1973). The cleavage and polyadenylation events happen at the 3′ end of the transcripts as a symbol of the end of transcriptional termination stage, so a poly(A) tail can be considered as a mark of mature message RNA. The discovery of poly(A) tail was strong enough to alter the method of mRNA purification. The oligo(dT) priming on the poly(A) tail and the usage of oligo(dT) beads are still the most common ways of isolating mRNAs nowadays. With the purified RNAs, the common sequence of ‘AAUAAA’ close to the 3′ end poly(A) sequence was also detected (Hamilton et al., 2019; Tabaska and Zhang, 1999). Later known as the poly(A) signal, the ‘AAUAAA’ sequence upstream of the cleavage is conserved and shown in the upstream region in more the 50% of the poly(A) site in human and mouse.

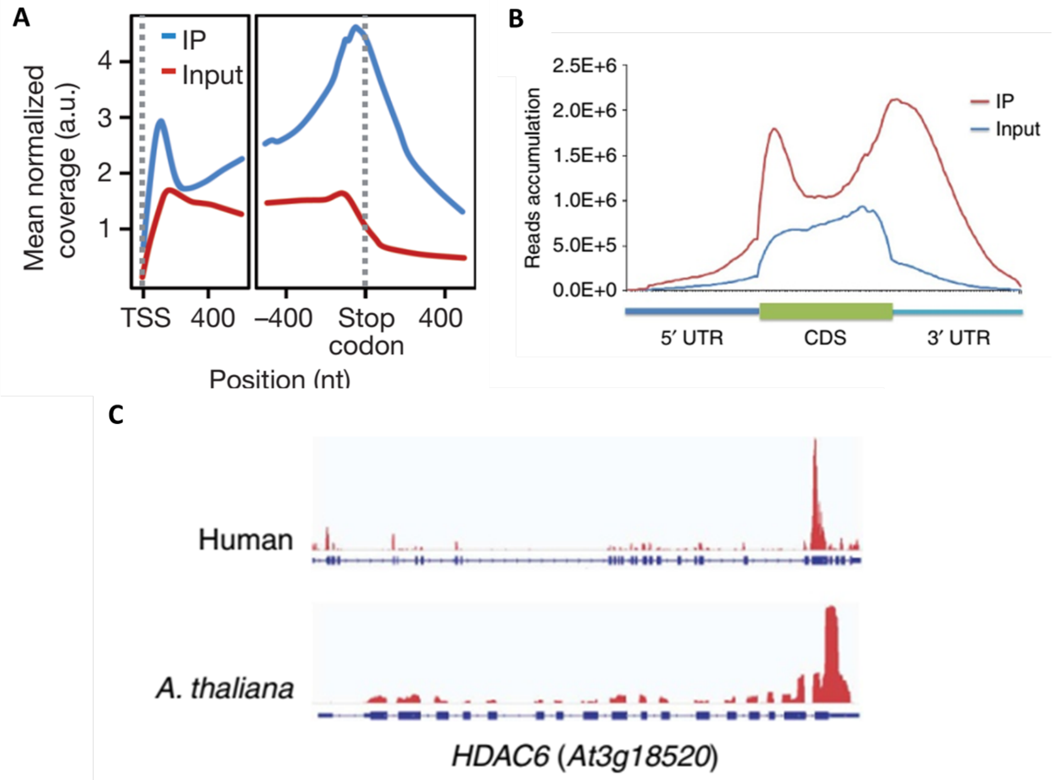
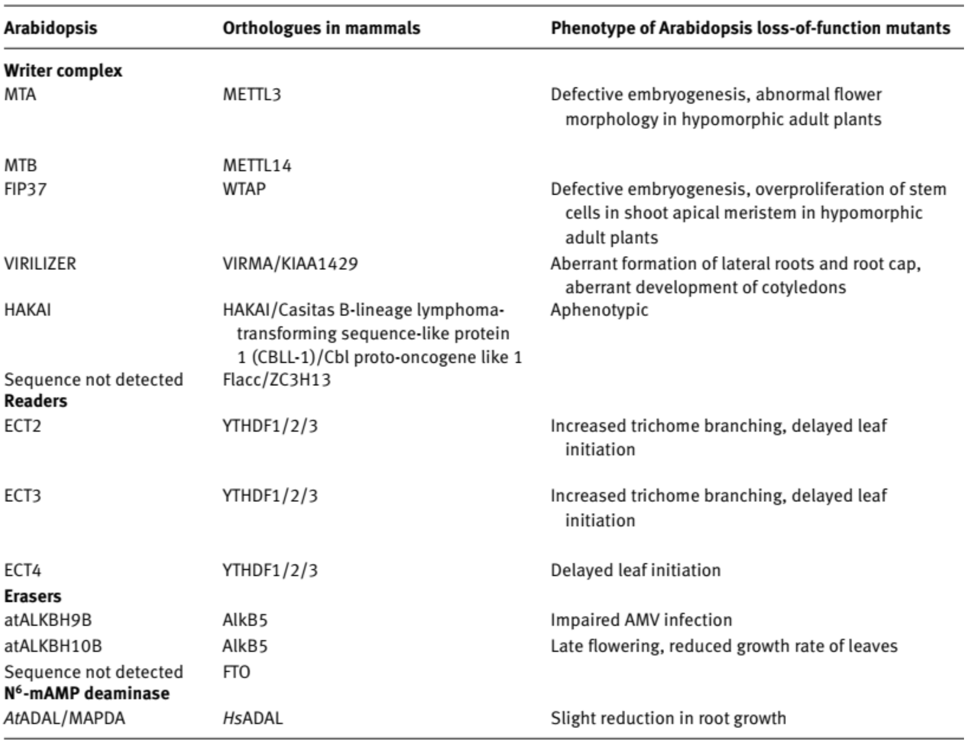


Figure 2.4: m^{6}A modification is conserved between human and Arabidopsis. A.Metagene profiles from Dominissini et al. (2012) depicting sequence coverage in windows surrounding the TSS (left) and stop codon (right); B. Accumulation of m^{6}A-IP reads along transcripts from Luo et al. (2014), C. An example of homologous genes with m^{6}A peaks conserved in human and *A. thaliana,* from Luo et al. (2014)

RNA modification N^6-methyladenosine (m^{6}A) was first detected in ploy(A) RNA fraction by Desrosiers et al. (1974). It is a significant and conserved modification found in both mRNAs and non-coding RNAs (ncRNAs) based on research of mammalian and plants. It is the most common internal mRNA modification, but the progress of deeper research on m^{6}A was delayed due to the lack of method in detecting m^{6}A sites. In 2012, Meyer et al. presented a new method to localize m^{6}A transcriptome-widely by combining m^{6}A-specific methylated RNA immunoprecipitation with next generation sequencing (MeRIP-seq). With this method, the researchers identified mRNAs of 7676 mammalian genes which contain m^{6}A modification. An enrichment of m^{6}A near stop codons and in $3'$ UTRs were detected (see Figure 2.4 A). In another paper published in the same year, a similar method called m^{6}A-seq (Dominissini et al., 2012) were presented. Besides the similar result of MeRIP-seq of enriched m^{6}A modification in near $3'$ stop codon, m^{6}A may also affect RNA splicing. Three RNA-binding proteins, YTHDF2, YTHDF3 and ELAVL1, may mediate novel connection between m^{6}A and cellular process. The case is similar in both animal (Kasowitz et al., 2018) and plants (Shen et al., 2016). Though little m^{6}A was detected at the intron regions of mRNAs, it has been confirmed before that m^{6}A modification resides in intronic region of pre-mRNA (Carroll et al., 1990). These results suggest that m^{6}A is likely to play a significant role in RNA metabolism. In plants, m^{6}A modification was detected in maize first (Nichols, 1979), and later in Arabidopsis(Zhong et al. (2008)). The first transcriptome-wide map of m^{6}A in Arabidopsis was obtained in 2014(Luo et al. (2014)), presenting that m^{6}A is highly conserved in *A. thaliana* mRNA. The topology of human and Arabidopsis m^{6}A RNA methylomes are similar, both showing enrichment near stop codon, but the peaks near $5'$ are different in position (see Figure 2.4 B,C). Just like other modifications, m^{6}A is deposited on RNAs by writers, removed by erasers and recognized by readers. These proteins are reserved between species, thus the m^{6}A writers, readers, and erasers in *Arabidopsis thalian* can find their orthologues in mammalian (See Table 2.1, reviewed by Reichel et al. (2019)).

Table 2.1: m^{6}A writers, readers and erasers in Arabidopsis reviewed by (Reichel et al., 2019)



Arabidopsis	Orthologues in mammals	Phenotype of Arabidopsis loss-of-function mutants
Writer complex		
MTA	METTL3	Defective embryogenesis, abnormal flower morphology in hypomorphic adult plants
MTB	METTL14	
FIP37	WTAP	Defective embryogenesis, overproliferation of stem cells in shoot apical meristem in hypomorphic adult plants
VIRILIZER	VIRMA/KIAA1429	Aberrant formation of lateral roots and root cap, aberrant development of cotyledons
HAKAI	HAKAI/Casitas B-lineage lymphoma-transforming sequence-like protein 1 (CBLL-1)/Cbl proto-oncogene like 1	Aphenotypic
Sequence not detected	Flacc/ZC3H13	
Readers		
ECT2	YTHDF1/2/3	Increased trichome branching, delayed leaf initiation
ECT3	YTHDF1/2/3	Increased trichome branching, delayed leaf initiation
ECT4	YTHDF1/2/3	Delayed leaf initiation
Erasers		
atALKBH9B	AlkB5	Impaired AMV infection
atALKBH10B	AlkB5	Late flowering, reduced growth rate of leaves
Sequence not detected	FTO	
N^6-mAMP deaminase		
*At*ADAL/MAPDA	*Hs*ADAL	Slight reduction in root growth

Loss of core components FKBP12 INTERACTING PROTEIN 37 KD (FIP37, Shen et al. (2016)) and MRNAADENOSINE METHYLASE (MTA, Zhong et al. (2008)) of the m^{6}A methyltransferase complex are lethal in Arabidopsis, indicating that m^{6}A may be indispensable for the successful reproduction or embryo development of plants. They have currently known to affect the shoot stem cell fate and inhibits local ribonucleolytic cleavage to stabilize mRNAs.

2.1.2.3 Functional Coupling between Co-transcriptional Splicing and Termination

One of the most important complex which functions in the coupling of splicing and termination is U1 snRNP. As part of the spliceosome, U1 snRNP can be recruited by the Pol II complex near 5′ splicing site and hold the newly transcribed RNA sequence resulting in the formation of splicing lariat. In the mutant of Arabidopsis LUC7, a U1 snRNP subunit, large number of alternative splicing events comparing to the wild type can be detected (Amorim et al., 2018). Yet, the complex also plays an important role in suppress of pre-mature termination. U1 snRNP knockdown results in increase in premature cleavage and polyadenylation, which frequently happen in introns near (5 kilobases) the start of the transcript (Kaida et al., 2010). By protecting RNAs from premature termination, U1 snRNP regulates the expression and length of mRNAs (Berg et al., 2012). Besides termination, U1 snRNP is also involved in TFIIH-dependent transcription initiation (Kwek et al., 2002). By long-read sequencing on nascent RNAs, it was also revealed that Co-transcriptional splicing efficiency influences $3'$ end cleavage efficiency (Reimer et al., 2021).

2.1.3 Difference in Transcription of Plants and Animals

While most of the proteins in charge of the transcription process are conserved, there still exist difference between species. Many pioneer works of this research was not carried on plants. It is worth looking into the difference in the current discovery in difference of transcription in plants and animals to see to what extent we can refer to the result in research in animals and whether we can apply the previous method in animal research on plants.

2.1.3.1 Differences in Pre-mRNA Splicing of Animals and Plants

As most of the works on pre-mRNA splicing were carried out on animals, there remains a large gap concerning relevant works in plants. Accumulating evidence shows plant and animals differ in many ways of RNA splicing.

Early research (Fedorov et al., 2002) shows that even though we cannot establish any differences based on the sequences, we could still find that some introns have the same position in separate species. The final estimate is that around 14% of animal introns match intron positions in plants, while the 30 genes with the highest match numbers have a 40% match. These results indicate that some part of the introns may be ancestral, existing before the separation into animal and plants.

Besides these similarities, we could find many differences between plants and animals. First and most obviously, the size of introns in the genome. Generally, mammals have longer introns than non-mammal animals, fungi and plants, for example, humans have the longest introns with an average length of 3.4k bps among the ten species studied (Long and Deutsch, 1999). The number is much smaller in plants, with 152 bps per gene in Arabidopsis and 387 in rice (Ren et al., 2006).

Besides, plants have a lower proportion of genes showing alternative splicing (AS) than animals. Interestingly, intron-retention is the dominant form of AS in many plants, contrasting with exon-skipping being dominant in animals. In intron retention, the intron remains in the mature RNA after splicing (Zhang et al., 2014). The intron retention rate ranges from ˜30%-64.1% in Arabidopsis but with a much lower rate of ˜5%-15% in fruit fly and human cells (Boothby et al., 2013). Therefore, it is possible that plants could produce functional proteins rapidly by fast and constitutive splicing when necessary, questioning the importance of alternative splicing in plants.

2.1.3.2 Difference in Poly(A) Signals in Plant and animals

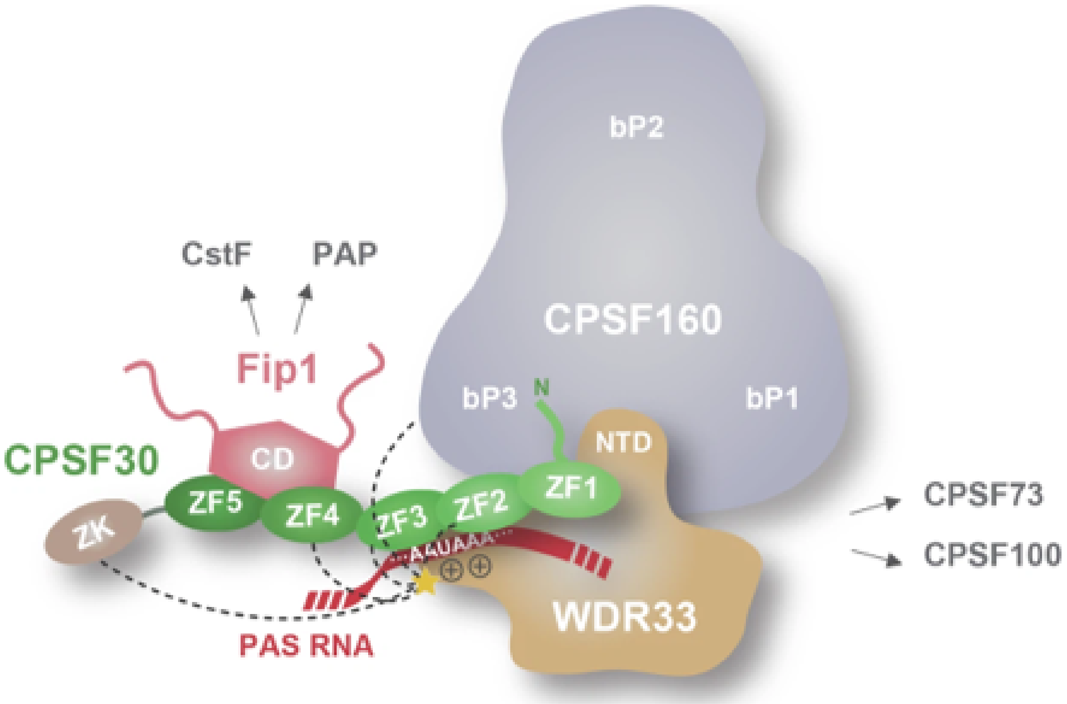


Figure 2.5: Schematic model of the mammalian core CPSF complex bound to the AS hexamer motif. Proteins involoved in this process including CPSF160, WDR33, CPSF30 and Fip1. This figure was summrized by Clerici et al. (2017).

The conserved poly(A) signal 'AAUAAA' was discovered in 1970s (Proudfoot and Brownlee, 1974). This hexamer was later detected in the upstream region of over 50% of polyadenylation sites. The proteins work to recognize the poly(A) signals were also discovered (Clerici et al., 2017) (see Figure 2.5). All these proteins are conserved between species (WDR33 was homolog of FY in human) some of them have been confirmed to recognize the 'AAUAAA' signal in Arabidopsis as well. But the chimeric genes carrying poly(A) signals from human genes introduced onto tobacco cells cannot be properly polyadenylated, indicating that the polyadenylation signal of plants might differ from those of human (Hunt et al., 1987). Meanwhile, only 10% of the PAS have AAUAAA-like signal in the upstream region was detected in Arabidopsis. Instead, it is possible that UGUA-like element, rather than

'AAUAAA', might serve as a possible signal of cleavage and polyadenylation events (Ye et al., 2021). There might exist some plant specific 3' processing factors and FCA might be one of them.

2.1.3.3 Comparison between Arabidopsis and Human Genome

The difference in transcription might be originated from the difference in the genomes themselves. Here, we compared genome of Arabidopsis and Human based on the annotation files (see Table 2.2). Human genome is larger in size compared to Arabidopsis, with the total number of bases over 50 times of the one of Arabidopsis. While the human genome seems to have much larger transcripts and introns comparing with Arabidopsis, we are surprised to see that the mean and median values of the exon size in the two groups are close. These differences might be the origination of some distinct result in the two species. For example, the most abundant form of alternative splicing in human is exon skipping, while in Arabidopsis, intron retention is more common. The large size of introns seems to be spliced more efficiently in plants.

Table 2.2: Comparison between Genome of Arabidopsis and Human

Feature		Arabidopsis (TAIR10)	Human (GRCh38)
Exon	Mean	330	376
	Median	164	155
Intron	Mean	171	9468
	Median	100	1987
Transcript	Mean	2212	37843
	Median	1921	10796

2.2 RNA-Binding Proteins

RNA-binding proteins (RBPs) are the key players in the regulation of gene expression, involved in almost every essential steps of transcription. RBPs contains RNA-binding Domains (RBDs), recognising different targets including single strand RNAs, double strand RNA and specific three-dimension structure of folded RNAs. In this section, we reviewed the role of RBPs in a few biological processes of RNA transcription, as well as the current methods in detecting the function of RBPs.

2.2.1 Introduction to RNA-Binding Proteins

The role of RBPs is essential in the whole life-span of an mRNA. For instance, RBPs can function in transcription, by modulating RNA polymerases basal activities, by providing specificity to gene transcription regulation, or by terminating transcript through the recognition of the polyadenylation signal on the nascent mRNA (Re et al., 2014).

RBPs bind to RNAs through the specific RBDs. RBPs are a diverse class of proteins defined by their ability to interact with RNA molecules. In the Arabidopsis genome, more than 800 RBPs have been identified based on sequence homology to RBPs known in other eukaryotes, including RBPs predicted to locate in organelles. Traditionally, RBPs are characterized by RBDs interacting with single-stranded or double-stranded RNA. The most-abundant RBD is the RNA-recognition motif (RRM), present in 197 Arabidopsis proteins. There are many other well-studied RBDs including KH domain, PUF, Zinc-Finger.

2.2.2 Capture of RNA Interactome

Research to elucidate RNA-RBP interaction can be divided into two kinds: RNA-centric and Protein-centric(see Figure 2.6). In RNA-centric approaches, proteins associated with RNAs are recovered by RNA pull-down methods and subsequently identified by Mass Spectrometry (MS) or Western blot. In protein-centric approaches, RNAs associated with RBPs are identified in cell lysates via immunoprecipitation of the RBPs and RT-qPCR or RNA-seq.

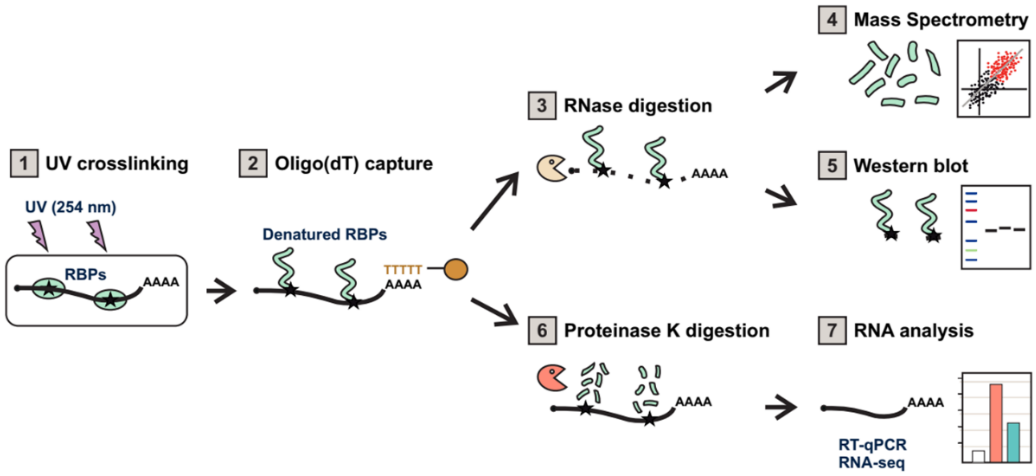


Figure 2.6: RNA interactome capture (Bach-Pages et al., 2017): (1) Cells or tissues are irradiated with UV light at 254 nm in order to promote crosslinking and form covalent bonds between interacting RNAs and proteins. (2) After cells lysis, mRNAs are pulled-down using oligo(dT) magnetic beads. The RNA-protein complexes are recovered and can be analysed using different techniques after stringent washes. Firstly, RNA can be enzymatically digested (3) and the proteins quantitatively analysed by Mass Spectrometry (4) or Western blotting (5). Alternatively, the protein fraction can be enzymatically digested (6) and the RNA analysed by RT-qPCR or RNA sequencing (7).

Protein-centric research of RNA-RBP interaction in Arabidopsis has been carried on leaf mesophyll protoplasts (Zhang et al., 2016), etiolated seedlings (Reichel et al., 2016), 4-week-old leaves (Marondedze et al., 2016). With the combination of in vivo UV-crosslinking of RNA to RBPs, oligo(dT) capture and mass spectrometry, the proteins binding with RNAs were detected. Not only RBPs with known RNA-binding domains but many novel RBPs were detected in the three experiments, but the validation of the discovery remains to be confirmed. The existence of tissue-specific RBPs was also revealed, as some know RBP was not detected in Arabidopsis leaves. The RBPs detected into different categories by Gene Ontology (GO) analysis¬(Zhang et al., 2016), supporting the significant role of RBPs in transcription metabolism.

The RNA-centric experiment of CLIP-seq was applied in this study and the method would be introduced in detail in Chapter 3.

2.3 Autonomous Flowering Pathway

Time of flowering is critical to the reproduction of plants. The rapid flowering of Arabidopsis is repressed by the floral repressor *FLOWERING LOCUS C (FLC)* through the activity of a group of proteins which belong to the autonomous pathway. Back in the 1980s, koornneef et al.((1982; 1991)) performed a genetic and physiological analysis of a series of late flowering mutants in *Arabidopsis thaliana*. In this seminal work, the researchers classified eleven different late-flowering mutants according to their physiological responses to day length and the long period of cold required to accelerate flowering which is known as vernalization. Among all the mutants, flowering time of the four, *fca*, *fve*, *fy* and *fpa*, appeared largely affected both under long day (LD) and short day (SD). These mutants were later grouped into the autonomous pathway (Levy and Dean, 1998), indicating their flowering time is independent of environmental cues (particularly day length in this case). Below, we will focus on the three genes of the pathway, *FCA*, *FPA* and *FY*, which have been found to affect the flowering process by cleavage and alternative polyadenylation of the long non-coding RNA *COOLAIR*.

2.3.1 Flowering Control Locus A (FCA)

FCA is considered to be a strong positive regulator of flowering by repression of *FLC* expression. The *FCA* gene was recovered to encode a protein with two RNA Recognition Motifs (RRMs), a WW domain which contributes to protein-protein interaction (Macknight et al., 1997) and two Prion-like Domains (PrLDs) by prediction later in 2016 (Chakrabortee et al., 2016). Besides, FCA is nuclear localized (Quesada et al., 2003) and exhibits the ability to form nuclear bodies (Fang et al., 2019).

The full-length FCA protein has a size of 747 amino acids. The two RRMs locates between amino acid residues 118-199 and 209-289 shows similarity to the RRMs of ELAV in fruit fly (*Drosophila melanogaster*) and its homologs in mammalians, the ELAV-like family. In the test of RNA-binding ability of FCA, it was found to bound to poly(G) and poly(U) sequences, but not to poly(A), poly(C) or DNAs (whether double-stranded or single-stranded) (Macknight et al., 1997). Assuming that the RNA-binding feature of FCA mainly originated from RRM domains, we should expect to observe similar binding preference in ELAVL proteins. Taking the probably best-known ELAVL protein Human antigen R (HuR) for example, it is a protein constituted by three RRMs (Colombrita et al., 2013) with two RRMs (RRM1 and RR2) at the N-terminus which are important for binding of target RNA sequence and one (RRM3) in the C-terminal separated from the other two. The two N-terminal RRMs of FCA show similarity with RRM1 and RRM2 of HuR. Previous research of HuR revealed its binding to poly(U) sequence and AU-rich elements in the $3'$ UTR region, so we may expect a similar binding feature in vivo of FCA. But the subcellar localizations of FCA and HuR are different, with FCA only located in nucleus especially in nuclear bodies (Fang et al., 2019) while localization of HuR in the cytoplasm was related to mRNA stabilization (Tran et al., 2003). The two might have a similar preference in binding AU-rich sequence but rather distinct in functions judging from the difference in subcellar localization.

The WW domain, located between amino acid 597 and 622, have the potential to interact with other protein, probably with a protein containing proline-rich sequence. The WW domain is required for auto-regulation of FCA. A transgenic-line with FCA-WF, of which a W to F within the WW domain, expressed in *fca-1*, lost feedback autoregulation (Simpson et al., 2003). Meanwhile, the late flowering genotype in *fca-1* cannot be rescued by FCA-WF but by FCA-WW. On the other hand, it was revealed that FY interacts with the WW domain of FCA through its proline-rich PPLPP-motifs (Simpson et al., 2003).

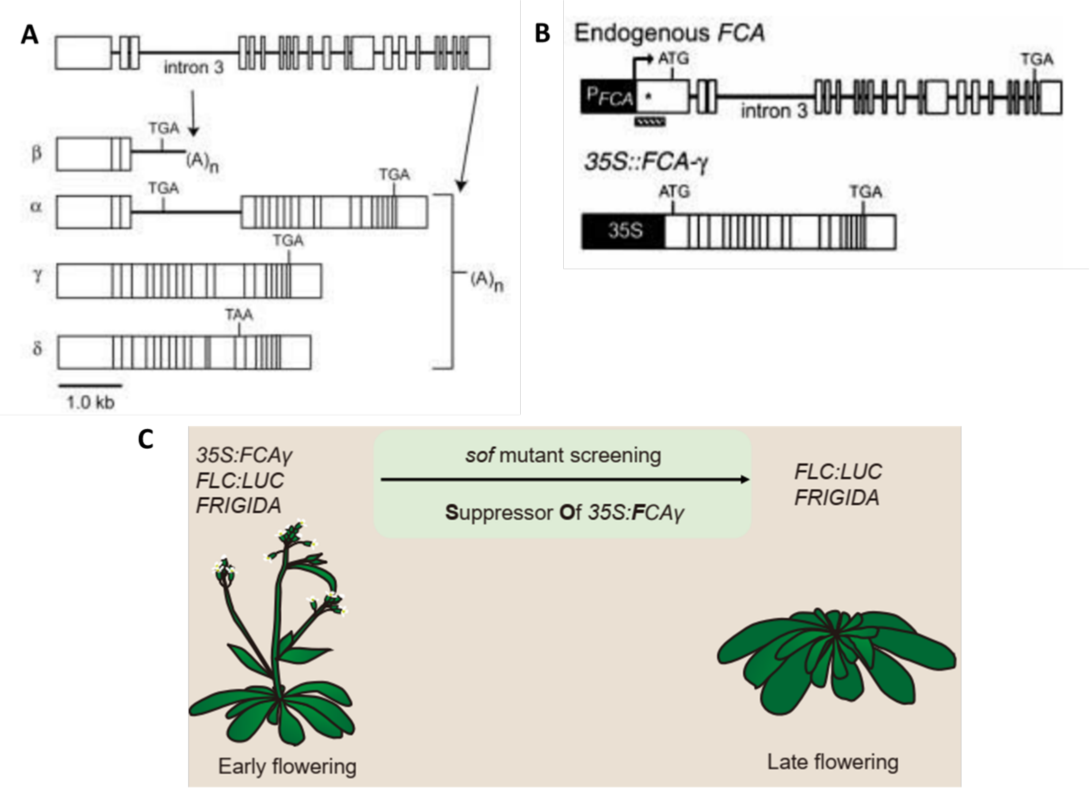


Figure 2.7: Different isoforms of *FCA* transcripts A. Representation of alternative processing of FCA pre-mRNA, Exons are represented as filled box and introns as lines (Quesada et al., 2003); B. Difference between endogenous FCA and the *35S::FCA-*γ (Quesada et al., 2003); C. Principle of *sof* mutant screening. FCA overexpression leads to early flowering in the presence of active FRIGIDA (FRI). Mutants suppressing the effect of *35S::FCA-*γ are late flowering (Wu et al., 2019)

In transcription, FCA transcripts are spliced and terminated alternatively. Four forms of *FCA* transcripts were detected (Macknight et al., 1997),see Figure 2.7 A. Of all the four transcripts, only *FCA-*γ encodes functional protein after translation. The shortest of the four, *FCA-*β, is the most abundant form (˜55%) of transcription in the wild type, which pre-terminated in the third (and longest) intron of *FCA*. Though *FCA-*β displays the highest level, even in the *35S::FCA* transgenic line of overexpressed *FCA* with an acceleration in flowering time, the line overexpressing *FCA-*β at *fca-1* background showed same flowering time with the *fca-1* mutant, indicating the inability of the transcript to promote flowering. The *fca-1* mutation caused premature termination just upstream of the WW domain (Simpson et al., 2003), suggesting that the WW domain is significant for the function of FCA. The *FCA-δ* with an unspliced intron 13 have the two RRMs preserved but lost the WW domain. The *FCA-α* is the longest but of the lowest level. The RRMs are affected in *FCA-α* by the retained intron 3. The changes in the sequence of functional protein domain regions in these transcripts could also result in the loss of function of the proteins they encode. While it seems unreasonable for a 'nonfunctional-protein-coding' transcript to be abundant, the *FCA-β* was later confirmed to play an important role in the regulation of *FCA* expression (Quesada et al., 2003). In the process, referred to as auto-regulation of *FCA*, *FCA* negatively regulate its own expression by promoting the usage of proximal poly(A) site located in intron 3 to generate the isoform of *FCA-β* instead of the functional *FCA-γ*.

FCA-mediated repression of *FLC* was studied through genetic screening of suppressors of over-expressed FCA (*sof*) (Liu et al., 2010). In this line, the functional FCA-γ isoform was overexpressed by *CaMV 35S* promoter under Col-*FRI* background (with activate *FRIGDIDA*, see Figure 2.7 B,C). The *35S::FCA-γ FRIGDIDA* represent early flowering phenotype and the mutations found to suppress the expression of *35S:FCAγ* would be considered to disrupt the function of FCA. Interestingly, both *fpa* and *fy* were recovered in the screening for several times (Liu et al., 2010), indicating they are required for FCA to repress *FLC*. Other factors involved in RNA transcriptions were also identified as *sof* mutants including $3'$ processing factors of CstF64 and CstF77 (Liu et al., 2010), a component of spliceosome (PRP8, Marquardt et al. (2014)) and an elongation factor Arabidopsis cyclin-dependent kinase C (CDKC;2, Wang et al. (2014)). Through the research above, it was discovered that the expression level of *FLC* was affected by the expression of its antisense non-coding transcript *COOLAIR* (Swiezewski et al., 2009).

As for the PrLDs, in the research of a prion-like protein FUS (Patel et al., 2015), it has been recovered that liquid compartment formation is dependent on PrLD. Later, Fang et al. tested whether FCA can form nuclear body of liquid-liquid phase separation. It has been confirmed that FCA phase-separate in vitro and shows features corresponding with phase separation in vivo, e.g., dynamicity of FCA body showed in fluorescence recovery after photobleaching (FRAP). It has also been recovered that FLL2, the mutant of which was also detected as *sof* mutant before, plays an essential role in the formation of FCA body. Taken together, FCA is a plant-specific RNA-binding protein with the ability to form nuclear bodies. It works in the nuclear and dynamically interacts with many other proteins to regulate the expression of genes including *FLC* and its own. It works closely with a series of proteins detected by forward genetic screening, including but not limited to 3′ processing factors, a member of spliceosome and an elongation factor. Currently, it is considered as a 3′ processing factor, but the function of FCA might not be limited to 3′ end only.

2.3.2 FLOWERING LOCUS PA (FPA)

FPA gene is composed of six exons that encode a 901–amino acid residue protein. *FPA* is expressed most strongly in developing tissues (Schomburg et al., 2001) like *FCA*. The predicted FPA protein contains three RRMs indicating that FPA may function as an RNA-binding protein. Unlike *35S::FCA*, overexpression of FPA causes precocious flowering in noninductive short days (Schomburg et al., 2001). Besides regulation of *FLC* expression, FPA is also believed to repress the expression of intergenic regions. Sonmez et al. (2011) characterized a subset of mis-expressed unannotated (UA) segment in double mutant *fca-9 fpa-7* with whole-genome tiling arrays (see an example in Figure 2.8). Similar result was also detected by (Duc et al., 2013) and the gene with extended transctiption to the downstream gene was named as ”Chimeric Transcripts”.

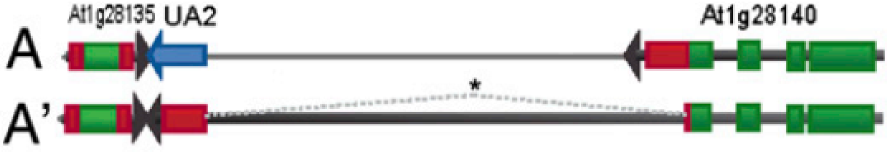


Figure 2.8: An example of unannotated segment with novel intronic region from (Duc et al., 2013). The expression level of intergenic region between AT1G28410 and AT1G28135 was raised in *fpa*. The transcription of upstream gene (AT1G28410) was affected by loss of FPA, extending to UA2 and resulting in splicing of a novel intron.

The majority of UA segments were mapped to the $3'$ ends of annotated genes. Not all the UA seemed to be explained by transcriptional read-through at first glance. In some cases, this should be attributed to the existence of novel, large introns within UA segments. FCA and FPA are involved in the interplay of $3'$ processing and chromatin changes. Since it has been revealed before that FCA and FPA promote asymmetric DNA methylation at some loci (Bäurle et al., 2007), an investigation in DNA methylation level was also carried out on UA segments. As the result, a significant increase in CG level within the UA segment of At1g55800 in *fca fpa*, a two-fold increase in CHH, and no change in CNG. While the CHH methylation within *AtSN1*, a gene with upregulated expression in *fpa* results from upstream read-through transcripts, increased in *fca* and *fca fpa* but not in *fpa*. These facts indicates that FCA and FPA might have altered impact of DNA methylation at different loci.

Later in 2013, Direct RNA Sequencing (DRS) was used to define the genome-wide pattern of cleavage and polyadenylation in *fpa* and found that FPA affects intronic cleavage site selection and intergenic read-through region (Duc et al., 2013). Moreover, a chimeric transcript of *PIF5–PA03* RNAs forms in *fpa-9* and the formation of this chimeric RNA does not depend on *FLC*. It was also detected that defective termination in *fpa* occurs at the same loci in *dicer-like 4* (*dcl4*), while expression of *DCL4* itself was unaffected in the absence of FPA, indicating that the two may share some common target regions (Duc et al., 2013). DCL4 is an RNase III-like protein which functions in processing small RNAs. This might indicate that FPA was involved in DCL-mediated silencing. Shifts in alternative polyadenylation at BONSAI Methylation 1 (IBM1), a gene encoding a histone demethylase specific for H3K9, was detected in *fpa-7*. FPA might control the expression of this gene by alternative polyadenylation. FPA was also related to the processing of mRNAs bearing heterochromatic marks, as *fpa* was screened as a suppressor of *ibm2* (Deremetz et al., 2019). Connection between FPA and H3K9me was shown in Figure 2.9. FPA might affect the self-reinforcing loop through antagonistical interaction with IBM2.

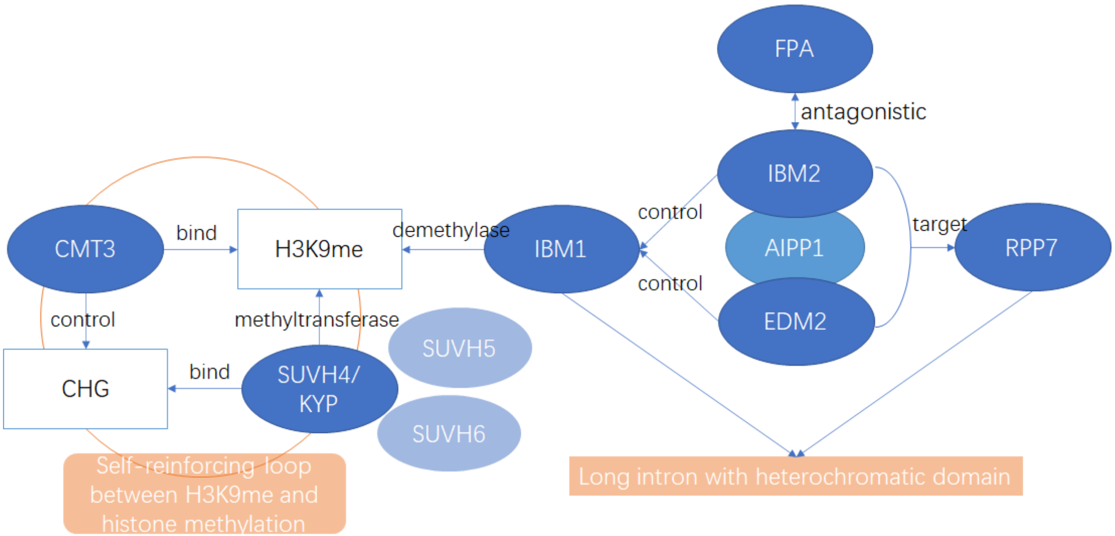


Figure 2.9: FPA antagonistically interact with histone demethylase IBM2. IBM2 was found to be in the same protein complex with ENHANCED DOWNY MILDEW2 (EDM2) and ASI1-IMMUNOPRECIPITATED PROTEIN1 (AIPP1). The three proteins share common targets of *IBM1* and *RPP7*, each contains a long intron with a heterochromatic domain, associated with H3K9me and DNA methylated. FPA was identified as a suppressor of *ibm2* mutation, promoting the usage of proximal poly(A) sites in genes targeted by IBM2. Through antagonistical interaction between IBM2, FPA might affect the self-reinforcing loop between H3K9me and histone methylation. This figure was summarized by the author base on Deremetz et al. (2019)

2.3.3 FLOWERING LOCUS Y (FY)

FY belongs to a highly conserved group of eukaryotic proteins represented in budding yeast (*Saccharomyces cerevisiae*) by the RNA 3′ end-processing factor, Pfs2p. The very typical structure of FY and its homologs are the seven WD40 domains. Unlike its homologs, FY in Arabidopsis has a disordered region in C-terminal with two Pro-rich motifs which interact with the plant specific RBP FCA.

The function of FY is so essential in plant that the null allele of FY is lethal. Therefore, mutants available nowadays lost only part but not all of FY function. The four *fy* mutants, along with the protein structure of each mutant after the mutation, were reviewed by Yu et al. (2019) (See Figure 2.10 A,B):

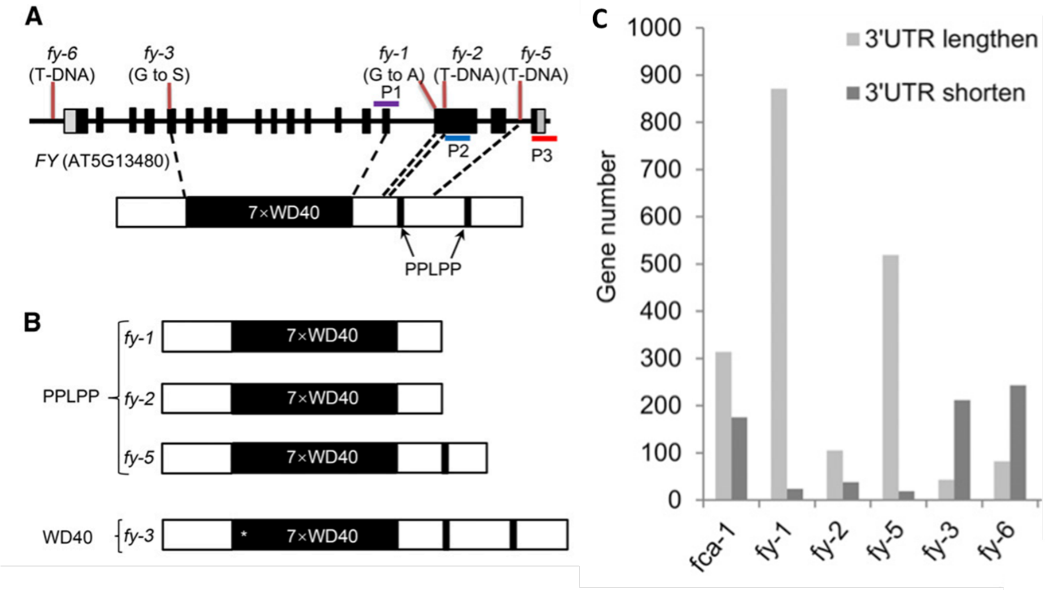


Figure 2.10: Different *fy* mutants and 3′ UTR APA Analysis of these mutants (Yu et al., 2019) A.Structure of *FY* gene (top) and FY protein (bottom); B: Schema of FY in mutants: *fy-1* and *fy-2* loss both two PPLPP motif, *fy-5* lost only PPLPP motif for T-DNA insertion, and *fy-3* has a point mutation on the WD40 domain; C. Comparison of 3′ UTR significantly lengthen or shorten genes

One of the interesting things about *fy* mutants is that the seriousness in the flowering phenotype does not correspond to the defect in termination (See Figure 2.10 C). The *fy-2* is the one which flowering the latest either in long day or short day, with or without vernalization. While *fy-3* is the one of the most altered poly(A) site usage of Col-0 ecotype. This might indicate that the seven WD domains are mainly responsible for 3′ end processing while the plant-specific PPLPP motifs are connected to control of flowering together with other components of the autonomous pathway. The *fy-2* mutant we used in this project has a T-DNA insertion in the exon 16 before both the PPLPPs and might result in a loss of ability to interact with FCA. The loss of FCA-FY interaction would cause a defect in the auto-regulation of FCA.

On the other hand, FY was grouped into different sub-pathway with FPA. The combination of *fy* and *fpa* mutation is lethal in *Ler* (but not lethal in Col-0), suggesting their possible redundancy in function. FY is a protein partner of the WW domain of FCA. The loss of

FY-FCA interaction by a change of the PPLPP motif in the C-terminal of FY would cause the loss of FCA auto-regulation as well. Besides FCA, FY also work together with CPSF proteins, including CPSF73-I, CPSF73-II, CPSF100, etc.

In this research targeted the autonomous pathway, we use only *fy-2* allele which displays the most severe late flowering phenotype.

3 High-Throughput Sequencing (HTS) Data Analysis

In this thesis, we mainly focus on transcription events, especially termination and $3'$ processing. To this end, HTS data are utilized to check the binding sites of RBPs and to study the differences between mutant and wild type at the RNA level.

Data used in this research can be generally divided into the following three parts according to the targeted RNAs:

- The first and most common target is messenger RNA (mRNA), spliced and polyadenylated, usually enriched with oligo(dT) beads or oligo(dT) priming. Understanding the expression level of mRNA is very important, because a significant change in mRNA level usually indicates a similar change in protein level, which in some cases directly results in phenotype changes detected in the mutant. Therefore, the detection of the changes in mRNA level is usually a good starting point for transcriptome-wide research. In addition to expression levels, the results of RNA processing steps such as alternative splicing (AS) and alternative polyadenylation (APA) are also detectable on mRNAs.
- The second target is precursor message RNA (pre-mRNA), which refers to RNA under the transcription process. They can be captured through their link with chromatin through RNA Polymerase II (Pol II). These pre-mRNAs are attached to the chromatin through Pol II, so they are also known as chromatin-associated or chromatin-bound RNAs. There also exist additional methods to measure them:

  - Application of Pol II antibody to enrich RNAs under transcription such as native elongating transcript sequencing (NET-seq). Furthermore, with antibodies for Pol II with differentially phosphorylated CTDs, it is possible to detect RNAs under different stage of transcription;
  - Separation of DNA template attached with Pol II from the complex such as Global run-on sequencing (GRO-seq).

  The separated DNA with Pol II will then be put into environment with radio-labelled bases to check the newly produced sequence. These data can be used to check the RNA processing steps including co-transcriptional splicing and Pol II pausing, which cannot be directly detected at the mRNA level.

- The last but not the least important target is protein-bound RNAs or RNAs with certain modification. In this case, typically only the part near the protein or subject to modification will be retained and amplified by PCR for later research. These experiments usually need strict condition for purification, more steps than common mRNA-seq and more risk of RNA decay before reverse transcription to cDNA.

The sequencing method in this project is short-read sequencing with Illumina Xten platform, paired-end for 150 bps on each direction (PE150).
The Chapter 3 has been divided into three sections by the three targets mentioned before. In each section, one or two main sequencing would be introduced with both library construction method and data analysis method.

- The first section is about sequencing methods for mRNAs, including mRNA-seq and $3'$mRNA-seq. We picked mRNA as the first section not only for the widely-use of mRNA-seq method but also for introduction of general-utility tools for mRNA-seq and other types of RNA-seq data at first. In this part, we also aimed to explain the difference between strand-specific and non-stranded mRNA-seq (see Figure 3.1 for comparison) to highlight the significance of strandness in RNA-sequencing. An mRNA-seq workflow will also be included. As for the $3'$mRNA part, a recommended analysis pipeline would be included.

- The second section is about Chromatin-Bound RNAs (CB-RNAs). The two main sequencing methods in this section are CB-RNA-seq and pNET-seq. In this research, only CB-RNA-seq data was actually used to calculate the efficiency of co-transcription splicing. We had some trial on pNET-seq analysis, but no data of satisfying quality has been produced yet.
- The last and the most important section is about protein-bound RNAs. The development of CLIP-seq is included in this section. CLIP-seq method has been improved by many researchers, making it an efficient and accurate method for RNA-protein interaction research. Though widely applied to mammalian cells, it has not yet been a major method in plant research. The difficulty of CLIP-seq application in plant would also be mentioned. As to the data analysis part, some tools especially designed for CLIP data analysis would be introduced.

3.1 Sequencing Methods for mRNAs

3.1.1 mRNA-seq

Traditional ways of RNA profiling include hybridization-based approaches like microarrays, and tag-based sequencing approaches such as cDNA-AFLP (Bachem et al. (1998)), serial analysis of gene expression (SAGE) (Velculescu et al. (1995)), and massively parallel signature sequencing (MPSS) (Crawford et al. (2006)). However, with the development of NGS, RNA-sequencing has become one of the most powerful techniques available. RNA-Seq not only helps us to identify new genes and alternative splicing but also allows us to compare the transcriptomes under different conditions (Pollier et al. (2013)), such as different developmental stages.

The materials used in this research are 10-day-old Arabidopsis seedlings, with Col-0 ecotype as the WT control. The total RNAs from 100 mg of ground seedlings were prepared with a DNase-treatment to avoid the impact of remain DNAs. In this research, we applied mRNA-seq to get differential expressed genes between wild type and *fca-9* mutant.

3.1.1.1 Library Construction

The construction of a cDNA library can be generally conducted by the following three steps:

1. Reverse transcription of the target RNA in to cDNAs, which is also known as first strand synthesis;
2. Generation of the complementary strand of the cDNA, also called second strand synthesis;
3. (optional) PCR amplification of cDNAs to ensure sufficient concentration for sequencing.

The actual steps of cDNA library construction are more complicated with purification, adaptor ligation steps, but the basic of library construction is to get cDNA with sequencing adaptors using target RNA fragments as templates. The methods can be divided into non-stranded methods and strand-specific methods according to whether the strand information (which strand of the DNA was used as template) can be kept.

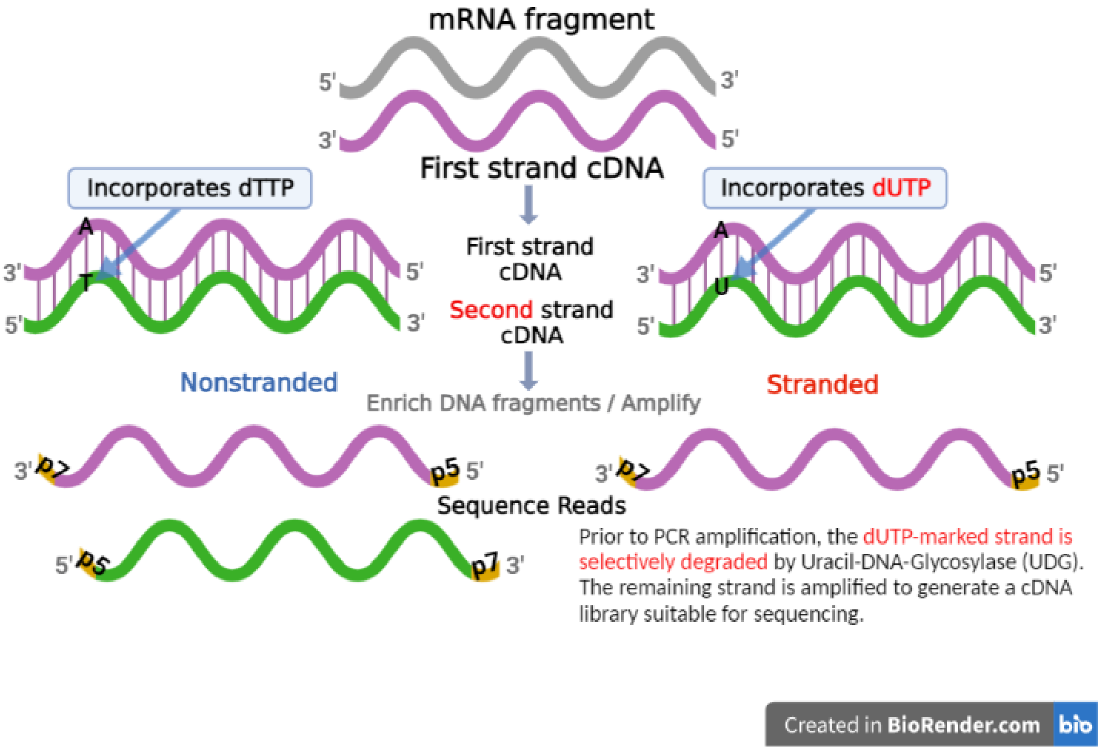

mRNA fragment
5'
3'
3'
5'
First strand cDNA
Incorporates dTTP
Incorporates dUTP
A
T
A
U
3'
5'
5'
3'
First strand cDNA
Second strand cDNA
3'
5'
5'
3'
Nonstranded
Stranded
Enrich DNA fragments / Amplify
3' p7
p5 5'
3' p7
p5 5'
Sequence Reads
5' p5
p7 3'
Prior to PCR amplification, the dUTP-marked strand is selectively degraded by Uracil-DNA-Glycosylase (UDG). The remaining strand is amplified to generate a cDNA library suitable for sequencing.
Created in BioRender.com bio

Figure 3.1: Comparison between Non-stranded and Stranded mRNA-seq. This figure was created based on Fig.1 in a review paper by Zhao et al. (2015) by BioRender (`https://Biorender.com`). In both Non-stranded and stranded method, the first strand of cDNA was synthesized through reverse transcription of a targeted RNA fragment. In synthesis of the second strand, chemical mark of dUTP was introduced in the stranded method so the newly synthesized strand with U could be degraded by Uracil-DNA-Glycosylase (UDG) to keep only the first strand for sequencing. Check 3.1.1.1 for other methods.

Non-Stranded Method

There are many commercial kit available for cDNA library construction of non-stranded method. Here are the brief steps of a commonly used kit TruSeq® RNA Sample Preparation v2.

1. **Purification and fragmentation of mRNAs:** In order to purify poly(A) tail-containing mRNAs, oligo(dT)-attached magnetic beads are used in two rounds of

purification. During the second elution, the polyadenylated RNAs are also fragmented and primed with random hexamers.

2. **Synthesis of First Strand cDNA:** The cleaved RNA fragments are then reverse transcribed into first strand cDNAs using random primers.

3. **Synthesis of Second Strand cDNA:** The RNA templates are removed and the replacement strand of the cDNAs are synthesized to form double-strand cDNAs.

4. **End Repair:** By now, the overhangs resulting from fragmentation have not yet been removed. The existence of overhangs will affect the efficiency of ligation thus the overhangs should be repaired. The End Repair Mix is used to perform $3'$ to $5'$ exonuclease activity and convert the $3'$ overhangs into blunt ends.

5. **Adenylation of $3'$ Ends:** To avoid the fragments being ligated to one another during adaptor ligation, a single "A" is attached to the $3'$ end as a complementary overhang for the T on the $3'$ end of the adaptors used in the next step.

6. **Ligation of Adaptors:** The adaptors with multiple indexes are ligated to the cDNAs, so the samples can be hybridized onto the same flow cell. The same adaptors are ligated to both of the cDNA strands, so for a single fragment, it is possible to get both the original sequence of it or the reverse complementary sequence, without strand-specificity.

7. **Enrichment of DNA Fragments:** The DNA fragments are amplified by PCR reaction.

8. **Library Validation:** The step is to ensure the quality of the library before sequencing.

9. **Normalization and Pooling of the Libraries:** Finally, in order to sequence multiple samples in a single flow cell, multiple samples with different index in the adaptor would be pooled together with equivalent RNAs from each sample.

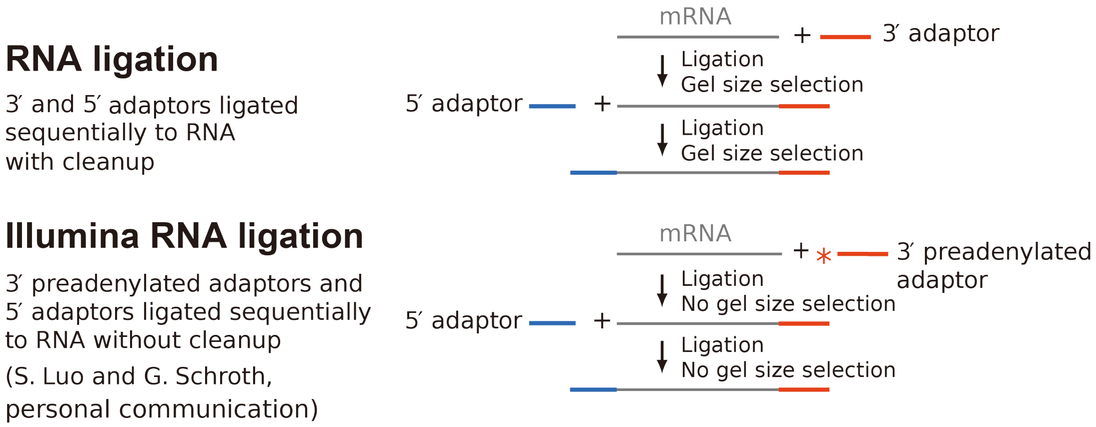


Figure 3.2: Methods for strand-specific RNA-seq (Levin et al. (2010)) Details for different adaptor methods (RNA ligation and Illumina RNA ligation) and mark method (dUTP second strand, see Figure 3.1). In different adaptor method, 3′ adaptor was added to the RNA fragment before synthesis of first strand cDNA and the 5′ adaptor would be later added to the newly synthesized DNA molecular.

Strand-specific Method

In transcription, one of the strands of the double-strand DNA (dsDNA) is taken as a template for the RNA. The strand information of RNA is important, especially to genes with anti-sense transcription. For example, the antisense non-coding transcript *COOLAIR* of *FLC* plays an important role in flowering control (Swiezewski et al., 2009). Also, there might exist genes with overlapped region transcribed from different strand. We cannot distinguish the origination of these overlapped regions in non-stranded sequencing but can easily manage to do so with strand-specific data in the later analysis. In summary, with the strand-specific method, more detailed information can be detected. The strand information can be reserved through different approaches including different adaptor methods and mark method (Levin et al. (2010)):

1. The first method uses ligation of adaptors to mRNA or first-strand cDNA. RNA ligation and illumine RNA ligation shown in Figure 3.2 are two typical methods of this approach. Unlike in the non-stranded method, the 3′ adaptors are ligated to the mRNA rather than dsDNAs to retain the strand-specificity. Similar method was applied in construction of eCLIP cDNA library.

2. The second approach is direct RNA sequencing (DRS), adding adaptors to the first-strand cDNA and sequencing without PCR amplification. Currently, DRS usually attach with long-read sequencing method like Oxford Nanopore, and is not limited to study of mRNAs, but also applied to chromatin-associated RNAs (Parker et al. (2020)). This approach was not applied in this project.

3. In the third method, the second strand is synthesized with chemical mark, like dUTP, and removed before amplification (see Figure 3.1). For the strand-specific mRNA-seq data in this project, dUTP method was used. In the synthesis of second strand cDNA, dUTP was used instead of dTTP. The strand with U can be later decayed with USER, leaving the first strand cDNA for amplification only. In material preparation, mRNA capture beads with Oligo(dT) were mixed with total RNAs of 10-day-old seedlings to capture mRNAs with poly(A) tails. The beads were then washed with Tris buffer to release the RNAs from the beads to the supernatant. The magnetic bead-based purification step is taken twice. The supernatant contained the mRNAs for library construction. Since short-read sequencing would be used, the transcripts were apparently too large to sequence. The first step in library construction was fragmentation of mRNAs by heating the mixture of mRNA, first strand synthesis buffer and random primers, followed by the synthesis of first-strand cDNA with random primer. The dUTP mix (along with second-strand synthesis buffer) was then added to the product for second strand synthesis. By now, double-strand cDNA (ds cDNA) had been synthesized, with the second strand marked by U instead of T. After purification of the dsDNA with purification beads, adaptors were ligated to both ends of these cDNAs. USER enzyme, which works on degradation of U-containing sequence, was added to remove the second strands. Extra adaptors in the tube were removed with purification beads. After PCR amplification, the library construction step was done. Samples with different sequencing index can be mixed and loaded on flow cells for sequencing.

With the mRNA-sequencing data gained from the constructed library, we can apply bioinformatics analysis to study the transcriptome-wide expression in materials of the target genotype.

3.1.1.2 mRNA-seq Analysis Workflow

In a typical analysis of RNA-Seq Data, the computational workflow includes quality control of the raw sequencing file in FastQ format, adaptor removal to generate clean reads, read alignment to map the reads to the reference genome (stored in BAM or SAM files), transcriptome reconstruction and expression quantification to get the expression matrix. Afterwards, downstream analysis can be applied to the aligned reads or the matrix.

There are some formats that are often used in analysis. FastQ format records each sequence with four lines. The first line starts with an '@', followed by sequence identifier; The second line contains the sequence; The third line starts with '+' and may contain the sequence identifier like in the first line (but usually not); The last line is composed of the quality score of the sequence, the length of which is the same to the sequence. The aligned file would be stored in SAM format or its binary equivalent BAM format. The format is flexible which can present not only the position where the sequence mapped to on the genome but also includes the quality of mapping, splicing sites and position of indels (insertions and deletions). The reference genome is usually stored in Fasta format, with two parts for each sequence, a line starts with '>' which bears the sequence identifier and the other lines for the sequence. The other commonly used formats include GFF and GTF for reference annotation file, BED files recording the position of peaks or target region, Wig or BigWig for visualization of the mapped reads, etc.

Currently, strand Specific library methods have become more popular. And most popular bioinformatic tools for RNA-seq data analysis have the corresponding strand-specific methods to make full use of the data. For those tools which disregard strand information (for example the old version of Deeptools which was firstly developed for ChIP-seq data processing targeting double-stranded DNAs rather than RNAs), we can still split the mapped reads into different files by strand and process the split files separately in downstream analysis.

Quality Control & Trimming The raw RNA-Seq data may contain reads of low quality. The quality control step should be applied before further analysis to make sure that the reads are good enough for use. In most QC tools, a report of read quality will be provided to visualize the overall quality of reads. The report would include GC content, sequencing quality, adaptor content et al. Some QC tools can also help us import the quality of the data by removing reads/bases failed in quality check.

Trimming can also be seen as a part of quality control. There are three kinds of trimming, defined by the sequence they trim:

1. **Adaptor trimming:**

   Paired-end sequencing of 150-nucleaotide-long reads (PE150) was applied for all cDNA library constructed in the project, but not all inserts are of the same length, some are under 150, while some are above. For those reads of length smaller than 150, the sequencing adaptor on the other side would be sequenced, which means the R2 adaptor would be sequenced in Read1, while the R1 adaptor would appear in Read2. One of the most important jobs for these tools is to remove these adaptor sequences.

2. **Quality trimming:**

   Sequencing quality might be bad at the beginning of the reads. In other cases, if a read contains a low-complexity sequence, Poly(A) tail for example, the sequencing quality would be affected. Low-quality bases at $5'$ or $3'$ end of the read could be trimmed for more precise alignment.

3. **Removal of barcodes/UMIs:**

Unique molecular identifiers (UMIs) are usually a short sequence added to the 5′ or 3′ of the reads. Reads mapped to the same start point and with the same UMI are considered as PCR duplicates. The UMI sequence is not part of the genome and should be removed before sequencing, but the sequence of UMI should be checked after mapping. Some trimming tools can remove the UMI while adding the sequence to read ID for later use. Quality trimming should not be applied to the side with UMIs. The process of UMI will be described in detail in the secession of CLIP.

Here we introduce some tools which enable us to check the quality of the data and to improve prove the quality if needed. Some of these QC tools can also be applied to aligned reads.

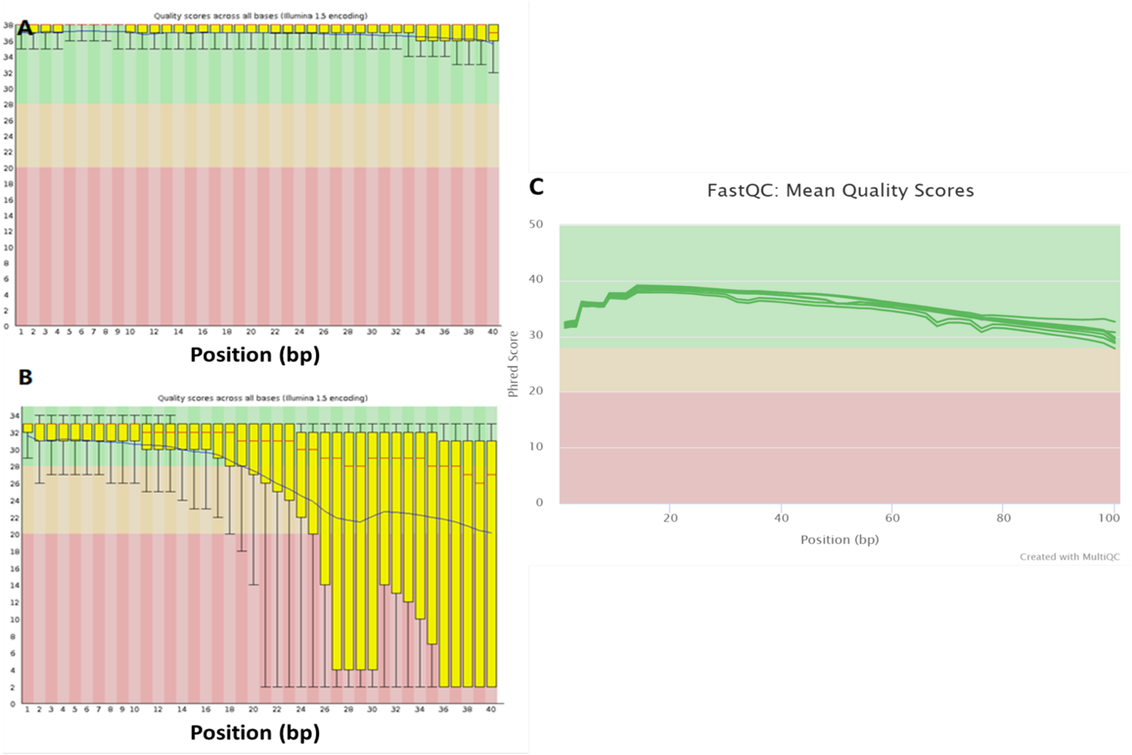


Figure 3.3: Examples of quality control report. A and B are figures generated by FastQC to present samples of before(B) and after(A) quality control. C. Report by MuiltiQC, showing merged FastQC report of multiple samples in a single report.

FastQC: a classic tool for quality control FastQC (Andrews et al. (2010)) is a tool to look into the quality of the NGS data. It can process several types of commonly used NGS datasets, including the files with format FASTQ/SAM/BAM to generate summary graphs and tables to provide a quick overview of the data. In Figure 3.3 A&B are two example reports of sequencing with good sequencing quality (A) and bad quality (B). The x-label are the position of the base in the reads, while y-label is the sequencing quality. The green region represents high quality score of the read, while red region means the sequence is of low quality. In the case of A, the lower quality bases have been trimmed, so most reads have good quality located in the green region, while B is an example of quality of raw reads without trimming, which shows bad quality.

MultiQC: summarizing QC report from result of multiple tools While MultiQC (Ewels et al. (2016)) can generate FastQC results of multiple datasets into one single report. In Figure 3.3 C is an example report of MulitQC. Unlike in FastQC, where one report figure shows the quality of base of a single sample, MultiQC report can merge these reports into a single report with information of multiple samples, which is very concenient.

Trimmomatic: a flexible read trimming tool Various software tools are available to trim the adaptors from raw data. One of the standard packages is FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). It is a collection of command line tools designed for short-reads FASTQ/FASTA file reprocessing to help to manipulate the sequences and to produce better alignment results later. As with clean reads after trimming the adaptor, there also exist some tools to help trim ambiguous (N) and low-quality residues from the ends of the reads, for example, fastq_quality_trimmer program, which is part of the FASTX-Toolkit, and Trimmomatic (Bolger et al. (2014)). In this tool, quality trimming is performed with Trimmomatic. Compared to other existing preprocessing tools designed previously, Trimmomatic shows significant advantages regarding flexibility (Bolger et al. (2014)), correct handling of pair-end data and high performance. Trimmomatic includes a variety of processing steps for reads trimming and filtering, but the main algorithmic innovations are related to the identification of adaptor sequences and quality filtering.

Cutadapt: removing unwanted sequence from your HTS reads The algorithm of aligning adaptors in Cutadapt (Martin (2011)) was based on semi-global alignment. Unlike in the global alignment, by which the sequences are compared base by base to count all the differences occurs, the sequences are allowed to shift freely and only penalized in difference in overlapping region. Meanwhile, instead of alignment scores, unit cost is used to describe the errors in the alignment. The alignment scores of matches are positive values while negative values are set for mismatches, insertions and deletions. The alignment with the maximal total score is used. It is intuitive for us to use the alignment score in define the error rate. The unit cost which treats all mismatches, insertions and deletions as errors is applied to get the adaptor sequence of the least error rate. With these improvements, Cutadapt is able to align the adaptors (or other sequence the users supplied) fast and remove them afterwards.

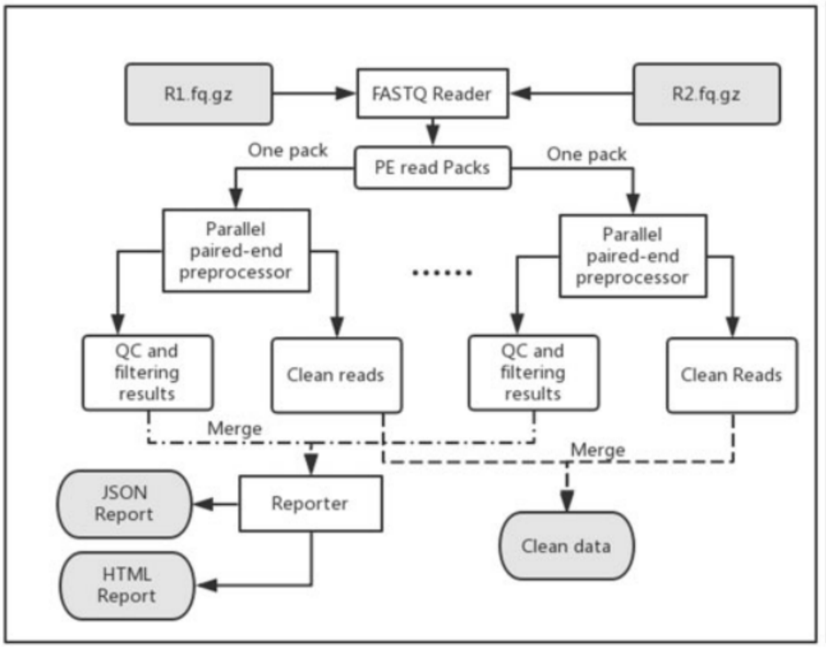


Figure 3.4: A workflow for quality control of paired-end sequencing data (Chen et al. (2018)) with Fastp to generate clean data without adaptor and quality control reports

fastp: a fast and handy tool for QC The tool fastp (Chen et al. (2018)) contains most function of FASTQC, Trimmomatic and Cutadapt but with much faster in speed (see Figure 3.4 for workflow). For Pair-end reads, it can automatically detect the adaptor sequence and correct bases of low sequencing quality by the overlapping region of the paired Read 1 and Read 2. It is also possible to merge corresponding Read1 and Read2 with overlap into a single insert by fastp. Another important feature of fastp is the interactive HTML report it generated for the users to check the quality of their data before and after processing. The content of the reports includes not only the basic information like GC-content, adaptor sequence, sequencing quality etc., but also an estimation of Duplication level and insert size distribution. The insert size distribution is useful in checking whether the fragments of RNAs are of the expected size. This tool can also deal with UMI-containing reads, to remove UMI sequence and add the sequence to read ID for later usage.

RSeQC: application of quality control after alignment RSeQC (Wang et al. (2012)) package comprehensively evaluate HTS data with a handful of modules, including format switch from BAM to FastQ or wig, plot profile of clipping, deletion, insertion, mismatch and coverage, annotation of junctions known or novel, summary for GC-content/duplication level etc. Unlike most quality control tools, RSeQC deal with aligned reads rather than raw reads.

Read Alignment

After quality control and trimming step, the reads are now clean reads ready for alignment. Alignment, also known as mapping, is the step to identify the position on the genome where the reads belong. Unlike DNAs, which would not be spliced, the alignment of RNA should be carried out in splicing-aware mode. For most popular aligner nowadays, an index should be built in advance for faster mapping with reference sequence (Fasta format) and annotation (GTF/GFF format). These files can be found online for model species, while denovo assembly should be applied instead of alignment if no reference is available.

Bowtie2: fast and sensitive read alignment TopHat2 (Kim et al. (2013)) can use either Bowtie (Langmead (2010) or Bowtie2 (Langmead and Salzberg (2012)) as its core read-alignment engine. Bowtie2 extends the full-text minute index-based approach of Bowtie to permit gapped alignment by dividing the algorithm broadly into two stages (Langmead and Salzberg (2012)). Firstly, it proceeds an initial seed-finding stage which benefits from the speed and memory efficiency of the full-text minute index. After that, a gapped extension stage with dynamic programming and benefits from the efficiency of single-instruction-multiple-data (SIMD) parallel processing available on modern processors. It is notable that the formats of the index used by Bowtie and Bowtie2 are different, so do notice to use the index of the proper format.

TopHat2: accurate alignment of transcriptomes All sequence mutations are explained as deletions and insertions(indel) of genetic material. TopHat has its indel-finding algorithm, which enhances indel-finding ability of Bowtie2 in the context of spliced alignments. TopHat2 incorporates many significant enhancements to TopHat1. To solve the problem lying with junction spanning reads, the software uses a two-step method. Similar to Tophat1, it detects potential splice site for introns in the first step and then uses these candidate splice sites in a subsequent step to correctly align multiexon-spanning reads. It also includes new algorithms to handle more diverse types of sequencing data, such as reads generated by ABI SOLiD technology (Life Technologies, Carlsbad, CA, USA).

Hisat2: hierarchical indexing for spliced alignment of transcripts 2 Hisat2 (Kim et al. (2017)) is one of the most popular aligners now days. It is faster than Bowtie2 and TopHat2, and is used to replace TopHat2 in many pipelines. It inherits the rapid and memory-saving feature of Bowtie. The memory needed is quite small that it is even possible for hisat2 to run on a laptop. Also, like many other mappers for RNA-seq, Hisat2 works in a transcriptome-aware mode, making splicing detectable. When mapping the stranded reads, the '--rna-strandness' parameter should be set to RF for library built with dUTP method.

Note that Read1 is actually complementary to the corresponding fragment of transcript, which is also a point that should be taken care of in later read counting steps.

Transcriptome Assembly

Unlike genome sequence coverage levels, which can vary randomly as a result of repeat content in non-coding intron regions of DNA, transcriptome sequence coverage levels can be directly indicative of gene expression levels. So, the assembly of transcriptome become extremely important. One of the most popular transcriptome assemblers is called Cufflinks. Cufflinks (Trapnell et al. (2010)) assembles transcripts estimates their abundances and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols. The counting of the mapped reads could also be performed with Cufflinks. Cufflinks can count the expression of each gene and report it in FPKM (fragments per kilobase of transcript per million fragments mapped), which measures the expression of a transcript, normalized by transcript length and the total number of fragments. As such, by comparison of the FPKM value, we can get the difference between the expression of the genes. Similarly, RPKM (Reads Per Kilobase Million) and TPM (Transcripts Per Kilobase Million) could be calculated for the same use. However, Cufflinks uses an annotation of the reference genome described in GFF (or GTF) format, which means that the results depend on the quality of the provided annotation. Correspondingly, another way which does not rely on denovo assembly. In model plants like Arabidopsis, the genome of which is well-annotated, the assembly step is not always necessary. But it might be helpful in detecting novel transcripts in some cases. For most RNA-seq data, the assembly step is not taken. But we tried to recover the full sequence of FPA with T-DNA insertion using Whole Genome Sequencing (WGS) and RNA-seq data, which will be described in Chapter 4.

Read Counting

Besides Cufflinks mentioned in the last section, there are many other tools which could count the read numbers on the genes or on each site. In some cases, raw count matrix is in need while the others the normalized ones.

featureCounts The featureCounts tool (Liao et al. (2014)) can summarize reads in a fast and accurate way. It can count the mapped reads for not only genes but also features like exons, promotors, genomic bins. It takes as input mapped reads of SAM/BAM files and annotation in GFT or simplified annotation format (SAF). For stranded reads, '-s' should be set. While '—countReadPairs' mode is also available for paired-end sequencing reads. The matrix from featureCounts can be used as an input for downstream analysis tools like DESeq2 (Love et al. (2014)). Note that featureCounts returns only matrix of raw read count. Further normalization can be made manually later.

TPMCalculater TPM is currently one of the most recommended methods of normalization for gene expression in RNA-seq data. And TPMCalculater (Vera Alvarez et al. (2019)) is a one-step software to get the read count along with TPM value for each gene from mapped BAM/SAM files. The generated matrix of expression can be applied in downstream analysis like differential expression analysis.

Other Downstream Analysis

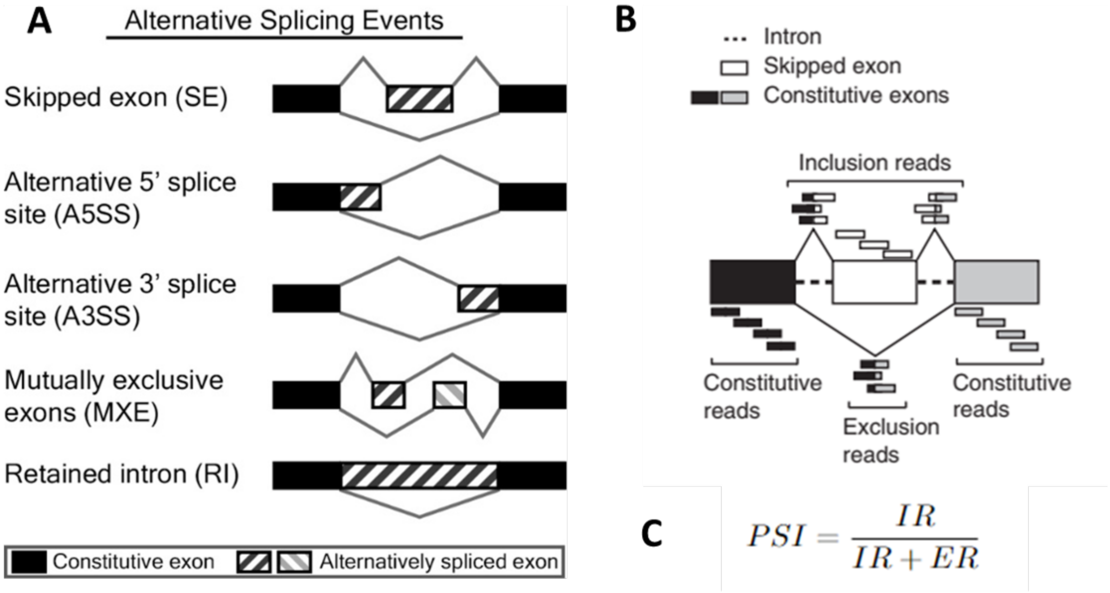


Figure 3.5: Detecting Alternative Splicing Events A. five kinds of splicing events: Skipped Exon(SE); Alternative 5′ splice site (A5SS); Alternative 3′ splice site (A3SS); Mutually exclusive exons (MXE); Retained Intron (RI); B. principle of counting inclusion reads and exclusion reads; take exon skipping as an example, inclusion reads are reads partially or fully located on the exon which might be skipped, indicating that exon-skipping does not happen; exclusion reads are junction reads, as the result of exon skipping;C. Formula of PSI (all these figure are from Park et al. (2013))

rMATS: detecting alternative splicing events rMATS (Park et al. (2013)) is a tool for detecting alternative splicing events between two conditions. There are five kinds of alternative splicing events (Figure 3.5 A), including skipped exon, alternative 5′ splice site, alternative 3′ splice site, mutually exclusive exon and retained intron. The most common AS event is intron retention in plant and exon skipping in human. This could probably be concluded to the difference in intron length. The number of genes in both Arabidopsis and human is around 30000, but the size of human genome is much larger (about 3GB) than the one of Arabidopsis (about 200M). Human genome contains many large introns while long introns are not common in plants. The rMATS tool takes annotation file in GTF format and sorted Bam with index as the input. The splicing events are mainly detected based on the annotation, while novel splicing events might also be detected. An index, Percent-splice-in

(PSI), is used to represent the AS event, where IR is inclusion reads and ER is exclusion reads (Figure 3.5 B, C).

DaPars: Dynamic analysis of alternative polyadenylation from RNA-seq DaPars (Xia et al. (2014)) is the first tools developed for analysis of APA events with normal mRNA-seq data. The algorithm and usage are described in detail in Chapter 5.

3.1.2 3′mRNA-seq

Lexogen's QuantSeq™ 3′ mRNA-seq focus on the 3′ end of each transcript. It can be considered as a supplementary to the mRNA-seq data in detecting position of poly(A) sites, quantifying the mRNA expression level and usage of alternative poly(A) sites. Since FCA has been reported to regulate the usage of proximal/ distal poly(A) site of itself, we hope to know whether this kind of regulation of PAS happens on only a few genes or on a transcriptome-wide base. The other important thing we would like to know is that whether the transcription extended to the downstream of the original PAS in the wild type.

3.1.2.1 Library Construction

Building cDNA library with QuantSeq 3′ mRNA-seq kit includes five steps and can be finished in four and a half hours.

- Step 1. Reverse Transcription: With total RNA as input and no need of rRNA removal, the generation of library starts with oligo(dT) priming containing the Read2 linker sequence.
- Step 2. Removal of RNA: The RNA template is removed after first strand synthesis.
- Step 3. Second Strand Synthesis: The second strand synthesis starts with random priming by primer with read1 linker sequence. After the synthesis, a magnetic bead-based purification step is taken.

- Step 4. Library Amplification: Then in the amplification step, sequences required for cluster generation are introduced. With different index applied, multiplexing of samples are made possible.

- Step 5. Sequencing: Finally, the library is built and to be sequenced. Single-end sequencing of longer read is recommend to pinpoint the exact $3'$ end. Pair-end sequencing is not recommended since read2 start with poly(T) stretch and have low sequencing quality (see Figure 3.6 B for structure of the cDNA library).

In summary, QuantSeq $3'$mRNA-seq can help us detect RNA sequences close to the $3'$ end of polyadenylated RNA.

Improvement: QuantSeq $3'$ mRNA Sequencing REV A method for solving the issue of low sequencing quality near the poly(A) tail is provided by Lexogen itself. Unlike QuantSeq $3'$ mRNA Sequencing FWD, QuantSeq $3'$ mRNA Sequencing REV include the poly(T) sequence as part of primer in sequencing, improving the sequencing quality of the poly(A) tail side. Unlike QuantSeq FWD, QuantSeq REV aims to get the extract the exact PAS, which might be more suitable for our research.

3.1.2.2 Data Analysis

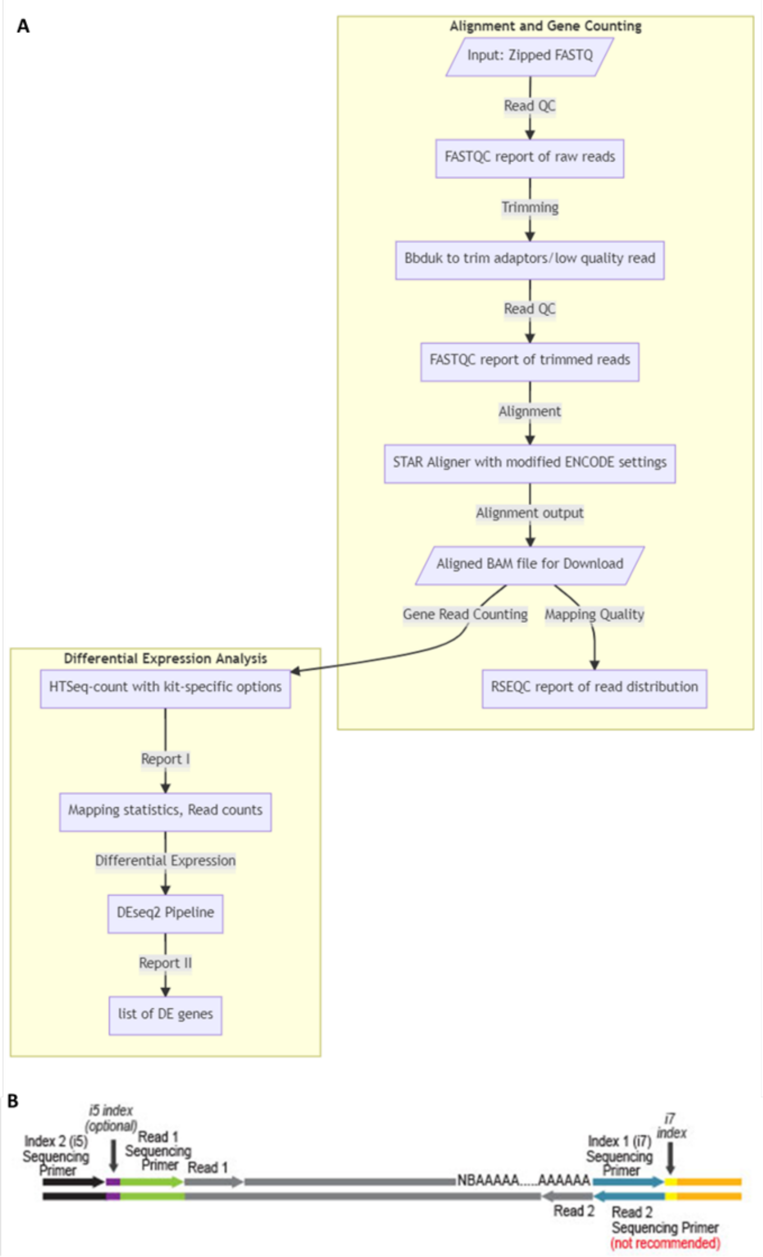

A
Alignment and Gene Counting
Input: Zipped FASTQ
Read QC
FASTQC report of raw reads
Trimming
Bbduk to trim adaptors/low quality read
Read QC
FASTQC report of trimmed reads
Alignment
STAR Aligner with modified ENCODE settings
Alignment output
Aligned BAM file for Download
Gene Read Counting
Mapping Quality
RSEQC report of read distribution
Differential Expression Analysis
HTSeq-count with kit-specific options
Report I
Mapping statistics, Read counts
Differential Expression
DEseq2 Pipeline
Report II
list of DE genes
B
i5 index (optional)
Index 2 (i5) Sequencing Primer
Read 1 Sequencing Primer
Read 1
NBAAAAA.....AAAAAA
Index 1 (i7) Sequencing Primer
i7 index
Read 2
Read 2 Sequencing Primer
(not recommended)

Figure 3.6: Data analysis workflow A. Recommended workflow for alignment and differential expression analysis; B. structure of the cDNA library constructed, Read2 would contain the poly(A) tail.

A workflow for data analysis is recommended in the user guide (see Figure 3.6). FASTQC (Andrews et al. (2010)) is used for quality control, bbduk (Bushnell (2018)) for trimming low-quality tails and adaptors, STAR (Dobin et al. (2013)) for alignment, RSeQC (Wang et al. (2012)) for after-alignment quality control, HTSeq (Anders et al. (2015)) for read counting, DESeq2 (Love et al. (2014)) for differential expression analysis. This workflow can help us get differentially expressed genes, but we expect more information from $3'$mRNA-seq data — the position of poly(A) sites. Thus, we develop a new workflow for getting the exact PAS which would be described in Chapter 5.

Besides the expression, the other unique analysis which can be applied to $3'$ mRNA-seq data is detection of change in poly(A) site usage. Unlike in mRNA-seq where the transcripts are fragmented and sequenced, $3'$ mRNA-seq only retains the $3'$ end of the mRNAs, making it possible to extract the exact poly(A) sites.

Due to the alternative polyadenylation, we can divide the poly(A) sites of a gene into the most commonly used a dominant poly(A) site and the less used weak poly(A) site by their usage. The weak poly(A) sites are hard to detect in low coverage genes. For example, a gene may have largely decreased expression in the target genotype comparing to the wild type. In this case, the weak poly(A) site might be undetectable in the mutant but detectable in the wild type, which would be considered as increased usage of the dominant poly(A) site in the mutant. Therefore, it is important to ensure that sequencing depth was sufficient for the analysis and remove the genes of low coverage in any sample during the analysis. On the other hand, the merge of poly(A) clusters (PAC) by distance makes the detection even harder. Many PACs are unreasonably large, which might be a combination of several smaller PACs. QuantifyPoly(A) (Ye et al. (2021)), a new tool developed is helpful in solving the problems above. We would describe the tool in detail in Chapter 5.

3.1.3 Section Summary

After several trials, we picked the analysis workflow for mRNA-seq data, from quality control step to alignment. The very quick and easy-to-use tool Fastp was applied for quality control and adaptor trimming. Since all sequencing data used in this project are paired-end and not extra sequence besides the adaptors should be removed from the read in mRNA-seq data, the autodetection of adaptors in Fastp was enough for trimming. With the clean reads, Hisat2 was used for alignment. One thing to mark in alignment is the strandness. Here, dUTP method was used in library construction, the strand of read2 rather than the one of read1 was the same with the original fragment. After the alignment with Hisat2, we would get a SAM file with all reads, whether they are mapped or not. So, an after-alignment quality control step would be applied with samtools to remove the reads unmapped and reads with mapping quality lower than 20 (considered as reads with secondary alignment). The final alignment would be stored in a sorted and indexed BAM file.

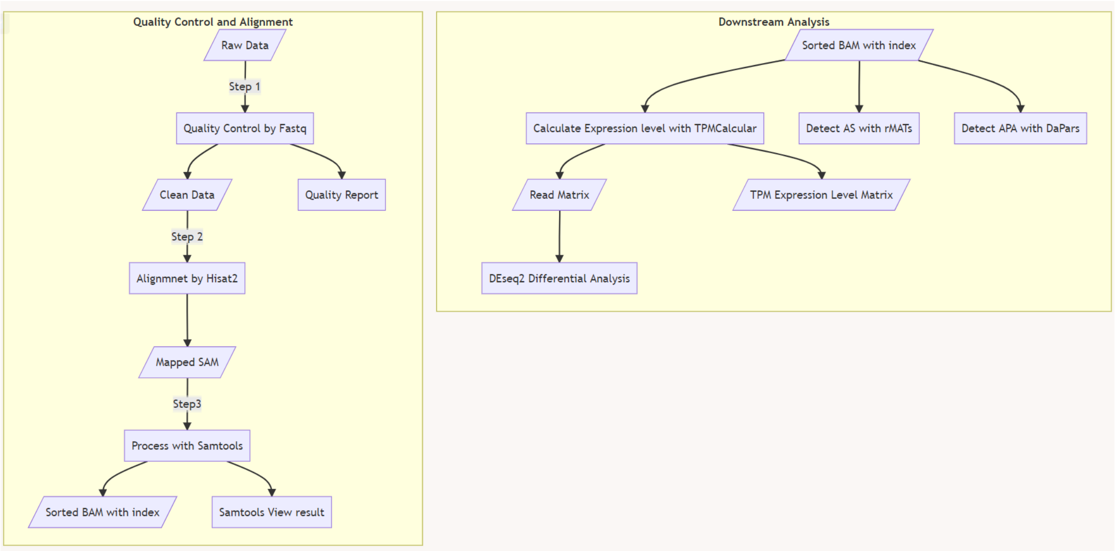


Figure 3.7: Data analysis workflow of mRNA-seq data used in this project: Quality control with Fastp, Hisat2 for alignment. The QC and alignment tools could also be used in CB-RNA-seq analysis. Downstream analysis after alignment were not limited to differential expression analysis. Analysis on alternative splicing and alternative polyadenylation would also be included.

The workflow was also suitable for analysis of CB-RNA-seq data mentioned in next section. But the analysis of 3'mRNA-seq data would be quite different. Read2 contains poly(T) sequences at the beginning, which made the sequencing quality low and not trust-worthy.

The normal alignment workflow was very similar to the one for mRNA-seq while the parameters should be adjusted for single-read. Unfortunately, we can not locate the PAS by read1 alone and the result might be affected by the existence of internal priming events. In order to get the exact position of PASs, we tried some new methods for read filtering, which would be further discussed in Chapter 5.

3.2 Sequencing Methods for pre-mRNAs

3.2.1 Chromatin-Bound RNA-seq (CB-RNA-seq)

The idea of separating nascent RNAs from nuclear RNAs could be traced back to 1994 when Wuarin and Schibler tried to figure out whether splicing happens co-transcriptionally or post-transcriptionally. This kind of RNA was named Chromatin-associated or Chromatin-bound RNA by its feature of binding to the chromatin, indicating that they are still under transcription and are attached to the chromatin through RNA polymerase (especially Pol II). The researchers developed a procedure to separate the chromatin-associated and released transcripts. This was made possible based on the steady connection between the DNA template and Pol II which resists even under some harsh treatments like high salt, detergent and urea. So, after the treatment, the released RNAs would be washed away, while the nascent RNAs remains attached to the chromatin through Pol II.

3.2.1.1 RNA Extraction and Library Construction

In this experiment, the Chromatin-bound RNAs (CB-RNAs) were extracted instead of total RNA or mRNA (with poly(A) tail and usually enriched with oligo(dT) beads). The DNAs were purified together with the CB-RNAs, so the DNAs will be removed before building the library, as well as rRNAs. Though it was usually believed that RNAs will be released from the chromosome after termination, we can see polyadenylated RNAs in the CB-RNAs. These transcripts remained after restricted washing steps, indicating that they are still attached to the chromosome. One of the reasons that they are still attached to the chromosome might be that introns of them have not been spliced co-transcriptionally and remain to be spliced after transcription. The library construction was generally based on method used in former research of co-transcriptional splicing (Zhu et al., 2020a). In this project, these polyadenylated RNAs were not removed. Here's the method for extraction of CB-RNAs:

1. **Extraction of nuclei pellet:**

   Nuclei from 1 g of ground seedlings were prepared using a Honda buffer supplement with the addition of RNase inhibitor and yeast tRNA, to protect the integrity of RNA. The nuclei pellet was first resuspended in an equal volume of resuspension buffer, followed by brief washing with two volumes of washing buffer to wash away the free RNAs in nucleoplasm. The chromatin was pelleted by centrifugation at 8000 g for 1 min, resuspended with an equal volume of resuspension buffer, followed by washing by using one volume of UREA wash buffer to further wash away the free RNAs near chromatin. With free RNAs removed and only chromatin-associated RNAs retain attached to the chromatin, the product is saved in Liquid nitrogen for later use.

2. **Get CB-RNA from nuclei pellet:**

   The resulting chromatin pellet after centrifugation was resuspended in 1 mL Trizol and followed by RNA extraction in combination with the RNeasy Mini Kit (Qiagen) for purification of total RNAs.

3. **Remove DNAs:**

   Chromatin DNAs co-purified with CB-RNAs should be removed with Turbo DNase to make sure only RNAs are retained.

4. **Quality Control of CB-RNA:**

The integrity of the resulting CB-RNA was checked by gel electrophoresis by comparing to the pattern of total RNA. The enrichment of CB-RNAs on certain genes can be tested with qRT-PCR as part of the QC test.

5. **Removal of rRNAs:**

   A large portion of nuclei RNAs are ribosomal RNAs (rRNAs). They can be removed with Ribosomal RNA depletion kit.

After the extraction of CB-RNAs, dUTP method of RNA-sequencing was used to build strand-specific library.

3.2.1.2 Data Analysis

The Upstream analysis of CB-RNA-seq is similar to that of mRNA-seq(see Figure 3.7). Expect for the expression level of CB-RNAs, we can also use the data to check the co-transcriptional splicing efficiency of the transcripts. In the previous research (Zhu et al. (2020a)), SS ratio (Figure 3.8) was applied to describe the splicing efficiency.

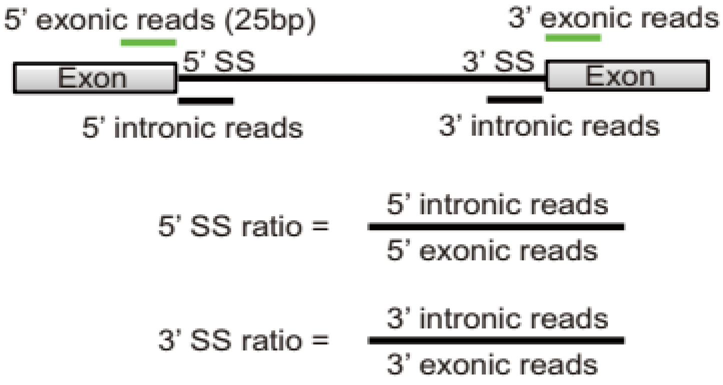

5’ exonic reads (25bp)
3’ exonic reads
5’ SS
3’ SS
Exon
Exon
5’ intronic reads
3’ intronic reads
5’ SS ratio = 5’ intronic reads / 5’ exonic reads
3’ SS ratio = 3’ intronic reads / 3’ exonic reads

Figure 3.8: Calculation of co-transcription splicing efficiency (from Zhu et al. (2020a): read numbers in a 25-base window on both flankings of the splicing site would be counted for the calculation of splicing index)

And the intron position was defined by TAIR10 annotation. Here we made some modification to the original method, defining the intronic regions based on not only annotation but also splicing efficiency at the mRNAs level with mRNA-seq. Basically, if an annotated intron region is not actually spliced at mRNA level, it is fair to believe that the intron was retained in the transcript and to classify the region as part of exon rather than an intronic region, while it is difficult to distinguish a post-transcriptionally spliced intron and an unspliced region in CB-RNA-seq. But together with mRNA-seq, we can pre-define the introns that would be at least partially spliced.

3.2.2 Plant Native Elongating Transcript Sequencing (pNET-seq)

Using Pol II antibody, pNET-seq can precisely define the position of Pol II on the transcript under transcription. The NET-seq technique was first developed in yeast, later in mammalians (Nojima et al. (2015)) and was not applied to plants until 2018 when Zhu et al. developed the pNET-seq method and applied to Arabidopsis successfully. The use of antibodies in the experiment would result in different patterns and the pattern in different species might differ (Figure 3.9 A). The pausing of Pol II would result in the enrichment of NET-seq peak in the region. For example, Pol II pausing near the $5'$ of the gene to transform from initiation to elongation stage of transcription.

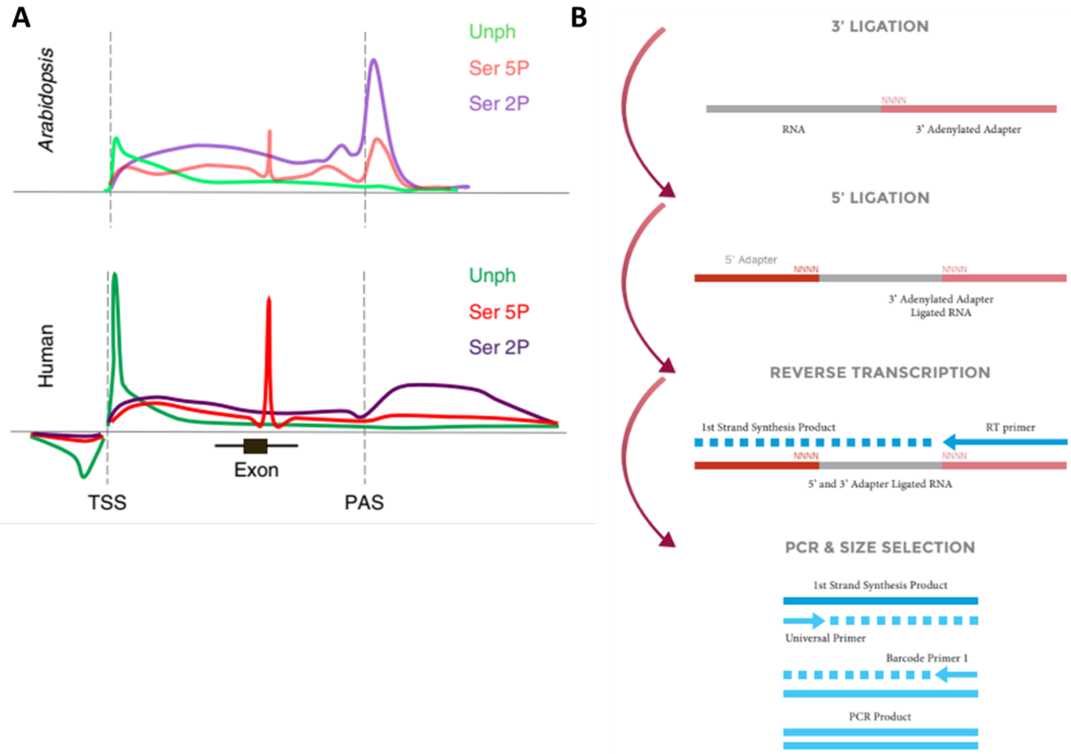


Figure 3.9: Introduction to pNET-seq (figures by Zhu et al. (2018)) A. Comparison between Arabidopsis and Human: a 5′ peak of Pol II Unph was detectable in both plant and animal, indicating that the unphosphrylated Pol II was enriched before transcriptional initialation; similiarly, Pol II pausing near splicing junction exist in both species; however, there exist a sharp 3′ peak of Pol II Ser2P in plant but not in human; B. Library construction workflow: the transcripts with Pol II was fragmented and enriched with Pol II antibody. The 3′ adaptor was first ligated and then the 5′. The fragments would be reverse trancribed, amplified by PCR and sent to sequecning after size selection.

In this research, we planned to use antibody for Pol II Ser2P, which was believed to be enriched in the termination process to check the activity of Pol II near the 3′ end of genes. Currently, no actual pNET-seq data were involved in this thesis, but the data would be applied in the future research.

3.2.2.1 Library Construction

As materials, grand seedlings of 2 g are mixed with lysis buffer and filtered to extract the nuclei. Only the sequence near Pol II will be detected later, MNase instead of heating was applied for fragmentation so only the RNAs/DNAs protected by proteins would retain in supernatant. To enrich Pol II-associated RNAs, beads with Pol II-antibody was used in immunoprecipitation. The Pol II-bound nascent RNA was treated with T4 PNK on beads and extracted with Trizol reagent. The IP RNA was resolved on an 8% TBE-urea polyacrylamide gel to select RNA with sizes from 35 to 100 nt. Note that the insert sizes of pNET reads were much smaller than the one of common mRNA fragments. Here, NEXTflex™ Small RNA-Seq Kit v3, a kit for small RNA library construction, was applied to these RNAs.

The strand-specific feature of the library originated from the 2-step adaptor ligation before first-strand synthesis, with the $3'$ adaptor first ligated to the $3'$ of the RNA first and the ligation of $5'$ adaptor following. Adaptor on both sides contain random sequences of 4 bases (NNNN) attached to the insert as unique molecular identifier (UMI). After reverse transcription, the product will be sequenced after PCR amplification. PCR duplicates can be recognized by the UMIs.

3.2.2.2 Data Analysis

The analysis workflows suggested in mammalian NET-seq (mNET-seq) article (Nojima et al. (2016)) and pNET-seq article (Zhu et al. (2018)) are different in several steps. Here's the two workflows.

mNET-seq data analysis pipeline

1. Trim the adaptor sequences from the $3'$ end using Cutadapt and meanwhile discard short reads and unpaired reads.
2. Align paired reads using TopHat2 and keep only one alignment to the reference genome allowed for each read.
3. Filter the BAM file with the aligned read pairs using Samtools to obtain those in which both elements properly aligned with the genome.
4. Identify the last base incorporated by the Pol II. According to the protocol, this will be the last base from read 2, but with the directionality of read 1.

This is a pipeline (Nojima et al. (2016)) referred to by many articles with mNET-seq data. Since the article introducing it was published in 2016, many tools used in the workflow have a later version or even substitution with better performance now. For example, Tophat2 is no longer a suggested tool for RNA mapping and should be replaced by Hisat2 or similar. Even Tophat2 mapping is much more time-consuming compared with the one performed with hisat2, the parameter '-g n' reporting no more than n alignments, doesn't seem replaceable at the moment. In HISAT2, a similar parameter is '-k 1', but the finial alignment reported may not be of the best scores if '-k 1' is specified. Currently, Tophat2 is still the first choice for mapping in pNET-seq analysis.

pNET-seq data processing workflow

1. Remove PCR duplications by clumpify.sh from BBMap:

   With the script, the reads will be sorted and clustered according to the sequence. The reads in the same clusters do not necessarily need to be PCR replicates but with common k-mers. This will not work well with high error rate data, so quality-trimming or error-correction may be prudent. However, it is carried out before quality control steps in this workflow.

2. adaptor trimming;
3. mapping to the reference genome;
4. and removing reads from rRNAs, tRNAs and snoRNAs were conducted as described above.
5. Since the 5′ end of read2 represented the last nucleotide incorporated by the polymerase, the aligned reads were trimmed to keep only the 5′ nucleotide of read2, with the directionality indicated by read1.

In this pipeline, PCR duplication removal step is performed before adaptor trimming. This is an important step, according to our preliminary data, the duplication levels were estimated to vary from 30% to 50% in the four samples by fastp. This might not be the best method to remove PCR duplicates, but since the UMIs are short in NET-seq data and located at the beginning of the read where the sequencing quality is lower than the one of middle part of the read, it is fair enough to compare not only the UMI sequence but also the sequence following for PCR duplicate removal.

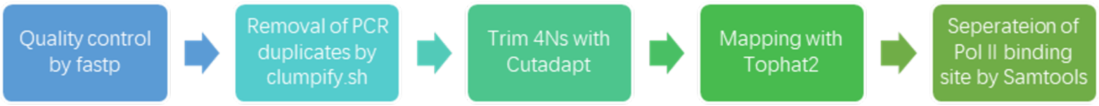

Quality control by fastp
Removal of PCR duplicates by clumpify.sh
Trim 4Ns with Cutadapt
Mapping with Tophat2
Seperateion of Pol II binding site by Samtools

Figure 3.10: Current Analysis workflow of pNET-seq modified based on previous workflows: fastp of quality control; clumpify.sh for removal of PCR duplicates based on UMIs; UMI removal with Cutadapt; alignment with TopHat2; and finial filtering with Samtools

We combined the two workflows and replaced several tools by updated analogues for better performance. Here (Figure 3.10) we may try to replace Cutadapt with fastp, since fastp have extra function of correcting low-quality bases in overlapping regions of paired reads.

And the PCR remove step was carried out after the correction in order to gain better performance. But Tophat2 is still our first choice in mapping for it would return the best alignment if multiple alignments are available.

3.3 Sequencing Methods for Protein-Bound RNAs

Since FCA is an RNA-binding protein, it is important for us to learn where on the transcriptome does FCA bind. Currently, one of the most common ways is through CLIP-seq. Though widely applied to human cells, it is not common for this technique to be applied to plant tissues. Dr Wu has successfully applied iCLIP-seq method in the research of another RBP called RZ-1C with GFP (Zhu et al., 2020a). In this project, another similar method call eCLIP is induced and performed with native antibody of FCA.

In this part, we will introduce the origination and the development of CLIP method, to show the readers what make CLIP a very precious method for learning RNA-RBP interaction, how do we process CLIP data and what's the defect in current methods.

Here, the wild type is not used as a control but as the target of research, while *fca-9* mutant without functional FCA protein was treated as negative control.

3.3.1 Birth of CLIP Method

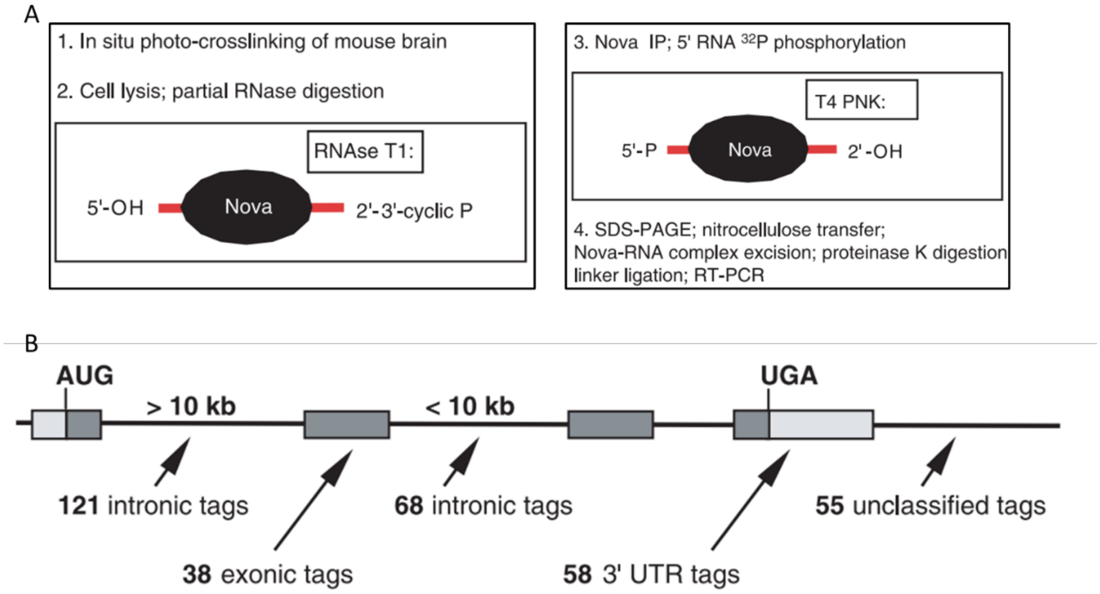

A
1. In situ photo-crosslinking of mouse brain
2. Cell lysis; partial RNase digestion
RNAse T1:
5'-OH
Nova
2'-3'-cyclic P
3. Nova IP; 5' RNA ^{32}P phosphorylation
T4 PNK:
5'-P
Nova
2'-OH
4. SDS-PAGE; nitrocellulose transfer;
Nova-RNA complex excision; proteinase K digestion
linker ligation; RT-PCR
B
AUG
> 10 kb
< 10 kb
UGA
121 intronic tags
38 exonic tags
68 intronic tags
58 3' UTR tags
55 unclassified tags

Figure 3.11: CLIP method and result described by Ule et al. (2003) A. Schematic of CLIP method: 1. In situ UV-crosslinking of the cells, 2. Cell lysis and RNase digestion,3. Immunoprecipitation, 4. Purification of target RNAs and RT-PCR; B. Genomic location of the tags. Tags belonging to genomic regions with no annotated transcripts were labeled as 'unclassified'.

The application of ultraviolet cross-linking and immunoprecipitation (CLIP) can be traced back to 2003 when Ule et al. first developed this new method (Figure 3.11 A) to identify the RNA targets of Nova, a mammalian tissue-specific splicing factor, in vivo. To get the transcripts which interact directly with the target protein Nova, brain tissue of mouse was irradiated by UV-B light in order to form covalent bonds between protein and RNA of direct contact. The target RNA, partially digested with RNase, was later co-purified with Nova by immunoprecipitation in the form of RNA-protein complex. Then, the Nova protein was removed by proteinase K so only the target RNA left. Finally, the RNA was cloned with linker ligation and reverse transcription (RT)-PCR.

1. **UV Crosslinking**

UV light radiation is used to 'freeze' the interaction between the protein and the RNA. It is considered as a 'Zero-length' crosslinking method. The in situ crosslinking by UV-B (later changed into UV-C) not only strengthens the connection between RNA and interacting protein in the formation of covalent bond, but also avoids the possible bias aroused from re-association of RNA-binding proteins after cell lysis (Mili and Steitz (2004)) like in RIP method (Keene et al. (2006)) without crosslinking. On the other hand, compared with chemical crosslinking, for example the widely-used formaldehyde cross-linking, it is unlikely that UV cross-linking will induce indirect binding event through protein-protein crosslinking. Meanwhile, chemical crosslinking might change the structural/folding of the protein and cause failure in capture of the RNA-RBP complex by the antibody. Meanwhile, UV-crosslinking have its own disadvantages of low efficiency and irreversible, but it is irreplaceable in CLIP method currently.

2. **Gel Purification**

   Another improvement on CLIP assay compared to RIP assay was the introduction of SDS-PAGE and membrane transfer. Unlike RIP which largely relay on the specificity of immunoprecipitation, only the RNA-protein complex of size similar to the target protein will be retained. In CLIP, Radioisotope-labeled RNAs in the protein-RNA complex were used to determine the position of the complex in radiography. But the application of the radioisotope also restricted the large-scale application of CLIP assay.

The UV-Crosslinking and SDS-PAGE was used together to determine the molecular weight of DNA-Binding Proteins (Wolf et al. (1995)). By crosslinking the radioactively labeled nucleic acid with the protein of interest, the position of the protein can be compared with the mark after running the gel.

Back to the research (Ule et al. (2003)), 340 Nova CLIP tags with average length of 71 nucleotides was sequenced by Sanger Sequencing. The overall tag number was quite small limited by the scale of Sanger Sequencing technique. The largest set of these CLIP tags (121) were within long introns over 10kb (Figure 3.11 B). Meanwhile, Tetramer frequency analysis comparing the Nova CLIP tag sequence to the known Nova binding element YCAY tetramer (Y = U or C), showing that each tag has 4.2 YCAY on average. The number in random sequence and another RNA binding protein was much smaller (Control=1.1, HU=1.7). The binding preference to region of multiple YCAY (especially the overexpressed YUCA) of Nova might be related to its regulation of alternative splicing.

3.3.2 Development of CLIP Method

3.3.2.1 Attached with High-Throughput Sequencing Method

The initial protocol of CLIP method attached with Sanger Sequencing was later replaced by high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP, Licatalosi et al. (2008)). Together with high-throughput sequencing, the CLIP method became even stronger. In five experiments by traditional CLIP strategies, 2481 Nova-bound tags in total were detected while 412,686 CLIP-tags were identified in HITS-CLIP, among which 168,632 unique tags left after filtering. The application of HITS-CLIP greatly increased the number of tags observed. With replicates and negative control, it was possible to distinguish between robust RNA-binding sites and noise. Nova HITS-CLIP extended not only the number of binding sites identified, but also the understanding of Nova-RNA interaction at transcriptome level. Some robust but atypical Nova-binding sites was detected unexpectedly, which point to the possibility of new mechanisms of Nova. For example, Nova binding near polyadenylation sites help recognize its role in regulation of alternative polyadenylation in brain tissue. Similar method named as CLIP-seq, coupling the modified CLIP workflow (Ule et al. (2005)) with HTS technique, was applied to another splicing regulator FOX2(RBM9) by Yeo et al. (2009).

In short, HITS-CLIP/CLIP-seq provide a general solution to the identification of direct protein-RNA binding sites transcriptome-widely in living cells, making it a powerful platform for research of RNA regulation in vivo.

3.3.2.2 Improvement in Efficiency and in Accuracy

Through the joint application of second-generation sequencing technology, researchers can quickly obtain large-scale data. Next, they turned to work on improving the efficiency and accuracy in the capture of RNA-protein interaction. With the spread of high-throughput sequencing, it cost less time and money to apply the technique. More improvements over the CLIP method appeared, and the growth of modified CLIP method spurt in 2015˜2017 (Imig et al. (2015); Sugimoto et al. (2015); Van Nostrand et al. (2016); Zarnegar et al. (2016); Rosenberg et al. (2017)).

The modification for better performance of CLIP method can be divided into two main genres, UVA-dependent PAR-CLIP (Hafner et al. (2010)) to improve the lower efficiency of crosslinking with UVC light and single-nucleotide resolution iCLIP (Huppertz et al. (2014)) to gain accurate crosslinking sites. On the other side, many CLIP-like methods are inspired by the CLIP method, targeting at other objects like RNA-RNA interaction or RNA modifications.

PAR-CLIP

Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) was first described by Hafner et al. (2010). The step of culture of cells with photoactivatable nucleoside 4SU was performed to facilitate RNA-RBP crosslinking. This step is much harder to apply in plant rather than in animal cells. Animal-related research often uses cell lines, such as the HeLa cell line, but plant-related experiments mostly use seedlings or tissues such as flowers and leaves. In vitro culture of plant cells is relatively rare. In some experiments, protoplasts are separated, that is, the cell wall of plant cells is removed with cellulase without harming other parts inside the cell wall, thus obtaining a structure similar to animal cells. Perhaps in the future, we will have the opportunity to see PAR-CLIP experiments with protoplasts. But so far, PAR-CLIP has not been applied to plants.

iCLIP

Here we use 'iCLIP' to represent all Single-nucleotide-resolution CLIP. In CLIP method, reverse transcription, which pass over the remained amino acids to the adaptor on the other side, is needed. However, premature termination of reverse transcription happens near the cross-linking nucleotide on the majority of cDNAs, making a large amount of crosslinking site indetectable in CLIP-seq result. For Improvement of this limitation and also to achieve single-nucleotide resolution by the truncated sites, a second adaptor ligation step via self-circularization was induced after the reverse transcription to the $3'$ end of cDNA (König et al., 2010). The end point of reverse transcription should be the 'crosslinking site' in most cases. To quantify the fragment detected, random barcodes were included in the adaptor, attached to the $3'$ of cDNA sequence in the second ligation. The barcode is used here as unique molecular identifier in removal of PCR duplicates in data analysis later. In conclusion, iCLIP method gains single-nucleotide-resolution through:

1. premature termination of reverse transcription
2. the two-step adaptor ligation:
    - first step: add adaptor to $3'$ end of RNA fragments;
    - second step: Circularization for adaptor ligation to the $3'$ end.

Like the original CLIP method, the wide application of iCLIP method was also limited by the radiography step. Data used in this project were produced by enhanced Crosslinking and Immunoprecipitation (eCLIP) method. It is a very comprehensive experiment with strict steps of purification. The RNAs should be protected with RNase inhibitor. Here's the pipeline modified by Qiqi based on the method applied in (Zhu et al., 2020a) and the principle of the experiment was shown in Figure 3.12.

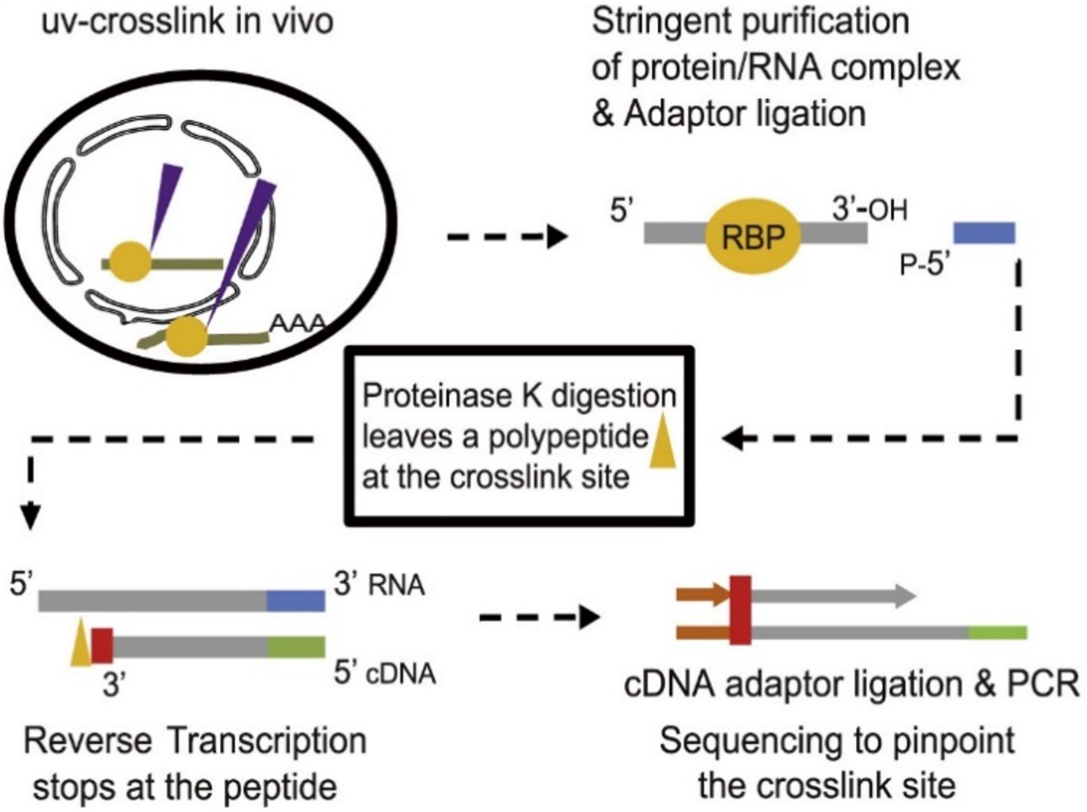

uv-crosslink in vivo
AAA
Stringent purification of protein/RNA complex & Adaptor ligation
5'
RBP
3'-OH
P-5'
Proteinase K digestion leaves a polypeptide at the crosslink site
5'
3' RNA
3'
5' cDNA
Reverse Transcription stops at the peptide
cDNA adaptor ligation & PCR
Sequencing to pinpoint the crosslink site

Figure 3.12: Diagram illustrating the principle of eCLIP-seq by (Zhu et al., 2020a): UV-crosslinking results in formation of covalent bonds between RNA and RBPs. The RNA-RBP complexes would go through stringent purification steps and the immunoprecipitation step to get the complex with target Protien and the RNA it binds. The adaptor of 3' end was ligated to the RNA before Proteinase K digestion of the protein. Due to the exsitance of polypepitide at the crosslinking site, the following reverse transcription would stop at the pepited. After cDNA adaptor ligation, the cDNAs would go through PCR and a final gel purification step and send to be seqeucned.

1. 10-day-old seedling are first radiated with UV-light for crosslinking, to form bonds between interacting RNA and proteins. The material is then frozen and grinded. After cell lysis, RNase I together with Turbo DNase is added to the product to remove the DNAs while cutting the RNAs into smaller fragment. The RNAs with protein attached are protected by the proteins, so they won't be easily digested. On the other hand, pieces too small in size are not easy to detect and might further degrade in the later steps, so the digestion should not happen for too long. After 3 min of RNase

treatment, the tube is put on ice and mixed with Ribolock RNase Inhibitor to protect RNAs. After centrifugalizing, the supernatant is removed to a new tube, with 70 μl separated for the input.

2. Next comes the immunoprecipitation (IP) step by the magnetic beads coupled with antibody. The Beads are incubated with the sample at 4 ℃ for 2h.

3. The beads are washed once with lysis buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 1% NP-40, 0.1% , 0.5% Sodium deoxycholate), twice with Cold High Salt Buffer (50 mM Tris-HCl pH 7.5, 1 M NaCl, 10 mM EDTA, 1% NP-40, 0.1% SDS, 0.5% Sodium deoxycholate), twice with Cold Wash Buffer (20 mM Tris-HCl pH 7.5, 10 mM $MgCl^2$, 0.2% Tween-20), once with Fast AP buffer (10 mM Tris-HCl pH 7.5, 5 mM $MgCl^2$,100 mM KCl, 0.02% Triton X-100). All these steps are carried out at 4 ℃, for each wash the tube should be rotated for 2 mins. After the wash, beads are transferred to a new tube.

4. The $3'$ adaptor is ligated on-beads with 5 µL adaptor added to the sample, incubating under 23 ℃ for 75 mins. The beads are further washes to remove the extra adaptors after $3'$ ligation. By now, the target RNAs are still attached to the beads.

5. Treated with heat, the protein will denaturalize and release from the beads. The sample will then be split with 26.5 µL loaded to the RNA gel and 3.5 µL for western blot. Run the two gels together at 150 V for 75 mins. The input without IP is also loaded, 50 µL to RNA gel and 20 µL for western blot. After the run, transmembrane is applied to the gel at 120 V for 90 mins. In western blot, bands are dyed to indicate the size of proteins by their position on the membrane. Running at the same time, we can locate the target protein by western blot and get the part on and above the corresponding band on the membrane transferred from RNA gel for both input and IP sample.

6. After cutting the target band from, Proteinase K is used to digest the protein and to release RNAs from the membrane. The released RNAs are first purified with a mix of phenol-chloroform, vortexed and centrifugalized. The supernatant is retained and transferred to a new tube. Further purification is carried out on Zymo column.
7. The purified RNAs are reverse transcribed to get cDNAs. MyOne™ Silane beads are applied to bind RNAs in the product thus cleanup cDNAs.
8. The $5'$ linker is ligated to the purified cDNAs. Again, MyOne Silane beads are used but this time to bind cDNAs. After washing beads with 75% EtOH, cDNAs are eluted from the beads and transferred to a new tube.
9. cDNAs, with both $5'$ and $3'$ linker ligated, then go through PCR amplification, followed by SPRI cleanup step to remove primer in the product.
10. The final step before sequencing is Gel-purification step along with quality check for cDNA library with Nanodrop for density and QSep for size distribution.

We could see from the protocol that the radioactive-label was replaced by western blotting to check the size of the protein in eCLIP, making the experiment easier to apply. Also, no circularization was included in the adaptor ligation, adding the risk of double ligation of the adaptor, which should be taken care of in the quality control of the reads.

Apart from the two types, there also appeared some other CLIP-like methods.

Other CLIP-like method

- Crosslinking and Affinity Purification (iCLAP): iCLAP was used parallel with iCLIP in the research of TIA-RNA interaction (Wang et al. (2010)). It is an antibody-free method to enrich RNA-protein complex by stringent affinity purification. It seems to have lower efficiency RNA-protein complex than iCLIP in purification of TIA-RNA complex.

- RNA hybrid and individual-nucleotide resolution ultraviolet crosslinking and immunoprecipitation (hiCLIP, Sugimoto et al. (2015)) is a CLIP method inspired by CLASH. Unlike most other CLIP which deals with the position on the transcriptome which protein binds, hiCLIP is a biochemical technique for transcriptome-wide identification of RNA secondary structures interacting with RBPs.
- miCLIP: With the application of m^6A-specific antibody, CLIP method can be used to get the exact position of RNA modification as well (George et al. (2017)).

3.3.3 Analysis of Different Types of CLIP-seq Data

Since we used eCLIP-seq data in this project, we will focus on the analysis of the two similar methods of iCLIP/eCLIP with single-nucleotide resolution in this part.
In a general CLIP-seq data analysis workflow, we usually need the steps bellow:

1. Quality control and adaptor trimming
2. Alignment to the reference genome
3. Removal of PCR replicates by the random barcode
4. Peak-calling: Define the significant binding sites with the mapped reads

After we get the peak information, we may further apply some downstream analysis, including:

1. Annotation: Define which gene, what kind of feature does the binding site belongs. The position of binding would largely reflect the possible function of the protein.
2. Motif finding: finding motif of binding region or around the binding site

The QC and alignment of CLIP-seq data is similar to the one of common RNA-seq, but the reads are smaller in size. Here, we used fastp for QC, Cutadapt for trimming. The alignment can be applied to the one side of read with crosslinking site or both to side of reads but with the usable read left for downstream analysis. We map reads of both side with hisat2 and keep only read2 of mapped pairs for eCLIP-seq by Samtools.

3.3.3.1 Differences in iCLIP and eCLIP Data Analysis

The analysis of iCLIP-seq and eCLIP-seq are similar but differ in several features:

- The start site of read1 is considered as RBP-binding site in iCLIP-seq while in eCLIP-seq the site is supposed to be the first base of read2
- The ligation of adaptor iCLIP is followed by a circularization step, but the end of eCLIP is still accessible after the ligation so the second (even third) ligation might happen. In the analysis of eCLIP data, Van Nostrand et al. (2016) suggested that the removal of adaptor by Cutadapt should be carried on twice to avoid the case of double ligation of the adaptors. In their workflow, they perform the command twice with each time removing the adaptor once. But we've found that if we set the parameter -n/--time to 2, which means remove the adaptors to up to twice, the result should be no different. We tested the method with our IP data of Col-0. The raw data has 9623420 pair of reads, and 7533432 pairs left after a primary PCR duplicate removal step by sequence. In the first method which applies Cutadapt twice, 7479219 pairs left after the first trimming, and 7463824 pairs left after the second trim. If we applied Cutadapt with -n 2, the final result is exactly the same of the one of performing the -n 1 twice. To summarize, parameter "-n 2" should be applied in Cutadapt to remove the potential double ligation.
- The last but not the least important, only eCLIP-seq have 'input' in the experiment.

  CLIP was considered as an experiment to identify RNA-RBP interaction sites with barely no background. Unlike RIP, which should work together with the corresponding input file to remove the background noise, CLIP method worked without input before the birth of eCLIP method. The Size-Matched Input (SMInput) is produced parallelly with the IP sample but without immunopre-

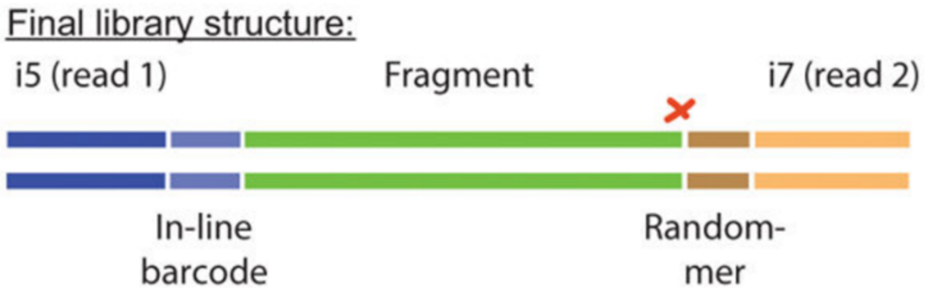


Figure 3.13: Structure of cDNA library in eCLIP experiment (from Van Nostrand et al. (2016)): RNA is prepared into a paired-end high-throughput sequencing library. In the final result of sequencing, Read1 should begin with the in-line barcode and Read 2 starts with a random-mer sequence which was added during the 3′ DNA adaptor ligation. Following the random-mer, these exists a sequence corresponding to the 5′ end of the original RNA fragment (which usually marks reverse transcriptase termination at the crosslinking site (red X)).

cipitation. In other word, the RNAs which the target protein binding are not enriched in SMInput. In the gel purification, the size of complex recycled is restricted to the region near the target protein in both IP and input samples, this is why it was called ‘Sized matched’ since the size of the protein in the input matched with the size of the target protein. When dealing with eCLIP-seq data, we can either use the input or not. Most tools originally designed for iCLIP analysis done not include SMInput in analysis, while some were developed for eCLIP like PureCLIP, so the input can be used.

3.3.3.2 iCLIP/eCLIP-seq Analysis Tools

Although the experimental procedures and the underlying principle are drastically different, CLIP-seq resembles common RNA-seq in many ways. We can learn from the workflow in the last section that, in most analysis steps, including quality control, adaptor removal and mapping, the process is the same between the two experiments. While in eCLIP (the structure of cDNA library shown in Figure 3.13), there exist some unique steps. We will talk about the process of random barcode, peak-calling and the downstream analysis of motif finding in this part.

Remove PCR duplicates by UMIs

A random barcode, within the DNA adaptor, is ligated to the $3'$ end of cDNA fragment. The barcode will not be considered as a part of adaptor in sequencing by synthesis, so it will remain in the finial FastQ read, at the beginning of Read2 (Figure 19). This adaptor ligation is done before PCR amplification, which means the barcode can be later used as a marker to identify PCR duplicates. These sequences actually serve as unique molecular identifier (UMI) of the reads. The removal of PCR replicates by UMIs is a very important step in the analysis of eCLIP. Generally speaking, there are two strategies of using the UMIs:

- The first one depends only on the sequence of the reads and should be applied before alignment. Theoretically, PCR replicates should be of same sequence and with same UMI, so we can detect them by sequence. But the truth is that mistakes induced in replication of cDNAs might cause the replicates to end up with sequences of slight difference. In comparison of the sequence, mistakes must be allowed. This method is recommended in processing NET-seq data with 4-nt random sequence on both size of the insert. The tool first merges similar reads into clusters and further remove reads with exactly same sequences as PCR duplicates and only the one of the best qualities would be kept for later alignment. After the removal of PCR duplicates, the random-mers have fulfilled their duty and will be removed by Cutadapt.
- In the second method, not only the sequence of UMIs but also the position of the read on the genome are taken into consideration. UMI is not part of the genome, so it should be removed before alignment. To retain the UMI sequence for later usage, the UMIs are removed but added to the read ID, the information of which would be fully remained even after alignment, so that we can identify the PCR duplicates with both

start position of the read and the sequence of UMI after alignment. This method can remove the replicates more efficiently than merely use the sequence. It is the more recommended way of UMI usage. We can use UMI-tools or Gencore to do it.

UMI-tools: UMI sequence, usually added to the start of reads, are of high error rate in sequencing. To account for the errors, UMI-tools introduce network-based methods in identification of PCR duplicates. It has been tested on both real iCLIP and single cell RNA-seq datasets.

Gencore: Work together with fastp (with parameter –umi), Gencore is a fast and easy-to-use tool written in python. The development team also provide the compiled version to make it user-friendly. We can apply the tool to mapped Bam files which was pre-processed with fastp, with UMI attached to read ID.

Peak-calling

Peak-calling is a computational method to identify areas with enriched aligned reads in ChIP-seq or similar methods. Here in eCLIP/iCLIP-seq, we use peak-calling tools to define significantly binding regions as peaks from mapped BAM/SAM.

PIPE-CLIP: PIPE-CLIP (Chen et al. (2014)) is one of the most commonly used CLIP-seq data analysis tool. It was designed for not only HITS-CLIP but also PAR-CLIP and iCLIP/eCLIP. It is available on galaxy for web-based analysis. To identify enriched peaks, the adjacent mapped reads are clustered together if they overlap each other by at least one nucleotide. The clusters are used for further analysis. Given that all clusters receive at least one read, researchers propose a model equipped with the zero-truncated negative binomial (ZTNB) likelihoods. It reports cross-linking regions with high reliability. A script (barcodeRemover) to remove barcode and PCR duplicates for iCLIP is also provided.

Piranha: Like PIPE-CLIP mention before, Piranha (Uren et al. (2012)) could also be used for most variation of CLIP experiments. For finding differentially used binding sites between samples, Piranha uses read counts in the first tissue or condition as a covariate of the second. It also uses a ZTNB regression model when covariates are provided and otherwise uses a ZTNB model for finding the enriched peaks. Binding motifs are identified using the distributed mutual exclusion (DME) algorithm.

iCount: The iCount tool (Curk et al. (2019)) was used in iCLIP-seq data analysis the RZ-1 paper by Zhu et al. (2020a). It is a Python module and associated command-line interface (CLI), which provides all the commands needed to process protein-RNA iCLIP interaction data and to identify and quantify sites of protein-RNA interactions on RNA. Unlike most peak-calling tools, main inputs of iCount are FASTQ files with iCLIP sequencing data, which means it can also do trimming and mapping. Its main output are BED files with identified and quantified cross-linked sites. But iCount was designed for iCLIP so it cannot process peak-calling step with SMInput.

CLIPper: CLIPper (Lovci et al. (2013)) was first introduced as a python tool for definition of HITS-CLIP peak. But it was later applied to newly developed eCLIP-seq data in the paper. Before using CLIPper, we'll have to generate mapped reads with an aligner which is tolerant of spliced reads. The developers recommend users to use GSNAP for this since all testing is done with GSNAP-derived mapped reads. Since the expression of RNA transcripts differs gene by gene, the significance of a peak will be defined based on the read number and length of the gene. They may also introduce module for different binding RNAs (mRNAs/pre-mature RNAs) which might be helpful. But the tool only has inbuilt genome of hg19(human) and mm10(mouse) so it cannot be applied to data of other species easily.

dCLIP: This tool helps users to find differentially bound regions in two HITS-CLIP/PAR-CLIP/iCLIP experiment (Rosenberg et al. (2017)). It is written in Perl to identify common or differentially bound regions with an MA-plot normalization. In detail, iCLIP datasets must be single-stranded with barcodes trimmed before alignment. The developers also provide the user a supplementary Perl script, remove_barcode.pl to trims barcodes.

PureCLIP: PureCLIP (Krakau et al. (2017)) is a tool developed for iCLIP and eCLIP data analysis. It takes reference genome along with sorted Bam files with index as inputs, and can process eCLIP together with its SMInput. The outputs of PureCLIP are BED files of crosslinking sites and binding regions. One of the important parameters in PureCLIP '-bc' (0 or 1). The default value is 0, indicating that the RBP binds to short specific motifs like PUM; otherwise, the value should be set to 1 with proteins of larger crosslink clusters and lower read count level. Setting '-bc' to 1 will increase the number of peaks detected. The one thing inconvenient is that PureCLIP only takes A/T/G/C/N as effective bases, so the degenerate bases in the reference genome should be replaced by Ns before use.

Motif-finding tools

Many RBPs have preference for binding certain sequences. We can discover the binding motif of a certain RBP by analysis on sequences near the peaks.

MEME: MEME Suite (Machanick and Bailey (2011)) is designed to discover novel motifs and to perform downstream analysis based on motif analysis. MEME-ChIP is recommended for comprehensive motif analysis of large datasets from CLIP-seq experiments. With extract sequences in fasta as input, the tool would help us discover the motifs in the central regions of the input sequence and compare the enriched sequences to known motifs. The MEME-ChIP supplies both web service and common-line tools. It employs two motif discovery algorithms, expectation maximization (EM) to discover possible binding model and TOMTOM algorithm to compare the discovered motif to a known motif dataset.

HOMER: HOMER (Hypergeometric Optimization of Motif EnRichment) is a suite of tools for Motif Discovery and next-generation sequencing systems written in Perl and C++ by Heinz et al. (2010). It was primarily written as a denovo motif discovery algorithm and is well suited for finding 8-20 bp motifs in large scale genomics data. HOMER contains many useful tools for analyzing ChIP-Seq, GRO-Seq, RNA-Seq and numerous other types of functional genomics sequencing data sets.

Zagros: Zagros (Bahrami-Samani et al. (2015)) is a motif discovery software for CLIP-Seq data. The tool can characterize the binding site for the given RBP within given regions of enriched reads. It also contains two additional programs one for calculation of the base pairing probabilities of the input sequences extracting experiment specific events to incorporate such information for an extremely accurate motif discovery.

Summary for eCLIP analysis

In our research, the tools used were fastp and Gencore for PCR duplicates removal, PureCLIP for peak-calling and Homer for motif finding (see Figure 3.14).

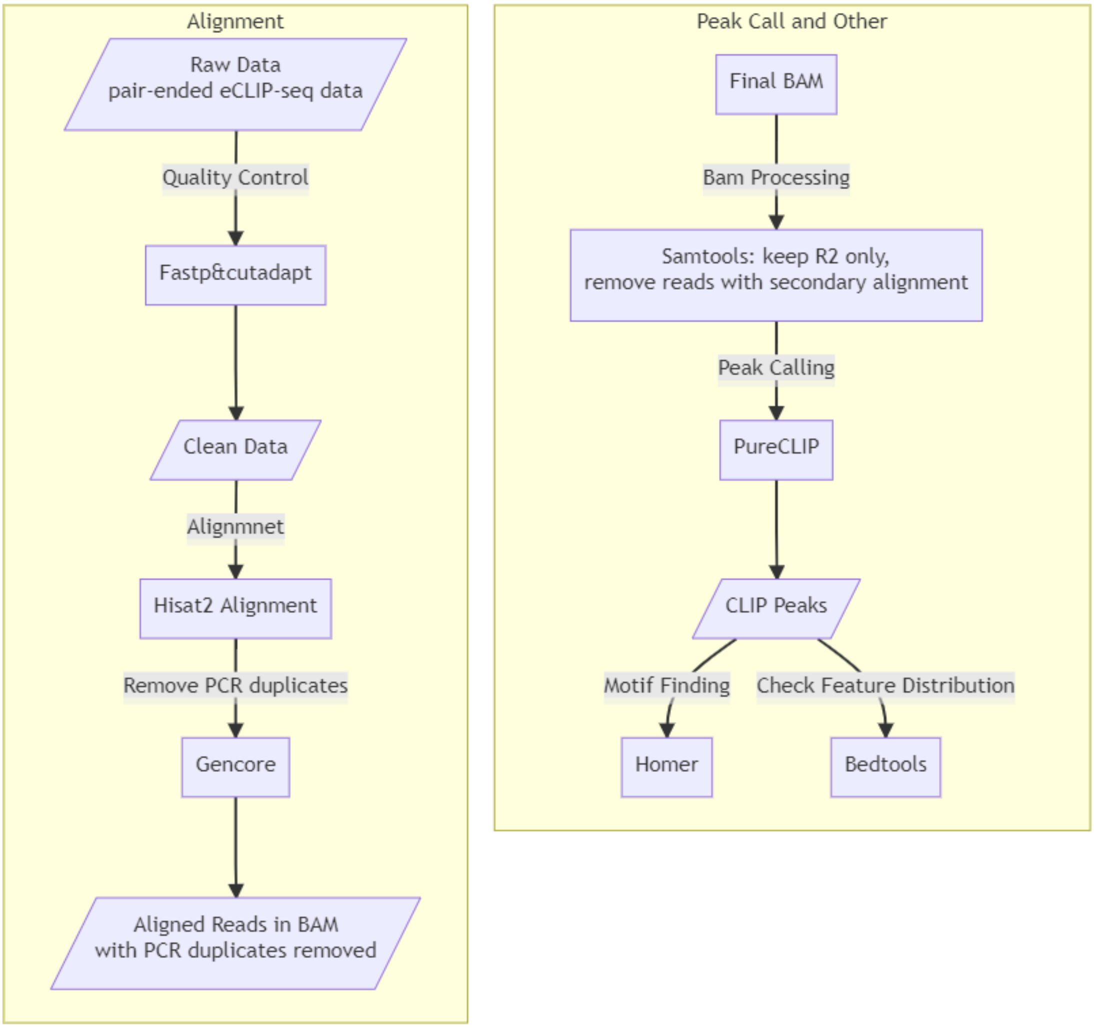


Figure 3.14: Summary for the current eCLIP data analysis workflow from raw data to FCA-binding peaks: Fastp&Cutadapt for quality control, Hisat2 for alignment, Gencore for UMI processing, PureCLIP for peak calling.

4 General Binding Feature of FCA as an RBP

This chapter presents the basic binding feature of FCA. It would start with the structure of FCA protein for the possible binding preference and the potential interaction between FCA and other proteins as revealed by IP-MS (ImmunoPrecipitation-Mass Spectrometry), a technique for detecting protein-protein interaction. To test the RNA-binding features of FCA in vivo, we used eCLIP data for identification of the binding site in not only wild type Col-0 but also the mutants of two other proteins which belong to the autonomous pathway, FPA and FY, to see the possible impact of these two proteins on FCA binding. The connection between FCA and the N^6-methyladenosine (m^{6}A) would also be investigated into by checking the binding of FCA under the mutant of a m6A reader *fip37-4*. To summarize, we would like to describe the overall binding feature of FCA as an RBP in vivo with CLIP-seq data in this chapter, including preference in genomic feature and binding motif. In addition, we would also describe the other possible factors of FCA recruiting like RNA modification and interaction with other proteins. The eCLIP cDNA library used in this chapter was constructed by Dr. Chen and Qiqi of Wu Lab (SUSTech), data analysis was completed by me. The mutants used are described in Appendix A

4.1 Structure of FCA and Other Interacting Proteins

As described in 2.3.1, FCA is a typical RNA binding protein with two RRMs and a WW domain. I tried Homology modelling with SWISS-MODEL (Waterhouse et al. (2018)) to check the structure of FCA. (see Figure 4.1 A). Both of the domains show a typical structure of RRM with two α-helices and four β-sheets ($\beta\alpha\beta\beta\alpha\beta$). As predicted, the two RRMs of FCA shown highly similarity with the RRMs in ELAV-like family member 2 (ELAVL2), which is consistent with the research before (Macknight et al. (1997)), indicating that we may get a similar binding motif from FCA-CLIP data to the one of ELAVL protein research. Unfortunately, the modelling only helps us get the possible structure of the RRMs, the structure of other region of FCA was not modelled.

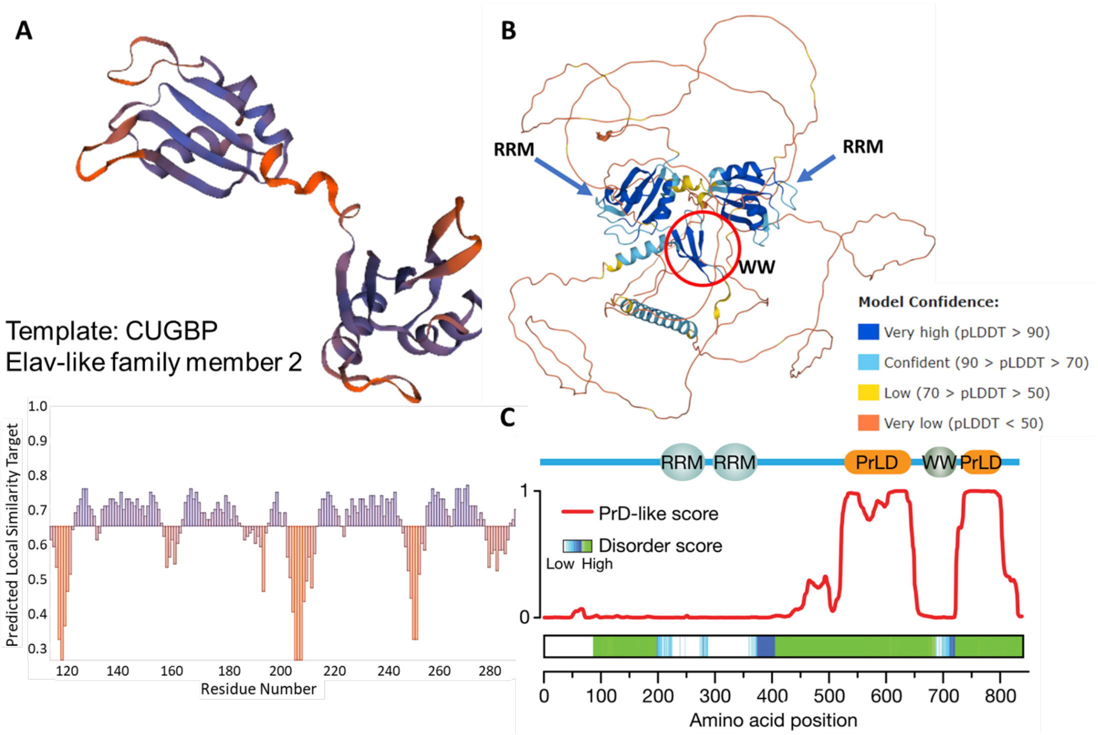


Figure 4.1: Structure of FCA protein: A. Predicted structure of FCA RRMs by SWISS-MODEL, taking ELAVL2 as template. The RRMs of FCA shows highly similarity to RRM1&2 of the template. Color of the residues represent the similarity score between the target and the template, with region of high confidence shown in blue and regions of lower confidence shown in red; B. Structure of full-length FCA predicted by AlphaFold2 (Jumper et al. (2021)), downloaded from the website(https://alphafold.ebi.ac.uk/entry/Q6K271); C. Position of the PrLD and disorder region in FCA (predicted in Fang et al. (2019)).

In 2021, the structure of full length FCA can also be predicted with AlphaFold2, thus it is available on the site (https://alphafold.ebi.ac.uk/entry/Q6K271). We downloaded the pridicted structure of full-length FCA from the website (see Figure 4.1 B). Judging from the model, the confidences of the two RRMs and the WW domain are very high, while the confidences of most other regions are low. The WW domain is short but the structure of 3 parallel *beta*-sheets are clearly shown in the model (Figure 4.1 B). On the other hand, C-terminal of FCA seems highly disordered. It was recovered in the previous research that the C-terminal region was composed of two prion-like domains (PrLD) of high disorder score (Figure 4.1 C), which might be a reason for the low prediction confidence of these regions. Despite low confidence in prediction of structure, we have confidence in the possibility that the low complexity sequences might trigger formation of nuclear bodies. In previous study, it has been proved that FCA function requires liquid-liquid phrase separation (Fang et al. (2019)). It is reasonable to assume that the multivalent interactions among FCA protein itself would affect the function of FCA, specially by affecting its binding with RNAs. Proteins interacting with FCA can be detected in the IP-MS experiment. In the experiment carried out by Fang et al. (Fang et al. (2019)), the unique peptides of 63 proteins were detected in the experiment, while only 12 of them (including FCA) appeared in both replicates. This indicated that the interaction between FCA and other proteins might be dynamic and transient, making it hard to detect by the traditional method.

To address this issue, a method called crosslinked nuclear immunoprecipitation and mass spectrometry (CLNIP-MS) was proposed in the same research. Unlike the traditional method without crosslinking, formaldehyde-crosslinking was applied before immunoprecipitation to fix the transient interaction of proteins within the nuclear body. To be more specific, the proteins detected in this CLNIP-MS might be indirectly linked to the target protein through their interaction with directly-binding proteins or nuclei acids. Another list of proteins interacting with FCA in vivo has been detected by CLNIP-MS. Not surprisingly, the CLNIP-MS captured more potential interacting proteins with unique peptides of 134 proteins detected and 92 of them appeared in all three replicates.

According to the published data (Fang et al. (2019)), the number of proteins detected in both IP-MS without crosslinking and CLNIP-MS was rather small. Theoretically, the proteins detectable in IP-MS should be also detected in CLNIP-MS, but it seems not the case in the two experiments with FCA. Besides FCA, only three other proteins were captured in both experiments. The genes are listed in Table 4.1 and proteins which were known to have interaction with FCA in research before, such as FY, were only detectable in CLNIP-MS. The result might indicate that the interaction between FCA and some proteins might be transient and hard to detect without crosslinking.

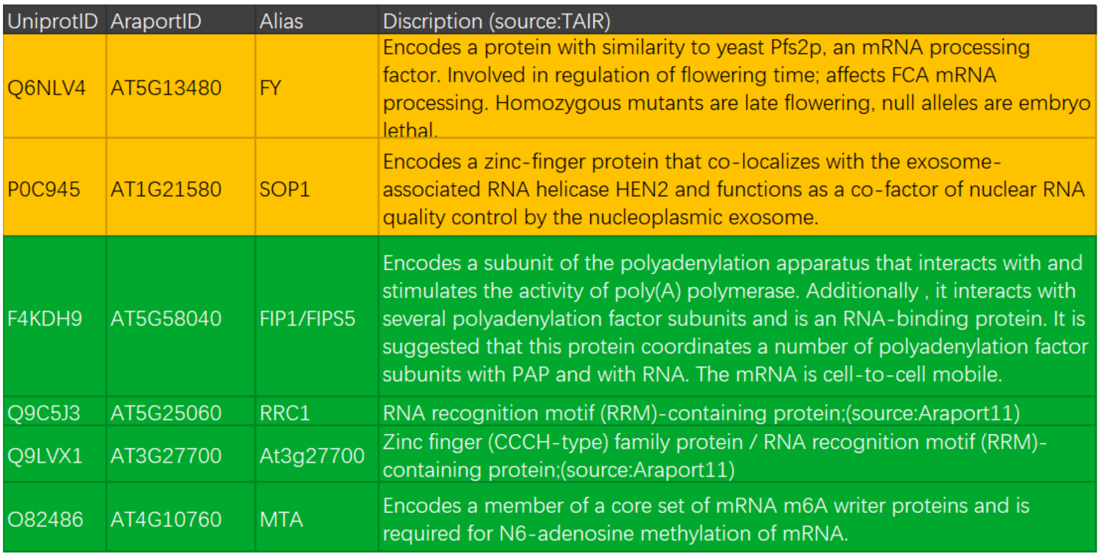
Table 4.1: Proteins detected in both IP-MS and CLNIP-MS experiments (Annotated by Metascape)

Gene ID	**Gene Symbol**	**Description**
AT4G38680	GRP2	glycine rich protein 2
AT5G52090	AT5G52090	P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT3G13470	Cpn60beta2	TCP-1/cpn60 chaperonin family protein

Besides the $3'$ processing factors (FY, FPA, CPSF100, etc), we noticed that homolog of a core subunit of yeast chromatin remodeling complex AtSWI3B was captured as well as AtSWI3A in the CLNIP-MS result. And the interaction between FCA and AtSWI3B has been recovered in the previous research (Sarnowski et al. (2002)). It might be an evidence of FCA's binding to pre-mRNA RNA as FCA seems to work at the place very close to the chromatin through interaction with AtSWI3B. It is hard to consider that it functions on mature RNAs that have been released from the chromatin.

Inspired by the interaction between FY PPLPP motif and FCA WW domain, we searched for all the proteins with PPLP/PPLPP motifs in Arabidopsis. The PPLP/PPLPP motif-containing proteins which also appear in CLNIP-MS of FCA might directly interact with FCA. Besides FY, we found five more proteins (see Table 4.2) with the motif and meanwhile detected in the MS result.

Table 4.2: PPLP/PPLPP-motif containing protein detected in CLNIP-MS, with PPLPP-containing protein marked yellow and other PPLP-containing proteins marked green



UniprotID	AraportID	Alias	Discription (source:TAIR)
Q6NLV4	AT5G13480	FY	Encodes a protein with similarity to yeast Pfs2p, an mRNA processing factor. Involved in regulation of flowering time; affects FCA mRNA processing. Homozygous mutants are late flowering, null alleles are embryo lethal.
P0C945	AT1G21580	SOP1	Encodes a zinc-finger protein that co-localizes with the exosome-associated RNA helicase HEN2 and functions as a co-factor of nuclear RNA quality control by the nucleoplasmic exosome.
F4KDH9	AT5G58040	FIP1/FIPS5	Encodes a subunit of the polyadenylation apparatus that interacts with and stimulates the activity of poly(A) polymerase. Additionally , it interacts with several polyadenylation factor subunits and is an RNA-binding protein. It is suggested that this protein coordinates a number of polyadenylation factor subunits with PAP and with RNA. The mRNA is cell-to-cell mobile.
Q9C5J3	AT5G25060	RRC1	RNA recognition motif (RRM)-containing protein;(source:Araport11)
Q9LVX1	AT3G27700	At3g27700	Zinc finger (CCCH-type) family protein / RNA recognition motif (RRM)-containing protein;(source:Araport11)
O82486	AT4G10760	MTA	Encodes a member of a core set of mRNA m6A writer proteins and is required for N6-adenosine methylation of mRNA.

There also exist other proteins without PPLP-motif but contain a Pro-rich subsequence, such as FPA and FIP1, which might also interact with WW domain of FCA. Whether these proteins interact with FCA directly in vivo remains to be tested.

To summarize, the focus of this research, FCA, is a typical RNA-binding protein with two RRMs. The binding of FCA in vivo might also be affected by the interaction between other proteins and its WW-domain, although these interactions might be transient. The function of FCA is related to the 3′ processing machinery, together with FY, FPA and many other proteins. To study the function of FCA in 3′ processing and termination as an RBP, FCA-CLIP was applied to explore the general binding feature of FCA. To further understand the connection between FCA and FY/FCA, we applied not only CLIP-seq in Col-0 background but also *fpa-7* and *fy-2*.

4.2 Transcriptome-wide Binding Feature of FCA in Col-0

To study the binding feature of FCA, we first carried out CLIP experiments under the Col-0 background, using *fca-9* mutant without functional FCA as the negative control. The binding in negative control sample is considered as non-specific binding with the corresponding regions in the Col-0 shaded in the later analysis. In CLIP-seq analysis of this section and the two sections following, we would first call peaks of significant binding and check the feature distribution the peaks as a start point of studying the general binding feature. Second, we looked into the typical genomic features of binding to further looked into the binding in detail.

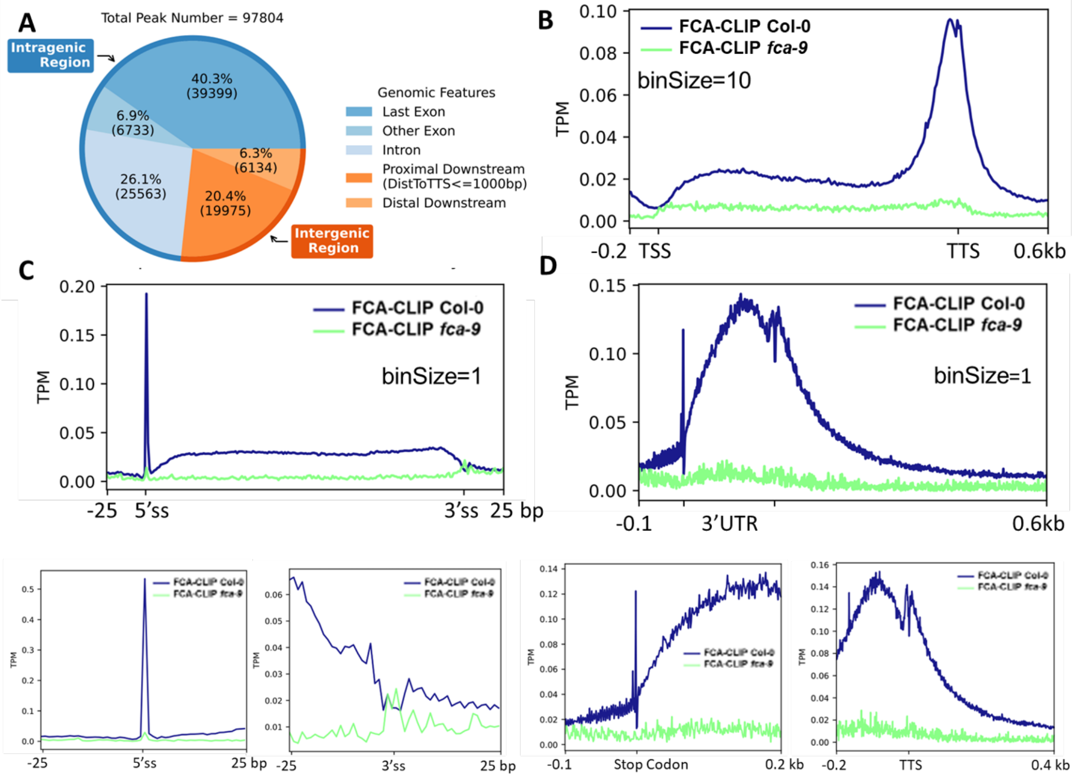


Figure 4.2: Binding Pattern of FCA in Col-0: A. Pie Chart of Genomic Feature Distribution of FCA-eCLIP Peak in Col-0; B. Meta-profile of FCA-binding on gene body; C. Meta-profile of FCA-binding on intronic region, up: with all introns scaled to the same length, down: the specific binding pattern around splicing sites; D. Meta-profile of FCA-binding near 3′ end.

4.2.1 Meta-Profile of FCA's Binding

To localize the genomic features of binding, we first applied peak-calling step on both Col-0 and *fca-9*. The peak regions in *fca-9* would be removed from Col-0. Seen as significant binding sites, the distribution of peaks on the genome could largely represent the overall binding preference of genomic features of FCA. With the genomic features of significant binding, we can further check the binding of FCA in detail.

4.2.1.1 Distribution of FCA-Binding Peaks

Peak-calling was carried out with PureCLIP as described before. We annotated the peaks with bedtools (Quinlan and Hall (2010)) and the distribution of peaks is shown in the pie chart above (Figure 4.2 A). To avoid counting a peak mapped to multiple features repeatedly, we set priority for annotation. Usually, the intragenic region is of higher expression level than intergenic region, while within intragenic region, the half-life of exons should be longer than the one of introns. If a peak is annotated to multiple genomic regions, the rule of precedence should be >intron >intergenic. In this figure, the blue parts represent intragenic region while the orange parts represent the intergenic region.

The genomic feature with the most abundant FCA-binding peaks was 'Last Exon'. In genes without intron, last exon is also the only exon. Among 8360 transcripts without intron, only 936 were detected to have FCA-binding peaks in Col-0 and the binding of FCA on these genes also tend to be near the $3'$ end. We can conclude that FCA-binding peaks mostly located on the $3'$ end of the genes.

Meanwhile, we were surprised to see 25.6% of the binding peaks are located in intergenic region. Usually, intergenic regions are covered with very few reads in mRNA-seq data. The binding to intergenic region might be an indication of FCA binding to pre-mRNAs rather than mRNAs.

Intron-binding feature of FCA was also very specific. Just like intergenic regions, reads in introns were rarely detected in mRNA-seq or even CB-RNA-seq as a result of efficient co-transcriptional splicing. This was another strong indication for FCA binding to pre-mRNAs. On the other hand, the existence of introns in mRNA are usually transient, which made us wonder whether these FCA-targeted introns are of lower splicing efficiency or the binding might be an indication of retained intron. We would try to answer this question in Chapter 5.

A possible explanation for binding to intergenic and intronic regions is that these FCA-targeted transcripts actually decay before polyadenylation, making them undectable at mRNA level. But the fate of these FCA-binding RNAs is unknown limited by current technique.

While peaks helped us localize FCA-binding genomic features, there existed certain limitations if only CLIP peaks are used in the analysis. First, peaks can be found only on genes with high CLIP read coverage. The restricted peak-calling method resulted in the small size of CLIP peak. To show the overall binding pattern of FCA on larger amount genes, we use filtered 'crosslinking sites' rather peaks. Here, crosslinking sites were defined as the first site of mapped Read 2, with mapping quality (MapQ) over 20 (to remove reads with secondary alignment) and with PCR duplicates with same start site and same barcode removed. As introduced in Chapter 3, reverse transcription of the RNA crosslinking with RBPs usually stopped at crosslinking site of the protein, whereas the beginning of Read2 can be considered as the crosslinking site of RNA to the binding protein. After we get the position, more detailed profiles can be plotted on the region with enriched FCA peaks. Secondly, raw reads are also informative in some cases, especially reads across splicing junctions. The raw reads might help us reveal whether the binding happens before or after splicing. Thus, we might apply raw CLIP read in the analysis of binding towards intron regions.

4.2.1.2 Binding of FCA on Gene Body

We first check the over-all binding pattern of FCA on gene body (Figure 4.2 B). The intronic binding of FCA is not likely to be revealed in this part. But the preference for the $3'$ end might be significant within gene body.

In this part as well as the other meta-profiles in this chapter, we used the deeptools for generation of BigWig files and plotting meta-profile with the files. Only the first base of the mapped read was used for read-counting with the parameter 'offset=1'. Since PCR duplicate removal had been applied in the previous steps, there was no need to perform the step again in counting. Reads were also detected in the negative control group, indicating that there might be non-specific binding of the antibody even in the absence of FCA. However, we noticed that only a few of regions on the chromosome are of high coverage in *fca-9*. So, instead of shading all the regions with reads in negative control, we remove only the outliers with extremely high coverage. These regions would be extracted and set as 'Blacklist' in plotting. We did not filter the regions by the coverage of FCA-CLIP read in Col-0, so most of the regions with binding would be covered.

Below are the technical details in using Deeptools (Ramírez et al. (2016)):

1. In generation of bigwig files by bamCoverage, parameter 'offset' was set (for CLIP offset=1, for $3'$ mRNA-seq offset=-1), to ensure the single-nucleotide within analysis.

2. Correct mode for different cases in computing of expression matrix should be picked. In the scale-region module, regions of unequal length are scaled to the same. This is useful in plotting CLIP reads on a certain genomic feature, for example gene body or intron. Here, we set the length by the median or average value if the length of region (See Table 2.2 for the values). For example, the average transcript length is 2212 and the median is 1921, then we picked 2000 as the length for scaling in plotting read density on gene bodies. As for introns, 200 was picked for similar reason. In some other cases, the reference point module might be applied. The region before and after the reference point would be used without scaling. Unlike the first mode, no normalization for region size is need, making the figure of higher

resolution. We would use this mode to look at the exact position of binding with single-nucleotide resolution.

3. Be careful with zero values. With '--skipZero' parameter, regions with no read covered in both experimental group and negative control are removed while regions with reads in negative control group only will still be retained. Even though we set a blacklist for regions with high coverage in negative control, it was not ensured that all the region plot were with no reads of FCA-CLIP reads in *fca-9*. Take gene bodies for example:

Table 4.3: Genes with FCA-CLIP peaks

Sample	**Total**	**Zero**	**%Zero**
Col-0	23742	1230	5.18%
fpa-7	25067	1010	4.03%
fy-2	24856	1055	4.24%
fip37-4	24646	1080	4.38%

About 4˜5% (see Table 4.3) gene body regions used was proved to have no reads covered in the sample. We can remove these regions, along with some regions of low expression (the lowest 25%), manually for clear pattern in the profile.

4. By default, strand information is not taken into consideration in generating the bigwig file. But we can manage to split the strands with --filterRNAstrand {forward,reverse}. A scaling factor or the parameter '—exactScaling' can be applied to fix this issue.

Figure above (Figure 4.2 B) helped us to discover the very specific pattern of binding on $3'$ region. The binding peak raised sharply before annotated polyadenylation site and reaches to the top near annotated PAS. Interestingly, the binding expended to the downstream of PAS to 300 bps or even further downstream regions, which was corresponding to the distribution of peaks in proximal downstream regions of genes. The pattern remains similar after min-max scaling, indicating that the binding near $3'$ is robust in regardless of the coverage of FCA-CLIP peaks. This is a strong indication of FCA's binding to pre-mRNAs rather than mature mRNAs. But the fate of these binding RNAs cannot be determined based on CLIP data. In the following parts, we would look at the feature of FCA-binding at the most enriched regions separately, $3'$ end and the introns respectively.

4.2.1.3 FCA's Binding on Introns

In the feature distribution of FCA-binding peak, a large portion of peaks are located in introns. To ensure whether such binding peaks appeared on intron alone or on the region of intron and the both of the flanking exons, we plot the meta-profile and heatmap of peaks around intron regions (Figure 4.2 C). Interestingly, the boundary of binding fit quite well with the boundary of introns.

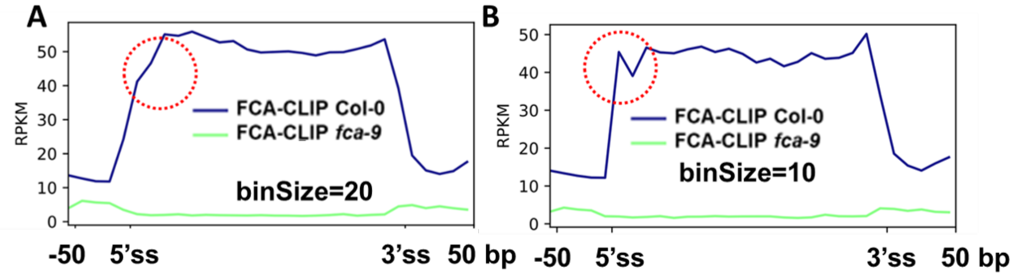


Figure 4.3: Preliminary Trial for Meta-profile of FCA's binding on introns, A: bin size 20 bp, B. bin size 10 bps. A sharp peak was detected near the $5'$ splicing site in B

Even though we used a bin size of 20 bps for a smooth profile, there appeared an unusual turning point inside the intron region right after the junction (see Figure 4.3 A). At first, we guess it might happen on the introns with imprecisely annotated $5'$ splicing site, with the actual $5'$ SS after the annotated one. To avoid cases like this, we filtered introns unspliced along with the introns spliced but overlapping with another spliced intron. This strategy helped us avoid most unused splicing sites. But unexpected, the drop become even more obvious with the filtered list of introns (Figure 4.3 B).

To plot the overall binding pattern of introns of various length, we normalized length of introns to the same length of 200 bp, which make it hard to detect the exact position of the

drop within intron region. To get the exact position of the drop, we take fully advantage of 'single-nucleotide resolution' feature of eCLIP data and plot the feature of a 50-nucleotide region centered on the exon-intron junction in the following figures. Interestingly, there exist a very sharp peak right after the junction on the first base of the intron.
This is a very specific binding feature. One of the possibilities is that FCA is recruited by the reformatted Pol II near splicing site as a part of the splicing process. It was also possible that FCA might be recruited right after the cleavage of the $5'$ splicing site or after the occupancy of other proteins to the splicing site.

We further wondered whether these introns were spliced or not. This question can be partially answered by looking at the raw reads. We detected CLIP reads on intron-exon junction, indicating that FCA are more likely to bind to the intron retained or not yet spliced. While there exists a strong binding peak on the beginning of the intron, we are not sure whether these introns were attached to the exon before or part of the splicing-intermediate. On the other hand, the 'spliced reads' in CLIP-seq were not really treatworthy. The UV crosslinking might form covalent bond between RNAs close in space after folding, resulting in many 'split reads', with two side on even separated genes on the chromosome. It is hard to identify split reads and exon-exon junction reads. Despite the splicing event itself, it is unclear whether the splicing of these FCA-binding introns was affected by the loss of FCA function or the splicing efficiency would affect the binding of FCA. To further explore these features, we also applied CB-RNA-seq of Col-0 and *fca-9* for the change of splicing efficiency and mRNA-seq of the two to check the intron retention level.

4.2.1.4 Binding Near $3'$ End

Here we checked in detail whether the binding accumulated near annotated PAS region. Since most FCA-binding genes are protein coding genes, the $3'$ untranslated region ($3'$ UTR) after the stop codon was used first for plotting, regardless of introns in these regions.

Interestingly, the binding seems to start near stop codon (Figure 4.2 D).

To test whether the binding start near stop codon, we next plot the binding pattern centered on the end of $3'$ UTR regions (Figure 4.2 D). Since FCA binding is likely to happen during transcription, it is hard to imagine that the binding was correlated with the stop codon which is usually recognized during translation.

On the other hand, unlike the binding in the intron which shows a specific binding site on the first base of intron, we can hardly find the strongest binding 'site' near annotated $3'$ end. The apparent bimodality appeared near PAS in both meta-profile (Figure 4.2 B) of gene body and a detailed profile of the $3'$ end (Figure 4.2 D). One of the concerns is that PAS might not be as precisely defined as the introns in annotation. Attached with other sequencing data, we would try to re-define the precise PAS to see if we can get the exact binding position comparing to the PAS in the next chapter.

4.2.1.5 Connection between Binding on Introns and on $3'$ UTRs

Since FCA shows very specific pattern of binding to $3'$ UTR and intron regions, we are curious about whether the binding to these two kinds of regions happen independently. We first get the introns and UTRs used in meta-profile plotting and exact the corresponding gene ID of these regions. Since we are considering 'intron' and 'UTR' region, the total gene set should satisfy the following standards:

- Expressed in Col-0 background
- Coding genes (with $3'$ UTR region)
- Containing at least one intron in CDS region

Here, the 'introns' should be regions between CDSs, outside UTR region.

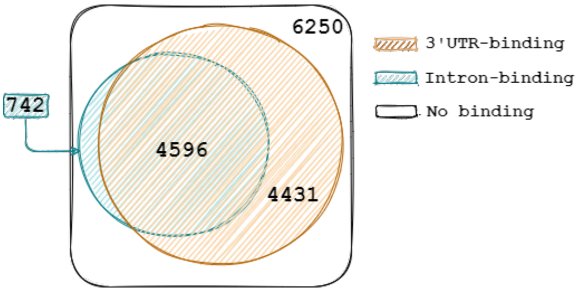


Figure 4.4: Venn Diagram for FCA binding on introns and 3′UTR. With in the frame are all the coding genes, blue one is gene with FCA's binding in introspection and the orange one represents genes with binding in 3'UTRs.

We then applied Chi-square test of independence to data shown in the Venn diagram Figure 4.4 and got a p-value less than 0.01, rejecting the null hypothesis of the binding to the two regions being independent. In other word, it is high possible that the binding of FCA to intron and to 3′ UTR are dependent. Such inner-dependence has been suggested by the auto-regulation of FCA, in which case FCA also binds to the intron where the 3′ end of FCA-beta is located. It is reasonable that the binding of FCA might be connected with the usage of proximal/distal poly(A) sites. The idea would be tested in the next chapter.

To summarize, typical FCA-binding regions include 3′ end of the gene (the last exon and proximal downstream of the gene) and the intronic regions. It is quite interesting since downstream and intronic regions are generally not presented in mature massager RNAs. From this aspect, we conclude that FCA binds pre-mRNA, or FCA binding happens co-transcriptionally rather than post-transcriptionally. The interaction between FCA and sub-unit of chromatin remodelling complex might confirm the hypothesis from another aspect. The binding of FCA to the 3′ end is consistent with the research before that FCA might be a 3′ processing factor, while the exact function of FCA remains to be explored in combination with other sequencing data (See Chapter 5).

4.2.2 Motifs of FCA's Binding

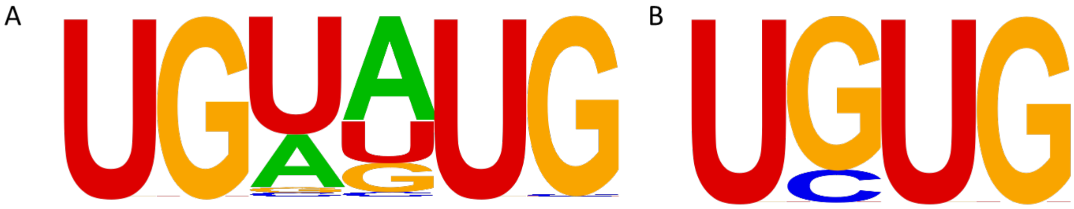


Figure 4.5: Top2 motifs enriched near FCA-CLIP peaks, both are U-rich sequences

CLIP binding peaks represent the sites of strong, typical binding of the RBP. By scanning the regions near CLIP peaks of enriched subsequences, we may find the sequence preference of FCA. We applied motif searching on a 21-bp window including the peak and 10nt on each flanking. The command line tool Homer is used in this analysis. We tried to find motif of length of 4, 6, and 8 nt. We got U-rich motifs like 'UGUG' and 'UGUAUG' (Figure 4.5), which are corresponding to the binding motif of RBPs with homological RRMs.

Since the $3'$ UTR was enriched with binding of FCA, it should be interesting to see whether the motif corresponds to the poly(A) signal. The most abundant PAS signals detected is 'AAUAAA' in animal but a possible poly(A) signal in Arabidopsis is 'UGUA', which is similar in sequence with the most enriched motif of FCA. The connection between FCA and FY, along with CPSF30 and CPSF100, can be seen from CLNIP-MS data, though the interaction might be transient. The CPSF proteins was believed to be involved in the recognition of poly(A) signal in human. It was also known that FCA is unlikely to interact with FY at the same time of CPSF proteins, competing with other CPSF factors to recruit FY (Yu et al. (2019)), which indicates that FCA might be working in transcriptional termination process but involved in a different mechanism with CPSF proteins.

4.3 Transcriptome-Wide Binding Feature of FCA in *fpa-7*

As a possible $3'$ processing factor, the loss of FPA function would cause defect in termination comparing to the wild type. Previously, it has been reported that transcripts in

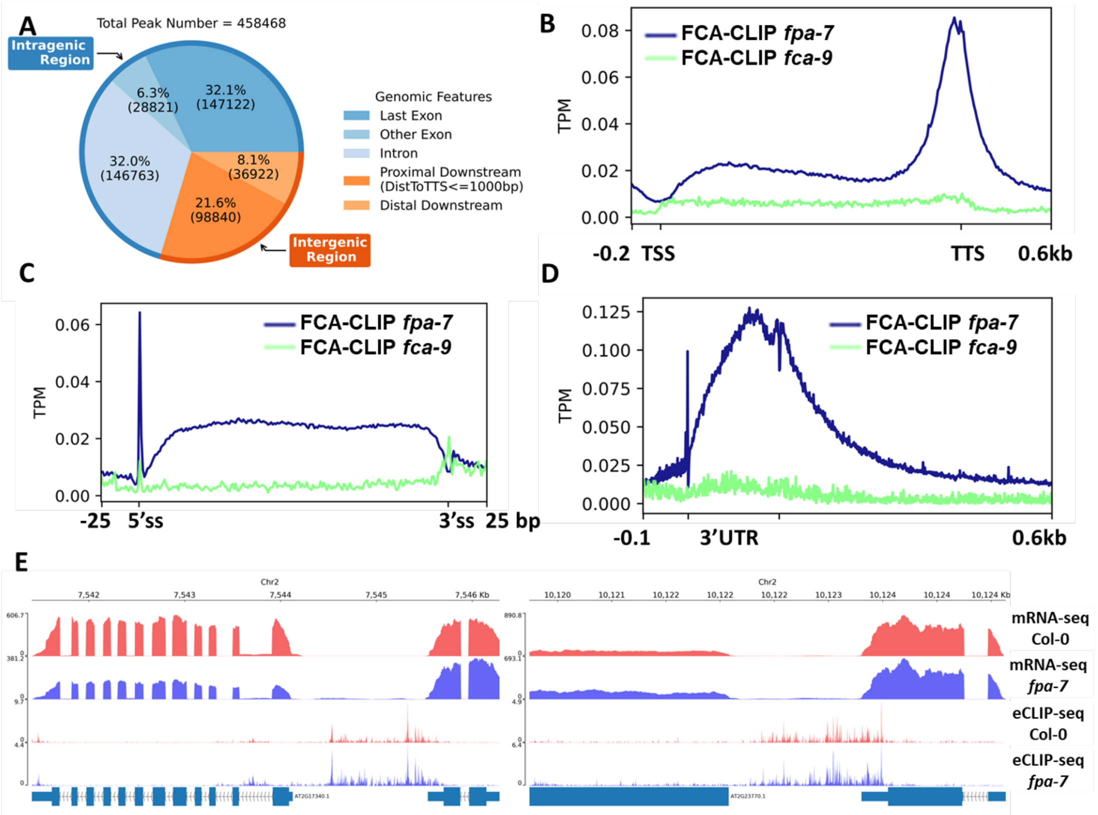


Figure 4.6: Overall binding feature of FCA in *fpa-7*. A. Pie Chart of Genomic Feature Distribution of FCA-eCLIP Peak in *fpa-7*; B. Meta-profile of FCA-binding on gene body; C. Meta-profile of FCA-binding on intronic regions; D. Meta-profile of FCA-binding near 3′ end; E. Examples of FCA's binding to the downstream regions of genes and increased expression of intergenic regions in *fpa-7*

intergenic region increased in *fca fpa* double mutant (Duc et al., 2013). In some extreme cases, the extension of transcription did not stop until reaching the termination sites of the downstream genes. Part of these read-though transcripts, also known as unannotated genes or chimeric reads, are also detectable in *fpa* mutants. Here, we checked the binding of FCA in *fpa-7*. Since FCA binds to the 3′ end and the downstream region, we wonder whether the binding pattern change in genes with read-through transcripts in *fpa*.

4.3.1 Overall Binding Feature of FCA in *fpa-7*

Firstly, we followed the workflow described before to check the overall binding pattern of FCA in *fpa-7*. The pie chart and meta-profile of FCA-binding are shown in Figure 4.6 A. To our surprise, the overall binding pattern of FCA in *fpa-7* seem not very different to the one in Col-0. In the distribution of peaks, the portion of peaks within introns and intergenic regions increased slightly, while the one of exons dropped a little. Meanwhile, the binding pattern shown in meta-profile of binding in gene body remain similar. These figures (4.6 A˜D) might indicate that FPA have weak impact on the overall binding pattern of FCA. However, we noticed that FCA binding extends to further downstream regions in *fpa-7* comparing to Col-0. Some examples are shown in Figure 4.6 E. Furthermore, the binding seems to correspond to the read-through in mRNA level. It might be interesting to find more examples of read-through transcripts to prove the idea. It might be more suitable to compare the binding position of FCA near $3'$ end to the $3'$ end of pre-mRNAs, but the CB-RNA-seq for *fpa-9* was not available yet.

4.3.2 FCA Binds to *FPA* with T-DNA Insertion in *fpa-7*

One quite unique and informative case of FCA-binding was observed at FPA locus in *fpa-7*. The *fpa-7* is a SALK T-DNA insertion line with T-DNA inserted to the first intron of *FPA*. Usually, insertion of a large fragment to the gene results in premature termination and cause reduction in expression of the gene. Surprisingly, even though the expression level of *FPA* did reduce in *fpa-7*, we still detected a very strong peak of FCA's binding just before the insertion site. Since the genome we used for mapping is the reference sequence without T-DNA insertion, the binding of FCA on RNAs transcribed from the inserted T-DNA cannot be detected. We wonder whether FCA binds to the insertion sequence, but we need to get the whole sequence of *FPA* first before alignment of RNAs to the region.

Though the sequence of the plasmid is known, it is possible that the sequence will be inserted to the same site multiple times, even with different direction. Here whole-genome sequencing (WGS) was produced to see the exact sequence of *FPA* with T-DNA insertion. The WGS library was contructed by Qiqi.

4.3.2.1 De novo Assembly of *fpa-7* Genome

The denovo assembly takes 4 steps:

1. Quality Control of the Sequencing Data and adaptor removal with fastp
2. Define best k-mer to use in assembly by kmergenie (Chikhi and Medvedev (2014))
3. Denovo assembly by SOAPdenovo (Li et al. (2010)) with the k number get in step2
4. Check the result by comparison with the reference genome by QUAST (Gurevich et al. (2013))

Probably due to the low coverage of our whole-genome sequencing data, the quality of the assembly was not very satisfying. But we have no plan to do extra DNA sequencing for this sample. Since we failed in assembly of the entire sequence of FPA with the insertion, we would like to check the boundary of the insertion.

4.3.3 Mapping with T-DNA Sequence Added to the Genome

Here we added the original insert sequence as a new dummy chromosome to the end of the reference sequence. The coverage of ‘T-DNA’ chromosome was much higher than the most part of the original genome. It is likely that there exist multiple insertion events. On the other hand, there must exist pairs on the edge of insertion, with one read mapped to the original sequence of *FPA* and the other to the insertion. Here we call these read pairs ‘chimeric reads’, for they are hybrids of the *FPA* and the inserted T-DNA. Of 3559816 pairs mapped, only 954 pairs are ‘chimeric’ (0.268 ‰). In the dummy chromosome of T-DNA, we found two chimeric read peaks, one near left border the paired reads of which directed to the first intron of *FPA* (AT2G43010), and the other near right border to the $3'$ UTR of *SR1* (AT2G47090).

In summary, it might be impossible to get the complete sequence of the insert in *FPA* due to multiple insertion and the limitation of short-read sequencing. But it is highly possible that the binding of FCA to *FPA* extended to the inserted T-DNA. The T-DNA insertion in *FPA* is aberrant to the plant which might be removed in the transcription. The binding of FCA to these "exotic transcripts" might have some connection with the function of FCA in termination or RNA quality control. If we can get the whole sequence of *FPA* with T-DNA insertion and align FCA-CLIP reads to this region for the binding pattern of FCA on T-DNAs, it would be helpful for us to understand the function of FCA. We might get the sequence with long-read sequencing method in the future (see discussion in Chapter 6).

4.4 Transcriptome-Wide Binding Feature of FCA in *fy-2*

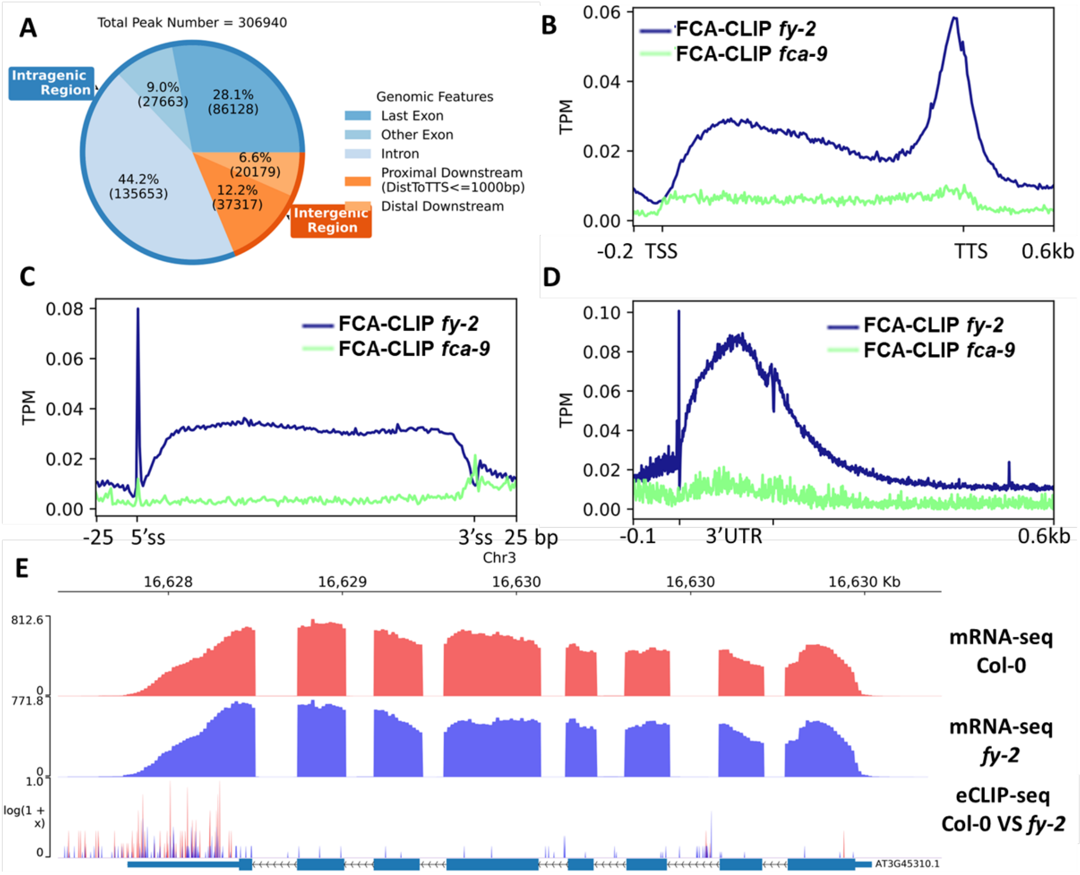


Figure 4.7: FCA's binding in *fy-2*: A. Pie Chart of Genomic Feature Distribution of FCA-eCLIP Peak in *fy-2*; B. Meta-profile of FCA-binding on gene body; C. Meta-profile of FCA-binding on intronic region; D. Meta-profile of FCA-binding near 3′ end; E. An example of increased binding in intron but decreased level near 3′ end, the mRNA expression level of this gene has no significant change in *fy-2* and Col-0.

While the meta-profile of FCA binding in *fy-2* (Figure 4.7 B˜D) seems similar to Col-0 at the first glance, it is not hard to tell the portion of intron-binding peaks increased in *fy-2* from the pie chart (Figure 4.7 A). If we plot FCA-binding of Col-0 and *fy-2* in the same figure, it would not hard to discovery that binding of FCA near 3′ ends of the genes decreases near 3′ ends in *fy-2* mutant, while an increase of binding was detected in the intron regions correspondingly. It is not surprising that *fy-2* could affect the binding of FCA. The two proteins interact through WW domain of FCA and two PPLPP motifs in the C-terminal of FY. Here, the mutant of *fy-2* was a SALK line with T-DNA inserted before both of the two PPLPP, resulting in loss of ability to interact with FCA. The drop in $3'$ end along with the increase in introns might indicate a change in the binding pattern, or it might be a result of change in expression level or poly(A) usage. An example of changed expression was shown in Figure 4.7 E.

4.5 Correlation between FCA-Binding and m^{6}A Modification in Position

The m^{6}A modification is wide-spread and plays an important role in the development of plants. In the preliminary research, the $3'$ UTR regions and stop codons with enriched m6A level were preferred by binding of FCA. Here we would like to see whether the loss of m6A would affect FCA's binding to check whether m^{6}A was the deciding feature of FCA recruitment.

FIP37 is a core component of the m^{6}A methyltransferase complex, also known as the 'Writer' of m^{6}A modification. The loss of function of FIP37 could cause a drop in m^{6}A level which is fatal to plants. In order to get usable material, the mutants studied are actually *fip37-4 LEC1:FIP37* (*fip37-4*, Shen et al. (2016)) with the promoters of embryo-specifically expressed proteins LEC1. The plants to produce the seed are heterozygous, and the useful seedling can be selected after germination. Homozygous of *fip37-4 LEC1:FIP37* would only have two cotyledons, but seeds of other genotype would have real leaves, so seedlings with true leaves would be removed. Through this way, a mutant that lack m^{6}A in post embryonic state can be obtained.

Since the m^{6}A peaks were found mostly near $3'$ end in $3'$ UTRs and near stop codon and the $3'$ end is also the place of great FCA abundance, we would like to see if the correlation was found in $3'$ only or in other features as well. According to the result, a peak of FCA-

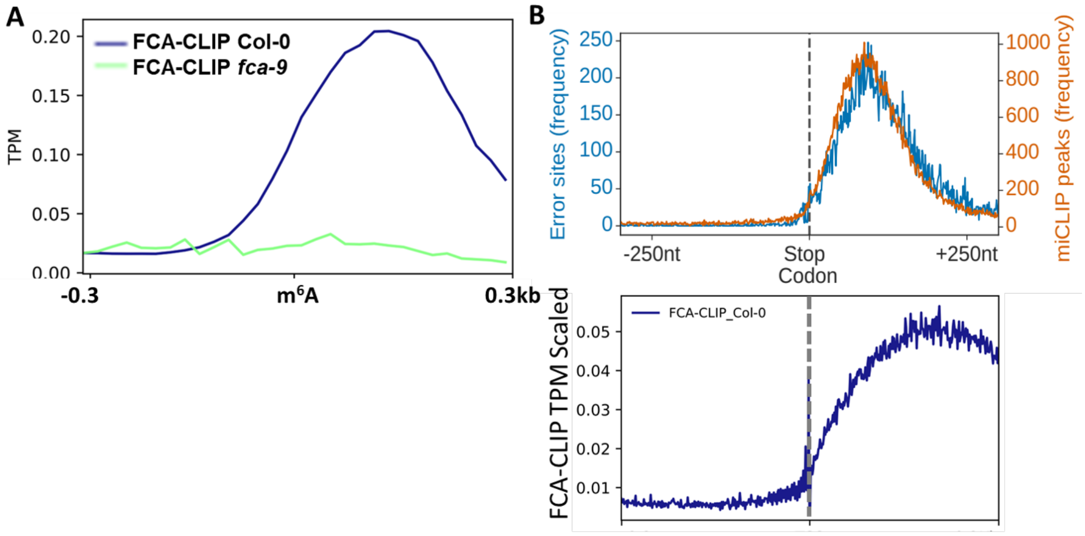


Figure 4.8: FCA binding near m^{6}A modification. A. FCA binding near m6A peaks (m6A position from *fip37-4* research [38]), bin size = 20 bp; B. Comparison between position of miCLIP peak around the stop codon (from George et al. (2017)) and the FCA binding near the same

binding was seen downstream of the center of the m^{6}A peaks. However, the m^{6}A detection was of low resolution, so we also compared the FCA-binding near stop codon with m^{6}A modification (Figure 4.8 B down) to the distribution of m^{6}A modification detected with miCLIP-seq of signal-nucleotide resolution (Figure 4.8 up, orange line). As is shown in the figure, even though the two was both enriched after the stop codon, FCA tends to extend to further downstream region while the m6A peak seems closer to the stop codon.

We also applied FCA-CLIP to *fip37-4* background. It turned out that the binding pattern of FCA is almost unaffected by m^{6}A modification (Figure 4.9). Along with the result before, we may conclude that FCA-binding might accumulate near m^{6}A modification of 3′ end, but the binding of FCA seems in general unaffected by this modification. But it is still possible that FCA was involved in the building of m^{6}A modification, which might be tested by looking into the change of m^{6}A level under *fca-9* background.

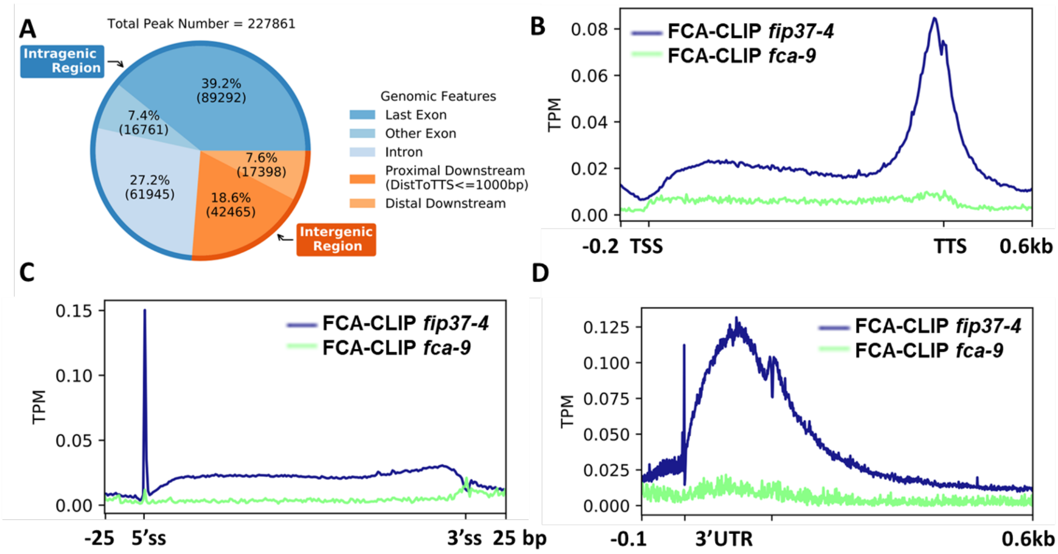


Figure 4.9: FCA's binding in *fip37-4*: A. Pie Chart of Genomic Feature Distribution of FCA-eCLIP Peak in *fip37-4*; B. Meta-profile of FCA-binding on gene body; C. Meta-profile of FCA-binding on intronic region; D. Meta-profile of FCA-binding near 3′ end

4.6 Comparison between FCA-Binding under Different Background

Table 4.4: Comparison of feature distributions of FCA-binding peaks between different genotypes

Sample	Last Exon	Intergenic Proximal	Intron
Col-0	46.39%	23.52%	30.10%
fip37-4	46.10%	21.92%	31.98%
fpa-7	37.46%	25.17%	37.37%
fy-2	33.24%	14.40%	52.36%

The total number of reads captured varies among samples due to technique limitations, but we can still compare the samples by the genomic feature distribution of the peaks under different background (Figure 4.10 A). Here we tried to use chi-squared test of Goodness of fit to test whether the distribution in certain mutant is corresponded to the one in the wild type. The null hypothesis is that the observed distribution is the same as the expected one.

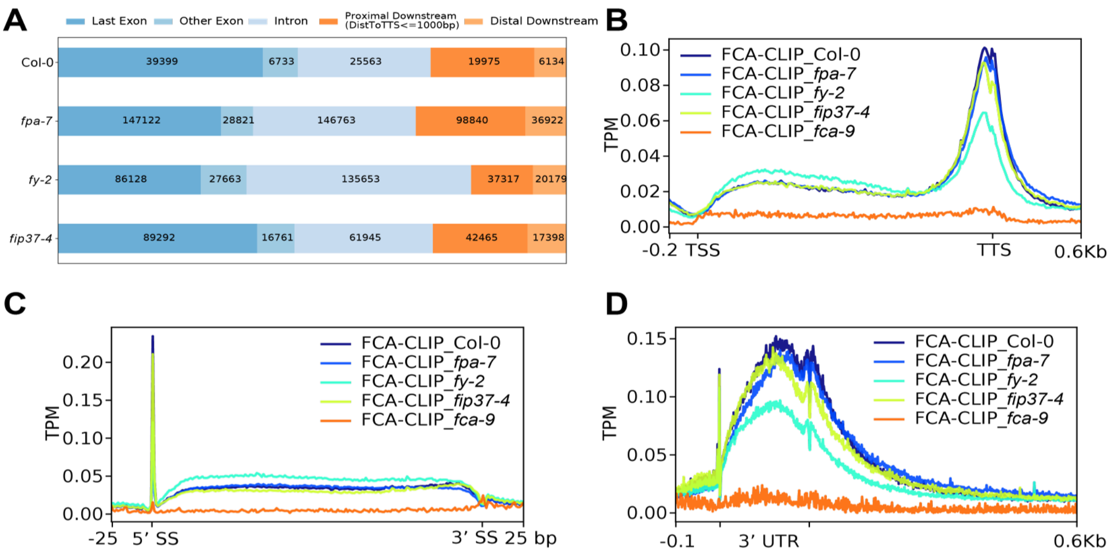


Figure 4.10: Comparison between different genotypes, A. bar chart of feature distribution; B. meta-profile of FCA-binding on gene body; C. meta-profile of FCA-binding on intronic region; D. meta-profile of FCA-binding near 3′ end.

The null hypothesis is only rejected when comparing *fy-2* and Col-0, indicating that only the distribution peaks in *fy-2* is significantly different from the wild type.

The meta-profile (Figure 4.10 B) showed a similar pattern. Here, we pick only the genes with binding in all samples. The curve for Col-0, *fpa-7* and *fip37-4* background are similar and very close in position. Meanwhile, binding of FCA in *fy-2* background shows a lower peak near 3′ end and higher coverage on gene body, in agree with the distribution of genomic features shown in the table.

Another defect on the CLIP data is that we cannot compare the absolute binding scale between different samples. The normalization method, whether RPKM or TPM, suggest that the overall amount of binding number is the same but the might change under different background. Like in the case of FCA binding in *fy-2*, we can see a drop in the 3′ but cannot say for sure that this is caused by the decrease of FCA-binding number on this region. But the impact of FY on FCA binding is obvious. Upon loss of interaction with FY, the most abundant region of FCA binding switched from 3′ end to the introns.

5 Functional Dissection of FCA Through Combined Analysis of HTS Data

In Chapter 4, we described the binding pattern of FCA as revealed by CLIP-seq data. Although the specific binding of FCA to the introns and to gene 3′ end is clear, the underlying mechanism as well as the biological consequence of such position specific binding remains elusive. Here in this chapter, more data, including mRNA-seq, 3′mRNA-seq and CB-RNA-seq, were obtained and used to study the feature of FCA-binding region and to further dissect the molecular function as well as the recruiting mechanism of FCA. Experiments for library construction of mRNA-seq, 3′mRNA-seq and CB-RNA-seq were all done by Qiqi Zhang, and I independently completed the data analysis part in this chapter.

5.1 Analysis of Differential Expression

As a starting point, we aimed to isolate the genes with significant changes in expression level after loss of FCA function. Such information may serve as an indication for the function of FCA at transcriptome level. After obtaining the differentially expressed genes, we may also investigate FCA's binding on these genes in order to build a link between FCA-binding and altered transcriptome.

For transcriptome analysis after mapping of mRNA-seq reads to the genome, we used the one-step software, TPMCalculator (Vera Alvarez et al. (2019)), to get the expression matrix of genes in each sample. Matrix for both TPM and raw read count can be gained by the tool. The TPM matrix was then applied in calculation of correlation coefficient between samples and principal content analysis (PCA), while the read count matrix was prepared for differential analysis by DESeq2 (Love et al. (2014)).

In this research, 10-day-old seedlings of six different genotypes were used as materials, including Col-0, *fca-9*, *fpa-7*, *fy-2*, *fca-9 fpa-7*, *fca-9 fy-2*, with three biological replicates sequenced for each genotype. As a part of the quality control process, Spearman Correlation coefficients between samples were calculated to check the repeatability of the experiments (Figure 5.1 A). Replicates of different samples were clustered into same groups, indicating that the biological replicates are of good repeatability. Similarly, the clusters in PCA were defined (Figure 5.1 B). Notably, the cluster of *fca-9 fy-2* was the most distant to that of wild type, indicating that the double mutant may have the most significant difference in mRNA expression comparing to the wild type.

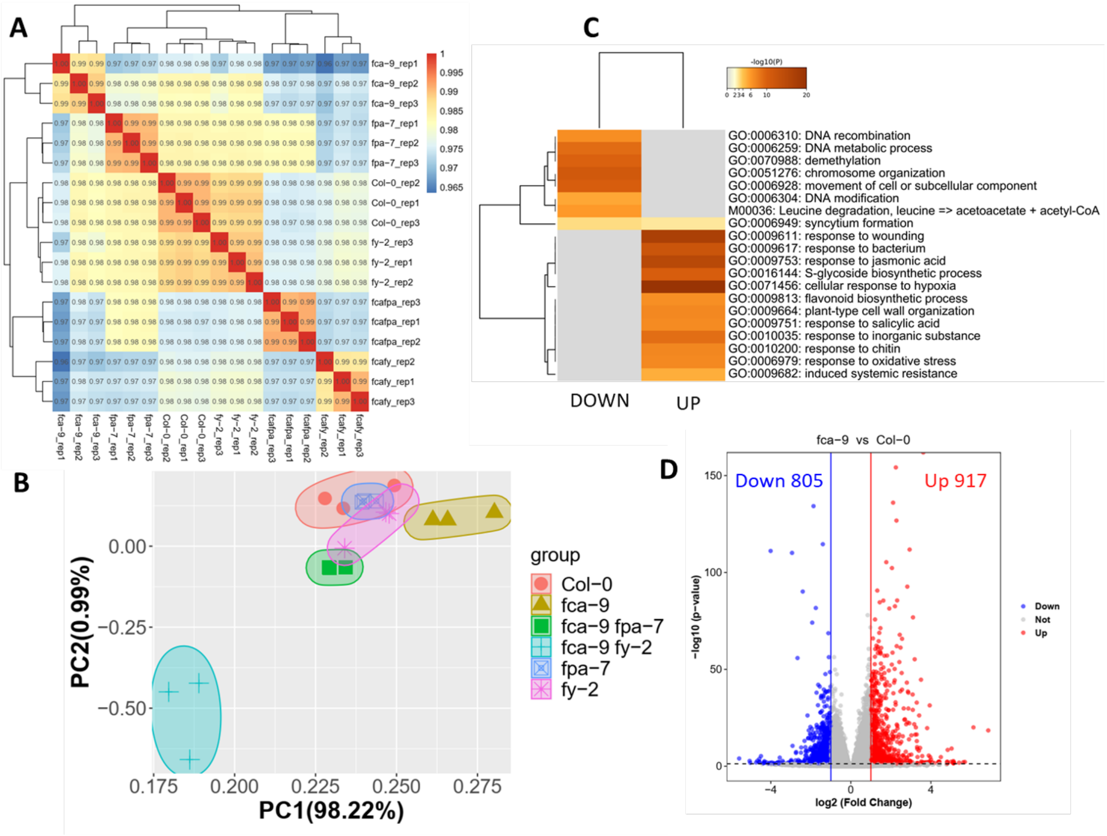


Figure 5.1: Quality Control and Differential expression analysis of mRNA-seq data: A. Spearman correlation coefficient calculated between individual samples; B. Principal Content Analysis, plotted on the PC1 and PC2; C. Gene Ontology (GO) analysis shown the terms enriched in differentially expressed genes in *fca-9* v.s. Col-0. Results were clustered based on analysis by Metascape; D. Volcano plot of DE genes in *fca-9* v.s. Col-0

5.1.1 Differentially Expressed Genes in *fca-9* Comparing to Col-0

In order to identify differentially expressed genes between *fca-9* and Col-0, I used the R package DESeq2 to perform differential expression analysis. DESeq2 takes matrix of non-negative integers as input, which means we should use raw count instead of normalized values like TPM. Among the 24230 genes I checked in comparison between *fca-9* and Col-0, 805 genes showed significant decreases in expression level in *fca-9*, while 917 genes displayed significant increase in expression level. Next, Gene ontology (GO) analysis was applied to classify these genes by their functions and to find the enriched pathways. For the genes of increased expression, several groups related to stimulation response were enriched, such as response to wounding, hypoxia and jasmonic acid. In comparison, genes with decreased expression were connected with chromosome, like chromosome organization and DNA metabolic process. Overall, altered expression of hundreds of genes in *fca-9* mutant was observed. However, whether these changes were a direct consequence of FCA-binding remains unknown.

Next, I tried to combine the FCA-binding data with mRNA-seq data in our analysis. I first compared the expression level of genes either with or without FCA binding to see whether binding of FCA would directly affect gene expression. I used the gene list gain in Chapter 4 for plotting the meta-profile of FCA binding in Col-0 (Figure 4.2 B) as 'FCA-binding genes'. The exact position of binding was not restricted to intron or $3'$ end for a broader vision. As a result, 17599 genes with binding were picked, with 560 genes of significantly higher expression level after mutant and 562 with lower level. To test whether there exist certain correlation between FCA-binding and change of in expression level, Chi-Square test was applied on the contingency table (Table 5.1):

Table 5.1: Contingency Table of FCA binding and Differential Expression

	padj>=0.05	**padj<0.05**
FCA-Binding	7870	9729
No/weak binding	4509	2122

Here in the columns, 'padj' refer to the adjusted student's t-test P-value by BH process in differential expression analysis. To be specific, padj<0.05 indicated that distribution of expression of the gene in the mutant is different to the one in the wild type, while padj>=0.05 mean no significant change detected between the two. In other word, the padj value was used to represent whether the expression level of the gene change after loss of FCA. As the value in each row, I marked genes in the binding gene list as 'FCA-binding genes', while the rest of the genes, which can be of very low or even no binding, are marked as 'No/weak binding gene'. In the Chi-Square test, I got a chi-square value of 1043.69 and a p-value of 5.73e-229. The p-value was much smaller than 0.01, indicating the null hypothesis, that the expression level is independent of FCA-binding, was rejected. Here, we concluded it is highly possible that the altered mRNA expression in *fca-9* mutant was correlated to loss of FCA binding, and FCA-binding genes are more likely to have changed expression in *fca-9* mutant.

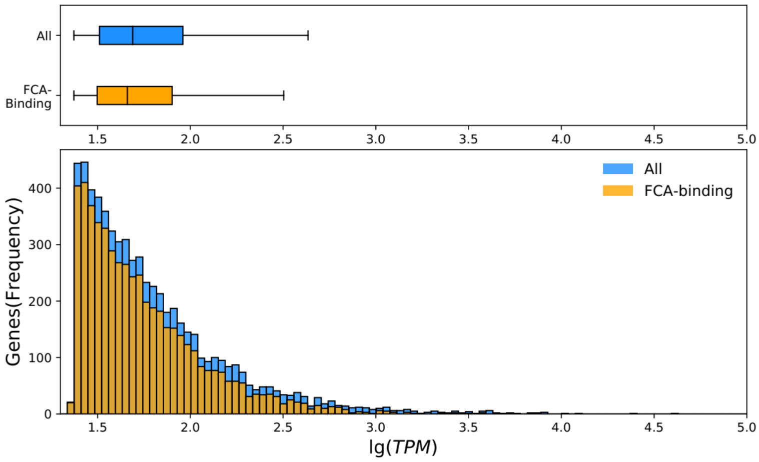


Figure 5.2: FCA-binding on genes with high expression level (top 25%): it might be easier to detect FCA-binding genes of high expression level, but not all highly-expressed genes have FCA's binding. There should be some specific feature which triggers FCA's binding.

Besides the differentially expressed genes, I also checked the original expression levels of FCA-binding genes in Col-0 to see if FCA-binding might be related to high expression level. Since genes with low expression might also be of low binding level, I picked only the genes with high expression to perform the analysis, avoiding the possibility of failure in detecting binding for low expression level. Here in the genes of top 25% expression level (normalized to TPM), FCA-binding gene somehow showed lower median value comparing to all the genes of high expression (see Figure 5.2), indicating that not all genes of high expression level would have FCA binding on it and there should exist some specific feature to trigger FCA binding.

5.1.2 Result of Interaction between FCA and FPA/FY at Transcriptome Level

In previous genetics research, the members of autonomous pathway showed non-linear interaction between each other. FCA and FPA both contain multiple RRM domains but share no other homology in sequence. The two RRM-containing proteins both suppress *FLC* expression partly through histone demethylase FLOWERING LOCUS D (FLD, Liu et al. (2007)), while FY is only required by FCA but not FPA in regulation of flowering time. Meanwhile, FPA over-expression reduced *FLC* expression in *fca* and *fy* (Bäurle and Dean (2008)), indicating that FCA and FY are not required in FPA-mediated repression of *FLC*.

In spite of these genetics evidence, we may also get some hints for explaining possible interaction between FCA and FPA/FY by comparing the differentially expressed genes. Take the expression of *FLC* for example, it is low in Col-0, but increased in all of *fca-9*, *fpa-7* and *fy-2* mutants, resulting in a late flowering phenotype. The severity of late flowering is positively correlated with *FLC* expression level, which means higher the expression level of FLC is, more time is in need for the plant to flower under the same condition (vernalization/day time/temperature). Among the three single mutants used in this research, *fy-2* has the lowest *FLC* level and the earliest flowering time. Like I've mentioned before, among the several *fy* mutants in Col-0 background (*fy-2/fy-3/fy-4/fy-5*), *fy-2* used in this project is the one with the latest flowering time (Henderson et al. (2005)) though it showed lower FLC level comparing to *fca-9* or *fpa-7*. And the flowering time of double mutant *fca-9 fy-2* does not increase comparing to *fca-9* single mutant, indicating that *fca-9* might be epistasis to *fy-2* in regulating *FLC* expression.

By comparing single mutants and double mutants to the wild type separately, we can check the genetic relationships between these proteins in regulating different genes. I get only the genes differentially expressed in all three mutants, and the change of expression is shown in the two tables (Table 5.2, Table 5.3)below:

Table 5.2: Number of Differentially Expressed Genes in *fca-9/fy-2/fca-9 fy-2*

fca-9 fy-2	*fy-2*	*fca-9*	Gene Number
UP	Up	Up	408
		Down	79
	Down	Up	67
		Down	48
Down	Up	Up	75
		Down	29
	Down	Up	206
		Down	365

Table 5.3: Number of Differentially Expressed Genes in *fca-9/fpa-7/fca-9 fpa-7*

fca-9 fpa-7	*fpa-7*	*fca-9*	Gene Number
UP	Up	Up	970
		Down	210
	Down	Up	60
		Down	29
Down	Up	Up	61
		Down	58
	Down	Up	340
		Down	667

We can conclude from the table above that the direction of change in expression (up-regulated or down-regulated) are the same in the two single mutants with the corresponding double mutants at majority genes (60.53% of DE genes in *fca-9 fy-2*, and 68.35% of *fca-9 fpa-7*). This is a strong indication of FY/FPA working in the same pathway with FCA, which might be not only in flowering control.

5.1.3 Section Summary

In this section, we mainly focused on the change of expression level caused by the loss of FCA function. The up-regulated and down-regulated genes in *fca-9* were enriched in different pathways, but it seems insufficient for us to conclude the function of FCA besides flowering by these DE genes only and further research from other aspects would be included in the following sections. In checking genes bound by FCA, I found that they have higher possibility to have altered expression level in loss of FCA function, which means FCA might affect the expression level of genes directly in some cases. We have also known that not all genes with high expression level have a detectable binding of FCA but what triggers FCA to bind on these genes are not very clear yet. The last but uncompleted research was about interactions between FCA and other members of autonomous pathway FY and FPA. It was shown that the direction of change in expression are the same in the two single mutants with the corresponding double mutants at majority gene, indicating that FY/FPA should be in the same pathway with FCA, not only in the regulation of *FLC*.

5.2 FCA's Binding at Intronic Regions

As demonstrated in the last section, changes in expression of genes in *fca-9* mutant is related to loss of FCA binding. Next, I would like to investigate into the influence of FCA's binding in intronic regions and the regions near $3'$ end separately to see if FCA's function is the same in different regions. In this section, we would study the feature of FCA-binding introns first.

Speaking of introns, it is natural to assume that an intron-binding protein might affect splicing through its binding. I found an enrichment of FCA near $5'$SS, which is also the target region of U1 snRNP of spliceosome. So, I first checked whether alternative splicing events happen in *fca-9* to see if FCA might be part of the splicing mechanism. Another significant feature of splicing is co-transcriptional splicing efficiency, percentage of the splicing events which happen during the synthesis of pre-mRNA. Besides mRNA-seq, chromatin-bond RNA-seq (CB-RNA-seq) was also applied to *fca-9* and Col-0 for calculation of splicing efficiency. I would check whether binding of FCA would affect splicing efficiency and whether

FCA prefers introns of lower co-transcriptional splicing efficiency for their longer half-life.

5.2.1 Few Alternative Splicing Events Detected in *fca-9*

Firstly, I used mRNA-seq of *fca-9* and Col-0 as input and rMATS (Park et al. (2013), see **rMATS: detecting alternative splicing events**) as a tool to get the alternatively spliced introns. The number of alternatively spliced introns of each type is shown in Figure 5.3 B. Take intron retention, the major form of AS in plant, for example, 36 introns were of higher retention level in the wild type and 48 are of higher level in introns. The result informs us that the number of alternative splicing events seems not large comparing to the number of introns FCA binds. On the other hand, in comparison with the mutant of the U1 snRNP subunit LUC7 (Figure 5.3 A), a member of the spliceosome, number of alternatively spliced introns in *fca-9* mutant was rather small. The results of analysis indicate that the loss of FCA function does affect splicing at some positions (See examples in Figure 5.3 C), but the total number of AS events happened in the mutant was not as huge as the one in *luc7* triple mutant, indicating that FCA is not as important as the U1 snRNP in regulating splicing.

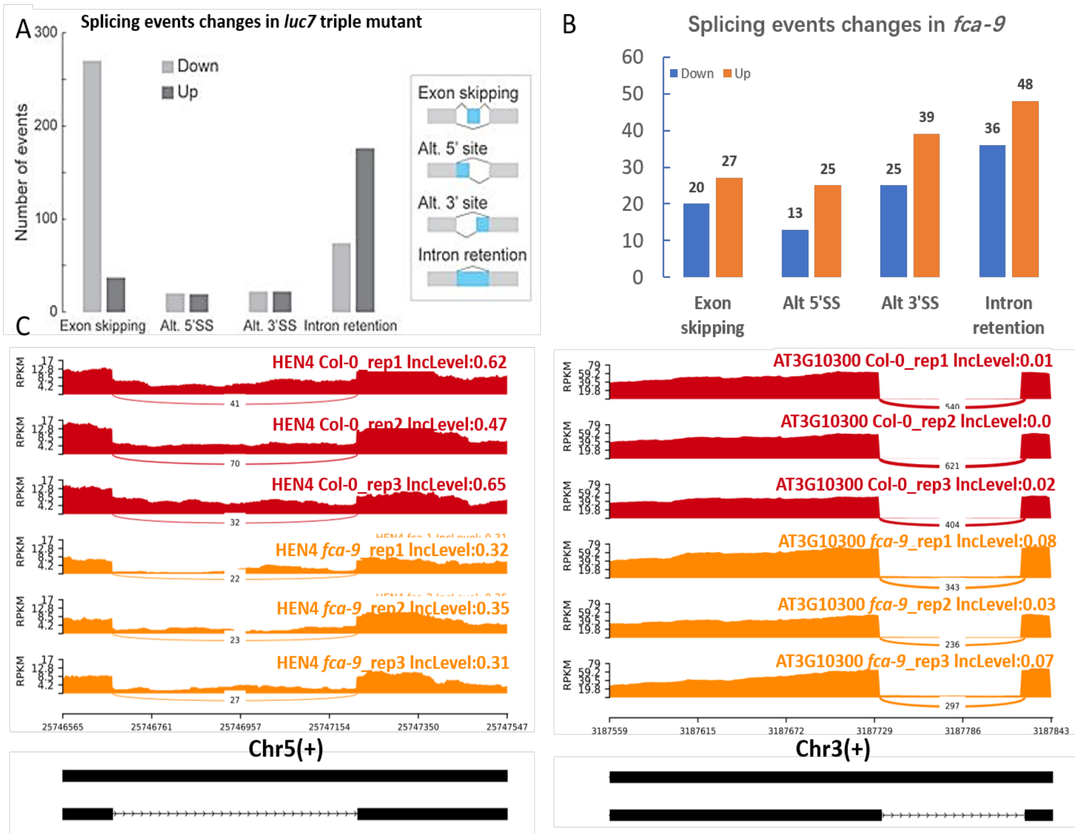


Figure 5.3: Alternative spliced introns in *fca-9*: A. Number of alternative splicing (AS) events detected of each type in *luc7* triple mutant, showing the AS event with the function of spliceosome affected; four types of AS detected: Exon skipping: the one of exon is skipped in splicing; Alt 5′ site: the position of 5′ splicing site is altered; Alt 3′ site: position of the 3′ splicing site is altered; Intron retention: the intron is not spliced, retained in the transcript; B. Number of AS events affected in *fca-9* mutant, which seems much smaller that the number in A; C. two examples of introns with 'Intron retention': left: retention level reduced in *fca-9*, right: retention increased in *fca-9*

5.2.2 FCA Affects Splicing Efficiency of Introns

Since the number of AS event was not very large in *fca-9*, I later turned to the splicing efficiency to see whether FCA affects co-transcription splicing. In previous research (Zhu et al. (2020a)), RNA-binding protein RZ-1B and RZ-1C were shown to affect the co-transcriptional splicing efficiency of introns. We wondered whether FCA would have similar effects on splicing.

To check FCA binding on introns, I first designed a method to check whether splicing event occurs at the annotated intron. In this step, we can identify all possible splicing events through exon-exon junction reads. But in the later calculation, we would use exon-intron/intron-exon junction reads. If the both splicing event and proximal polyadenylation events happen in the same intronic region annotated, it might hard to tell whether the exon-intron junction reads should be counted as "unspliced" reads which has not been spliced co-transcriptionally or part of the polyadenylated transcripts which would not be spliced. Here, I meant to avoid dealing with reads located in region with multiple processing results, such as alternative splicing site or proximal polyadenylation within introns. In the case of FCA, the proximal PAS should be taken extra care since $3'$ end also seems a significant target of FCA and the binding in different regions might carry out under distinct mechanisms. In this section, a series of custom methods were used to locate intron region based on the annotation file and experimental sequencing data of the wild type.

After locating introns, I would use the RZ1B/C research as a reference for studying the change of splicing efficiency in *fca-9* compared with the wild type. Not only the splicing site (SS) ratio in wild type would be calculated, but also the change of SS ratio after mutation. We were interested in finding genes with significant change in splicing efficiency. By these results, we aimed to obtain further insights into features of FCA-binding introns.

5.2.2.1 Re-define Intronic Regions Based on $5'$SS and $3'$SS Ratio

Though few AS events were detected in *fca-9* mutant at steady state, it is still possible that FCA might affect the co-transcriptional splicing (CTS) efficiency of the introns. Here I used the method defined before (Zhu et al. (2020a)) for computing CTS efficiency:

At each boundary between exon and intron, the splicing site ratio (SS ratio) was calculated with the formula below:

$$5' \text{ SS Ratio} = \frac{5' \text{ intronic reads}}{5' \text{ exonic reads}}$$

$$3' \text{ SS Ratio} = \frac{3' \text{ intronic reads}}{3' \text{ exonic reads}}$$

Here, the number '$5'$ intronic reads' is the count of reads in a 25-bp window next to the $5'$ splicing site in the intron, and the number '$5'$ exonic reads' is the read count in the 25-bp window in the exon upstream to the splicing site. Similarly, '$3'$intronic reads' and '$3'$ exonic reads' are calculated on the flanking region of $3'$ splicing site. Intronic reads are usually the reads not yet spliced co-transcriptional, while the exonic reads are consist of both spliced reads (also known as splicing junction reads or junction reads) and unspliced reads. By dividing the intronic read number to the exonic read number, I can get the ratio of unspliced on the junction. An SS ratio of zero indicates that the intron is fully spliced, while a ratio of 1 appears when the splicing site is not used for splicing at all. A ratio over 1 means intronic counts is larger than the exonic ones, which indicates abnormality and hence the corresponding intron should be removed from the list for analysis.

In the original method, the intron regions were extracted based on annotation file, with the ones of SS ratio over 1 filtered. But the strategy seems not good enough when dealing with special cases such as introns with poly(A) sites and alternatively spliced introns. In the four kinds of AS mentioned above, the three of them, including exon skipping, Alt $5'$ SS and Alt $3'$ SS, show multiple types of splicing events over one junction. Limited by short-read sequencing technique, the transcripts were fragmented to lengths of 200˜500 bps in library construction. Currently, it is not possible for us to define which splicing isoform does a non-junction read belong with our sequencing data, making the calculation of splicing efficiency inaccurate in these positions. (However, this issue might be solved in the future with Nanopore Sequencing, see Chapter 6). As for the intronic poly(A) site, it is hard to distinguish the premature termination events from the newly transcribed RNA which have not been spliced in CB-RNA-seq.

To remove these complicated cases, I would look at splicing events in mRNA-seq first. Compared to CB-RNA-seq data, mRNA-seq provides more spliced reads. Similarly, I can calculate SS ratio with mRNA-seq of wild type Col-0. There exist regions annotated as introns but are not spliced in any transcript, which should not be considered as introns since there's actually no splicing event occurred. Therefore, a reasonable SS ratio should belong to the region [0, 1). Another feature of SS ratio of mRNA-seq is that it should be balanced on both splicing sites of a single intron. For mature RNAs, a single splicing event can be simply considered as an independent event with only two outputs, spliced or unspliced. But most genes are transcribed multiple times, resulting in the final mRNA set which may have multiple splicing events at one intron. In the extreme case, all the introns are spliced, which is quite common in mRNA-seq. For an intron without alternative $5'$ SS or $3'$ SS, the whole intron region should be spliced at the same time so that the $5'$ SS ratio and $3'$ SS ratio of a certain intron should be the same. Even though these two values are not necessarily the same in many cases, possibly due to bias in the experiment, they should not be of significant difference.

To avoid repeated calculation of the same region, I first merged the introns in different isoforms but of same start and end position into a single intron. No matter which transcript the intron belongs, the splicing event of the same place is considered as the same event. Meanwhile, result calculated is not reliable when the coverage of the gene is too low. Therefore, lower coverage genes with lower than 10 reads are removed from calculation to gain trust-worthy intron regions.

After getting reliable $5'$SS and $3'$SS ratio of all the introns in expressed genes, I can then classify the introns into three groups by the value of both $5'$SS and $3'$SS ratio:

1. Fully spliced introns, both the $5'$SS and $3'$SS ratio of which are 0;
2. Partially spliced introns, with intron retained in at least one biological replicates, both $5'$SS and $3'$SS are within the range [0, 1);
3. Unspliced introns, with any of the $5'$SS/$3'$SS over 1, implicating the intron is not spliced or has a $5'$SS/$3'$SS which was barely used.

Here, we mainly focus on FCA-binding introns in Col-0. The introns annotated but unspliced will be discarded.

In the practice, I found it difficult to distinguish Alt 5′SS/Alt 3′SS by SS ratio only. To solve this issue, I could refer to the direct evidence of splicing event, exon-exon junction read in mRNA-seq data. The introns annotated but with no supporting reads in any replicates of Col-0 mRNA-seq would be considered as an exonic region instead. Thanks to Pysam (Li et al. (2009)) module, it is convenient to process BAM/SAM file with Python. I first get the introns with junction reads in all samples and removed ones with overlapping to avoid Alt 5′SS/Alt 3′SS. I got 97734 introns of 16909 genes. After filtering, the read number in the 25-base window, before and after the splicing site, would be counted. Note that the 'bamfile.fetch' function of Pysam would include the spliced reads overlapped with the target region, but I only expected unspliced reads in intronic regions. So, the spliced reads should be removed when counting reads in introns.

While I removed most cases Alt 5′SS/Alt 3′SS with the previous process with Pysam, there still remained novel splicing site or splicing sites with premature termination sites. To further remove these cases, I apply t-test to compare between the 5′SS and 3′SS ratio of the three repeats, followed by a B-H process to get the adjusted P-value (False Discovery Rate, FDR). All the introns with FDR<0.05, rejecting the null hypothesis of no difference between 5′SS and 3′SS ratio, were removed. I could later check the introns with higher 5′SS ratio for existence of proximal PASs.

After the removal of unspliced introns and introns with unequal 5′ SS and 3′ SS ratio, 91007 introns was left. With the remaining introns, I plotted the distribution of the SS ratio (in Figure 5.4 A). Judging from the histogram, the distribution of SS ratio of these introns did not follow a normal distribution. The idea was later confirmed by KS test, rejecting the null hypothesis of distribution of SS ratio and a normal distribution are identical. Here, I calculated the Non-parametric Spearman correlation coefficient (Spearman r) rather than Pearson correlation coefficient (Pearson r) for underlying normal distribution. The Spearman r between 5′ SS ratio and 3′ SS ratio is 0.79, indicating a strong positive correlation between the two.

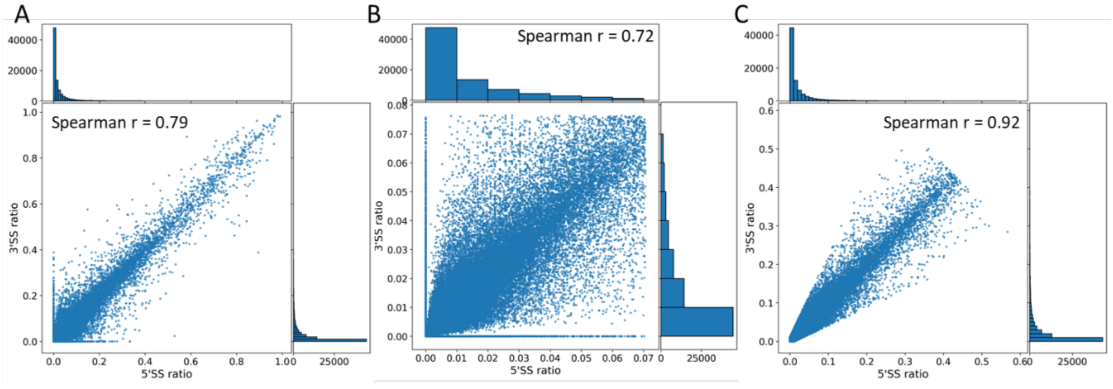


Figure 5.4: Distribution of SS ratio and the scatter plot for 5′SS ratio and 3′SS ratio in Col-0: A. distribution of SS ratio after removal of possible alternative spliced introns and introns with poly(A) sites; B. a trail in removal of outliers; C. the finial distribution of SS ratio of Col0-0 at mRNA level

While the two groups are of high correlation, I can still judge from the scatter plot that there exist introns of high retention rate and of unequal SS ratio on 5′ site and 3′ site. Here, I tried to remove the outliers to see if that can make any difference and help improve the correlation coefficient. I first applied the Box-Cox transform to the non-norm SS ratio dataset to make it close to normal. After transformation, I still got a non-norm distribution. So, the original data was used, with 12347 (about 13% of the total) introns with SS ratio over the threshold (Q3+1.5*(Q3-Q1)) to be removed. And I get a new distribution shown in Figure 5.4 B. However, the filtering step did not help us get a higher efficiency but only removed genes which might be unspliced. I can still see introns with the SS ratio on one side close to zero, but much higher on the other. To remove these introns, I applied the Coefficient of Variation (CV) of mean value of 5′ SS and 3′ SS ratio:

$$CV = \frac{|5'\text{SS ratio} - 3'\text{SS ratio}|}{5'\text{SS ratio} + 3'\text{SS ratio} + 0.01}$$

Here I added 0.01 to the sum of the two values due to the following consideration. For an intron with one SS ratio equals zero, and the other over 0, if no extra constant value is added to the denominator, the CV value would always be 1. In this case, for points on the scatter plot (x-5′SS ratio, y-3′SS ratio), there would be no difference between (0,0.01) and (0, 0.99) in CV value, but the first one should be kept while the second one should be removed for the later analysis. By adding a constant value of 0.01, I could make the CV of the two sites different and removed the second point.
After several trails, I finally set the rule for filtering:

1. Remove the introns of the top 1% of SS ratio, after this step the maximum of average SS ratio is 0.43, which means over half of the transcripts with the intron was spliced;
2. Filter the introns with outliers of CV. This helped increased the correlation efficiency to 0.92, and a reasonable number of 83354 introns remained at last.

This distribution (Figure 5.4 C) probably can be seen as a superposition of three distributions, one for fully spliced introns, one for partially spliced introns and one for (nearly) unspliced introns, but there showed no extra peaks near the ratio of 1. After the process, SS ratio of introns remained were no larger than 0.43, indicating over half of the transcripts had the intron spliced, which seems reasonable at mRNA level. The final set of introns was composed of two parts, the fully spliced introns and the partially spliced introns. This intron set would be used in later analysis of CB-RNA-seq.

5.2.2.2 Change of Splicing Efficiency in *fca-9*

First, I checked whether splicing efficiency is changed in *fca-9* using SS ratios of CB-RNA-Seq of *fca-9* and Col-0 of the selected intron set, especially in the FCA-binding introns.
Even with restricted filtering steps mentioned in this section, we noticed that there were some introns with unreasonable SS ratios over 1. After removal of intron with no expression on the flanking exons, I get 82662 introns remained. Since these were all spliced introns at

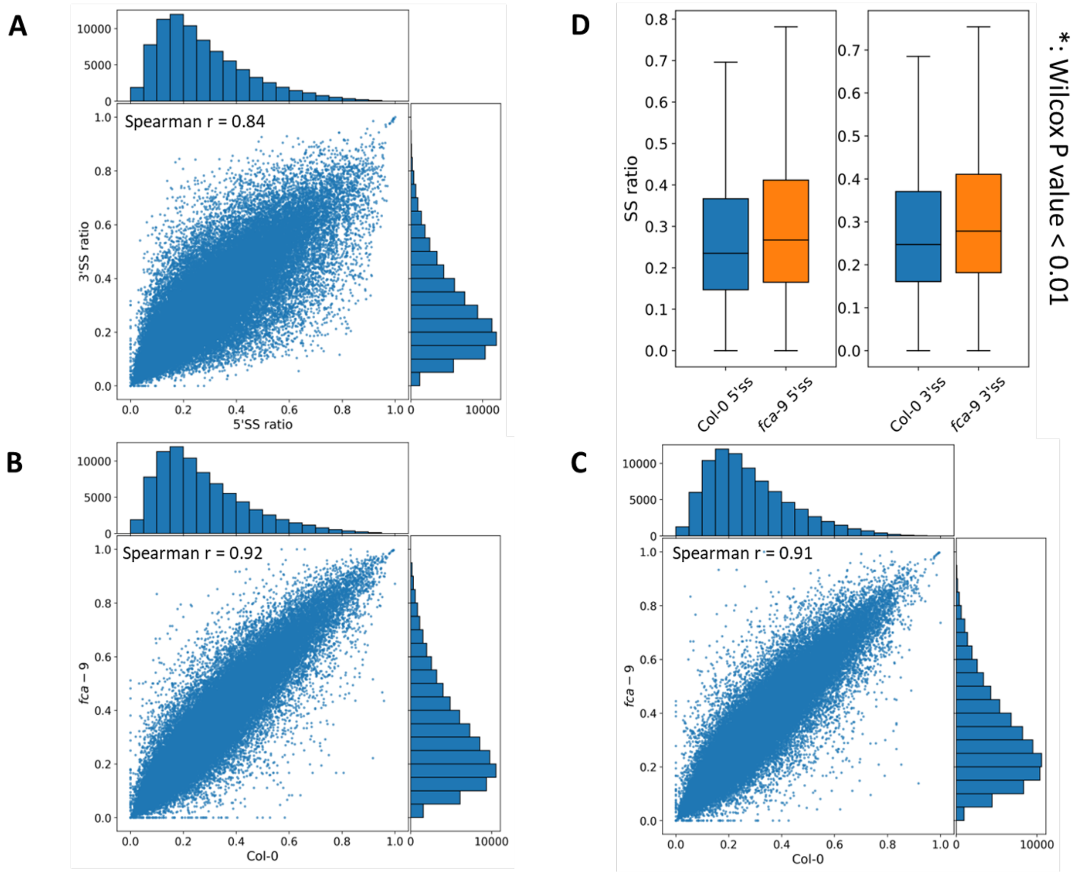


Figure 5.5: Distribution of SS ratio (CB-RNA-seq) A. Comparison between 5′ and 3′SS ratio; B. Comparison between 5′SS ratio of Col-0 and *fca-9* shows high correlation efficiency; C. Comparison between 3′SS ratio of Col-0 and *fca-9*; D. Boxplot of SS ratio (blue: Col-0; orange: *fca-9*), '*' means Wilcox P value < 0.01
, the distribution of SS ratio in *fca-9* is different to the one in Col-0. Co-transcription termination efficiency might be affected by loss of FCA function.

the steady state, I did not remove the values over 1 as they might indicate post-transcription splicing events, but only SS ratios under 1 were presented in the figures below.

Similar to the SS ratio in mRNA-Seq, the values in CB-RNA-Seq were not normally distributed. Again, Spearman Correlation Coefficient was used to see if the values of the two group have a positive correlation. As seen from Figure 5.5 A, B and C, the correlation coefficient between 5′SS ratio and 3′SS ratio of the wild type, and both 5′SS ratio and 3′SS ratio of Col-0 and *fca-9* were high, indicating that these values are highly correlated.

A plausible explanation was that loss of FCA function had little impact on the overall splicing efficiency. But it did not mean that the SS ratio showed no change after loss of function of FCA. In the boxplot of SS ratio, we can see that the median value in *fca-9* is slightly higher than the one of Col-0. Afterward, I used the Wilcoxon signed-rank test to test the null hypothesis that two related paired samples come from the same distribution. As a result, the null hypothesis was rejected. While the SS ratio in *fca-9* and *Col-0* are highly correlated, the ratio in *fca-9* is slightly higher, indicating that the splicing efficiency decreased in *fca-9*.

Next, I compared the intron regions one by one to find introns of changed splicing efficiency. By comparing the SS ratio of the wild type with that of the mutant, I can get introns of significant changes near both splicing sites. 4080 genes was identified to have significant change of SS ratio (p-value¡0.05, t-test attached with B-H process) on both $5'$ and $3'$ site in comparison between *fca-9* and *Col-0*, which 3887 introns of significantly increased SS ratio and 193 genes of decreased ratio. The result here corresponded to the boxplot and again confirmed the idea that FCA would affect the splicing efficiency of introns.

Since it is likely that FCA promote the efficiency of splicing, we were interested in whether the impact directly relies on the binding of FCA. Here, I would check whether these introns of significant change in the SS ratio were also FCA-binding introns.

5.2.2.3 Splicing Efficiency of FCA-binding Introns

While the loss of FCA caused a decrease in the overall splicing efficiency, I next tried to narrow down the range of intron for research to those with FCA-binding, checking if loss of FCA-binding would affect the splicing of these introns. As discussed in Chapter 4, it is believed that FCA binds to RNA under transcription. Many FCA-binding reads overlapped with the exon-intron junction, indicating they had not been spliced by the moment when FCA bound to the region. Meanwhile, the strong binding on the first base of the intronic region indicated that FCA binding might happened after cleavage of the $5'$SS or might be blocked by the U1 snRNP or other protein complex binding to the 5′SS. No matter under which mechanism FCA binds to the introns, it should be easier for the protein to bind if the intron was retained or need more time for splicing. The intron retention level (5′SS ratio and 3′SS ratio of mRNA-Seq) and the splicing efficiency (5′SS ratio and 3′SS ratio of CB-RNA-Seq) of FCA-binding introns can be checked to test the hypothesis. Interestingly, FCA-binding introns showed slightly larger ratios of splicing in both mRNA-seq and CB-RNA-seq (Figure 5.6 A, B).

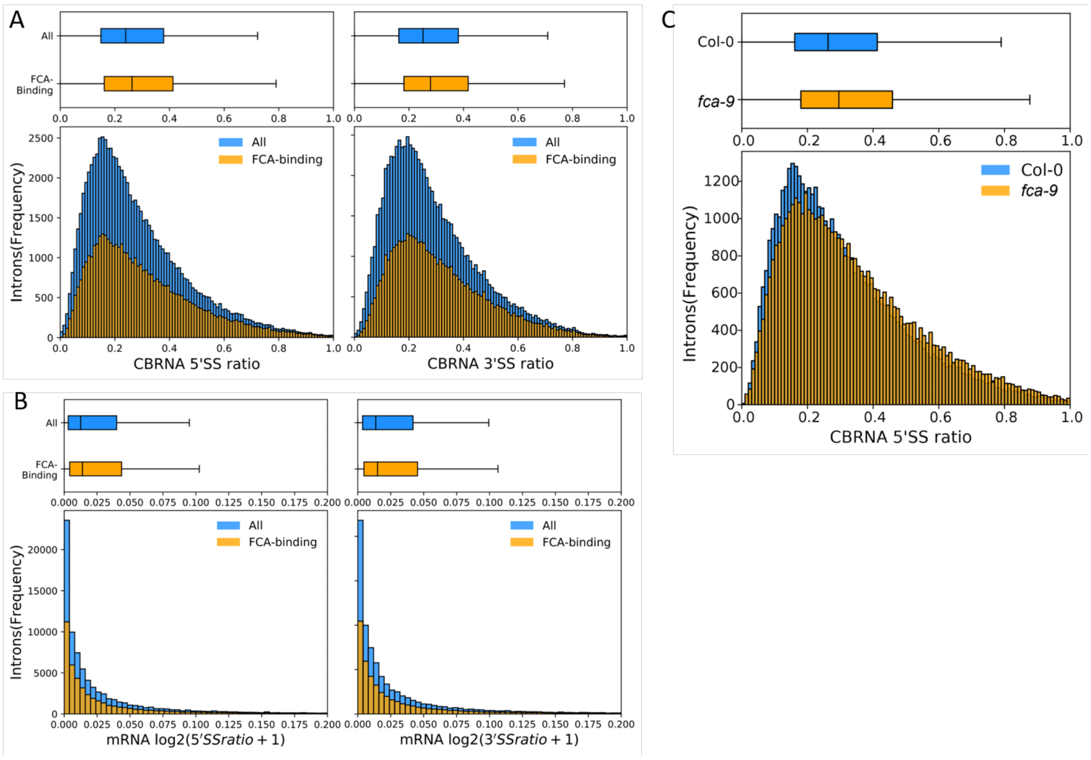


Figure 5.6: SS ratio of FCA-binding introns A. Comparison of SS ratio at CB-RNA level between FCA-binding introns and all introns in Col-0; B. Comparison of SS ratio at mRNA level between FCA-binding introns and all introns in Col-0; C. Comparison between 5′SS ratio of FCA-binding introns in *fca-9* and Col-0

The other comparison was carried out between *fca-9* mutant and Col-0, in order to see whether the splicing efficiency decreased without FCA (see Figure 5.6 C). It turns out that the median SS ratio of FCA-binding introns raised in *fca-9*, indicating a decrease in splicing efficiency without FCA-binding. We are not surprised to see the difference in the distribution of SS ratio, since 2355 of the 3886 introns with significantly increased SS ratio are FCA-binding introns. These FCA-binding introns showed a higher possibility to have up-regulated SS ratio compared to the expected (all introns included regardless of whether binding or not).

Table 5.4: Number of genes with significant change in SS ratio

	Total	**UP**	**Down**
All	**83663**	**3886**	**193**
Binding	**46137**	**2355**	**99**

5.2.3 FCA-Binding Introns Tend to be Longer and with Higher G-content

FCA showed a very specific binding to intronic regions. We can see a clear boundary of binding near both splicing sites. In other word, the binding is enriched in the intron region but rarely extend to the flanking exons. FCA bound to many introns, with binding of FCA detected on over half of the re-defined introns. Since FCA binding on introns are specific and wide-spreading, the feature of these intron should be worth looking at to see the possible trigger of FCA-binding. I checked whether these introns have common features in length and sequence.

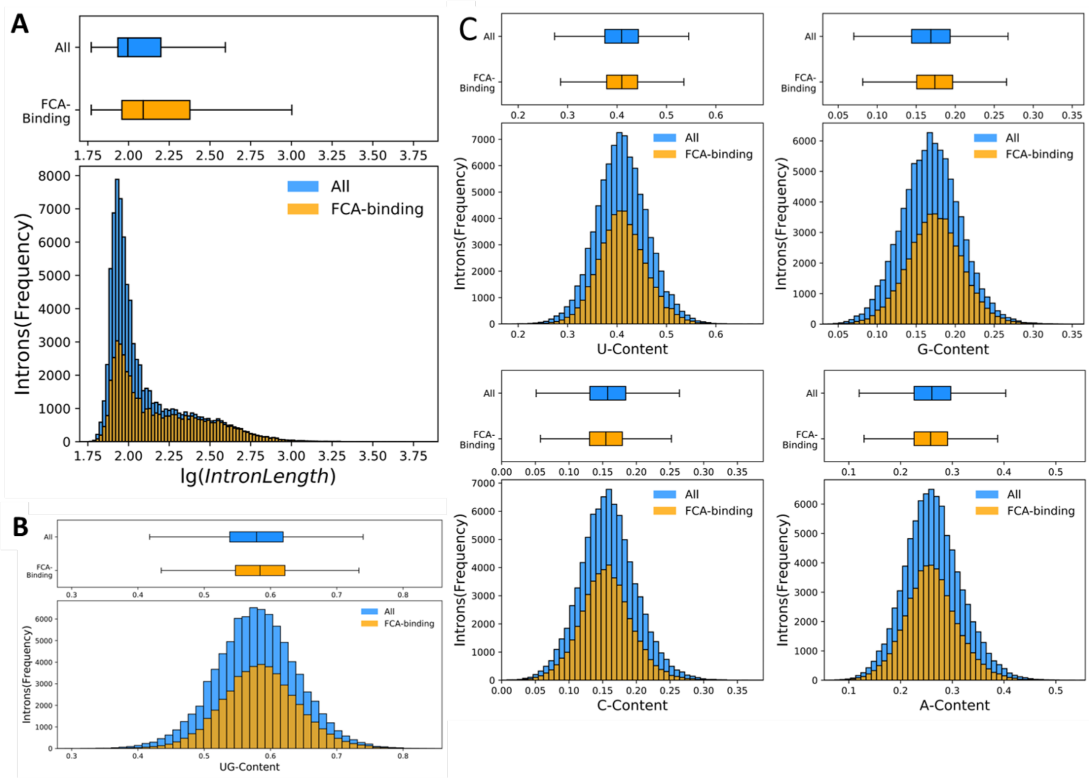


Figure 5.7: Comparison between FCA-binding intron and all introns in distribution of intron length and base content A. intron length; B. UG content C. base content (up left: U content, up right: G content, down left: C content, down right: A content)

5.2.3.1 FCA-Binding Introns Have Longer Length

Before analysis, we noticed that FCA bond to the largest intron of its own transcript. But that might a special case since a proximal poly(A) site was also detected in this intron. Here I plotted the distribution of intron length with histogram and boxplot (Figure 5.7 A). The lengths of introns varied from 59 to 7384, so here I used the log value of the length for plotting. To further investigate whether the length distribution of FCA-binding introns was different from that of all introns, Mann-Whitney U-test was applied to see whether the two sets of intron lengths (Blue: All introns; Orange: FCA-binding Introns) were from the same distribution. The null hypothesis was that they are from the same distribution. As a result, the null hypothesis was rejected (p<0.01). It is relatively easy to find that FCA-binding introns were of larger size.

On the other hand, I had noticed that introns in Arabidopsis are typically short, with the median length around 100 bps. According to the boxplot (Figure 5.7 A), introns of length over 266 are outliers, and hence I define these introns as ‘extremely long introns’. Interestingly, 9819 of 10687 extremely long introns are bound by FCA, while the rest 72667 introns had only 36399 with FCA binding. It seems that FCA had a strong preference for these very long introns. I next compared the length of FCA-binding intron to other introns in the same gene. Among 9302 genes with at least three introns left after filtering, 8320 of them have the largest intron bound by FCA while FCA may also bind other shorter introns of the gene. These observations all served as evidence for FCA’s preference for long introns.

5.2.3.2 FCA-Binding Introns Have Slightly Higher G-content

FCA binds poly(U) and poly(G) sequence in vitro. We have already known that the U-content in introns is usually higher than that in exons, which might trigger FCA to ‘recognize’ intronic regions rather than exonic regions. It is also highly plausible that the introns with FCA binding are of higher U or G level than non-binding ones. Here I first checked UG-content (the content of the base of U and G of all bases) of the introns (Figure 5.7 B). Mann-Whitney U-test was applied to check whether the FCA-binding group and the group of all introns belongs to the same distribution. It turned out that FCA-binding introns had slightly higher UG content. Next, I looked at the content of each base (Figure 5.7 C) to see whether U-content and G-content were higher in FCA-binding introns. Interestingly, I found both U- and G-content higher, while C- and A-content lower in binding introns (U-test p-value¡0.01). We can also see that introns are of high U-content with a median of 0.41 while G-content in introns is relevantly low with a median of 0.17.

We next checked whether the G-content was relevant to the intron length. For example, the number of Gs in introns were linearly correlated, which means the longer an intron is, the higher its G-content. Thus, these two features are dependent. While in the case of G content, it showed very weak correlation with intron length and can be considered as independent of intron length.

The difference of G content in the FCA-binding group and the group of total introns was not as significant as the one in intron length. Meanwhile, many of the introns with high G content were not bound by FCA. These clues indicated that G content might not the deciding feature of FCA binding.

5.2.4 Section Summary

In this section, we study the impact of FCA on splicing events. The splicing of mRNA was weakly affected by the loss of FCA, judging from the number of alternative splicing event detected. Comparing to the subunit of U1 snRNP, member of spliceosome, FCA alone plays a role of less importance. But it does not mean that FCA have no impact to splicing co-transcriptionally. We have found from the CB-RNA-Seq data that SS ratio increased in *fca-9*, which means FCA might promote splicing co-transcriptionally. Another interesting observation about SS ratio was that the median SS ratio of FCA-binding introns was higher than the expected ratio among all introns, indicating FCA is more likely to bind to introns with low splicing efficiency. We have also observed that FCA-binding introns are typically larger in size and have slightly higher G-content.

5.3 FCA's Binding to $3'$ End and Proximal Downstream of Genes

In this section, we turned to another region with enriched FCA binding, the $3'$ end. Unlike in introns, we cannot find a very clear boundary for binding near $3'$ end. Not only $3'$ of the gene body but also the downstream are bound by FCA. Moreover, the binding of FCA seems to extend with the transcription, even to the further downstream region in *fca-9* mutants with read-through transcripts. I showed some examples of FCA-binding to read-through transcripts in last chapter(see Figure 4.6). Here I would like to see if the binding of FCA to the read-through is a common phenomenon.

5.3.1 FCA's Binding Extends with Transcription of Read-through Transcripts in *fpa-7*

In the last chapter, I found that FCA binds to further downstream region in *fpa-7* which might correspond to the read-through transcripts in the mutant. Here, I tried some different methods to find the exact binding region transcriptome-wide.

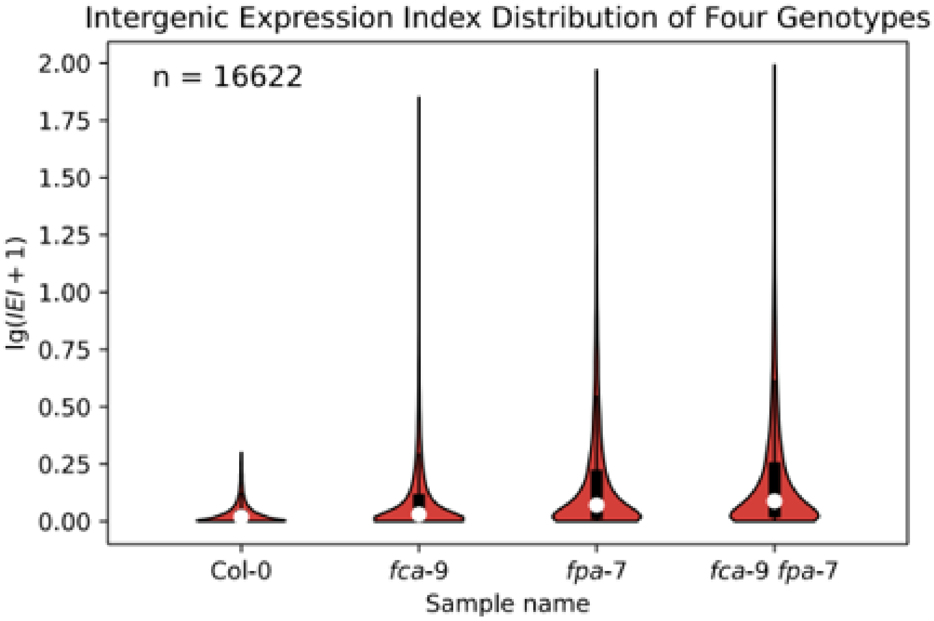


Figure 5.8: Increased Intergenic Expression Index (IEI) in *fca-9*/*fpa-7*/*fca-9 fpa-7* mutants, which might indicate the existence of read-through transcripts in these mutants.

Intergenic Expression Index (IEI) increased in *fpa-7*

To make sure that the read-through transcripts exist, we first check whether the expression of intergenic region increased in the mutants (see Figure 5.8). To this end, we introduced an index called intergenic expression index (IEI) as below (5.1):

$$\begin{aligned} IEI(\text{GeneA}) = \frac{ReadCount(\text{Downstream}) \times \text{Exonic Length}}{ReadCount(\text{Exonic Regions}) \times 200nt} \times 100\% \\ IEI \in [0\%, 100\%] \end{aligned} \tag{5.1}$$

$ReadCount$(Downstream) is the read count in 200-nt region downstream of GeneA. IEI represents the expression level of proximal downstream region, normalized to the expression level of exonic region of the gene. Note that the expression in intergenic regions is not necessarily equal to read-through because one cannot be sure whether the transcription is continuous from the upstream gene. Nevertheless, the IEI can still serve as a useful signal for the possibility of read-through. Here, we calculated IEI with mRNA-seq data from Col-0, *fca-9*, *fpa-7*, and *fca-9 fpa-7*.

To calculate the index, we need to get the read count of the gene body and that of the intergenic region, normalize by the length of the region and divide the normalized intergenic expression level by the level in the intragenic region upstream. Before calculation, gene that are too small in size, low in expression and too close to the downstream gene would be removed from the target gene list.

As preparation, I first calculated the length of each transcript. Transcripts of small size (size < 300 bps) was removed. To make sure that the genes were expressed and get ready for the next step, I get the reads count of each base to calculate the average number of reads per base pair of all transcripts in the list. Transcripts of low expression (< 1 read/bp) was removed. The following step was designed to get the exact PAS in Col-0. The actual PAS is sometimes different from the one annotated, but usually located in the $3'$ UTR region. Limited by the mRNA-sequencing technique, I cannot get the PAS of single-nucleotide resolution. Instead, I would scan the last exon of each transcript with a sliding window of 20 nucleotides and stop when the expression in the region is lower than 1% exonic expression, taking the 11th base of the region as the modified PAS. This step was designed to make sure that the region downstream of the modified PAS is of low expression level, which would help to increase the sensitivity in getting possible read-through transcripts. After the modification, I ensured that 99% of the transcripts have been terminated by the re-defined PAS. Afterwards, multiple transcription isoforms of the same gene would be merged to form one and the only transcript with the most distal modified PAS. Before the formal calculation, the last step was to calculate the Intergenic Expression Level of the wild type to remove genes of which the IEI were lager than resulting from expression of unannotated gene in the downstream region of 200 bps. With the preparation above, I've got the target transcripts and the modified PAS where most of the transcripts have terminated in wild type.

I next calculated the IEI of all the samples and compared the IEI of mutants to the wild type. It can be told from Figure 5.8 that transcription of intergenic regions increased in mutants, especially in *fpa-7* and the double mutant of *fca-9 fpa-7.* Through the rough method, I ensured that the expression level in intergenic region was raised in the mutants of *fca-9*, *fpa-7*, and the double mutant *fca-9 fpa-7.* In the future research, improvements are in need to make it possible to recognize novel introns.

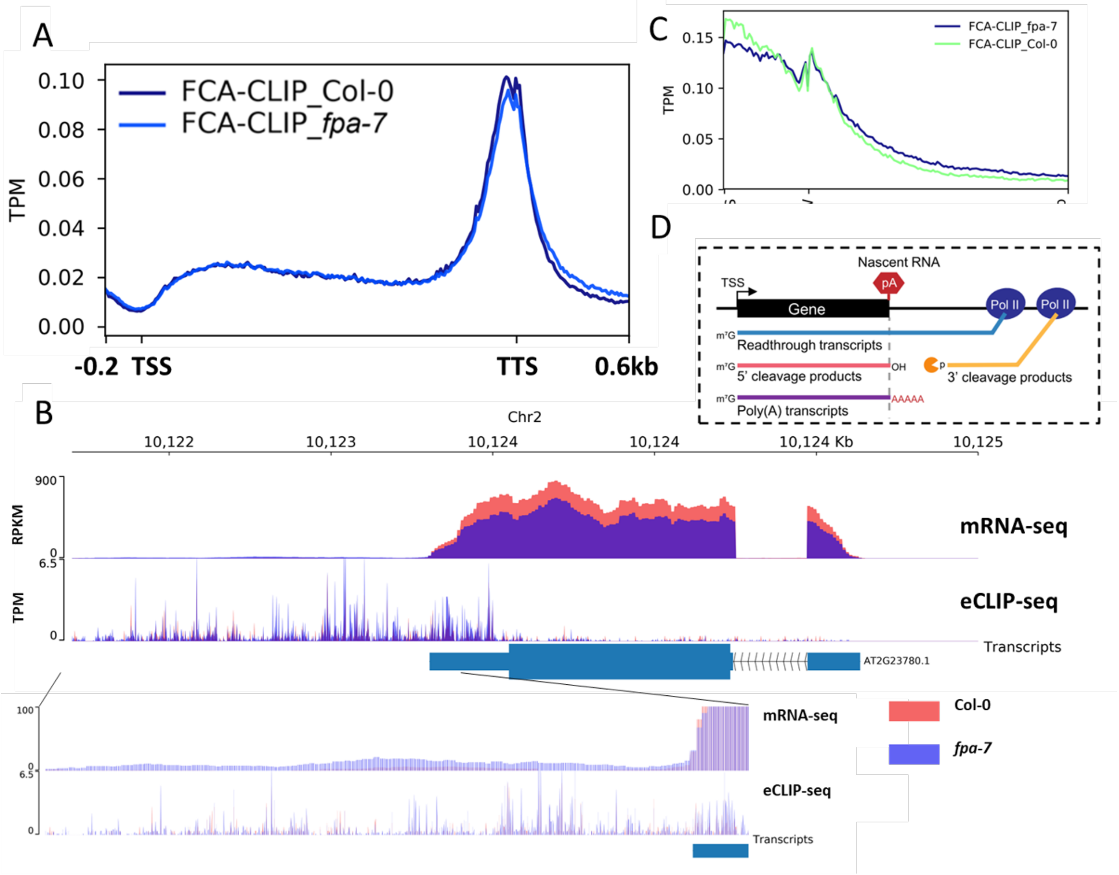


Figure 5.9: Read-through in *fpa-7* A. The overall binding pattern of FCA in Col-0 and *fpa-7* mutant; B. An example of FCA's binding to read-through transcripts. The expression level of AT2G23780 decreased in *fpa-7* but the read-through level increased. FCA's binding seems strengthened with read-through level. C.increased FCA binding after termination window; D. the four types of transcripts detected by FLEP-seq (described in Mo (2021)),;

Increased binding of FCA after Termination Window (TW) in *fpa-7*

It can be seen from the meta-profile that FCA's binding increased in *fpa-7* comparing to the wild type from Figure 5.9 A. An example shown in Figure 5.9 B indicates that the increase of binding corresponds with the increased level of read-through transcripts in *fpa-7*. The difference seem to appear during transcriptional termination stage. To describe the termination more precisely, here I used the 'PAS' and 'Termination Window' from the paper (Mo (2021)), generated from FLEP-seq with nanopore sequencing. In this paper, researchers used nuclear RNAs after rRNA removal as material and get four kinds of reads according to their state of termination (see Figure 5.9 C): read-through reads which have not been cleaved, $5'$ termination intermediates after cleavage but have not been polyadenylated, $3'$ termination intermediates not yet degraded, and the reads with poly(A) tails. The PAS information was gained from the last kind of transcripts while the termination window (TW) was defined as the median value of read-through length.

The binding of FCA extends with transcribing of RNA to the downstream of genes as well, but I were not sure where the transcription of pre-mRNAs ends even in wild type, making it hard to find ‘extended’ binding of FCA under *fpa-7* mutant background. But the FLEP-seq data provided the ‘Termination window’ which corresponds to the read-through of nascent RNAs. We plotted the meta-profile of FCA-binding in Col-0 and *fpa-7* mutant (Figure 5.9 C). It is interesting to see that binding of FCA increased in *fpa-7* after the termination window comparing to the one in Col-0. Given the termination window is an estimate of where the majority of transcription events terminate, and also given that the read-through transcripts are generally increased in *fpa-7* mutant, the increased binding of FCA downstream of termination window support our hypothesis that FCA's binding is sensitive and specific to the read through transcripts.

However, single read-through transcripts are not easy to detect at mRNA level. First, the read-through transcripts are of relevantly low expression level comparing to the normal transcripts. Secondly, novel splicing events happen in these transcripts which is unexpected. Especially in the case with very low coverage, where junction reads of the novel intron cannot be detected, it would be almost impossible to locate the exact position of read-through. We are planning to design a method for detecting the exact genes with read-through later.

5.3.2 Loss of FCA Affects Termination Transcriptome-Widely

In this part, we used two kinds of sequencing data, mRNA-seq and $3'$mRNA-seq, to check whether loss of FCA function would cause terminational defects especially switch of poly(A) site usage transcriptome-widely.

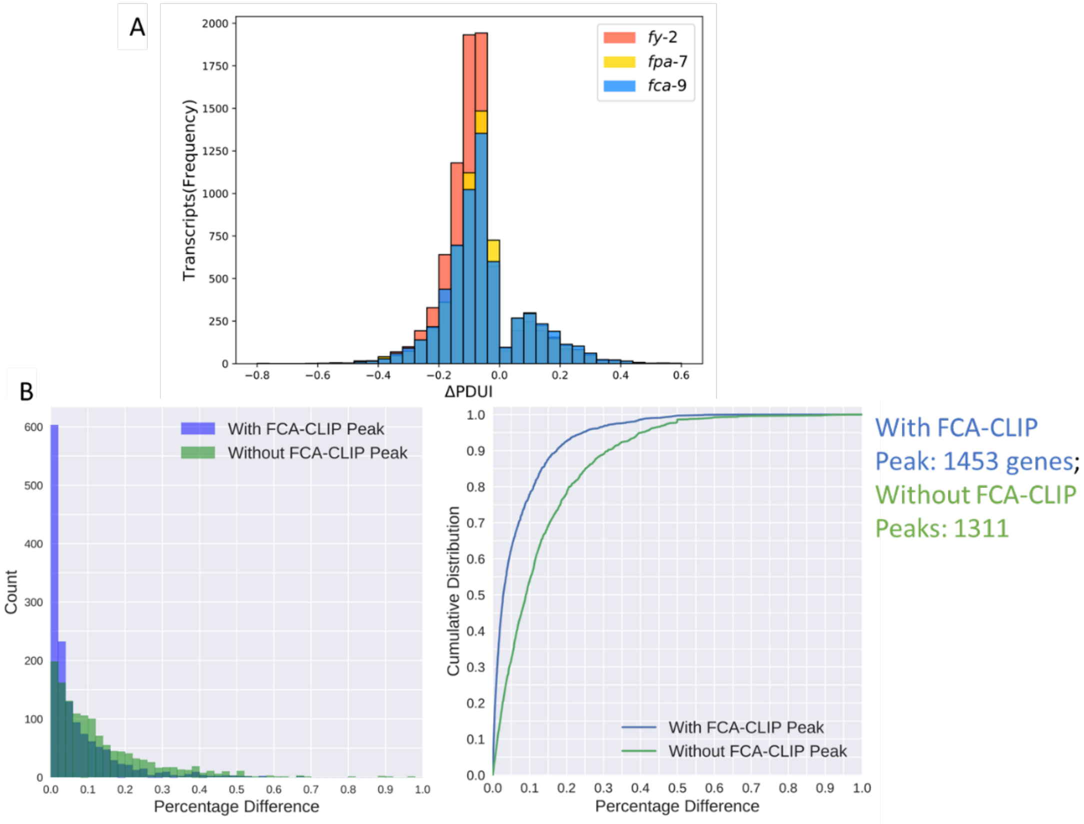


Figure 5.10: FCA affect transcriptional terminataion directly and indirectly: A. Analysis of DaPars result, genes with changed PDUI in the three mutants of *fy-2*, *fpa-7* and *fca-9*. In all the mutants, we can detect changes in length of $3'$ UTR in a bunch of genes; B. left: The distribution of Percentage Difference (PD) in genes with and without FCA-CLIP peak, right: the cumulative curve of the PD distribution; PD represents the change in Poly(A) site usage; Changes in usage of poly(A) site can be detected in both FCA-binding genes and genes without FCA's binding.

5.3.2.1 Detect Alternative Polyadenylation Events with mRNA-seq Data

DaPars (Xia et al. (2014)) is a tool developed for identification of distal Poly(A) sites usage from mRNA-seq data. The exact position of the distal and proximal PAS is detected by regression model below:

$$(\omega_L^{1*}, \omega_L^{2*}, \omega_S^{1*}, \omega_S^{2*}, P^*) = \underset{\omega_L^1, \omega_L^2, \omega_S^1, \omega_S^2 \geq 0, 1<P<L_{i=1}}{\arg\min} \sum_{i=1}^{2} \parallel \boldsymbol{C}_i - (\omega_L^i \boldsymbol{I}_L + \omega_S^i \boldsymbol{I}_P) \parallel_2^2$$

ω_L^i and ω_S^i are the abundance of transcripts with distal and proximal poly(A) sites for sample i; i; $\boldsymbol{C}_i = [\boldsymbol{C}_{i1}, ..., \boldsymbol{C}_{ij}, ..., \boldsymbol{C}_{iL}]^T$ is the read coverage of sample i at single nucleotide resolution normalized by total sequencing depth; L is the longest $3'$UTR from previous step; P is the length of proximal $3'$UTR to be estimated; $\boldsymbol{I}_L$ and $\boldsymbol{I}_P$ are the indicator functions such that $\boldsymbol{I}_L = [\underbrace{1, ..., 1}_{L}]$ and $\boldsymbol{I}_P = [\underbrace{1, ..., 1}_{P}, \underbrace{0, ..., 0}_{L-P}]$. For each P, the expression levels of each transcript in both samples can be estimated by optimizing this linear regression model by quadratic programming. The optimal novel proximal poly(A) site P^* minimal objective function value. Distal PAS Usage Index ($PDUI$) for sample i as following:

$$\text{PDUI} = \frac{\omega_L^{i*}}{\omega_L^{i*} + \omega_S^{i*}}$$

In summary, the greater the $PDUI$ is, the more distal poly(A) site is used. To compare each gene between samples, ΔPDUI = PDUI(WildType) − PDUI(Mutant) was used. If ΔPDUI>0, the gene in the mutant use less percentage of distal PAS than the one in wild type; while ΔPDUI<0 indicates acceleration of distal poly(A) site usage.

ΔPDUI is a reasonable index for describing the change in distal poly(A) site usage, but the tool was originally designed for RNA-seq data of human tissues, so it only takes chromosomes in the format ‘chrN’ (N refers to chromosome number). But the genome of Arabidopsis was formatted as ‘ChrN’. I have to change the symble of chromosome in the source code before application of the tool. Meanwhile, DaPars does not support strand-specific analysis. To make full use of our strand-specific data, I split the mapped reads into two groups by their first-in-pair strand and applied the tool separately.

As a result, I compared the ΔPDUI between the three groups, *fca-9* VS Col-0, *fpa-7* VS Col-0, *fy-2* VS Col-0 (Figure 5.10 A). I found that the usage of distal poly(A) site increased in most genes of the three groups, and the loss of FY caused the most severe defect in termination. The loss of FCA would also result in the increased usage of distal poly(A) site, but the number was not as large as the one in *fy-2*. However, in the case of genes with decreased usage of distal poly(A) site, FCA affected more genes to use a proximal PAS than the other two genes.

While most genes with $3'$ UTR shortening/lengthening seems to have FCA-binding in the altered $3'$ UTR, it does not necessarily mean the switch of PAS is directly aroused by the loss of FCA binding. Here I tested the dependency between FCA-binding and PAS usage switch with Pearson's chi-square test of independence. The P value over 0.05 indicates that whether the usage of distal poly(A) site increased or decreased in *fca-9* is independent of binding of FCA on the gene.

Table 5.5: Contingency table of change in usage of distal poly(A) site and binding of FCA

Sample	To_distal	To_proximal
FCA-binding	562	274
No binding	110	74

5.3.2.2 Analysis Switch in Proximal/Distal Poly(A) Site Usage with $3'$mRNA-seq Data

Besides mRNA-seq, the information of poly(A) site usage can also be extracted from the $3'$mRNA-seq data. To compare the usage of different genotypes, I need to get the exact position of poly(A), cluster the PAS into poly(A) clusters and count the number of polyadenylation events in each cluster for comparison.

Extract Poly(A) Site Information

To get the exact poly(A) sites, reads should be aligned to the genome but only Read 1 of $3'$mRNA-seq data can be used, while Read 2 was of low sequencing quality with the poly(A) sequence at the beginning of the reads. For each fragment, the poly(A) site was only detectable when the insert size was less than the length sequenced. To maximize utilization of the reads, longer reads may be required. Here we use PE150 (SE150 is enough, but single-end sequencing is rarely used and may cost more money) and take the Read 1 as input data.

Instead of filtering the reads by length, I took another strategy described by Wallace et al. and considered only the reads that aligned to the genomes after poly(A) trimming but not before the trimming. In a common protocol of read alignment, quality control step is taken first to remove low-quality reads and adaptors. If QC is not processed, and raw data used for mapping, it is still possible for many reads to be mapped to the genome. In the case if $3'$mRNA-seq data, the reads should be like:

1. Reads originated from long insertions, with no poly(A) tail or adaptor sequenced in read1; in this case, the poly(A) site is not detectable;

2. Reads with poly(A) tail sequenced, but the region mapped on reference genome happens to end with A-rich sequence which means we still cannot detect the poly(A) site;

3. The reads can be mapped even with the poly(A)/poly(A)+adaptor, which is quite weird.

From information above, we can see that the reads mapped in the raw sequencing dataset cannot give us the exact position of poly(A) sites. So here I aligned both raw sequences (Alignment1, the yellow circle in Figure 5.11 A) and trimmed sequences (Alignment2, the blue circle in Figure 5.11 A) to the genome, and keep only read mapped in Alignment2 but not Alignment1 (the blue part which does not overlap with the yellow circle in Figure 5.11 A) for later usage.

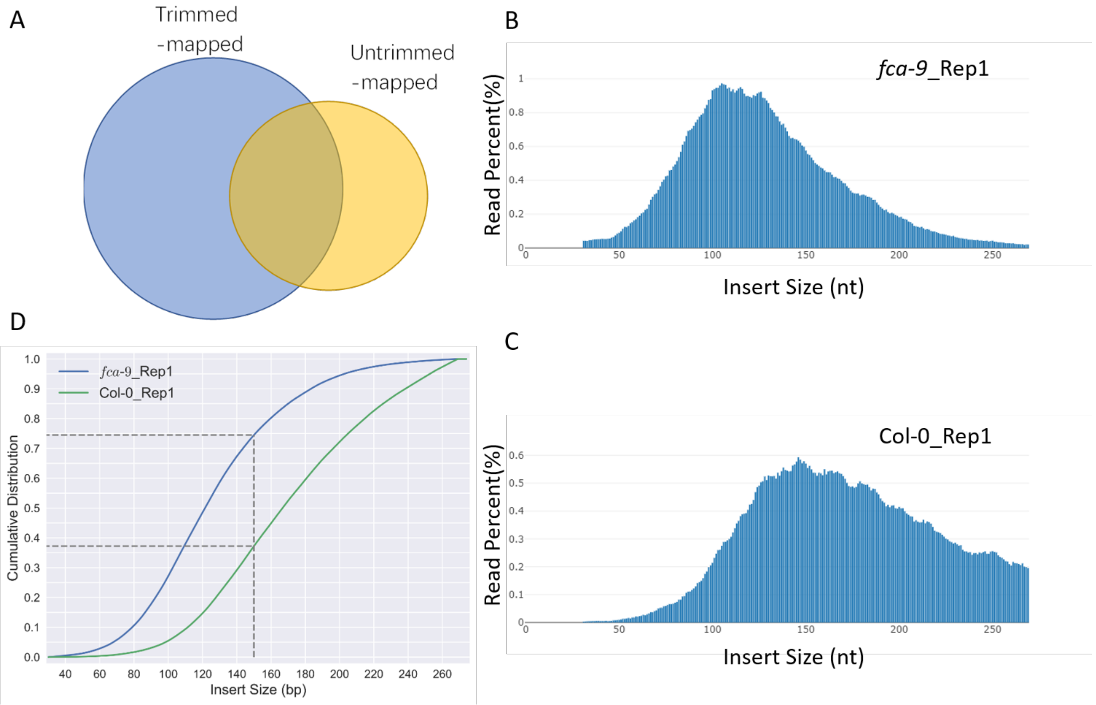
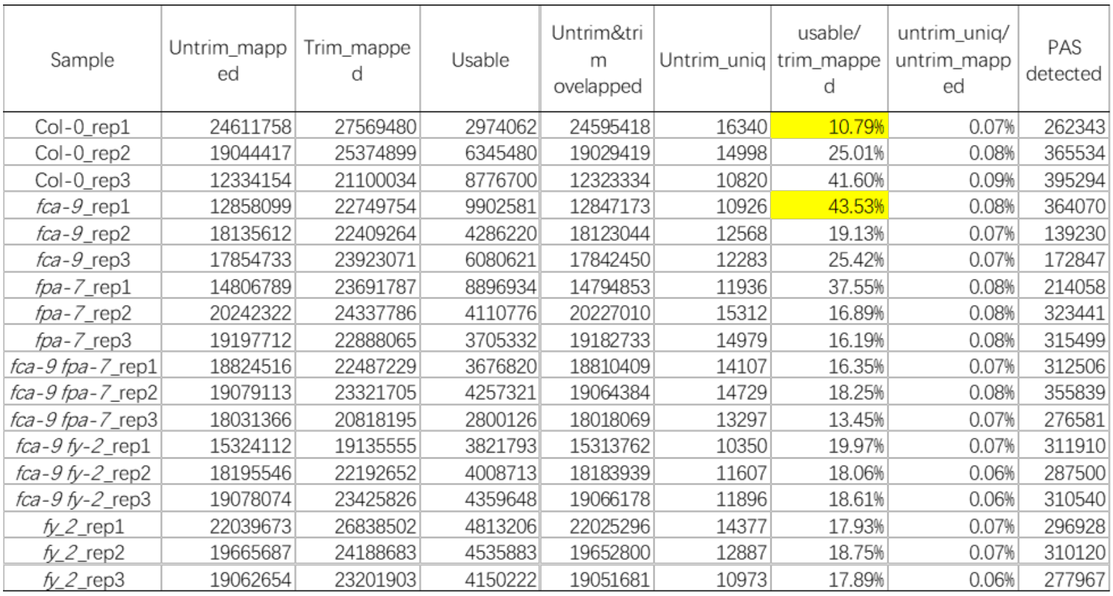


Figure 5.11: A. Method to filter usable reads, with blue circle represents mapped reads after adaptor trimming while yellow circle represents reads aligned without trimming. The reads in blue circle but not in the yellow would be used; B&C. two distinct insert size distributions of two sample *fca-9* and Col-0; D. Cumulative distribution curves of the insert size in B and C. The size distribution of the two samples are quite different. The distribution of insert size would affect usable reads ratio in identification of PAS.

For the biological replicates of each genotype, I retained the poly(A) sites in single-nucleotide resolution and keep the sites appeared in at least two replicates for later usage. In order to get trust-worthy poly(A) sites, the experimental sites are compared to the poly(A) clusters (PACs) in the PlantAPAdb (Zhu et al. (2020b)). The Result is shown as blow:

Table 5.6: Summary for 3′mRNA-seq Alignment



Sample	Untrim_mapped	Trim_mapped	Usable	Untrim&trim ovelapped	Untrim_uniq	usable/ trim_mapped	untrim_uniq/ untrim_mapped	PAS detected
Col-0_rep1	24611758	27569480	2974062	24595418	16340	10.79%	0.07%	262343
Col-0_rep2	19044417	25374899	6345480	19029419	14998	25.01%	0.08%	365534
Col-0_rep3	12334154	21100034	8776700	12323334	10820	41.60%	0.09%	395294
*fca-9*_rep1	12858099	22749754	9902581	12847173	10926	43.53%	0.08%	364070
*fca-9*_rep2	18135612	22409264	4286220	18123044	12568	19.13%	0.07%	139230
*fca-9*_rep3	17854733	23923071	6080621	17842450	12283	25.42%	0.07%	172847
*fpa-7*_rep1	14806789	23691787	8896934	14794853	11936	37.55%	0.08%	214058
*fpa-7*_rep2	20242322	24337786	4110776	20227010	15312	16.89%	0.08%	323441
*fpa-7*_rep3	19197712	22888065	3705332	19182733	14979	16.19%	0.08%	315499
*fca-9 fpa-7*_rep1	18824516	22487229	3676820	18810409	14107	16.35%	0.07%	312506
*fca-9 fpa-7*_rep2	19079113	23321705	4257321	19064384	14729	18.25%	0.08%	355839
*fca-9 fpa-7*_rep3	18031366	20818195	2800126	18018069	13297	13.45%	0.07%	276581
*fca-9 fy-2*_rep1	15324112	19135555	3821793	15313762	10350	19.97%	0.07%	311910
*fca-9 fy-2*_rep2	18195546	22192652	4008713	18183939	11607	18.06%	0.06%	287500
*fca-9 fy-2*_rep3	19078074	23425826	4359648	19066178	11896	18.61%	0.06%	310540
*fy_2*_rep1	22039673	26838502	4813206	22025296	14377	17.93%	0.07%	296928
*fy_2*_rep2	19665687	24188683	4535883	19652800	12887	18.75%	0.07%	310120
*fy_2*_rep3	19062654	23201903	4150222	19051681	10973	17.89%	0.06%	277967

It was shown in the Table 5.6 that only 18.43% (medium, 10.79%˜43.53%) reads were usable in detection of poly(A) sites. The number of usable reads is largely affected by the distribution of insertion size. Here, I checked samples with the largest and smallest proportion of usable reads and detected a great difference in the insert size (Figure 5.11 B&C), which might be the reason for the difference in portion of usable reads. The insert size can be detected by overlap between the two reads. In PE150, the maximum size of fragment to be fully sequenced should be 300 (150*2) bps, but the overlapping region of the two reads were set as 30 bps, so the finial insert length were within a range of 30˜270 bps, and most fragment have the insert size within this region. The cumulative distribution curves of the two samples (Figure 5.11 D) presented the difference in a more apparent way. It was shown that the insert with length no larger than 150 bps was around 75% in sample *fca-9* rep1 but less than 40% in Col-0.

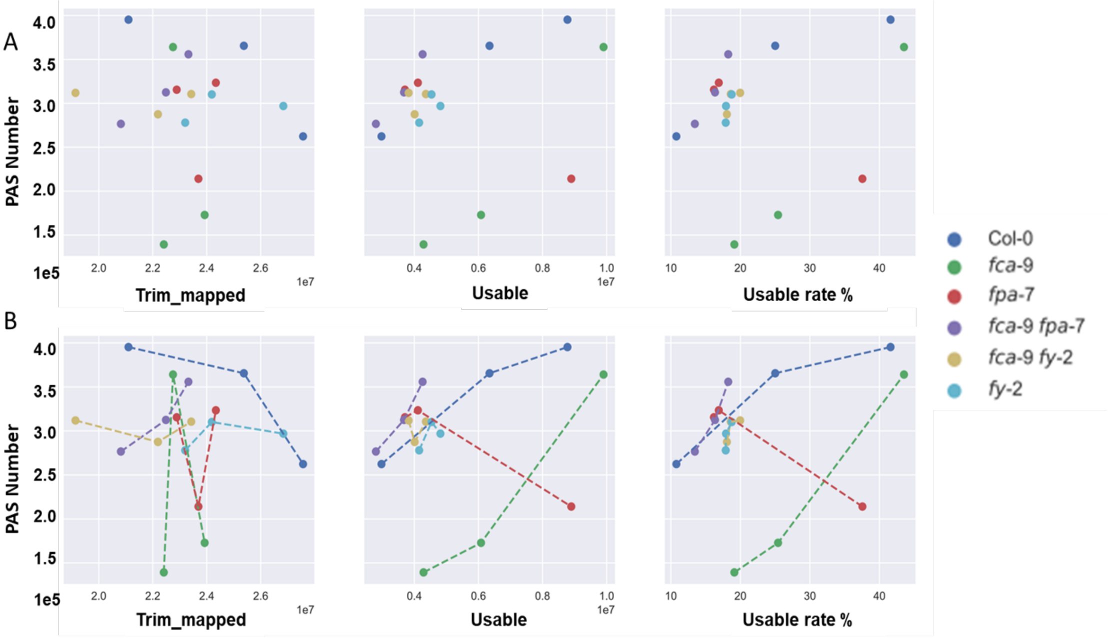
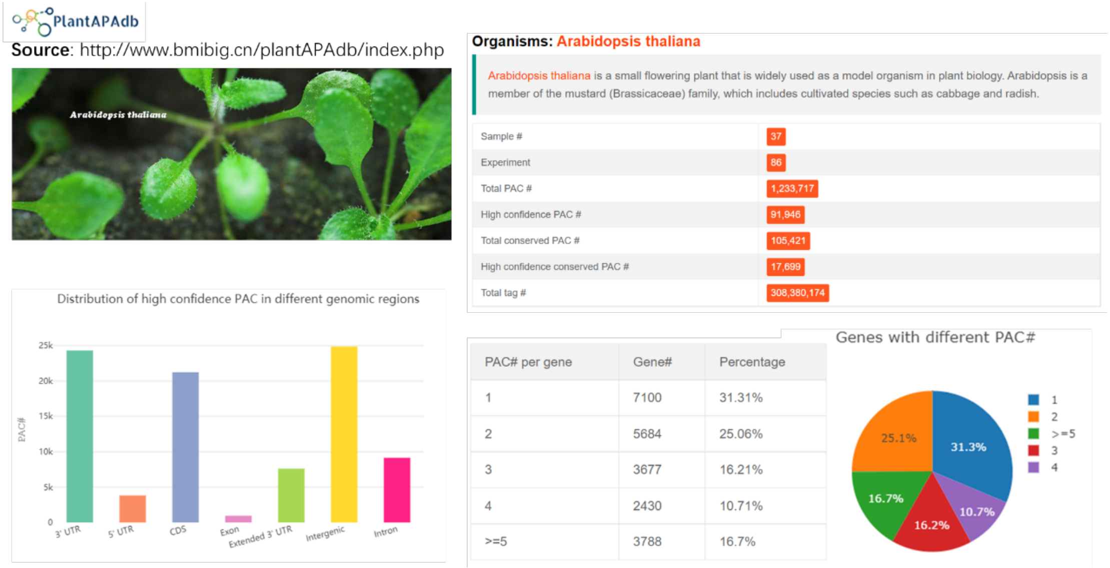


Figure 5.12: Connection between total read number/usable read number/usable number content and PAS number, A. Scatter plot, B. Scatter plot with replicates of the same sample attached by lines, with each

Even though the insert size distribution is important for PAS detection, it is actually out of control since second strand synthesis is initiated by random priming and no step for size selection like gel purification has been taken in this experimental workflow. After filtering, we consider the end of each read as the position of the PAS. Reads with end mapped to the same point are considered as originated from transcripts with the same PAS. We next checked if the number PAS detected increases with the scale of the reads mapped. Scatter plots (Figure 5.12 A) showed relationship between mapped reads number after adaptor trimming, usable read number, the percentage of reads usable and PAS number. None of the three factors show strong correlation with PAS number (Pearson's correlation coefficient <0.2). In comparison, between replicates of a same genotype (Figure 5.12 B), while in most genotype PAS number detected increased with usable read number, it is not the case in *fpa-7* mutant, indicating the number of usable reads is not the only factor to decide the PAS number, while the increase of usable read number does help to gain more PASs.

With the PAS information extracted from the usable read, we merged the three PAS replicates of each genotype into one file. The PAS appeared in at least two replicates were remained for later usage.



Source: http://www.bmibig.cn/plantAPAdb/index.php

Organisms: Arabidopsis thaliana

Arabidopsis thaliana is a small flowering plant that is widely used as a model organism in plant biology. Arabidopsis is a member of the mustard (Brassicaceae) family, which includes cultivated species such as cabbage and radish.

Sample #	37
Experiment	86
Total PAC #	1,233,717
High confidence PAC #	91,946
Total conserved PAC #	105,421
High confidence conserved PAC #	17,699
Total tag #	308,380,174

PAC# per gene	Gene#	Percentage
1	7100	31.31%
2	5684	25.06%
3	3677	16.21%
4	2430	10.71%
>=5	3788	16.7%

Figure 5.13: Introduction to PlantAPAdb (`http://www.bmibig.cn/plantAPAdb/index.php`), a database for polyadenylation sites in plants. The poly(A) clusters (PACs) available on the website was processed based on public data including PAT-seq and PAS-seq. More than one PAC was revealed on the majority of the genes, indicating the alternative splicing events widely present in Arabidpsis.

PlantAPAdb (See Figure 5.13) is a database containing APA sites of 6 species including *Oryza sativa* (*japonica and indica*), *Arabidopsis (Arabidopsis thaliana)*, *Medicago truncatula*, *Trifolium pratense*, *Phyllostachys edulis*, and *Chlamydomonas reinhardtii.* In the Arabidopsis part, 86 experiments of 37 samples in total are included. Details of samples from different tissues and under different conditions are also available. The database provides not only comprehensive information of APA events in Plant, but is also easy to use. On its homepage, result of certain target can be presented with one click. Besides the trust-worthy APA sites, the workflow provided is also useful. Since the APA sites vary between samples, it should be better if we can gain our own set of APA sites based on the certain experiments of ourselves.

Here, we first referred to the polyadenylation clusterss(PACs) contained in the database to check the feature of obtained PASs. If most PASs are within the known PACs (since an *fpa* mutant is also included in the dataset, we may expect most PASs of *fpa* located in these PACs), we may take the result with litter modification. But if number of novel PASs is large, we may next consider using the PASs to form new PACs based on workflow provided in the dataset.

To see whether our data reflected poly(A) site usages without bias, we checked the distribution of these PACs on different features and calculate the number of PAS on each gene. In fact, very few alternative polyadenylation events are detected and most of the PACs are on the $3'$ UTR region of the gene. This indicates that many proximal PACs with low coverage cannot be detected in the filtered reads.

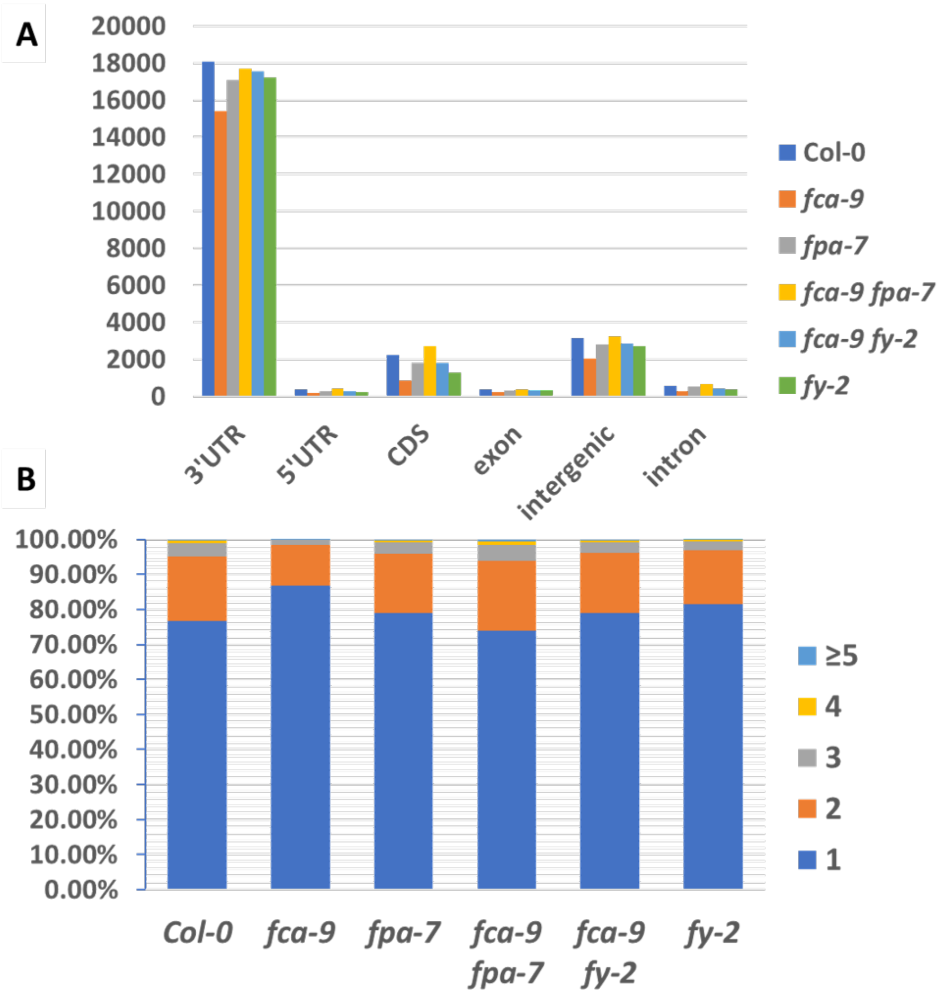


Figure 5.14: Poly(A) clusters detected by 3′mRNA-seq. Comparing to the distribution of PACs from PlantAPAdb in Figure 5.13, the number of PACs outside of 3′ UTR region and the number of AS events detected in our data was rather small.

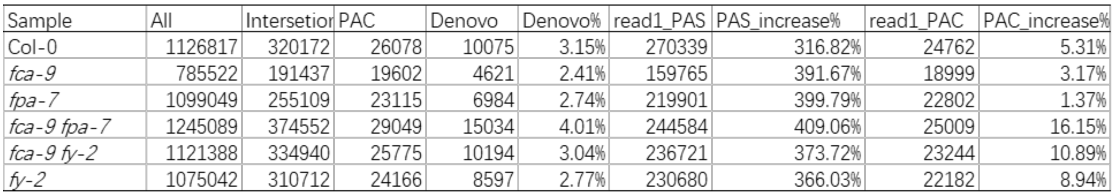
It can be seen from Figure 5.14 that the number of PACs detected was not satisfying. To increase usable reads number for quantification, we next tried to merge read1 and read2 to extend the length of the reads. The last nucleotides of read1 are considered to be the poly(A) site in usable reads and the closest to the poly(A) sites in other reads.

Here we used 'fastp -m' to merge read1 and read2, generating a new Fastq file with merged reads, and two other Fastq files for unmerged parts in read1 and read2. Read1 is usually considered of better quality than read2, especially in this case where read2 is of very low quality. The correction function of fastp was also used to ensure that bases of better quality is used in merged reads. It is not surprising to see that the sequencing quality of the merged reads was largely improved comparing to the one of read2. With reads of improved average mapping quality, we made it possible to use the information of read2 and get the poly(A) sites of numerous reads comparing to using read1 along. We also activated the polyX removal function of fastp, but it turned out that many reads with poly(A) tail escaped the step. So, we applied bbduc of BBTools (Bushnell, 2018) to those merged reads to further remove the A-rich sequences. The average poly(A) tail length removed was around 19 bps. Another step for quality control is to remove the low-quality reads on the 3′ end, only very few reads have been trimmed on the 3′, and the average number of reads removed varies from 5˜7 bps.

Since the quality removal step was taken, the reads trimmed cannot be located to the right PAS cluster in quantification of poly(A) site usage. The average bases removed in each read is around 6, which is not a big number, and only very small portion of reads were trimmed. It is fair to say that the ends of merged reads are very close to the poly(A) sites used and can be considered trust-worthy in quantification of poly(A) site usage.

In further calculation, we will only use the intersection of the three replicates of each genotype, which could filter out the reads of which the 3′ are not likely to be true PAS. Also, we may extend the PAC for 8 bps at 5′ flanking to gain larger region (and to include the reads with 3′ end trimming more or less). The total reads number mapped to the extended PAC can be considered as number of transcripts terminated within the cluster. For genes with multiple PACs, we can also calculate the usage of each PAC (check Table 5.7).

Table 5.7: Summary for PACs in Col-0/*fca-9*/*fpa-7*/*fy-2*/ *fca-9 fpa-7*/*fca-9 fy-2*



Sample	All	Intersetior	PAC	Denovo	Denovo%	read1_PAS	PAS_increase%	read1_PAC	PAC_increase%
Col-0	1126817	320172	26078	10075	3.15%	270339	316.82%	24762	5.31%
fca-9	785522	191437	19602	4621	2.41%	159765	391.67%	18999	3.17%
fpa-7	1099049	255109	23115	6984	2.74%	219901	399.79%	22802	1.37%
fca-9 fpa-7	1245089	374552	29049	15034	4.01%	244584	409.06%	25009	16.15%
fca-9 fy-2	1121388	334940	25775	10194	3.04%	236721	373.72%	23244	10.89%
fy-2	1075042	310712	24166	8597	2.77%	230680	366.03%	22182	8.94%

Surprisingly, even the number of usable reads largely increased in merged read data (over 300% PAS are detected), the increase in PAC number is not as big as the one in read number. This is later approved to be connected with bad data quality. Number of the denovo peaks are not large. But the annotation in the database is not satisfying in some sites. To improve the quality of annotation, here we try to re-cluster the poly(A) sites and annotation these newly formed sites.

Without PACs from the dataset, internal priming should be taken seriously. Unlike in the method with read1 only, where internal priming events are removed, here we may include some non-PA events with A-rich sequence on the $3'$ side. To avoid this, we will check the sequence on the $3'$ of the cluster. The definition of the internal priming events can be hard. In most cases, researcher use poly(T) sequence with 10 Ts or more, for example 18Ts are used in our experiments. But it does not necessarily mean that 18 continuous As is in need to start the priming. To get the minimum length of internal poly(A) sequence to make the binding of target RNA and the oligo(dT) primer spontaneous, we tried to calculate the lowest hybrid free energy of 18T and sequences with NAs(N=2˜8) within. The calculation is performed with online tool bifold (`http://rna.urmc.rochester.edu/RNAstructureWeb/Servers/bifold/Example.php`). As a result, at least 7As are necessary to make the free energy below 0. Another interesting thing is that, if the 7As are followed by a non-A base and then some other continuous As, it won't be helpful to lower the lowest free energy. So here we conclude that if a 'PAS' is followed by a sequence with seven or more continuous As, it is possible that it might be a result of internal priming rather than a real cleavage site. Here we get the 10 bases after the end of the read, and check the number of sub-sequences of 7As. If these exist such sub-sequence, it should be considered as a possible internal priming site.

To improve the reliability of Clustering, we finally used an R package QuantifyPoly(A) (Ye et al. (2021)) developed for poly(A) site usage analysis, with functions including removal of internal primed sites, PAS clustering, poly(A) cluster annotation, quantification and visualization of poly(A) sites usage etc.

a. Input preparing

The tool can be applied to poly(A) site information files originated from different types of sequencing data, including but not limited to PAT-seq, PAS-seq, DRS and 3′mRNA-seq etc. Pre-processing should be applied to get the poly(A) sites (of single-base resolution) in bed format before running the workflow. As is shown in the test dataset provided, the input required is not in standard bed formation, with the strand information in the second column and the position information in the third column. We can easily get the required format from a standard BED6 file.

b. Internal priming artifacts removing

The application of priming with oligo(dT) instead of purification with oligo(dT) beads like in mRNA-seq increased the possibility of internal poly(A) priming. As a result, it would form truncated cDNAs (Nam et al. (2002)), with the 5′ truncated cDNA end near the internal poly(A) sequence and the 3′ end near the real poly(A) tail (Figure 5.15 A). The end of 5′ truncated cDNA would be also recognized as poly(A) site if no extra processing is applied. In some case, the number of internal primed cDNAs can be very large (see Figure 5.15 B).

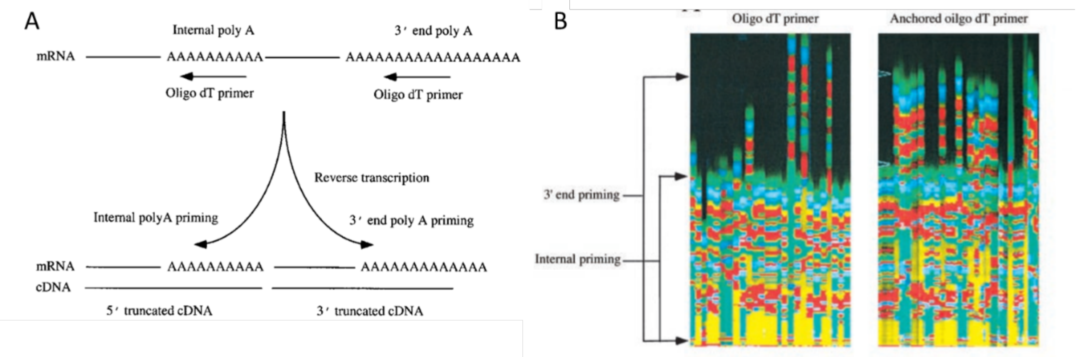


Figure 5.15: Internal poly(A) priming (Nam et al. (2002)): A. internal priming would produce truncated cDNAs; B. the number of internal priming can be large in some case

In experiment, this can be partially resecured by using Anchored oligo(dT). But the composite of Anchored oligo(dT) is more complicated and it was not included in the kit of $3'$mRNA-Seq we used in the experiment. Here, we can only remove the possible internal priming product after alignment. To recognize the artificial poly(A) site originated from internal priming, subsequences around the raw sites (-10~10 nt) are extracted. The sites with 6 continuous As or no less than 8 As in a 10-nt window (Beaudoing et al. (2000)) are considered as possible artifacts and removed.

c. Poly(A) site clustering

The most significant part of this tool is the poly(A) site clustering method. The common way is clustering by distances between poly(A) sites to merge poly(A) sites that are close in position. The traditional method usually results in clusters of a large variation in size. It has been a consensus view that polyadenylation happens on a window of size between a few to tens, which means poly(A) clusters with large size are not likely to be real cluster but a combination of several small clusters. Here in QuantifyPoly(A), a two-step clustering method is processed. In the first step, the poly(A) sites are clustered by distance. Followed by the second step, where a weighted peak density clustering method is applied to split the large clusters into sub-clusters.

d. Feature annotation

After we get the Poly(A) clusters (PACs), we may annotate these PACs for APA dynamic detection. Here, the developers defined an extended $3'$UTR region (‘ext_3UTR’) as the downstream region of $3'$ UTR within the length of twice $3'$ UTR length. The original $3'$ UTR region as well as the ext_3UTR are classified into canonical PAS region, so the PAS located in other genomic features ($5'$UTR, CDS, Intron, exon) are considered as non-canonical ones.

e. APA dynamic detection

Many genes have multiple PASs. It is not rare to see the usage of PASs switched among different genotypes/tissues/conditions. Comparison between samples can be applied in this step. The package provided four different comparing modes:

1. Quantify.CanonicalAPA: Only compare the canonical PACs between samples
2. Quantify.CNCAPA: compare the percentage of non-canonical and canonical PACs used
3. Quantify.GeneAPA: compare usage of all the PACs in the gene
4. Quantify.SplitAPA: compare the usage of split sub-PACs

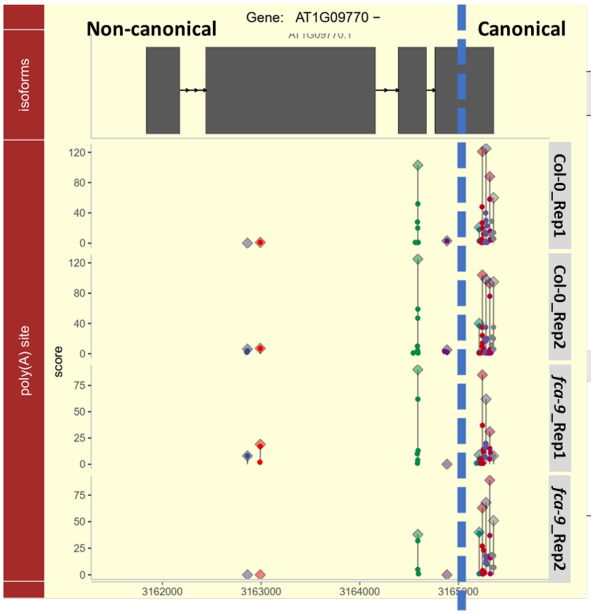
With each function, a Matrix of 4 columns will return, with:

1. Gene ID: the ID of the gene
2. Percentage Difference: A percentage difference metric measures the dynamics induced by biological condition changes.
3. Pearson correlation coefficient r: an averaged Pearson correlation coefficient r between every two replicates
4. Chi-square test p: a χ^2 test with a null hypothesis that the usage of PAS in mutant has no statistically significant difference with the usage of the wild type. Attached with Benjamin–Hochberg method to correct P-values in multiple testing. (P<0.05 for the significance cutoff).

$$\mathrm{PD} = \frac{\sum_{i,j,k} |p_{i,k} - q_{j,k}|}{2 \times m \times n}$$

where $p_{i,k}$ and $q_{j,k}$ represent the usage percentages of the kth PAC of a specific gene, and i and j denote the replicates of two samples p and q, respectively. The m and n denote the total number of replicates in each sample.

Here, we choose the ‘Quantify.CNCAPA’ mode to get the change in usage of canonical (near 3′) and non-canonical (proximal, can be within introns) PACs. An example of calculation is shown in Figure 5.16.



	type	Col-0 Rep1	Col-0 Rep2	*fca-9* Rep1	*fca-9* Rep2
PA298	CDS	0	6	8	0
PA299	CDS	1	7	19	0
PA300	CDS	103	125	90	38
PA301	CDS	3	5	0	0
PA16492	3UTR	21	40	9	40
PA16493	3UTR	121	104	85	63
PA16494	3UTR	125	98	62	68
PA16495	3UTR	88	93	31	89
PA16496	3UTR	60	95	8	51
SUM		522	573	312	349
Canonical VS Non-canonical					
Canonical		415	430	195	311
Non-canonical		107	143	117	38
%Canonical		79.50%	75.04%	62.50%	89.11%
%Non-canonical		20.50%	24.96%	37.50%	10.89%

Figure 5.16: An example of PD calculation and APA visualization: The figure was composed of lollipop chart, which indicate the position of the PAC center and total read count in the cluster, and scatter plot, showing position and read count of each single PAS. The calcultion of PD was shown in the table on the left. The PACs was classfied into two groups according tho the position, the canonial PACs and non-canonical PACs. Percentage Difference was calcualted for quantification of the difference in PAC usage between FCA

f. APA visualization

With the visualization module, we can plot the figure as shown in Figure 5.16 An example of PD calculation and APA visualization above. The figure mainly contains two parts (or two sub-figures): first the genomic structure, showing the position of exons of all the isoforms; second, a combination of lollipop chart (indicating the position of the PAC center and total read count in the cluster) and scatter plot (position and read count of each single PAS), with x-axis represent the position on the genome and y-axis coverage of the site. It helps us visualize the difference in usage of PACs of a certain gene between samples.

In summary, we can merge poly(A) sites in to poly(A) clusters and compare the usage of these poly(A) clusters between different samples with QuantifyPoly(A). It is a very convenient tool for identification of the PACs and comparison of PAC usage between samples. FCA-binding to intron may be connected to the proximal poly(A) site within intron regions. Loss of FCA has been proved to arouse change of poly(A) site usage in some sites, but whether the effect can be detected at genome-wide level or whether the result is caused by the loss of FCA binding directly is not yet known. Here, we checked the distribution of Percentage Difference (PD) between wild type Col-0 and *fca-9* of genes with both canonical and non-canonical PACs in both FCA-binding genes and non-binding genes. It turns out that FCA might affect the usage of poly(A) site both directly and indirectly (Figure 5.10 B).

5.3.3 Feature of Length and Sequence of FCA-binding 3′ UTRs

Another interesting discovery about FCA-binding 3′ UTRs is that they are of larger size, just like introns. But the increase in UG-content is not detectable in 3′ UTR region.

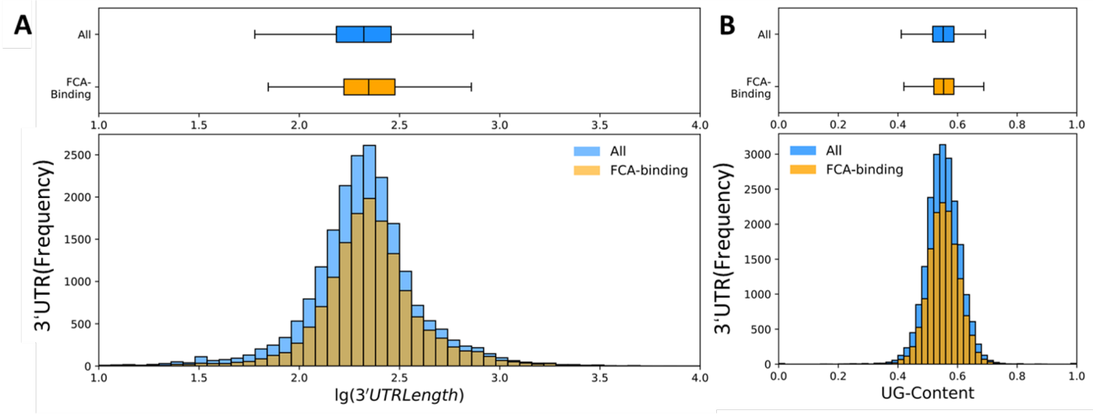


Figure 5.17: Feature of 3′ UTR in A. FCA-binding UTRs are longer in length; B. difference in UG-content was not detected.

5.4 Summary for FCA Function

In this research, multiple kinds of next-generation sequencing data were applied in FCA-centric way. We defined the binding site of FCA with eCLIP and found the protein specifically binds to intronic region and 3′ end of the genes. As for the change in the expression level, mRNA-seq data of *Col-0*, *fca-9*, *fpa-7*, *fy-2,* and the double mutants of *fca-9 fpa-7*, *fca-9 fy-2* were included. Meanwhile, 3′mRNA-seq of the mutants above were sequenced, focusing on the 3′ polyadenylation of these mutants. Last but not least, as FCA seems to work co-transcriptionally, it might be worthy checking the change in nascent RNA level, and only chromatin-bond RNA of *Col-0* and *fca-9* was sequenced currently.

To summarize, we found function of FCA a little different to research before. Traditionally, FCA was considered as a 3′ processing factor which would affect the choice of proximal/distal poly(A) site usage. But in this transcriptome-wide research, we found FCA itself has weak impact on the termination event. As an RNA-binding protein, it showed a strong preference for intronic regions and 3′ UTRs, but seems to have weak impact to the result of splicing and polyadenylation. With FCA binding to introns co-transcriptionally, it might help improve splicing efficiency. Since the splicing and termination event are related, FCA might also promote termination at the same time. Meanwhile, the binding of FCA to intron and the 3′ end are not independent. Most genes with intronic binding are also bond by FCA near the 3′ end, indicating the connection between the binding in these two regions. This might indicate the correlation between co-transcriptional splicing and termination events. The exact function of FCA near 3′ end is not very clear yet, but it seems extent with the transcription and in the *fpa-7* mutant background, even to the further downstream regions following the read-through transcripts. It is possible that FCA might be recruited by Pol II, since the region of FCA's binding is in conformity with the read-through transcripts in position. Currently, no report on FCA interaction with Pol II has been available yet, while it is possible that FCA cooperate with Pol II to promote splicing thus promote 3′ cleavage and termination in the end.

From other aspect, the regions of FCA binding seem ‘abnormal’ in transcription process. Some unspliced per-mRNA would be rapidly degraded, and the ‘read-through transcripts’ would also be cleaved and the part after PAS would be sent to degradation. The binding of FCA to these regions termed to be degraded might indicate that it is part of the quality control system of the plant and might be connected with decay of abnormal transcripts.

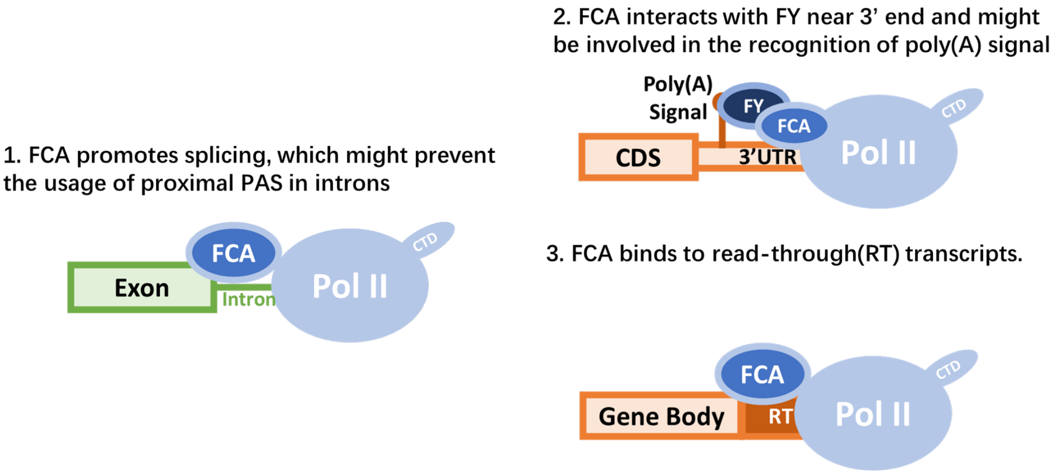

2. FCA interacts with FY near 3’ end and might be involved in the recognition of poly(A) signal
Poly(A) Signal
FY
FCA
CDS
3’UTR
Pol II
CTD
1. FCA promotes splicing, which might prevent the usage of proximal PAS in introns
FCA
Exon
Intron
Pol II
CTD
3. FCA binds to read-through(RT) transcripts.
FCA
Gene Body
RT
Pol II
CTD

Figure 5.18: FCA function model: In intronic regions, FCA’s binding promotes splicing and prevent premature termination while near the $3'$ end, FCA might be recruited by FY after the recognition of Poly(A) signal during termination.

According to results before, I proposed a model for FCA function as an anti-terminator (Figure 5.18). In introns, FCA’s binding stops the transcripts from entering pre-mature termination, which would result in increase in splicing efficiency. On the other hand, FCA binding near $3'$ end might be related to the recognition of poly(A) signal. First, binding motif of FCA (‘UGUAUG’) is similar to the potential poly(A) signal of plant (‘UGUA’). Second, FCA interacts with FY, the homology of which was known to be involved in recognition of poly(A) signal in human. Also, the loss of interaction between FCA and FY causes change in binding pattern of FCA. The decrease in FCA’s binding to $3'$ end of the transcripts in *fy-2* indicates that the recruitment of FCA to the $3'$ end rely on its interaction with FY. Meanwhile, binding of FCA near $3'$ end extends with transcription and the binding to downstream of the PAS increased in *fpa-7* with increase read-through level, indicating that it might be also correlated with Pol II activity. The recognition of PAS by FY may result in conformational change or switch in phosphorylation states of Pol II CTD, which could be a trigger of FCA's binding. In this model, FCA is involved in poly(A) signal-dependent termination (PAST) mechanism and acts to prevent the Pol II from getting into termination release of FCA. After the recognition of poly(A) signal, Pol II enters the terminational stage, thus some anti-termination factors would be recruited to prevent pre-mature termination. Thus, the binding of FCA would extend with Pol II transcript. To summarize, FCA might work with Pol II as an anti-terminator in preventing pre-mature termination.

6 Discussion and Future Work

As summarize in last chapter, we discovered the function of FCA in promoting splicing events, affecting alternative splicing and binding to abnormal read-through transcripts through application of multiple HTS data. But there also exist unsolved issues due to the limitation of current techniques as well as some problems that deserve further research in the future.

In this chapter, I would discuss the possible improvements in data analysis, including quantification of binding strength and revealing of secondary structure of RNAs with CLIP-seq data. I am also interested in comparison between mRNA-seq and $3'$mRNA-seq results which targeted at the same kind of RNA, the mature mRNAs with poly(A) tails. In mRNA-seq, oligo(dT) beads was used in the enrichment of mRNAs while oligo(dT) priming was used in $3'$mRNA-seq. Differences between the two methods might affect the result of the expression level detected in the experiments. By comparison between the two method, we might be able to decide which is the better way for sequencing mRNAs. In the second section, I would talk about the long-read sequencing data and how it can be important for FCA research. Last but not least, some unsolved issues would also be discussed in this chapter including the fate of FCA-binding transcripts and other untested features of FCA's binding.

6.1 Improvement in Data Analysis

6.1.1 Quantification of Binding Strength of RBPs in eCLIP-seq

In the original CLIP-seq method, no input was designed. In the lately developed method eCLIP, a Size-Matched Input (SMInput) is included to measure the enrichment of the target protein among proteins similar in size.

We failed in applying the SMInput to calculate the enrichment of binding in research of FCA. Actually, the SMInput of FCA-CLIP was sequenced, but we found the coverage of read unsatisfying. At first, we assigned this to the poor quality of experiment. In quality control, we found that the mapping quality of our SMInput sample is reasonable, so as the PCR duplicate rate. But not all the FCA-binding region is covered by reads in the SMInput. With the project moving on, we started to realize that this might be caused by the feature of FCA itself. The current hypothesis is that FCA binds to 'aberrant' RNAs, which is of low expression level comparing to the majority of normally transcribed RNAs. Without immunoprecipitation, FCA-binding aberrant RNAs would not be enriched, so most detected reads in SMInput should be fragmented from normal RNAs with RBPs. From this aspect, it is not surprising that even in regions of high FCA binding coverage, we might fail to get reads in input samples correspondingly. For FCA and other proteins alike, the SMInput might not be the best choice in calculation of enrichment.

Besides the enrichment of target RBP comparing to other proteins, we are also interested in the 'binding strength' of the RBP on a certain gene comparing to the binding on the others. This might be helpful in defining a gene group which are the most likely to be bound and affected by the target RBP. A gene of high CLIP read coverage, but low in expression level are more likely to be affected by the target RBP than a gene with much higher expression level of expression but similar in coverage of CLIP reads. A rough method is to normalize the CLIP read number to the number of mRNA-seq reads pairs which can roughly represent the expression level. The calculation can remove the difference in expression level between genes and length and make comparison between genes possible. The current issue unsolved is that for RBPs like FCA which tend to bind pre-mRNA, can we use the mRNA to represent the expression level of RNAs to bind? If we use the CB-RNA data, where many of the genes are not fully transcribed how to calculate the level of expression? It would be interesting to do some further study on the calculation later.

We believe that the CLIP experiment should be designed based on the feature of protein. Whether to use the SMInput or not and the method of binding strength calculation should be adjusted by the nature of target protein. And we may make more trials to improve the current pipeline for both experiment and analysis.

6.1.2 Chimeric Reads in CLIP as Indications of RNA Structure

In CLIP data, some 'Chimeric reads' was detected. The distance between two sides of the reads can be over 2000 bp and the long distance was not resulted from any known intron. This might indicate that the sequences are closely in position in the folded RNA. In fact, such chimeric reads might provide some hints of the structure of the RNA they belong. In fact, eCLIP data has been used in building structure-function model for the XIST RNA-protein complex in mammals (Lu et al. (2020)). Similarly, eCLIP of FCA might be applied to study the structure of long-noncoding RNA *COOLAIR*, antisense of *FLC* in the future.

6.1.3 Comparison Between $3'$mRNA-seq and mRNA-seq

Both $3'$mRNA-seq and mRNA-seq can be used to quantify gene expression level, but the meanings of the data are different. In $3'$mRNA-seq, one read represents a single mRNA molecular; while in mRNA-seq, the transcripts are fragmented before sequencing, so the number of the reads are not only connected with transcript number but also length of the transcript.

Since both the methods are applied to polyadenylated RNAs, we can compare the sequencing data of the same sample in these two methods to see whether the differential expression results are similar. Our expectation was that most of the change in expression should be consistent while there may exist some special cases. For example, if a gene has altered poly(A) usage, from $3'$ UTR to a site close to transcription start site, but change is detected in transcript number, a huge decrease should be detected in mRNA level but no significant change in $3'$mRNA read counts.

But when we compared our $3'$mRNA-seq with mRNA-seq dataset, things seem quite different.

1. The Correlation coefficient of the two are not very high
2. The overlapping differentially expressed gene number is not as large as expected
3. If we check the differentially expressed genes of mRNA-seq in $3'$mRNA-seq dataset, we can see genes show different direction in changing (for example, up-regulated gene in expression with down-regulated expression detected in $3'$mRNA-seq)

These results all lead to a simple conclusion: the result of mRNA-seq and $3'$mRNA-seq is not consistent. This can be aroused by the experiments themselves or the analysis.
The most significant difference in construction of library is that mRNA-seq use oligo(dT) beads to enrich and purify the transcripts while the $3'$mRNA use oligo(dT) priming. As discussed before, oligo(dT) priming would induce internal priming in the region with multiple As. We can remove the internal primed reads later in the analysis, but the bias introduced by this feature is hard to avoid. It is hard to tell which one of the experiments can present the polyadenylation of the sample more precisely, which might be worth looking into by experimental method in the future. This might help us define the better way for DE analysis and for deciding the poly(A) site for the later research.

6.2 Application of Long-read Sequencing in Future Research

In the last chapter, we've mentioned some disadvantages of short-read sequencing, and it might be helpful if long-read sequencing can be applied in our research.

6.2.1 Introduction to Long-read Sequencing

Currently, there exist two major long-read sequencing platforms, Oxford Nanopore and PacBio Single Molecule Real-Time(SMRT) Sequencing.

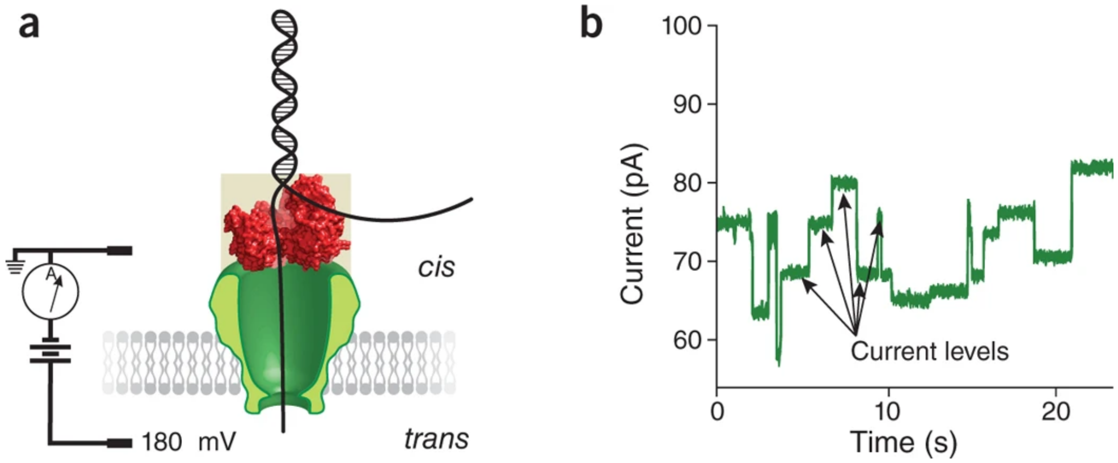


Figure 6.1: The basics of Nanopore sequencing (Deamer et al. (2016)). a. The single stranded (black) polynucleotide is driven through a Nanopore (green) in electrophoresis; b. Different current level is detected when different nucleotides passing the chamber.

The principle of Nanopore sequencing is shown in (Figure 6.1 A). The single stranded (black) polynucleotide is driven through a Nanopore (green) in electrophoresis. With the voltage also comes an ionic current, which is expected to be affected in a base-specific way. Different current level is detected when different nucleotides passing the chamber (Figure 6.1 B). The speed of movement through Nanopore is controlled by an enzyme (red).

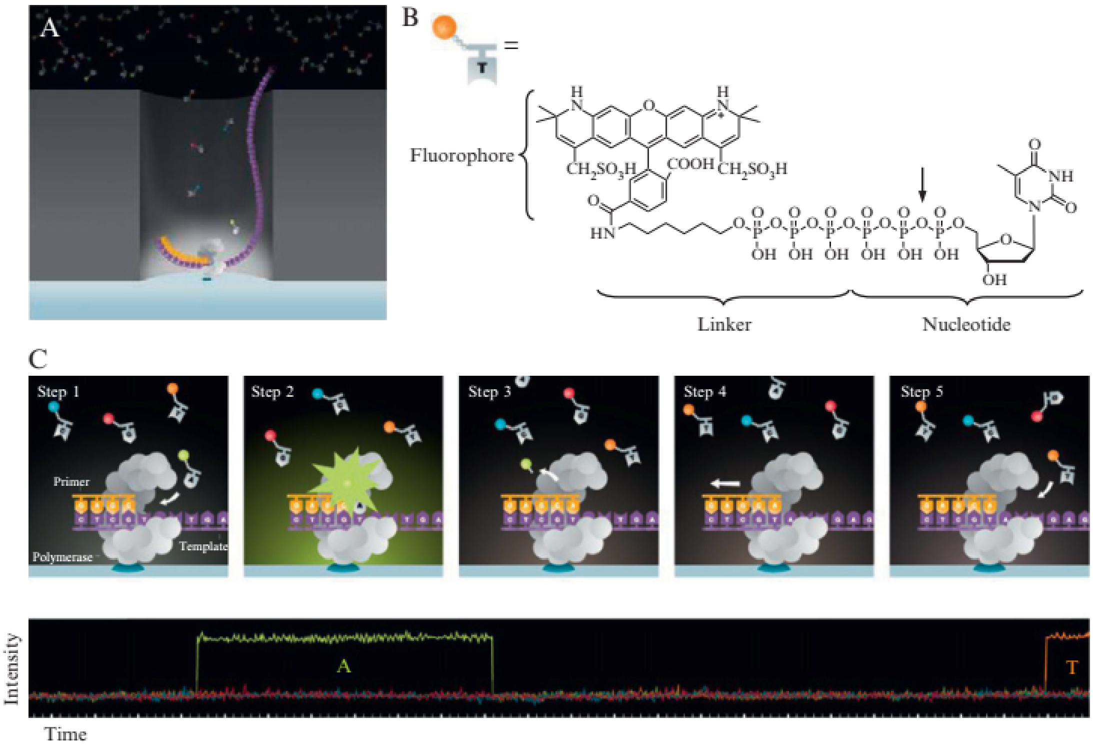


Figure 6.2: Principle of single-molecule, real-time (SMRT) DNA sequencing (Korlach et al. (2010)). A; B. Molecular structure of phospholinked nucleotides.

The zero-mode waveguide (ZMW) (Figure 6.2 A) and fluorescence-labeled, phospholinked nucleotides (Figure 6.2 B) are the two principal components to facilitate SMRT sequencing. Similar to Illumina short-read sequencing, PacBio SMRT sequencing is also based on sequencing-by-synthesis principle. On the bottom of each ZWM, DNA polymerase with bound DNA template is immobilized, so that ZWM can work as a nanophotonic visualization chamber to record the polymerization of the complementary DNA strand in real time by detecting enzymatic processing of fluorescent phospholinked nucleotide.

The two techniques have distinct principles. In Nanopore sequencing, a nucleotide was distinguished by the current levels generated when passing through the chamber of the Nanopore, while fluorescence signals are the key in SMRT sequencing. Despite the difference in principle, both of the method can help us get very long reads.

6.2.2 Possible Application in Our Future Research

One of the issues we faced in this project is that we confirmed the raised expression in the intergenic region of *fca-9*, *fpa-7* and *fca-9 fpa-7*, there is no direct evidence to say that these reads were continuous downstream the annotated PAS. But things would be different with the full-length mRNA sequenced in a single read. Long-read direct RNA sequencing (DRS) with Nanopores detected read-through at some poly(A) sites in the absence of m^{6}A (Parker et al. (2020)). In another related research, FLEP-seq was applied to characterize the *fpa* (Mo (2021)). Prolonged 3$'$ end distribution compared to the wild type was detected in *fpa*. Currently, no published data of *fca* by long-read sequencing is available yet.

The next advantage of long-read sequencing is that we can use the data to re-calculate the SS ratio. Since a single read represent the whole and even be able to distinguish the alternative splicing events (Reimer et al. (2021)).

But in some experiments where the reads are fragmented on purpose, like CLIP, there is no need to use the long-read sequencing. But in the future, it might be possible to specifically separate those transcripts with target protein and sequence them with the long-read sequencing technique to figure out which isoform does the protein bind.

6.3 Future Work on Binding of FCA

6.3.1 The Fate of FCA-binding Transcripts

Though we are quite confident that FCA binds to pre-mRNA rather than mRNA, it is not yet very clear what happened to these FCA-binding transcripts later, whether they are processed to form mRNA or decay before polyadenylation. If we can prove the that these FCA-binding transcripts would be cleared by RNA surveillance mechanism, it would be a strong evidence for us to say that FCA might be part of the mechanism.

Currently, we have no experimental method to look into the fate of these transcripts. To further understand the target transcripts of FCA, we designed the following model to describe the binding event in a simplified way:

The transcripts generated are split into two types: aberrant transcripts which would be degraded before polyadenylation and the final polyadenylated mRNAs. A single RBP may bind to both kind of the transcripts during processing, or just to one kind.

On the other side, the number of protein bond to RNAs should be correlated with the number of binding motifs. In the case of FCA, the number of possible binding feature is related to the length of binding region. We can then check the correlation between CLIP read number, which was normalized to read per kilo base (RPK), and the mRNA expression level in RPKM.

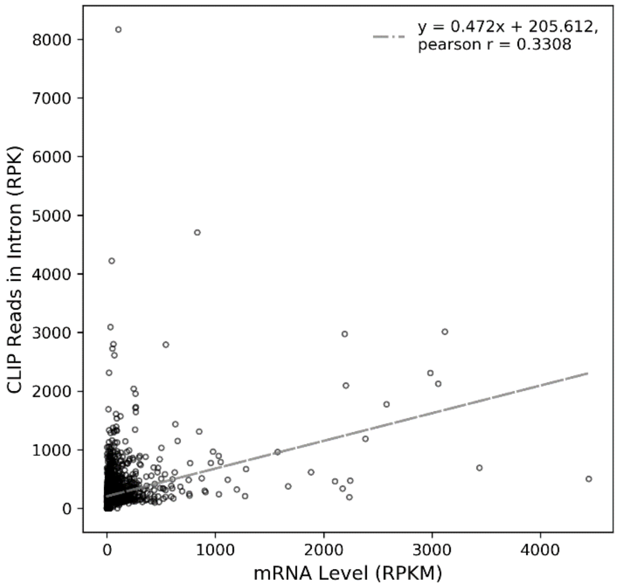


Figure 6.3: Correlation between binding of FCA and expression level of the transcripts. The binding of FCA and mRNA expression in Col-0 have weak correlation.

It seems that the coverage of CLIP reads is not always consistent with the expression level detected in mRNA-seq (see Figure 6.3). In the other word, it is possible that FCA-binding transcripts are not always mRNAs or about to be processed to mRNAs depending on the function of the RBP. If a protein is responsible for decay, the transcript it binds would be later transported out of the nuclear, there should not exist a strong correspond between the mRNA-level and the binding level.

Currently, it is a very rough model building on uncertain deals. For example, the mRNA sequenced were extracted from the total RNA, but FCA only works in the nucleus, it might be more suitable to use the expression level of nuclear mRNAs. The other thing is that the stoichiometric of RBP and RNA. For RBP which binding specific sequence, the number of RBP binding might be highly correlated with the length of transcript, but for protein recognize certain structure, the case might be different. Despite all these difficulties, the fate of FCA-binding RNA is still significant and interesting to look into in the later research.

6.3.2 Other Possible Binding Feature of FCA

6.3.2.1 Other Interacting Protein with FCA

Besides FY, there exist many other proteins which might be closely associated with FCA. In the future research, they might the target in defining the function of FCA. For example, the interaction between HLP1 and FCA protein was detected by Mass Spectrometry. HLP1 directly binds to intron 3 and $3'$ UTR of FCA, around the poly(A) sites of the two abundant isoforms of transcription FCA-β and FCA-γ. In *hlp1*, the usage of proximal PAS increased compared with the one in the wild type and resulted an increase in *FLC* level with delayed flowering. The similar binding region on *FCA* and the similar impact in *FLC* expression (Zhang et al., 2015b) makes HLP1 a possible target research in the future.

6.3.2.2 FCA-binding to Genes Encoding MADS-box Proteins

Since *FLC* encodes a MADS-box protein, it might be interesting to look into whether FCA prefer binding to genes encode MADS-box containing proteins. A group of MADS-box genes are named after AGAMOUS(AG) as AG-like (AGL) genes. *FLC* is also a member of this family. Many of them contain large introns, which seems to be favored by FCA. However, the majority of genes in AGL-family are not expressed in 10-day-old seedlings, but in other tissues of different developmental stage like seeds. FCA might also play different roles in other developmental stage by binding to the expressed AGL protein coding genes. The binding of FCA to these AGL family members might be the start point of study FCA function in other organs.

6.3.2.3 FCA-binding to 5′ UTR Introns

It has been reported that introns with certain sequence might increase the expression of the gene, known as intron-mediated enhancement (IME, Parra et al. (2011)). The IME signals are especially enriched in introns of 5′ UTR. We noticed that part of FCA-binding introns located in the 5′ UTR region. Meanwhile, loss of FCA-binding may affect the expression level. It might be interesting to check whether FCA-binding introns contain the element. And this might be a possible explanation for the FCA's effect on the expression.

6.3.2.4 Identification of FCA-binding in Proximal Poly(A) Sites inside Introns

FCA auto-regulates its own expression by usage of proximal PAS. Interestingly, the intron 3 with proximal PAS was also a target of FCA-binding. It is highly suspicious that Introns with proximal poly(A) sites might be bound by FCA. But in the research before, we removed introns with possible PAS inside (unbalanced 5′SS and 3′SS) in intron analysis, paying no extra attention to the these intronic PAS than other proximal Poly(A) sites. But it does not mean impact of FCA on these intronic PAS are not worth looking into.

There are multiple ways to find these intron PAS. The simplest way is to use the annotation file for annotated proximal PAS located in the intron. The proximal PAS of *FLC* and *FCA* can be found in the annotation file. We can than refer to mRNA-seq data to check whether both splicing and termination events happen within the same region. We can also get the intronic region with $5'$SS ratio significantly larger than the $3'$SS ratio or get the proximal PAC from $3'$mRNA-seq and whether the PAC is inside an intron.

A Material Used

Plants used includes single mutants of *fca-9*, *fpa-7*, *fy-2* and double mutants obtained through crossing of these single mutants, *fca-9 fpa-7* and *fca-9 fy-2*. All the mutants used are Col alleles so the Col-0 ecotype of *Arabidopsis thaliana* was used as the wild-type control in this study.

fca-9 is null allele of FCA isolated by C.-H.Y. whilst in Dr R. Sung's laboratory (University of California, Berkeley, USA) (Page et al., 1999). It is a point mutation line with severe late-flowering phenotype. *fpa-7* is an SALK line (SALK_138449) with T-DNA insertion in the first intron. FY function was partial loss due to the T-DNA insertion before the two PPLPP motif in *fy-2* (see Figure 2.10). All mutants above have late flowering phenotype.

fip37-4 was a SALK line with T-DNA insertion (SALK_018636). In this research we used a very special mutant of *fip37-4 LEC1:FIP37*, which failed to produce viable homozygous seeds after self-pollination. In this line, *fip37-4* was partially complemented with *LEC1:FIP37* transgene, so that the *FIP37* coding region was driven by the promoter of *LEAFY COTYLEDON 1* (*LEC1*) during embryogenesis as described by Shen et al. (2016).

All experiments used seedlings grown on 1/2 Murashige and Skoog (MS) medium for 10 days at 22 °C under 16h light/8h dark cycles.

B Abbreviation Used

AS: Alternative Splicing

ATP: Adenosine Triphosphate

CLNIP-MS: CrossLinked Nuclear ImmunoPrecipitation and Mass Spectrometry

CPA: Cleavage and PolyAdenylation

FCA: FLOWERING LOCUS CA

FLC: FLOWERING LOCUS C

FPA: FLOWERING LOCUS PA

FY: FLOWERING LOCUS Y

IP-MS: ImmunoPrecipitation-Mass Spectrometry

mRNA: messenger RNA

ncRNA: non-coding RNA

PAC: PolyAdenylation Cluster

PADT: PAS-Dependent Termination

PAS: PolyAdenylation Site

PCR: Polymerase Chain Reaction

PIC: Pre-Initiation Complex

Pol II: RNA Polymerase II

Poly(A): PolyAdenylation

RBD: RNA-Binding Domain

RBP: RNA-Binding Protein

RRM: RNA Recognition Motif

snRNP: Small Nuclear Ribonucleoprotein

SS: Splicing Site

TPM: Transcripts Per Kilobase of exon model per Million mapped reads

UTR: UnTranslated Region

C Bibliography

Amorim, M. d. F., Willing, E.-M., Szabo, E. X., Droste-Borel, I., Maček, B., Schneeberger, K., and Laubinger, S. (2018). The u1 snrnp subunit luc7 modulates plant development and stress responses via regulation of alternative splicing. *The Plant Cell*, 30(11):2838–2854.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.

Andrews, S. et al. (2010). Fastqc: a quality control tool for high throughput sequence data.

Bach-Pages, M., Castello, A., and Preston, G. M. (2017). Plant RNA Interactome Capture: Revealing the Plant RBPome. *Trends Plant Sci*, 22(6):449–451.

Bachem, C. W., Oomen, R. J., and Visser, R. G. (1998). Transcript Imaging with cDNA-AFLP: A Step-by-Step Protocol. *Plant Molecular Biology Reporter*, 16(2):157–157.

Bahrami-Samani, E., Penalva, L. O., Smith, A. D., and Uren, P. J. (2015). Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Research*, 43(1):95–103.

Bäurle, I. and Dean, C. (2008). Differential Interactions of the Autonomous Pathway RRM Proteins and Chromatin Regulators in the Silencing of Arabidopsis Targets. *PLOS ONE*, 3(7):e2733.

Bäurle, I., Smith, L., Baulcombe, D. C., and Dean, C. (2007). Widespread Role for the Flowering-Time Regulators FCA and FPA in RNA-Mediated Chromatin Silencing. *Science*, 318(5847):109–112.

Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J.-M., and Gautheret, D. (2000). Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Res.*, 10(7):1001–1010.

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(1):53–64.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Boothby, T. C., Zipper, R. S., van der Weele, C. M., and Wolniak, S. M. (2013). Removal of Retained Introns Regulates Translation in the Rapidly Developing Gametophyte of Marsilea vestita. *Developmental Cell*, 24(5):517–529.

Bregman, A., Avraham-Kelbert, M., Barkai, O., Duek, L., Guterman, A., and Choder, M. (2011). Promoter Elements Regulate Cytoplasmic mRNA Decay. *Cell*, 147(7):1473–1483.

Bushnell, B. (2018). BBTools: A suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data.

Carroll, S. M., Narayan, P., and Rottman, F. M. (1990). N6-methyladenosine residues in an intron-specific region of prolactin pre-mRNA. *Molecular and Cellular Biology*, 10(9):4456–4465.

Chakrabortee, S., Kayatekin, C., Newby, G. A., Mendillo, M. L., Lancaster, A., and Lindquist, S. (2016). Luminidependens (LD) is an Arabidopsis protein with prion behavior. *PNAS*, 113(21):6065–6070.

Chapman, R. D., Heidemann, M., Hintermair, C., and Eick, D. (2008). Molecular evolution of the RNA polymerase II CTD. *Trends in Genetics*, 24(6):289–296.

Chen, B., Yun, J., Kim, M. S., Mendell, J. T., and Xie, Y. (2014). PIPE-CLIP: A comprehensive online tool for CLIP-seq data analysis. *Genome Biology*, 15(1):R18.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890.

Chikhi, R. and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1):31–37.

Cho, E.-J., Takagi, T., Moore, C. R., and Buratowski, S. (1997). mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. *Genes Dev.*, 11(24):3319–3326.

Clerici, M., Faini, M., Aebersold, R., and Jinek, M. (2017). Structural insights into the assembly and polyA signal recognition mechanism of the human CPSF complex. *eLife*, 6:e33111.

Colombrita, C., Silani, V., and Ratti, A. (2013). ELAV proteins along evolution: Back to the nucleus? *Molecular and Cellular Neuroscience*, 56:447–455.

Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D., Zhou, D., Luo, S., Vasicek, T. J., Daly, M. J., Wolfsberg, T. G., and Collins, F. S. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.*, 16(1):123–131.

Curk, T., Rot, G., Gorup, C., Ruiz de los Mozos, I., Konig, J., Zmrzlikar, J., Sugimoto, Y., Haberman, N., Bobojevic, G., Hauer, C., et al. (2019). iCount: Protein-RNA interaction iCLIP data analysis (in preparation). `https://github.com/tomazc/iCount`.

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. *Nat Biotechnol*, 34(5):518–524.

Deremetz, A., Le Roux, C., Idir, Y., Brousse, C., Agorio, A., Gy, I., Parker, J. E., and Bouché, N. (2019). Antagonistic Actions of FPA and IBM2 Regulate Transcript Processing from Genes Containing Heterochromatin. *Plant Physiol.*, 180(1):392–403.

Desrosiers, R., Friderici, K., and Rottman, F. (1974). Identification of Methylated Nucleosides in Messenger RNA from Novikoff Hepatoma Cells. *PNAS*, 71(10):3971–3975.

Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M., and Pagano, A. (2007). The expanding RNA polymerase III transcriptome. *Trends in Genetics*, 23(12):614–622.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., Sorek, R., and Rechavi, G. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, 485(7397):201–206.

Du, T.-G., Jellbauer, S., Müller, M., Schmid, M., Niessing, D., and Jansen, R.-P. (2008). Nuclear transit of the RNA-binding protein She2 is required for translational control of localized *ASH1* mRNA. *EMBO Rep*, 9(8):781–787.

Duc, C., Sherstnev, A., Cole, C., Barton, G. J., and Simpson, G. G. (2013). Transcription termination and chimeric RNA formation controlled by Arabidopsis thaliana FPA. *PLoS Genet.*, 9(10):e1003867.

Eaton, J. D. and West, S. (2020). Termination of transcription by rna polymerase ii: Boom! *Trends in Genetics*, 36(9):664–675.

Edmonds, M., Vaughan, M. H., 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. *PNAS*, 68(6):1336–1340.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048.

Fang, X., Wang, L., Ishikawa, R., Li, Y., Fiedler, M., Liu, F., Calder, G., Rowan, B., Weigel, D., Li, P., and Dean, C. (2019). Arabidopsis FLL2 promotes liquid–liquid phase separation of polyadenylation complexes. *Nature*, 569(7755):265.

Fedorov, A., Merican, A. F., and Gilbert, W. (2002). Large-scale comparison of intron positions among animal, plant, and fungal genes. *PNAS*, 99(25):16128–16133.

Fish, R. N. and Kane, C. M. (2002). Promoting elongation with transcript cleavage stimulatory factors. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1577(2):287–307.

George, H., Ule, J., and Hussain, S. (2017). *Illustrating the Epitranscriptome at Nucleotide Resolution Using Methylation-iCLIP (miCLIP)*, volume 1562. Springer.

Goodfellow, S. J. and Zomerdijk, J. C. B. M. (2013). Basic Mechanisms in RNA Polymerase I Transcription of the Ribosomal RNA Genes. In Kundu, T. K., editor, *Epigenetics: Development and Disease*, Subcellular Biochemistry, pages 211–236. Springer Netherlands, Dordrecht.

Gregersen, L. H., Mitter, R., Ugalde, A. P., Nojima, T., Proudfoot, N. J., Agami, R., Stewart, A., and Svejstrup, J. Q. (2019). SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. *Cell*, 177(7):1797–1813.e18.

Guo, Y. E., Manteiga, J. C., Henninger, J. E., Sabari, B. R., Dall'Agnese, A., Hannett, N. M., Spille, J.-H., Afeyan, L. K., Zamudio, A. V., Shrinivas, K., Abraham, B. J., Boija, A., Decker, T.-M., Rimel, J. K., Fant, C. B., Lee, T. I., Cisse, I. I., Sharp, P. A., Taatjes, D. J., and Young, R. A. (2019). Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature*, 572(7770):543–548.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141.

Hamilton, K., Sun, Y., and Tong, L. (2019). Biophysical characterizations of the recognition of the AAUAAA polyadenylation signal. *RNA*, 25(12):1673–1680.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4):576–589.

Henderson, I. R., Liu, F., Drea, S., Simpson, G. G., and Dean, C. (2005). An allelic series reveals essential roles for FY in plant development in addition to flowering-time control. *Development*, 132(16):3597–3607.

Hsin, J.-P. and Manley, J. L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.*, 26(19):2119–2137.

Hunt, A. G., Chu, N. M., Odell, J. T., Nagy, F., and Chua, N.-H. (1987). Plant cells do not properly recognize animal gene polyadenylation signals. *Plant Mol Biol*, 8(1):23–35.

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L. E., Sibley, C. R., Sugimoto, Y., Tajnik, M., Koenig, J., and Ule, J. (2014). iCLIP: Protein-RNA interactions at nucleotide resolution. *Methods*, 65(3):274–287.

Imig, J., Brunschweiger, A., Bruemmer, A., Guennewig, B., Mittal, N., Kishore, S., Tsikrika, P., Gerber, A. P., Zavolan, M., and Hall, J. (2015). miR-CLIP capture of a miRNA targetome uncovers a lincRNA H19-miR-106a interaction. *Nature Chemical Biology*, 11(2):107–U43.

John S.Jacobs Anderson, R. P. (1998). The $3'$ to $5'$ degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3prime to 5prime exonucleases of the exosome complex. *The EMBO Journal*, 17(5):1497–1506.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

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(7324):664–668.

Kasowitz, S. D., Ma, J., Anderson, S. J., Leu, N. A., Xu, Y., Gregory, B. D., Schultz, R. M., and Wang, P. J. (2018). Nuclear m6A reader YTHDC1 regulates alternative polyadenylation and splicing during mouse oocyte development. *PLOS Genetics*, 14(5):e1007412.

Keene, J. D., Komisarow, J. M., and Friedersdorf, M. B. (2006). RIP-Chip: The isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nature Protocols*, 1(1):302–307.

Kim, D., Langmead, B., and Salzberg, S. (2017). HISAT2: Graph-based alignment of next-generation sequencing reads to a population of genomes.

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 biology*, 14(4):1–13.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 17(7):909–915.

Koornneef, M., Dellaert, L. W. M., and van der Veen, J. H. (1982). EMS- and relation-induced mutation frequencies at individual loci in Arabidopsis thaliana (L.) Heynh. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 93(1):109–123.

Koornneef, M., Hanhart, C. J., and van der Veen, J. H. (1991). A genetic and physiological analysis of late flowering mutants in Arabidopsis thaliana. *Molec. Gen. Genet.*, 229(1):57–66.

Korlach, J., Bjornson, K. P., Chaudhuri, B. P., Cicero, R. L., Flusberg, B. A., Gray, J. J., Holden, D., Saxena, R., Wegener, J., and Turner, S. W. (2010). Real-Time DNA Sequencing from Single Polymerase Molecules. In *Methods in Enzymology*, volume 472, pages 431–455. Elsevier.

Krakau, S., Richard, H., and Marsico, A. (2017). PureCLIP: Capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biol*, 18(1):1–17.

Kuehner, J. N., Pearson, E. L., and Moore, C. (2011). Unravelling the means to an end: RNA polymerase II transcription termination. *Nat Rev Mol Cell Biol*, 12(5):283–294.

Kurosaki, T. (2019). Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nat Rev Mol Cell Biol*, 20:406–420.

Kwek, K. Y., Murphy, S., Furger, A., Thomas, B., O'Gorman, W., Kimura, H., Proudfoot, N. J., and Akoulitchev, A. (2002). U1 snRNA associates with TFIIH and regulates transcriptional initiation. *Nat. Struct. Biol.*, 9(11):800–805.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics*, 32(11):7.1–7.14.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359.

Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods*, 7(9):709–715.

Levy, Y. Y. and Dean, C. (1998). The Transition to Flowering. *The Plant Cell*, 10(12):1973–1989.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, 20(2):265–272.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., and Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469.

Lim, L. and Canellakis, E. S. (1970). Adenine-rich Polymer associated with Rabbit Reticulocyte Messenger RNA. *Nature*, 227(5259):710–712.

Liu, F., Marquardt, S., Lister, C., Swiezewski, S., and Dean, C. (2010). Targeted $3'$ Processing of Antisense Transcripts Triggers Arabidopsis FLC Chromatin Silencing. *Science*, 327(5961):94–97.

Liu, F., Quesada, V., Crevillén, P., Bäurle, I., Swiezewski, S., and Dean, C. (2007). The Arabidopsis RNA-binding protein FCA requires a lysine-specific demethylase 1 homolog to downregulate FLC. *Mol. Cell*, 28(3):398–407.

Long, M. and Deutsch, M. (1999). Intron—exon structures of eukaryotic model organisms. *Nucleic Acids Research*, 27(15):3219–3228.

Lovci, M. T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T. Y., Stark, T. J., Gehman, L. T., Hoon, S., Massirer, K. B., Pratt, G. A., Black, D. L., Gray, J. W., Conboy, J. G., and Yeo, G. W. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature Structural & Molecular Biology*, 20(12):1434–1442.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550.

Lu, Z., Guo, J. K., Wei, Y., Dou, D. R., Zarnegar, B., Ma, Q., Li, R., Zhao, Y., Liu, F., Choudhry, H., Khavari, P. A., and Chang, H. Y. (2020). Structural modularity of the XIST ribonucleoprotein complex. *Nat Commun*, 11(1):6163.

Luo, G.-Z., MacQueen, A., Zheng, G., Duan, H., Dore, L. C., Lu, Z., Liu, J., Chen, K., Jia, G., Bergelson, J., and He, C. (2014). Unique features of the m6A methylome in Arabidopsis thaliana. *Nat Commun*, 5(1):5630.

Luse, D. S. (2013). Promoter clearance by RNA polymerase II. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(1):63–68.

Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–1697.

Macknight, R., Bancroft, I., Page, T., Lister, C., Schmidt, R., Love, K., Westphal, L., Murphy, G., Sherson, S., Cobbett, C., and Dean, C. (1997). FCA, a Gene Controlling

Flowering Time in Arabidopsis, Encodes a Protein Containing RNA-Binding Domains. *Cell*, 89(5):737–745.

Marondedze, C., Thomas, L., Serrano, N. L., Lilley, K. S., and Gehring, C. (2016). The RNA-binding protein repertoire of Arabidopsis thaliana. *Sci Rep*, 6(1):29766.

Marquardt, S., Raitskin, O., Wu, Z., Liu, F., Sun, Q., and Dean, C. (2014). Functional consequences of splicing of the antisense transcript COOLAIR on FLC transcription. *Mol. Cell*, 54(1):156–165.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12.

McCracken, S., Fong, N., Rosonina, E., Yankulov, K., Brothers, G., Siderovski, D., Hessel, A., Foster, S., Program, A. E., Shuman, S., and Bentley, D. L. (1997). $5'$-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes Dev.*, 11(24):3306–3318.

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive Analysis of mRNA Methylation Reveals Enrichment in $3'$ UTRs and near Stop Codons. *Cell*, 149(7):1635–1646.

Mili, S. and Steitz, J. A. (2004). Evidence for reassociation of RNA-binding proteins after cell lysis: Implications for the interpretation of immunoprecipitation analyses. *RNA*, 10(11):1692–1694.

Mo, W. (2021). Landscape of transcription termination in Arabidopsis revealed by single-molecule nascent RNA sequencing — Genome Biology — Full Text. *Genome Biology*, 22(322).

Nam, D. K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., Chen, J., Rowley, J. D., and Wang, S. M. (2002). Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proceedings of the National Academy of Sciences*, 99(9):6152–6156.

Nichols, J. L. (1979). N6-methyladenosine in maize poly(A)-containing RNA. *Plant Science Letters*, 15(4):357–361.

Nojima, T., Gomes, T., Carmo-Fonseca, M., and Proudfoot, N. J. (2016). Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nat Protoc*, 11(3):413–428.

Nojima, T., Gomes, T., Grosso, A. R. F., Kimura, H., Dye, M. J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N. J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*, 161(3):526–540.

Page, T., Macknight, R., Yang, C.-H., and Dean, C. (1999). Genetic interactions of the arabidopsis flowering time gene fca, with genes regulating floral initiation. *The Plant Journal*, 17(3):231–239.

Park, J. W., Tokheim, C., Shen, S., and Xing, Y. (2013). Identifying Differential Alternative Splicing Events from RNA Sequencing Data Using RNASeq-MATS. In Shomron, N., editor, *Deep Sequencing Data Analysis*, Methods in Molecular Biology, pages 171–179. Humana Press, Totowa, NJ.

Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., Hall, A. J., Barton, G. J., and Simpson, G. G. (2020). Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *eLife*, 9:e49658.

Parra, G., Bradnam, K., Rose, A. B., and Korf, I. (2011). Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. *Nucleic Acids Research*, 39(13):5328–5337.

Patel, A., Lee, H. O., Jawerth, L., Maharana, S., Jahnel, M., Hein, M. Y., Stoynov, S., Mahamid, J., Saha, S., Franzmann, T. M., Pozniakovski, A., Poser, I., Maghelli, N., Royer, L. A., Weigert, M., Myers, E. W., Grill, S., Drechsel, D., Hyman, A. A., and Alberti, S. (2015). A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell*, 162(5):1066–1077.

Pollier, J., Rombauts, S., and Goossens, A. (2013). Analysis of RNA-Seq Data with TopHat and Cufflinks for Genome-Wide Expression Analysis of Jasmonate-Treated Plants and Plant Cultures. In Goossens, A. and Pauwels, L., editors, *Jasmonate Signaling*, volume 1011, pages 305–315. Humana Press, Totowa, NJ.

Price, D. H., Sluder, A. E., and Greenleaf, A. L. (1989). Dynamic interaction between a Drosophila transcription factor and RNA polymerase II. *MOL. CELL. BIOL.*, 9:11.

Proudfoot, N. J. (2016). Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science*, 352(6291):1291.

Proudfoot, N. J. and Brownlee, G. G. (1974). Sequence at the 3' end of globin mRNA shows homology with immunoglobulin light chain mRNA. *Nature*, 252(5482):359–362.

Quesada, V., Macknight, R., Dean, C., and Simpson, G. G. (2003). Autoregulation of FCA pre-mRNA processing controls Arabidopsis flowering time. *EMBO J.*, 22(12):3142–3152.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165.

Re, A., Joshi, T., Kulberkyte, E., Morris, Q., and Workman, C. T. (2014). RNA–Protein Interactions: An Overview. In Gorodkin, J. and Ruzzo, W. L., editors, *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, Methods in Molecular Biology, pages 491–521. Humana Press, Totowa, NJ.

Ream, T. S., Haag, J. R., Wierzbicki, A. T., Nicora, C. D., Norbeck, A. D., Zhu, J.-K., Hagen, G., Guilfoyle, T. J., Pasa-Tolić, L., and Pikaard, C. S. (2009). Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. *Mol. Cell*, 33(2):192–203.

Reichel, M., Köster, T., and Staiger, D. (2019). Marking RNA: m6A writers, readers, and functions in Arabidopsis. *Journal of Molecular Cell Biology*, 11(10):899–910.

Reichel, M., Liao, Y., Rettel, M., Ragan, C., Evers, M., Alleaume, A.-M., Horos, R., Hentze, M. W., Preiss, T., and Millar, A. A. (2016). In Planta Determination of the mRNA-Binding Proteome of Arabidopsis Etiolated Seedlings. *The Plant Cell*, 28(10):2435–2452.

Reimer, K. A., Mimoso, C. A., Adelman, K., and Neugebauer, K. M. (2021). Co-transcriptional splicing regulates $3'$ end cleavage during mammalian erythropoiesis. *Molecular Cell*, 81(5):998–1012.e7.

Ren, X., Vorst, O., Fiers, M., Stiekema, W., and Nap, J. (2006). In plants, highly expressed genes are the least compact. *Trends in Genetics*, 22(10):528–532.

Roeder, R. G. (1996). [14] Nuclear RNA polymerases: Role of general initiation factors and cofactors in eukaryotic transcription. In *Methods in Enzymology*, volume 273 of *RNA Polymerase and Associated Factors Part A*, pages 165–171. Academic Press.

Roeder, R. G. and Rutter, W. J. (1969). Multiple Forms of DNA-dependent RNA Polymerase in Eukaryotic Organisms. *Nature*, 224(5216):234–237.

Rosenberg, M., Blum, R., Kesner, B., Maier, V. K., Szanto, A., and Lee, J. T. (2017). Denaturing CLIP, dCLIP, Pipeline Identifies Discrete RNA Footprints on Chromatin-Associated Proteins and Reveals that CBX7 Targets 3 ' UTRs to Regulate mRNA Expression. *Cell Systems*, 5(4):368–385.

Sarnowski, T. J., Swiezewski, S., Pawlikowska, K., Kaczanowski, S., and Jerzmanowski, A. (2002). AtSWI3B, an Arabidopsis homolog of SWI3, a core subunit of yeast Swi/Snf chromatin remodeling complex, interacts with FCA, a regulator of flowering time. *Nucleic Acids Res.*, 30(15):3412–3421.

Saunders, A., Core, L. J., and Lis, J. T. (2006). Breaking barriers to transcription elongation. *Nat Rev Mol Cell Biol*, 7(8):557–567.

Schomburg, F. M., Patton, D. A., Meinke, D. W., and Amasino, R. M. (2001). FPA, a Gene Involved in Floral Induction in Arabidopsis, Encodes a Protein Containing RNA-Recognition Motifs. *The Plant Cell*, 13(6):1427–1436.

Shen, L., Liang, Z., Gu, X., Chen, Y., Teo, Z. W. N., Hou, X., Cai, W. M., Dedon, P. C., Liu, L., and Yu, H. (2016). N6-Methyladenosine RNA Modification Regulates Shoot Stem Cell Fate in Arabidopsis. *Developmental Cell*, 38(2):186–200.

Shen, Z., St-Denis, A., and Chartrand, P. (2010). Cotranscriptional recruitment of She2p by RNA pol II elongation factor Spt4–Spt5/DSIF promotes mRNA localization to the yeast bud. *Genes Dev.*, 24(17):1914–1926.

Simpson, G. G., Dijkwel, P. P., Quesada, V., Henderson, I., and Dean, C. (2003). FY is an RNA 3' end-processing factor that interacts with FCA to control the Arabidopsis floral transition. *Cell*, 113(6):777–787.

Sims, R. J., Belotserkovskaya, R., and Reinberg, D. (2004). Elongation by RNA polymerase II: The short and long of it. *Genes Dev.*, 18(20):2437–2468.

Sonmez, C., Bäurle, I., Magusin, A., Dreos, R., Laubinger, S., Weigel, D., and Dean, C. (2011). RNA 3' processing functions of Arabidopsis FCA and FPA limit intergenic transcription. *Proc. Natl. Acad. Sci. U.S.A.*, 108(20):8508–8513.

Sugimoto, Y., Vigilante, A., Darbo, E., Zirra, A., Militti, C., D'Ambrogio, A., Luscombe, N. M., and Ule, J. (2015). hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. *Nature*, 519(7544):491–494.

Swiezewski, S., Liu, F., Magusin, A., and Dean, C. (2009). Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature*, 462(7274):799–802.

Tabaska, J. E. and Zhang, M. Q. (1999). Detection of polyadenylation signals in human DNA sequences. *Gene*, 231(1):77–86.

Tran, H., Maurer, F., and Nagamine, Y. (2003). Stabilization of Urokinase and Urokinase Receptor mRNAs by HuR Is Linked to Its Cytoplasmic Accumulation Induced by Activated Mitogen-Activated Protein Kinase-Activated Protein Kinase 2. *Molecular and Cellular Biology*, 23(20):7177–7188.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515.

Ule, J., Jensen, K., Mele, A., and Darnell, R. B. (2005). CLIP: A method for identifying protein-RNA interaction sites in living cells. *Methods*, 37(4):376–386.

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B. (2003). CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science*, 302(5648):1212–1215.

Uren, P. J., Bahrami-Samani, E., Burns, S. C., Qiao, M., Karginov, F. V., Hodges, E., Hannon, G. J., Sanford, J. R., Penalva, L. O. F., and Smith, A. D. (2012). Site identification in high-throughput RNA–protein interaction data. *Bioinformatics*, 28(23):3013–3020.

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M., and Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial Analysis of Gene Expression. *Science*, 270:484–487.

Vera Alvarez, R., Pongor, L. S., Mariño-Ramírez, L., and Landsman, D. (2019). TPMCalculator: One-step software to quantify mRNA abundance of genomic features. *Bioinformatics*, 35(11):1960–1962.

Wade, J. T. and Struhl, K. (2008). The transition from transcriptional initiation to elongation. *Current Opinion in Genetics & Development*, 18(2):130–136.

Wallace, E. W. J., Maufrais, C., Sales-Lee, J., Tuck, L. R., de Oliveira, L., Feuerbach, F., Moyrand, F., Natarajan, P., Madhani, H. D., and Janbon, G. (2020). Quantitative global studies reveal differential translational control by start codon context across the fungal kingdom. *Nucleic Acids Research*, 48(5):2312–2331.

Wang, L., Wang, S., and Li, W. (2012). RSeQC: Quality control of RNA-seq experiments. *Bioinformatics*, 28(16):2184–2185.

Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N. M., Rot, G., Zupan, B., Curk, T., and Ule, J. (2010). iCLIP Predicts the Dual Splicing Effects of TIA-RNA Interactions. *Plos Biology*, 8(10):e1000530.

Wang, Z.-W., Wu, Z., Raitskin, O., Sun, Q., and Dean, C. (2014). Antisense-mediated FLC transcriptional repression requires the P-TEFb transcription elongation factor. *Proceedings of the National Academy of Sciences*, 111(20):7468–7473.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., and Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1):W296–W303.

Winters, M. A. and Edmonds, M. (1973). A Poly(A) Polymerase from Calf Thymus: CHARACTERIZATION OF THE REACTION PRODUCT AND THE PRIMER REQUIREMENT. *Journal of Biological Chemistry*, 248(13):4763–4768.

Wolf, S. S., Roder, K., and Schweizer, M. (1995). Determination of the molecular weight of DNA-binding proteins using UV-crosslinking and SDS-PAGE. *Mol Biotechnol*, 4(3):269–273.

Wong, K. H., Jin, Y., and Struhl, K. (2014). TFIIH Phosphorylation of the Pol II CTD Stimulates Mediator Dissociation from the Preinitiation Complex and Promoter Escape. *Molecular Cell*, 54(4):601–612.

Woychik, N. A. and Hampsey, M. (2002). The RNA Polymerase II Machinery: Structure Illuminates Function. *Cell*, 108(4):453–463.

Wu, Z., Fang, X., Zhu, D., and Dean, C. (2019). Autonomous pathway: FLOWERING LOCUS C repression through an antisense-mediated chromatin silencing mechanism. *Plant Physiol.*, 182:27–37.

Wuarin, J. and Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: Evidence for cotranscriptional splicing. *MOL. CELL. BIOL.*, 14:7.

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 $3'$-UTR Landscape Across 7 Tumor Types. *Nat Commun*, 5:5274.

Yamaguchi, Y., Shibata, H., and Handa, H. (2013). Transcription elongation factors DSIF and NELF: Promoter-proximal pausing and beyond. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(1):98–104.

Ye, C., Zhao, D., Ye, W., Wu, X., Ji, G., Li, Q. Q., and Lin, J. (2021). QuantifyPoly(A): Reshaping alternative polyadenylation landscapes of eukaryotes with weighted density peak clustering. *Briefings in Bioinformatics*, 22(6):1–14.

Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X.-D., and Gage, F. H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature Structural & Molecular Biology*, 16(2):130–137.

Yu, Z., Lin, J., and Li, Q. Q. (2019). Transcriptome Analyses of FY Mutants Reveal its Role in mRNA Alternative Polyadenylation. *The Plant Cell*, 31:2332–2352.

Zarnegar, B. J., Flynn, R. A., Shen, Y., Do, B. T., Chang, H. Y., and Khavari, P. A. (2016). irCLIP platform for efficient characterization of protein-RNA interactions. *Nature Methods*, 13(6):489–492.

Zhang, C., Gschwend, A. R., Ouyang, Y., and Long, M. (2014). Evolution of Gene Structural Complexity: An Alternative-Splicing-Based Model Accounts for Intron-Containing Retrogenes. *Plant Physiology*, 165(1):412–423.

Zhang, H., Rigo, F., and Martinson, H. G. (2015a). Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. *Molecular Cell*, 59(3):437–448.

Zhang, S., Aibara, S., Vos, S. M., Agafonov, D. E., Lührmann, R., and Cramer, P. (2021). Structure of a transcribing RNA polymerase II-U1 snRNP complex. *Science*, 371(6526):305–309.

Zhang, Y., Gu, L., Hou, Y., Wang, L., Deng, X., Hang, R., Chen, D., Zhang, X., Zhang, Y., Liu, C., and Cao, X. (2015b). Integrative genome-wide analysis reveals HLP1, a novel RNA-binding protein, regulates plant flowering by targeting alternative polyadenylation. *Cell Research*, 25(7):864–876.

Zhang, Y., Sun, Y., Shi, Y., Walz, T., and Tong, L. (2020). Structural Insights into the Human Pre-mRNA 3 '-End Processing Machinery. *Molecular Cell*, 77(4):800–809.

Zhang, Z., Boonen, K., Ferrari, P., Schoofs, L., Janssens, E., van Noort, V., Rolland, F., and Geuten, K. (2016). UV crosslinked mRNA-binding proteins captured from leaf mesophyll protoplasts. *Plant Methods*, 12(1):42.

Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., von Schack, D., and Zhang, B. (2015). Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics*, 16(675):1–14.

Zhong, S., Li, H., Bodi, Z., Button, J., Vespa, L., Herzog, M., and Fray, R. G. (2008). MTA Is an *Arabidopsis* Messenger RNA Adenosine Methylase and Interacts with a Homolog of a Sex-Specific Splicing Factor. *The Plant Cell*, 20(5):1278–1288.

Zhou, Q., Li, T., and Price, D. H. (2012). RNA Polymerase II Elongation Control. *Annu. Rev. Biochem.*, 81(1):119–143.

Zhu, D., Mao, F., Tian, Y., Lin, X., Gu, L., Gu, H., Qu, L.-j., Wu, Y., and Wu, Z. (2020a). The Features and Regulation of Co-transcriptional Splicing in Arabidopsis. *Molecular Plant*, 13(2):278–294.

Zhu, J., Liu, M., Liu, X., and Dong, Z. (2018). RNA polymerase II activity revealed by GRO-seq and pNET-seq in Arabidopsis. *Nature Plants*, 4(12):1112–1123.

Zhu, S., Ye, W., Ye, L., Fu, H., Ye, C., Xiao, X., Ji, Y., Lin, W., Ji, G., and Wu, X. (2020b). PlantAPAdb: A Comprehensive Database for Alternative Polyadenylation Sites in Plants. *Plant Physiol.*, 182(1):228–242.