Identification and functional characterization of a *de novo* point mutation identified in a patient with non-syndromal microcephaly and intellectual disability

Dissertation

zur Erlangung des akademischen Grades des Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im

Fachbereich Biologie, Chemie, und Pharmazie

der

Freien Universität Berlin

Vorgelegt von

Tao Chen (陈 涛)

aus Sichuan, China

Mar 2016

Die vorliegende Arbeit wurde von Januar 2012 bis März 2016 am Max-Delbrück-Centrum für Molekulare Medizin (MDC) unter der Anleitung von Prof. Wei Chen angefertigt.

1. Gutachter: Prof. Dr. Constance Scharff

2. Gutachter: Prof. Dr. Wei Chen

Datum der Disputation: July 14th, 2016

Table of Contents

DECLARATION ..... **I**

ACKNOWLEGEMENT ..... **II**

SUMMARY ..... **IV**

ZUSAMMENFASSUNG ..... **VI**

1 INTRODUCTION **1**

1.1 Pre-mRNA splicing **1**

1.2 The stepwise assembly of spliceosome **2**

1.3 Biogenesis of spliceosomal RNPs **5**

1.3.1 Nuclear transcription and exportation of snRNAs 6

1.3.1.1 Nuclear transcription of snRNAs 6

1.3.1.2 Nuclear exportation of newly synthesized snRNAs 6

1.3.2 Cytoplasmic biogenesis of snRNPs 7

1.3.2.1 The Sm core assembly of snRNPs 7

1.3.2.2 Hypermethylation of cap structure and 3' end trimming 10

1.3.3 Nuclear importation and snRNPs remodeling 10

1.4 Alternative splicing (AS) **11**

1.4.1 AS is a crucial step of gene expression regulation 11

1.4.2 Types of AS 12

1.4.3 The mechanism of AS 13

1.4.3.1 AS regulated by SS recognition 14

1.4.3.2 AS regulated by transition from exon definition to intron definition 16

1.4.3.3 AS affected by transcription and chromatin status 17

1.5 Splicing defect and human disease **20**

1.5.1 *Cis*-effect: splicing code disruption 20

1.5.2 *Trans*-effects: defects of splicing machinery or splicing regulators 21

1.5.2.1 Disruption of splicing regulator 21

1.5.2.2 Depletion of splicing factors by pathogenic RNA 22

1.5.2.3 Disruption of core spliceosomal component 22

1.5.3 Aberrant splicing contributes to cancer development 23

1.6 Previous studies about Sm proteins **24**

1.7 The goal of this study **26**

2 MATERIALS AND METHODS **27**

2.1 Materials **27**

2.1.1 Cell lines 27

Table of Contents

2.1.2 Vectors 27
2.1.3 Cell culture mediums 28
2.1.3.1 Maintainance medium for HEK293 Flp-In T-Rex cell line 28
2.1.3.2 Antibiotic-free medium for transfection 28
2.1.3.3 Stable cell line selection medium 28
2.1.3.4 Freezing medium 28
2.1.4 Enzymes 28
2.1.5 Kits 29
2.1.6 Chemicals 29
2.1.7 Other reagents 29
2.1.8 Major Equipments 30
2.2 EXPERIMENTAL METHODS 30
2.2.1 Cloning of exogenous SmE into mammalian expression Vector 30
2.2.1.1 Total RNA extraction 30
2.2.1.2 Reverse transcription (RT) and PCR 30
2.2.1.3 Cloning of snrpE cDNA into pENTR-D-TOPO vector 31
2.2.1.4 Construction of HA-tagged wild-type or mutant SmE expression pcDNA5/FRT/TO vector 32
2.2.1.5 Construction of wild-type and mutant SmE co-expression pcDNA5/FRT/TO vector 33
2.2.2 Construction of single guide RNA (sgRNA) expression vector 34
2.2.3 Construction of homologous recombination donor vector 35
2.2.4 Establishing exogenous SmE stable expression HEK293 cell lines 37
2.2.5 T7 endonuclease assay 38
2.2.6 Establishing HEK293 cell line with point-mutation in *snrpE* 39
2.2.7 Immunoprecipitation (IP) 40
2.2.8 Western blot (WB) 41
2.2.9 Immunofluorescence 42
2.3 COMPUTATIONAL METHODS 43
2.3.1 Exome sequencing and bioinformatics 43
2.3.2 Analysis of splicing by RNA-seq 43
2.3.3 Gene expression quantification 45
3 RESULTS 46
3.1 Identification of a *de novo* mutation from the patient with non-syndromal microcephaly and intellectual disability by whole exome sequencing 46
3.2 The integration capacity and subcellular distribution of mutant SmE protein is affected 49
3.3 The early assembly phase of snRNPs is affected by this mutation 52
3.4 The general mRNA splicing in the patient is affected 54
3.5 The deficiency of SmE protein in HEK293 cells causes abnormal splicing 56

Table of Contents

3.6 The mutation site can be successfully introduced into HEK293 cells by the combinational application of CRISPR/Cas9 and *piggyBac* transposon system .. 58

3.7 The HEK293 cells with point mutation showed the abnormal splicing pattern as observed in the patient fibroblasts 61

4 DISCUSSION 64

4.1 Exome sequencing is a powerful tool for identifying the potential causative mutations or genes 65

4.2 The mutation affects the function of SmE protein by interfering the early assembly phase 66

4.3 The heterozygous missense mutation causes the aberrant mRNA splicing ... 69

4.4 Aberrant splicing and human diseases 70

4.5 Perspective 74

5 REFERENCES 75

PUBLICATION 93

CURRICULUM VITAE 94

Appendix 95

DECLARATION

I hereby declare that this thesis is the results of my own original research work and has not been submitted in any form for another degree of diploma at any university or other institute of education. Contributions from others have been clearly and specifically indicated in the acknowledgement. This thesis is submitted to Department of Biology, Chemistry and Pharmacy of Freie Universität Berlin to obtain the academic degree of Doctor rerum naturalium (Dr. rer. nat.).

Tao Chen

2016-03-23

ACKNOWLEGEMENT

For me, five year's PhD study is an invaluable experience. During the study, I have learned a lot of knowledge and skills that serve as the lifetime benefit. I am very grateful to all of the people who provided the supporting hands during my PhD study.

I would like to thank…

…my supervisor Prof. Dr. Wei Chen for offering me the great opportunity to join his research at Max-Delbrück-Center for Molecular Medicine (MDC). For me, he is an excellent mentor and supervisor. He guides me into the field of "Systems Biology", "Medical and Functional Genomics" and "Splicing Regulation and Genetic Diseases". He also teaches me how to conceive and perform the scientific research in a rigorous and reliable way. I am really grateful for his brilliant supervisions and suggestions as well as the encouragement, which will benefit my future career.

...my university supervisor Prof. Dr. Constance Scharff for official registration at the Free University of Berlin and the critical review of my thesis. Her generous support has been of great help during my PhD stay.

...all of the lab members from Chen lab. I really enjoy the wonderful time we had together. I would like to specially thank Dr. Na Li, Dr. Xintian You (Arthur), Dr. Sebastian Fröhler and Bin Zhang for our collaborations in different projects. I also would like to thank Dr. Yuhui Hu, Dr. Yongbo Wang, Dr. Xi Wang, Dr. Kun Song, Dr. Xiao Cui, Dr. Jieyi Xiong, Wei Sun (Sunny), Hang Du, Qingsong Gao, Jingyi Hou, Meisheng Xiao, Yisheng Li for fruitful discussions and constitutive suggestions. I also want to express my acknowledgement to Claudia Quedenau, Mirjam Feldkamp, Madlen Sohn, Sarah Nathalie Vitcetz and Claudia Langnick for their technical support in my projects.

ACKNOWLEGEMENT

...my collaborators: Dr. Angela M. Kaindl for providing the clinical samples; Prof. Dr. Utz Fischer, Dr Archana Prusty, Thomas Ziegenhals for helping with biochemical experiments.

...Jennifer Stewart, Sabrina Deter and Sylvia Sibilak for their kind help with the registration in Free Univeristy of Berlin and dealing with my working contract and visa prolongation.

...all of BIMSB members for creating the open and cooperative working atmosphere.

...the China Scholarship Council (CSC) for supporting me with a four year scholarship.

...Finally, I want to devote my sincere thanks to my parents, grandma and younger sister for their encouragement and support. I also faithfully dedicate this dissertation to my beloved girlfriend (Yan Wen) who has always been making her efforts to support me and patiently waiting for me in Hangzhou, China. Without their sustained support, it is impossible for me to finish this journey abroad.

SUMMARY

Higher eukaryotes take advantage of alternative splicing to gain and expand the transcriptome complexity and proteome diversity without increasing the number of genes. In comparison to other organisms, the humans utilize the most complex alternative splicing events. With the increase of complexity along the evolution, the susceptibility of splicing to mis-regulation also increases.

Although lots of mutations within splicing factors have been found to associate with various human diseases so far, only very few have been identified in the basal components of spliceosome as causing factors of human diseases. Consequently, it leads to a severe shortage of established human disease models to investigate the functions of these basal components.

This thesis aims to identify and functionally characterize a key *de novo* mutation in a basal component of spliceosome putatively associated with a human disease. In this study, a missense heterozygous mutation in the gene *snrpE* was identified by the whole exome sequencing from a family with a child suffering from non-syndromal microcephaly and intellectual disability. The *snrpE* gene encodes the SmE protein, an indispensible spliceosomal component. The gene as well, the mutated site is deeply conserved in different species. The available 3D structure data indicates that the point mutation might affect the assembly of snRNPs, which could be supported by our biochemical and immunofluorecent results showing the decreased integration capacity and aberrant cytoplasmic enrichment of the mutant protein.

To further investigate the functional relevance of the identified mutation on splicing pattern at the whole transcriptome level, the RNA-seq was applied and the results showed that the mRNA splicing is abnormal in the patient derived fibroblasts and increased intron retention is the major defect. Similar effect on mRNA splicing can be recapitulated in SmE deficient HEK293 cells induced by

siRNA knock down approach, indicating the pivotal role of this mutation in the splicing function of SmE.

To clearly pinpoint the function of this mutation, we combined the recently developed CRISPR/Cas9 and *piggyBac* transposon system to introduce the mutation into the genomic loci in HEK293 cells. The RNA-seq data showed again that the mutation can induce a similar splicing defect as observed in the patient derived fibroblasts.

Taken together, our results showed that the *de novo* missense mutation we identified affects the functional amount of snRNPs by interfering the early assembly phase, leading to the transcriptome-wide mRNA aberrant splicing.

ZUSAMMENFASSUNG

Nahezu alle protein-codierenden Gene höherer Eukaryonten enthalten nicht-codierende Sequenzabschnitte (Introns), welche die codierenden Abschnitte (Exons) unterbrechen. Für die Ausbildung des offenen Leserahmens werden die Introns durch einen komplexen Prozessierungsschritt, dem sog. Spleißen, entfernt und die Introns präzise zur reifen mRNA vereinigt. Dieser Prozess wird durch das Spleißosom katalysiert, einer makromolekularen Maschine bestehend aus kleinen RNA-Proteinkomplexen (snRNPs) und einer großen Anzahl von Proteinfaktoren.

Speziell bei höheren Eukaryonten findet man neben dem kanonischen Spleißen auch das sog. alternative Spleißen, d.h. die wahlweise Inklusion bzw. Exclusion von Exons in einem Transkript. Dieser Prozess ist hoch-reguliert und erlaubt die massive Ausweitung der Transkriptom- bzw. Proteom diversität, ohne dabei die Anzahl ihrer Gene zu erhöhen.

Genetische Krankheiten des Menschen werden häufig durch Mutationen in den Spleißsignalen auf der mRNA oder in Spleißfaktoren gefunden. Jedoch sind bislang nur sehr wenige Fälle bekannt, bei denen Mutationen in den Hauptbestandteilen der Spleißmaschinerie menschliche Krankheiten hervorrufen. Dies führt zu einem Engpass an etablierten Modellkrankheiten, die zur Erforschung grundlegender Faktoren und deren Funktionen dienen könnten.

Ziel der Dissertation war die Identifizierung und funktionelle Charakterisierung einer *de novo* Mutation in einem konstitutiven Bestandteil des Spleißosoms, welche vermeintlich in Verbindung zu einer menschlichen Erkrankung steht. Hierbei handelt es sich um eine Missense-Mutation im *snrpE* Gen. Diese wurde mittels Exomsequenzierung in einer Familie identifiziert, deren Nachkommen Mirkozephalie und geistige Retardation aufweist. Das *snrpE* Gen codiert für das evolutionär hoch-konservierte SmE Protein, einem essentiellen Grundbaustein des Spleißosoms. Kristallstrukturen des Proteins weisen darauf hin, dass die

Punktmutation die Inkorporation des Proteins in das Spleißosom beeinflussen könnte. Diese Vermutung wurde durch biochemische und Zellbiologische Untersuchungen in dieser Arbeit bestätigt. Zum einen zeigen Immunofluoreszenz Untersuchungen die verminderte Integrationsfähigkeiten und anomale zytoplasmatische Anreicherung des Mutierten SmE Proteins. Zum anderen zeigen biochemische Untersuchungen, dass die Zusammenlagerung der snRNPs gestört ist.

Da das SmE Protein ein Grundbaustein des Spleißosoms ist, lag die Annahme nahe, dass die mRNA Prozessierung ebenfalls betroffen ist. Um dies zu Untersuchen, wurden Fibroblasten des Patienten hergenommen um deren RNA zu Sequenzieren. Hierbei wurde deutlich, dass einbehaltende Introns den Haupteffekt darstellen und somit das Spleißen gestört ist.

Um die Effekte der Mutation Patientenunabhängig betrachten zu können, wurden zusätzliche Experimente in der Zellkultur an HEK293 Zellen durchgeführt. Hierfür wurden zum einen mittels siRNA die Menge an SmE vermindert und zum anderen mit Hilfe der kürzlich entwickelten CRISPR/Cas9 und *piggyBAC* Transposon Systeme die Mutation des Proteins künstlich in die Zellen eingebracht. Sowohl die Verminderung von SmE als auch die künstliche Mutation zeigten die gleichen Auffälligkeiten wie die Spleißanomalien des Patienten.

Zusammenfassend zeigten unsere Experimente, dass die von uns identifizierte *de novo* Missense-Mutation, die Funktionalität des Spleißosoms erheblich einschränkt. Ausgelöst wird dies durch ein eingeschränktes Zusammenlagern der kleineren snRNP Proteinkomplexe. Diese Einschränkung führt zu transkriptomweiten abnormen Spleißen.

1 INTRODUCTION

1.1 Pre-mRNA splicing

Messenger RNA (mRNA), as one of the essential macromolecules of life, carries the information of genomic DNA and is used as the blueprint to guide the synthesis of proteins. To get mature mRNA, the nascent precursor mRNA (pre-mRNA) needs to go through the complex process, including 5' capping of 7-methylguanosine, polyadenylation at the 3' end of transcripts and removal of intervening sequence (intron) called splicing. The process of pre-mRNA splicing is catalyzed by a multi-megadalton ribonucleoprotein complex (RNP) named spliceosome [1]. In most eukaryotes, there coexist two types of unique spliceosomes: the U2-dependent spliceosome catalyzing the removal of U2-type introns, and the U12-dependent spliceosome responsible for removing rare U12-type introns [2].

Information within pre-mRNA contributes to define the intron, including the short, conserved 5' splice site (SS), 3' SS and branch point site (BS) [3] (Figure 1.1A). In yeast, the BS is typically located 18-40 nucleotide upstream from the 3' SS, while, in the higher eukaryotes the BS is followed by polypyrimidine tract (PPT) (Figure 1.1A).

The intron removal of pre-mRNA is catalyzed by a sequential two-step transesterification reaction (Figure 1.1B) [4]. First, the 2' OH group of BS, adenosine, carries out the nucleophilic attack to the 5' SS which results in the cleavage at 5' exon-intron boundary and the ligation of 5' end of intron with branch adenosine to form a lariat structure. Second, the 3' OH group of 5' exon attacks the 3' SS, leading to the ligation of 5' and 3' exons and release of the intron lariat (Figure 1.1B).

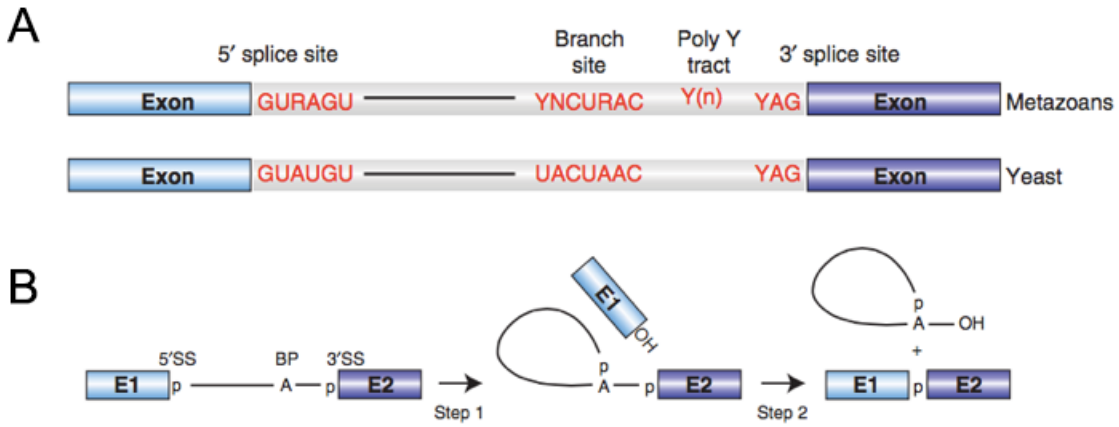

Figure 1.1: Illustration of pre-mRNA splicing. (A) The sequence for 5', 3' SS and BS are conserved between metazoans and budding yeast (*S. cerevisiae*) whereas the PPT marked by Y(n) is found only in higher eukaryotes. Y= pyrimidine, R= purine. (B) Schematic illustration of the two-step mechanism of pre-mRNA splicing. The colored boxes and lines represent the exons and introns, respectively. The branch point site, adenosine, is indicated by A, and the phosphate group (p) at 5' and 3' SS is also shown. Adapted from [5].

1.2 The stepwise assembly of spliceosome

The assembly of spliceosome is a dynamic process, including the dynamic conformation and composition change that confers accuracy and flexibility on the machinery.

In the eukaryotes, the U2-type spliceosome is composed by U1, U2, U4, U5 and U6 small nuclear ribonucleoproteins (snRNPs) and numerous non-snRNP proteins whereas the main subunits of U12-type spliceosome are U11, U12, U4atac, U5 and U6atac snRNPs [2]. The U5 snRNP is shared by both U2-and U12-type spliceosome. It is believed that the general architecture and biogenesis of the U snRNPs of the major and minor spliceosomes are similar [6]. In addition to U6 and U6atac snRNPs, each other snRNP contains a uridine-enriched snRNA, a common set of seven Sm proteins (B/B', D3, D1, D2, E, F and G) and a variable number of snRNP specific proteins [7].

The assembly of spliceosome is orchestrated in an ordered step-wise manner [8, 9] (Figure 1.2). For the short intron (200-250 nt), the spliceosome is assembled across the intron, called intron definition [10]. First, the earliest cross–intron complex, called E complex, is formed. The U1 snRNP interacts with 5' SS of intron through base pairing 5' end of U1 snRNA with 5' SS of intron in an ATP-independent way, and the non-snRNP proteins such as SF1/BBP and U2AF (U2 auxiliary factor) interact with BS and PPT, respectively. U2AF is composed by two subunits, U2AF35 and U2AF65. The U2AF65 binds to the PPT whereas the U2AF35 interacts with 3' SS. In a subsequent ATP-dependent process catalyzed by the DExD/H helicases pre-mRNA processing 5 (Prp5) and Sub2, U2 snRNP is recruited and SF1 is replaced by U2 snRNP. U2 snRNA recognizes sequences around the branch point adenosine and stably associates with BS to form the pre-spliceosome, also called A complex.

After the assembly of A complex, the U4/U6.U5 tri-snRNP pre-assembled from the U5 and U4/U6 snRNPs is then recruited to generate the pre-catalytic B complex in a reaction catalyzed by DExD/H helicase Prp28. The resulted B complex goes through a series of conformational and compositional rearrangements leading to form the catalytically active B complex (B^{act} complex). Multiple RNA helicases (Brr2, Snu114) are required for rearrangements leading to form the interaction between U2-U6 snRNA that catalyzes the splicing reaction afterwards [11]. The RNA helicase mediated rearrangements also leads to the dissociation of U1 and U4 snRNPs from the complex [12], which is thought to unmask the 5' end of U6 snRNA. Bact complex is further activated by DEAH-box RNA helicase Prp2, which leads to the generation of $B^{\star}$ complex. $B^{\star}$ complex then executes the first step of splicing, leading to the formation of C complex, which is composed by free exon 1 and the intron-exon 2 lariat intermediate.

Before carrying out the second step of splicing, C complex further undergoes additional ATP-dependent rearrangements that are dependent on Prp8, Prp16 and synthetic lethal with U5 snRNA 7 (Slu7). It results in a post-spliceosomal complex that contains the lariat intron and ligated exons. Finally, the U2, U5 and U6 snRNPs are released from the mRNP particle and recycled for additional rounds of splicing. The release of spliced product from spliceosome is catalyzed by the DExD/H helicase Prp22 [13, 14]. Post-catalytic spliceosome dissociation is also carried out by several RNA helicases (such as, Brr2, Snu114 and Prp43) in an ATP dependent way [15].

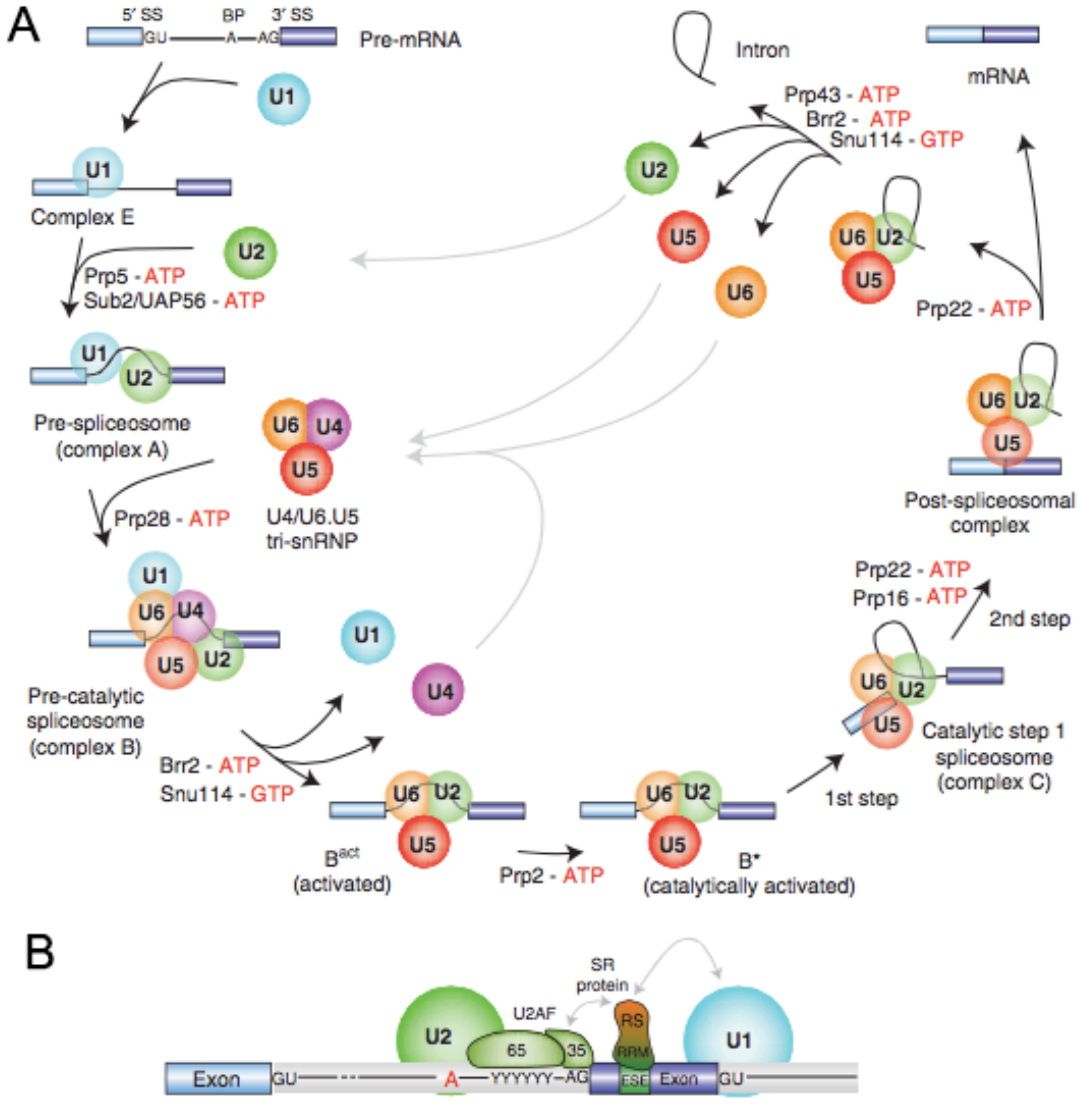

Figure 1.2: Pre-mRNA splicing by major spliceosome. (A) Canonical cross-intron assembly, catalysis and disassembly pathway of U2-type spliceosome. Boxes and lines indicate the exon and intron of pre-mRNA, respectively. The stepwise interactions of snRNPs (indicated by colored circles), but not those of non-snRNP proteins and the formed various spliceosomal complexes are indicated. For each step, the evolutionarily conserved DExH/D-box RNA ATPase/helicases are also shown. (B) Illustration of cross-exon spliceosome assembly for long introns. The mechanism for the transition from exon definition complex to intron definition complex is till poorly understood. Adapted from [5].

For the introns with length more than 250 nt, the spliceosome complex is first assembled across the exon, called exon definition [10] (Figure 1.2B). For exon definition, the U1 snRNP first binds to the 5' SS down-stream of target exon whereas the SF1/BBP and U2AF interact with BS and 3' SS up-stream of target exon, respectively. This in turn leads to the binding of U2 snRNP to the BS up-stream of target exon. The regulatory sequence located within the target exon recruits protein of SR protein family to establish the protein-protein interaction across the target exon. These interactions can help to stabilize the exon definition complex [16, 17]. Since the catalytic step of splicing occurs through the intron, a switch is required from the exon definition complex to intron definition complex. However, the mechanism of this process still remains elusive [5].

1.3 Biogenesis of spliceosomal RNPs

As the major building blocks of spliceosome, the biogenesis of snRNPs is pivotal to the assembly of spliceosome and RNA processing. The biogenesis of snRNPs is classified into three steps: nuclear transcription and exportation of snRNAs, cytoplasmic assembly of Sm core and nuclear importation for further maturation.

1.3.1 Nuclear transcription and exportation of snRNAs

1.3.1.1 Nuclear transcription of snRNAs

The nuclear transcription and exportation of snRNAs is considered as the starting point of biogenesis of spliceosomal snRNPs. The snRNAs are a group of abundant, non-coding, non-polyadenylated transcripts. Based on the common sequence features and protein cofactors, the snRNAs can be divided into two classes: Sm class snRNAs (U1, U2, U4, U5, U11, U12 and U4atac) and Sm-like class snRNAs (U6 and U6atac). Among the snRNAs, Sm class snRNAs are transcribed by RNA polymerase II whereas Sm-like class snRNAs are generated by RNA polymerase III [18, 19]. Although the different transcription systems are used for transcribing different snRNA genes, the regulatory sequences in these genes are remarkably similar.

1.3.1.2 Nuclear exportation of newly synthesized snRNAs

After transcription, all of the newly synthesized Sm class snRNAs first need to be transferred into cytoplasm to assemble the Sm core onto the Sm binding site of snRNAs. Given the biogenesis of Sm-like snRNP is distinct from that of Sm class snRNPs, we only focus on the biogenesis of Sm class snRNPs here.

To successful assemble the Sm class snRNPs, the newly synthesized snRNAs are actively exported into cytoplasm in a receptor-mediated transportation way (Figure 1.3). The heterodimeric cap binding complex (CBC) consisting of two subunits, CBP20 and CBP80, recognizes the m^7G-cap structure of newly synthesized Sm class snRNAs [20, 21]. After CBC binding, the phosphorylated adaptor of export protein (PHAX) [21, 22] is recruited, and the nuclear export signal (NES) located within PHAX interacts with export receptor CRM1 when CRM1 is bound by small GTPase Ran in its GTP-bound form [21, 23]. As a result, the snRNA export complex is formed. After transportation into the cytoplasm, the GTP bound to RAN is hydrolyzed by RanGAP1, which leads to

the decreased affinity of CRM1. The dephosphorylaton of PHAX triggers the dissociation of exportation complex and the cargo snRNAs are released into the cytoplasm [21].

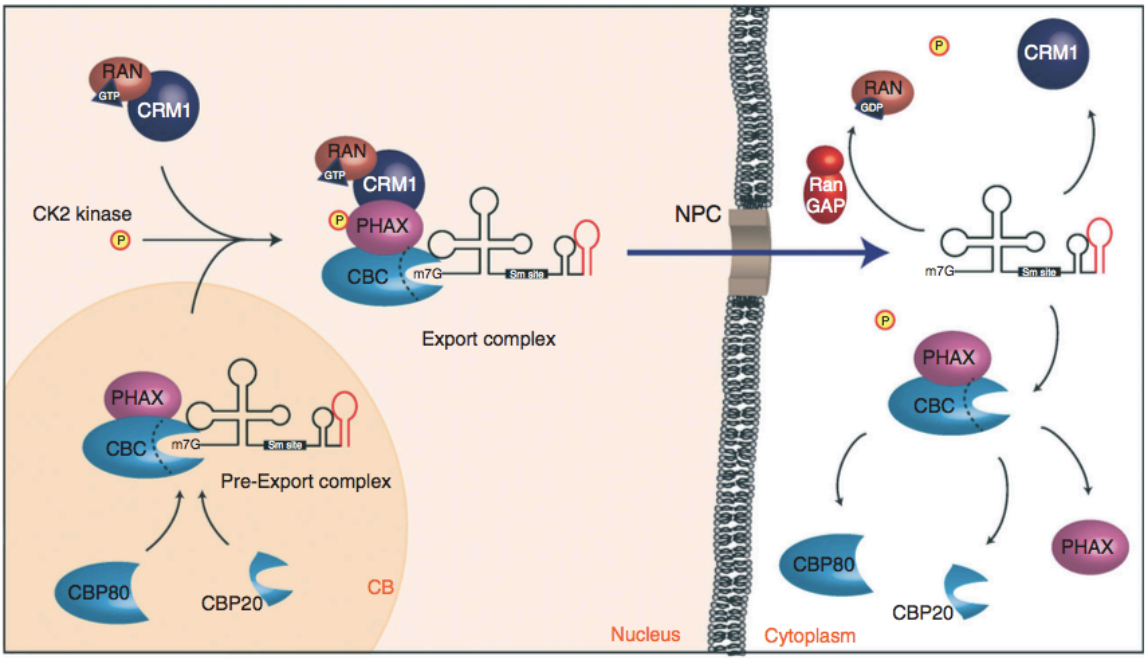

Figure 1.3: The exportation of newly transcribed pre-snRNA into cytoplasm. The CBC consisting of two subunits CBP20 and CBP80, recognizes the m^7G-cap structure and binds to it. Upon binding, the PHAX is recruited. Then, CRM1 together with GTP-bound Ran interact with phosphorylated PHAX in the neucleoplasm to form the exporting complex. After transportation, Ran-GTP is hydrolyzed to GDP by Ran GTPase-activating protein, RanGAP1. The hydrolysis of GTP results in the dissociation of exporting complex and snRNA is released into the cytoplasm. Adapted from [24].

1.3.2 Cytoplasmic biogenesis of snRNPs

After the dissociation of nuclear exportation complex, the cytoplasmic assembly process is initiated, including Sm core assembly, hypermethylation of cap structure and 3' end trimming of snRNAs.

1.3.2.1 The Sm core assembly of snRNPs

Once the newly synthesized snRNAs export into the cytoplasm, the Sm core assembly is started （Figure 1.4). In vitro studies indicated that purified Sm

hetero-oligomers could spontaneously form the Sm core in the presence of appropriate RNA target [25, 26]. However, the *in vivo* Sm core assembly is strictly scrutinized to prevent loading the Sm proteins onto unspecific RNAs. It is believed that *in vivo* the SMN complex，consisting of SMN and Gemin 2-8, confers the specificity of the reaction to avoid the assembly of Sm proteins onto non-target RNA [27, 28], and to accelerate the reaction by lowering the energy requirement [29].

Before interacting with SMN complex, the Sm proteins interact with PRMT5 complex consisting of WD45/MEP5, pICln and PRMT5 (protein arginine methyltransferase 5) to form two distinct higher order Sm subcomplexes and a subset of Sm proteins (B/B', D1 and D3) are symmetrically dimethylated on designated arginine residues by PRMT5 complex [30-32] to increase the binding affinity. The first higher order Sm subcomplex is ring shaped and composed by pICln, SmD1, SmD2, SmE, SmF and SmG. Once it is formed, it dissociates from PRMT5 complex and is called as 6S complex, because of its sedimentation coefficient [29]. The second higher order subcomplex contains pICln, SmB/B' and SmD3 and it stably associates with PRMT5 complex after formation [29]. Instead of promoting assembly, the interactions between pICln and Sm proteins efficiently prevent Sm proteins binding onto snRNAs [29, 33]. Hence, pICln is considered as assembly chaperone for snRNP assembly to prevent the premature interactions between Sm proteins and snRNAs.

After the formation of two higher order subcomplexes, the SMN complex recruits and interacts with 6S complex on its outer surface [29]. The interaction between 6S complex and SMN complex, results in the dissociation of pICln and opening of 6S-ring. The recent crystallographic research indicated that Gemin2, a conserved component of SMN complex, binds directly to the five of Sm proteins of 6S complex and holds them in the proper 'horseshoe' orientation for subsequent snRNAs binding and ring closure [34]. The recruitment of snRNAs to

SMN complex is mediated by Gemin5, another component of SMN complex, which recognizes the stem-loop structure of snRNAs located in the down-stream of Sm binding site [28, 35]. Under the help of SMN complex, snRNAs are loaded into the inner of Sm ring and the Sm ring is closed to form the Sm core.

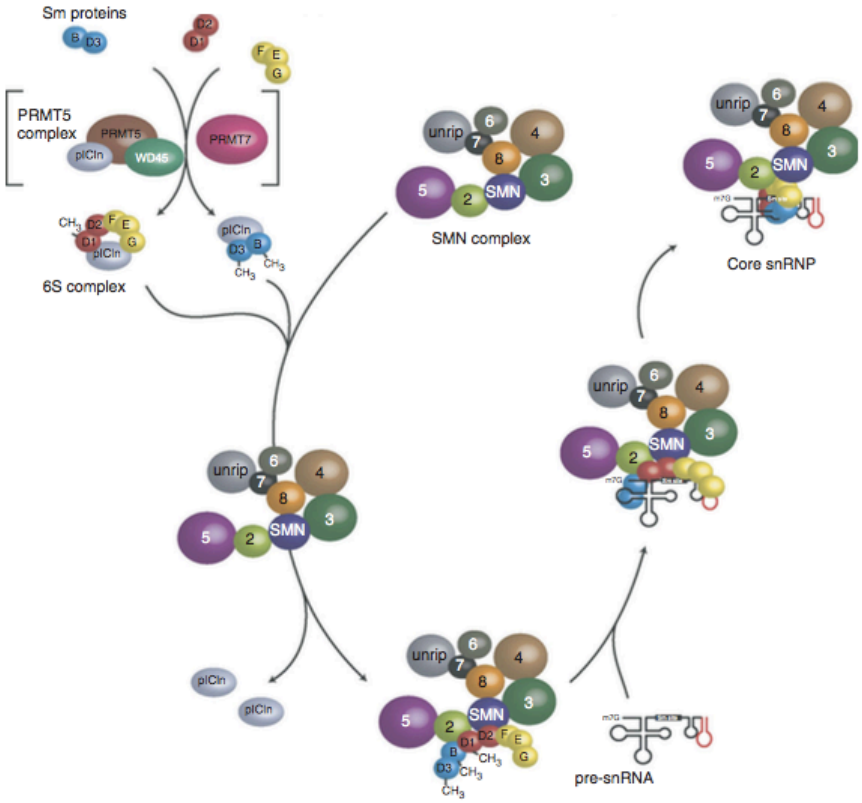

Figure 1.4: Cytoplasmic assembly of snRNPs. In the early phase of assembly, the Sm proteins interact with PRMT5 complex to form two higher order subcomplex (termed 6S complex and pICln-SmB-SmD3) and some Sm proteins (SmB, SmD1 and SmD3) are symmetrically dimethylated on designated arginine residues by PRMT5 complex to increase the affinity. Then, pre-formed subcomplexes are recruited onto the SMN complex. Upon binding onto the SMN complex, the pICln protein dissociates from the complex. snRNAs are recruited by Gemin5, a component of SMN complex. SMN complex loads the snRNAs into the inner of Sm subring and finishes the closure of ring to form the Sm core. D1, SmD1; D2, SmD2; D3, SmD3; E, SmE; F, SmF; G, SmG; B, SmB; 2-8, Gemin2-8. Adapted from [24].

1.3.2.2 Hypermethylation of cap structure and 3' end trimming

After the Sm core formation, another two processing steps, namely hypermethylation of cap structure and trimming of the 3'-extension ends, need to be finished prior to nuclear import (Figure 1.5). For the 5' cap structure hypermethylation, the Sm core is used as the platform for docking the methyltransferase Tgs1 (also termed PIMT in higher eukaryotes) [36]. Upon binding to the Sm core, Tgs1 catalyzes the reaction to convert the m7G-cap structure of snRNA into the trimethyguanosine $m_3^{2,2,7}$ G-cap structure [36, 37]. After hypermethylation, the Tgs1 will dissociate from the complex. Cocomitant with (or subsequent to) these 5' events, the extended 3' end of snRNAs is degraded by a yet unknown exo-ribonuclease. In yeast, it has been shown that exosome is involved in 3' end trimming. Whether these enzymes also play the same role in the snRNP maturation in higher eukaryotes is still unclear [38].

1.3.3 Nuclear importation and snRNPs remodeling

Once the cytoplasmic assembly is finished, the snRNP is actively imported into the nucleus (Figure 1.5). *In vitro* studies indicated that 2,2,7-trimethylguanosine (TMG) and Sm core function as nuclear localization signal [39]. Only the Sm core is strictly required for import, whereas the cap-dependence is varied among different snRNPs and cells. After dissociation of Tgs1 from the snRNP complex, the binding of snurportin 1, the snRNP specific import adaptor, to hypermethylated cap structure is followed. Then, the importin β binding domain (IBB domain) of snurportin 1 directly interacts with importin β to promote import [40]. The SMN complex accompanies the newly assembled snRNPs into the nucleus [41]. When the snRNPs have been successfully imported into the nucleus, SMN complex dissociates from the snRNPs immediately. Newly imported Sm-class snRNPs are transiently accumulated in the Cajal bodies (CBs) [42] for additional RNP remodeling and RNA-processing steps, including non-coding RNA guided covalent modification of snRNAs [43] and binding of

snRNPs specific proteins [44] to finish the maturation. Furthermore, the CBs are also thought to facilitate the *de novo* assembly and post-splicing reassembly of U4-U6 di-snRNP and U4-U6.U5 tri-snRNP [45, 46].

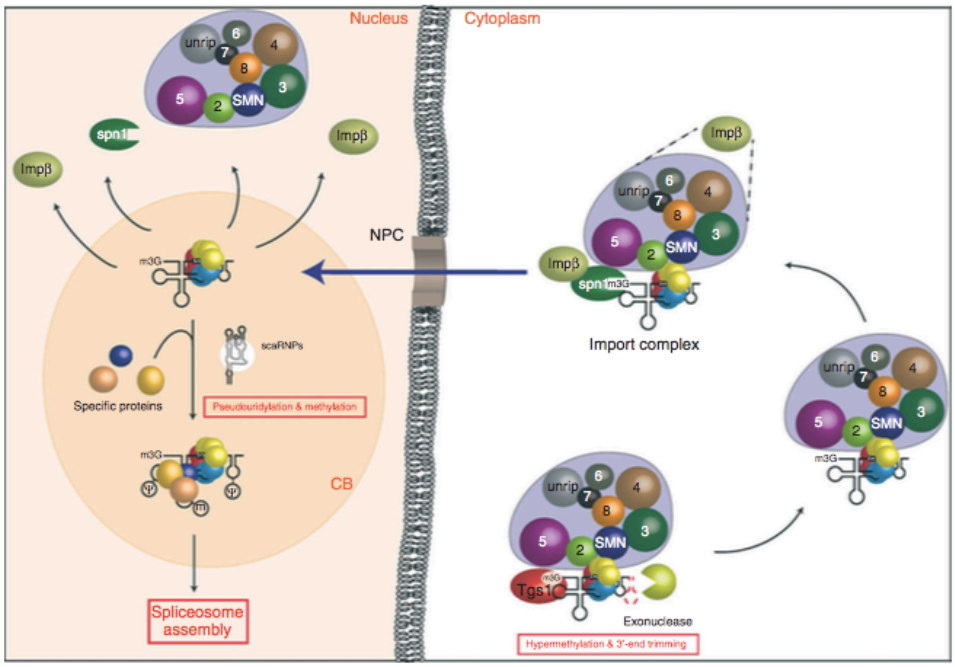

Figure 1.5: The nuclear import of snRNPs and final maturation. After the hyermethylation of cap structure and 3' end trimming, the hypermethylated cap structure is used as nuclear localization signal and bound by snurportin 1. The IBB domain of snurportin 1 directly interacts with importin β to form the import complex. The Sm core is also used as the nuclear localization signal. After nuclear import, the import complex is dissociated and the snRNPs accumulates in the CBs for further additional processing to finish the maturation, including adding snRNPs specific proteins and nucleotide modification. Ψ= pseudouridylation, m=2'-O-ribose methylation. Mature snRNPs are ready to engage into the splicing machinery. Adapted from [24].

1.4 Alternative splicing (AS)

1.4.1 AS is a crucial step of gene expression regulation

AS, the process by which the exons of pre-mRNAs from genes are spliced in different arrangements to produce structurally and functionally distinct mRNA

and protein variants [47], is one of the most extensively used mechanisms that accounts for the greater macromolecular and cellular complexity of higher eukaryotic organisms [48]. AS is a crucial mechanism for gene expression regulation and for generating proteomic diversity. Recent studies indicated that nearly 95% of human multi-exon genes are alternatively spliced [49, 50]. More importantly, AS plays numerous critical roles in the specification of cell fates [51, 52], tissue types [53, 54], development process [55, 56], sex determination [57, 58] and stimulation responses [59, 60].

1.4.2 Types of AS

Based on the pattern of AS, it can be classified into several types: (a) cassette exon splicing, (b) alternative 5' SS, (c) alternative 3' SS, (d) mutually exclusive alternative exons, (e) intron retention (IR), (f) alternative promoter and first exon, and (g) alternative poly A site and last exon (Figure 1.6). The cassette exon splicing accounts for at least one-third of known alternative splicing events, whereas the alternative 5' and 3' SS events together account for at least one-quarter [48]. These types of events not only have the potential to change the protein coding sequence, but they also have the capacity to generate the alternative 3' untranslated regions (UTRs) to affect the stability or translation of mRNA through the interaction between 3' UTR and micro-RNAs or RNA-binding proteins [61]. It is estimated that 80% of AS will affect the open reading frames (ORFs), leading to the amino acid deletion or insertion, or even frameshift [62]. The rest 20% located within the UTR affects the RNA stability, localization and translational efficiency [63, 64]. Additionally, a proportion of AS events will introduce the premature termination codon (PTC) triggering the mRNA degradation via nonsense mediated decay (NMD) pathway [65].

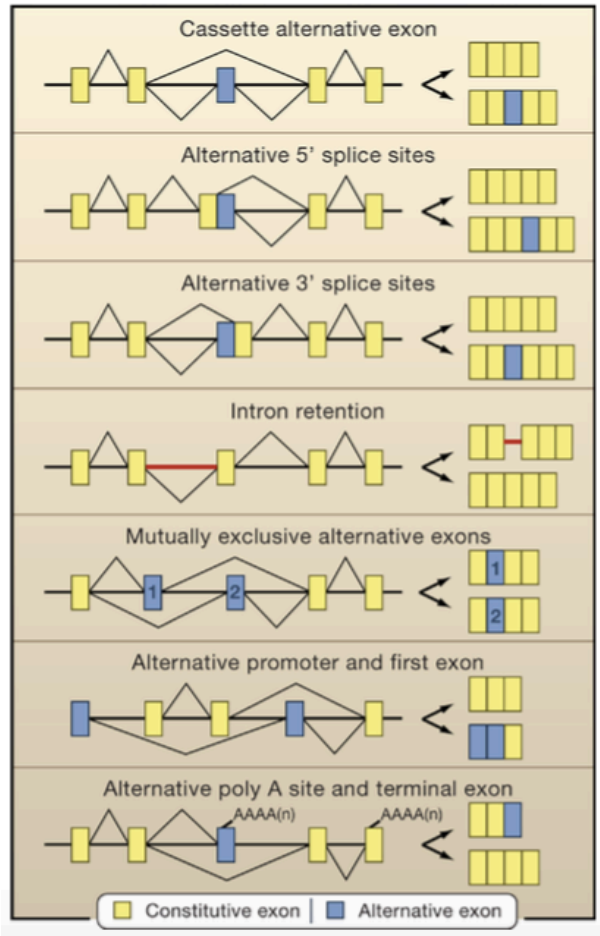

Figure 1.6: Different types of AS events. Yellow and blue boxes represent the constitutive and alternative exons, respectively. Black lines mean the introns to be spliced out, whereas the red line represents the retained introns. Adapted from [48].

1.4.3 The mechanism of AS

The decision of AS is made by the combinatorial interaction between *cis*-regulatory elements, *trans*-splicing factors and basal splicing apparatus. Based on the location and effect of *cis*-regulatory elements, they are classified into four categories: exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs). ESEs usually interact with SR family proteins [66-68] while the ISSs and ESSs are commonly bound by heterogeneous nuclear ribonucleoproteins (hnRNPs) [69, 70]. Compared to the other three types of *cis*-regulatory elements, the ISEs are

not well characterized. Only several studies have reported that some proteins, such as hnRNP F, hnRNP H, neuro-oncological ventral antigen 1 (NOVA1), NOVA2, FOX1 and FOX2 (also known as RBM9), can bind to ISEs and stimulate the splicing [71-74].

In general, the choice of AS is usually made at the stages of SS recognition and early spliceosome assembly [49]. However, in contrast to the general view, several studies reported that it could be made at different stages of spliceosome assembly, or even at the conformational change step [75-77].

1.4.3.1 AS regulated by SS recognition

The best-studied mechanism of AS regulation is the interference of the SS recognition. SR proteins usually promote the recognition of SS. After binding to ESE, SR proteins recruit the U1 snRNP to 5' SS and U2AF complex to 3' SS [78-80] (Figure 1.7A), respectively. In contrast, inhibition of SS recognition can be achieved in various ways. First, if the location of SS is close enough to splicing enhancers, the access of snRNPs or positive regulators will be blocked, leading to the inhibition of SS recognition [81-83]. For instance, binding of FOX1 and FOX2 onto intronic sequence blocks the formation of E complex by preventing SF1 from binding to the BS of CALCA (calcitonin-related polypeptide-α) pre-mRNA [83] (Figure 1.7B). Second, the splicing inhibitor can also sterically block the loading of activators onto enhancer sequence. The binding of FOX1 and FOX2 to exonic sequence in the CALCA pre-mRNA which is close to ESE for TRA2 and SRp55 binding, leads to the inhibition of E complex formation by preventing U2AF complex recruitment [83] (Figure 1.7B). hnRNP A1 binds an ESS upstream of the TRA2-dependent ESE in exon 7 in the SMN2 pre-mRNA, inhibiting the formation or stabilization of U2 snRNP complex [84, 85] (Figure 1.7C). Since the distance between the silencers and enhancers is sometimes over 100-200 bp, the simple 'bind and block' theory can't explain such kind of cases.

For such cases, one possible explanation is that these inhibitors function through multimerization along the RNA to mask the SS recognition [81].

Although the activator or inhibitor can facilitate or inhibit the recognition of SS, the final decision of alternative exon inclusion is determined by the combinatorial effect of these regulators, such as SR proteins and hnRNPs [86, 87]. For example, the SR protein 9G8 (also known as SFRS7), hnRNP F and hnRNP H regulate the splicing of exon 2 of α-tropomyosin through competitive binding to the same elements [88]. However, in some cases the effect of *cis*-regulatory elements and their cognate binding proteins totally depends on their position relative to the regulated exons. Several proteins, such as NOVA1, NOVA2, FOX1, FOX2, hnRNP L, hnRNP L-like, hnRNP F and hnRNP H, have been shown that different binding location leads to totally different effect [71, 73, 74, 89-91].

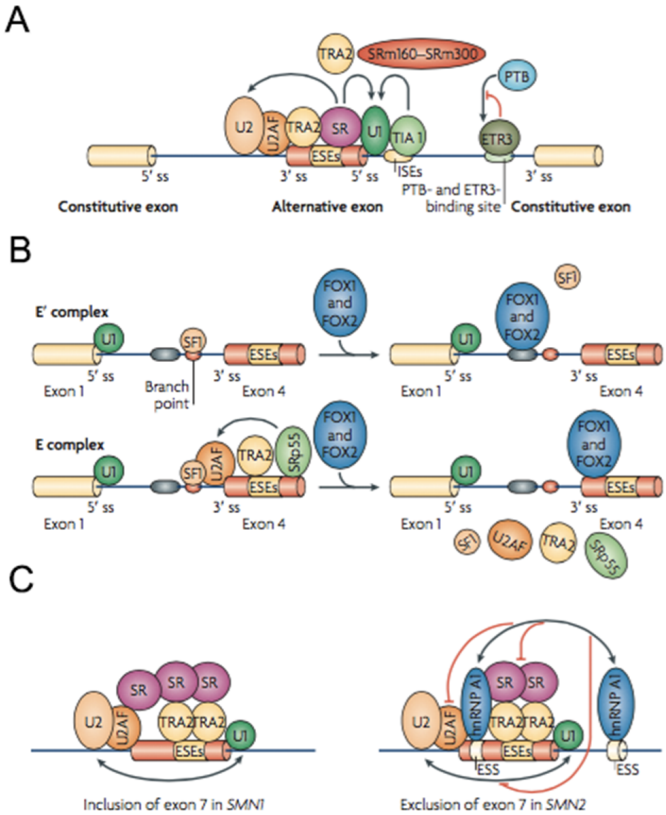

Figure 1.7: Mechanism of AS by SS selection. A, The binding of SR proteins to ESEs stimulates the binding of U1 snRNPs and U2AF to downstream 5' SS and upstream 3' SS, respectively. B, The binding of FOX1 and FOX2 inhibits the inclusion of exon4 of CALCA through blocking the binding of SF1 to the BS (top panel) or inhibiting the loading of TRA2 and SRp55 onto ESEs (lower panel), thereby arresting the spliceosome at two stages, E' and E complex. The arrow means that the binding of TRA2 and SRp55 promote the binding of U2AF complex. C, Single nucleotide variant creates the binding site for hnRNP A1 in the exon 7 of SMN2 and in the downstream intron. The binding of hnRNP A1 to the newly formed sites inhibits the formation or stabilization of U2 snRNP complex through direct or indirect interfering the activity of downstream TRA2-dependent ESE. Adapted from [92].

1.4.3.2 AS regulated by transition from exon definition to intron definition

In addition to regulating SS recognition, the transition from exon definition to intron definition can also affect AS. After the recognition of 5' and 3' SS, the exon definition must be converted into intron definition to form the functional spliceosome, involving the U1 and U2 snRNPs crossing intron interaction. A typical example is that the splicing factor RBM5 mediates exon 6 exclusion of the gene CD95 (also known as FAS) [93] (Figure 1.8). The binding of RBM5 to exon 6 doesn't affect the association of U1 and U2 snRNPs to adjacent SS. In contrast, it interferes the loading of U4/U6-U5 tri-snRNPs. Consequently, the transition from exon definition to intron definition is blocked.

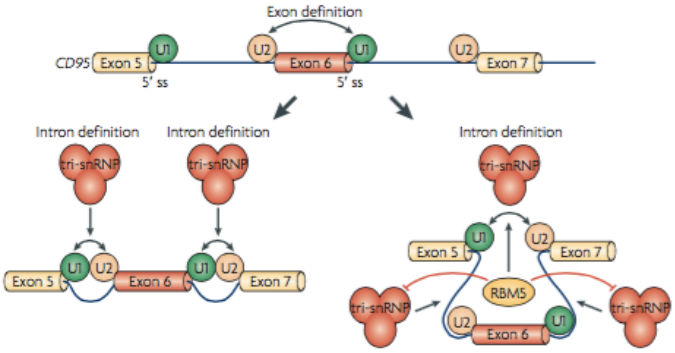

Figure 1.8: Mechanism of AS regulation at the transition from exon definition to intron definition. RBM5 disturbs the splicing of CD95 exon 6 by interfering the loading of U4/U6-U5 tri-snRNPs for further maturation of spliceosome. Double-head arrows represent intron and exon definition.

1.4.3.3 AS affected by transcription and chromatin status

Additionally, the decision of AS can also be affected by transcription [94-97] and chromatin status [98-101]. In the aspect of transcription, two models (Figure 1.9) have been proposed to explain the role of RNA polymerase II in the regulation of AS. The first one is called as recruitment model (Figure 1.9A), in which RNA polymerase II and transcription factors interact with splicing factors directly or indirectly [97, 102, 103], leading to the change of splicing efficiency. The second one is called as kinetic model (Figure 1.9B), which proposes that the rate of transcription affects the final decision of splicing. In brief, the transcription elongation rate could affect the efficiency of the recruitment of the splicing machinery, which subsequently influences the final splicing.

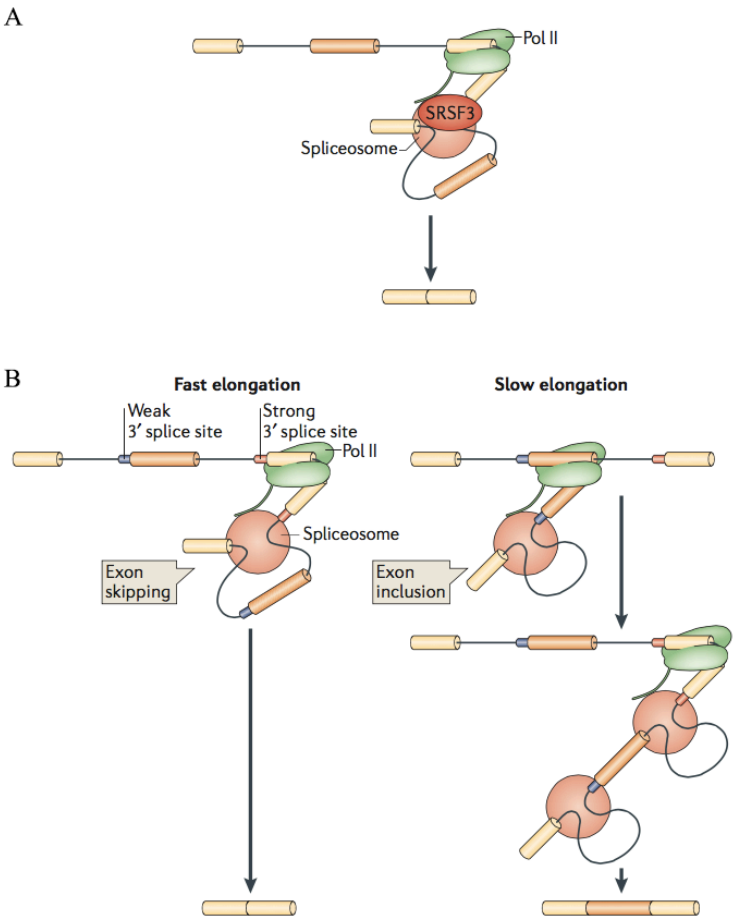

Figure 1.9: Two mechanisms of transcription coupling with AS. A, Recruitment model. The carboxy-terminal domain (CTD) of RNA polymerase II is used as platform to recruit the splicing factors affecting the splicing efficiency. B, Kinetic model. During the RNA polymerase elongation, the elongation speed affects the selection decision of SS. Fast elongation rate (left) favors the assembly of spliceosome to strong 3' SS downstream intron leading to exon skipping whereas the slow elongation (right) results in the recruitment of spliceosome components to the upstream intron. Adapted from [104].

On the other hand, increasing evidence demonstrates that the chromatin can also affect the decision of AS through histone modification (Figure 1.10) and nucleosome positioning [98]. Based on the relationship between histone modification and transcription status, the histone modifications can be classified

into two categories: one is associated with active transcription (such as: histone H3 Lys36 trimethylation (H3K36me3), H3K4me2, H3K4me3 and H3K9 acetylation (H3K9ac)) whereas the other one associates with transcriptional silencing (for example, H3K9me2, H3K9me3 and H3K27me3). In addition to histone modifications, nucleosome position also contributes to alternative splicing. It has been proposed that nucleosome position may help the splicing machinery to "find" exons since the nucleosomes would create transient pauses to polymerase II mediated elongation by "marking" the beginning of each exon and providing extra tie for the recognition of 3' SS by the auxiliary factor U2AF and other factors.

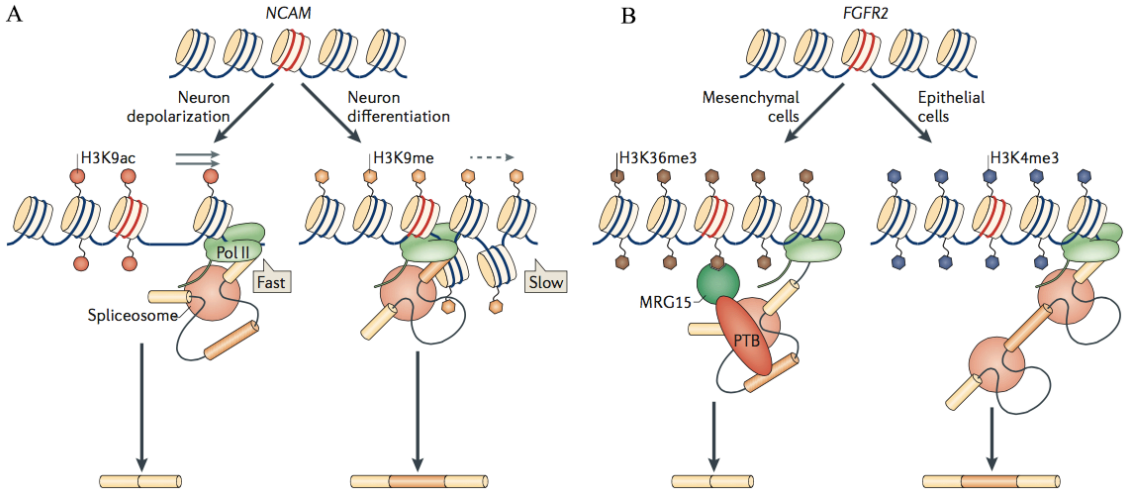

Figure 1.10: Two alternative mechanisms of histone modifications influences AS. A, The AS decision can be affected by the nature of histone modifications of chromatin in response to stimuli or cell differentiation. Neuron depolarization triggers the H3K9ac modification leading to exon skipping by increasing transcription elongation whereas neuron differentiation induces the formation of H3K9me resulting in exon inclusion through decreasing the rate of transcription elongation. B, Histone modification interferes the AS through a recruitment coupling mechanism. For instance, in mesenchymal cells, the intragenic H3K9me3 in fibroblast growth factor receptor 2 (FGFR2) recruits the negative splicing factor PTB leading to exon skipping whereas the H3K4me3 in epithelial cells increases the exon inclusion. Adapted from [104].

1.5 Splicing defect and human disease

Given that splicing is a compulsory step of mRNA maturation and 90% of multi-exon genes are alternatively spliced, it is not surprising that mis-regulation of splicing underlies many human diseases. According to the different underlying mechanism of mis-splicing, it can be classified into two main types: (1) disruption of *cis*- elements, such as the SS and the splicing *cis*-regulatory elements; (2) dysfunction of *trans*-acting factors affecting the basal splicing or splicing regulation, such as core spliceosomal components and splicing regulators.

1.5.1 *Cis*-effect: splicing code disruption

It has been reported that mutations located within the SS and splicing regulatory elements account for almost 50% of known human inherited disorders [105, 106]. The mutations lead to disease via changing protein function or imbalance the ratio of natural isoforms. One representative example of *cis*-effect induced disease is Frontotemporal dementia with Parkinsonism linked to chromosome 17 (FTDP-17), an autosomal dominant disorder caused by mutations in the *MAPT* gene encoding protein tau [107]. For FTDP-17, one of well-characterized mutations is N279K, which is caused by changing the TAAGAAG into GAAGAAG [108]. The GAAGAAG forms the core binding site for Tra2-β1 which acts as positive regulator of splicing and promotes the inclusion of exon 10 of tau gene. Since the exon 10 of tau gene encodes one of binding domains for microtubule, the introduction of mutation (N279K) could lead to the change of ratio between the isoforms with/without exon 10. To test such hypothesis that the change of the ratio is responsible for this disease, a transgenic mouse model is generated, which hosts a minigene of human tau, containing the promoter and all exons flanked with shortened introns. When the mutation (N279K) is introduced into the minigene, the mice shows the similar pathological phenotype as human with the same mutation [109]. Therefore, the imbalance ratio of two isoforms for tau

serves as the causal factor for FTDP-17.

Another example is Familial Dysautonomia (FD). FD as a recessive disease is caused by the loss function of IKBKAP (i-kappa-B kinase complex associated protein). More than 99.5% of FD patients are caused by a T>C mutation which happens at the position 6 of intron 20 [110]. Since U1 snRNA interacts with the last three nucleotides of the exon and the first six nucleotides of the downstream intron, this mutation disturbs the interaction between U1 snRNA and 5' SS by decreasing the strength of 5' SS of exon 20, and leads to the skipping of exon 20. Array data indicated that the expression of genes involved in oligodendrocyte and myelin formation is promoted by IKBKAP, which could perfectly explain the demyelination phenotype of FD [111].

1.5.2 *Trans*-effects: defects of splicing machinery or splicing regulators

1.5.2.1 Disruption of splicing regulator

Compared to *cis*-effects that normally act locally, the mutations located within the components of spliceosome or splicing regulators can globally disturb the splicing of a large number of genes. A diverse cohort of splicing regulators has been linked with diseases, such as TDP43, QKI, NOVA1, RBM10, RBM20, SRSF1 and RBFOX1 and so on. For instance, TDP43 (TAR DNA binding protein 43kDa) is a member of hnRNP family and shares the similar protein structure as hnRNP A1 and A2. TDP43 was originally identified as a transcriptional repressor binding to the HIV *trans*-acting response elements (TAR). It is reported that TDP43 is linked with many different human diseases. First, the TDP43 binds to expanded UG repeats in the intron 8 of CFTR gene, leading to exon skipping and cystic fibrosis [112]. In addition, the mutations in TDP43 are also found to associate with both sporadic and familial amyotrophic lateral sclerosis (ALS). TDP43 is a nuclear protein, however, cytoplasmic aggregation of TDP43 will lead to Frontotemporal dementia (FTD) and ALS

[113].

1.5.2.2 Depletion of splicing factors by pathogenic RNA

In addition to the mutations located in the *cis*-elements and splicing factors affecting splicing, the pathological expansion of repeats can also disturb the splicing pattern of the genes and lead to human diseases. Numerous exons contain short repeat sequences to increase the recognition possibility by a certain splicing factors [114, 115]. For instance, the number of intronic CA repeats in endothelia NO synthase gene correlates with the risk of coronary artery disease and hyperhomocysteinemie in a sex-specific manner [116]. The CA repeats interact with hnRNP L, and the action of hnRNP L is affected by the number of CA repeats. Besides, myotonic dystrophy (DM), the most common form of muscular dystrophy in adults, is another example of diseases caused by changing the number of repeat sequence. This disease is caused by extension of two types repeats: the CUG repeat extension in the 3' region of DMPK gene (dystrophia myotonica protein kinase) leading to DM1; the extension of CCUG repeat in the intron of ZNF9 accounting for DM2 [117]. For DM1, the CUG element usually interacts with two groups of RNA binding proteins, MBNL1 (muscleblind-like 1) and CUG binding protein. With the increasing the number of CUG repeats, more MBNL1 and CUG binding proteins are recruited and sequestered in the nuclear foci. The sequestration of MBNL1 leads to the decreased cellular available concentration and affects the AS of the genes that require the MBNL1. The CCUG repeats extension leads to DM2 also through sequestering the MBNL1 protein.

1.5.2.3 Disruption of core spliceosomal component

In addition to mutations within splicing regulator and expanding repeat sequence leading to human diseases, there are reports of mutations located in core component of splicing machinery resulting in human diseases. One notable

example is autosomal dominant forms of retinitis pigmentosa (RP), a photoreceptor neuronal degenerative syndrome, caused by mutations in the splicing factors: PRPF3, PRPF8, PRPF31, RP9 and SNRNP200 [118]. Another example is spinal muscular atrophy (SMA) caused by mutations in the SMN1 gene encoding SMN protein. Although the SMN is not splicing factor, SMN complex plays an essential role in the biogenesis of U snRNPs. It has been reported that each snRNA assembly is not uniformly affected [119].

Although many mutations within core component of spliceosome and splicing regulator are reported linked to diseases, only few mutations affecting basal components of spliceosome related to human disease have been identified so far. One remarkable example is that the mutations in U4atac snRNA that affect the formation of minor spliceosome, cause Taybi-Linder syndrome/microcephalic osteodysplastic primordial dwarfism type 1 (TALS/MOPD1), an autosomal recessive genetic and severe developmental disorder characterized by extreme intrauterine growth retardation, neurological and skeleton muscle abnormalities [120, 121]. The other two examples are: the mutations within the alternative exon of SmB/B' are linked with Cerebro-Costo-Mandibular Syndrome (CCMS) and the mutations of SmE are identified from the patient of Hypotrichosis Simplex (HS).

These examples raise the largely unanswered questions: (1) how mutations within these core components, that are expected to affect splicing of most introns, lead to organ or cell type specific phenotype; (2) which downstream target genes are responsible for the organ or cell type specific phenotypes.

1.5.3 Aberrant splicing contributes to cancer development

A lot of reports have shown that splicing patterns in cancer have been changed [122]. The aberrantly expressed transcripts contribute to tumor cell survival, proliferation, invasion and metastasis. Since the changes of the abundance,

localization or activity of splicing regulator (such as hnRNP and SR proteins) can affect the splicing pattern, more and more studies demonstrated the specific alterations in the expression of splicing factors in cancer. Among splicing factors implicating in caner, the SF2/ASF (also known as SRSF1) is the most extensively studied. The overexpression of SF2/ASF is found in different types of human tumors, which indicates that SF2/ASF can function as a proto-oncogene [123, 124]. SF2/ASF regulates the splicing of numerous transcripts that are related to cancer, including BIN1, MNK2, S6K1, BIM, BCL2L11 and RON [125, 126]. The SF2/ASF induces the inclusion of exon 12a of BIN1, an inhibitor of c-Myc. The resulted protein isoform loses the ability to interact with and suppresses c-Myc function. c-Myc can in turn transcriptionally increase the expression of SF2/ASF. Moreover, a novel SF2/ASF-induced isoform of S6K1 is sufficient to cause the transformation phenotype [127]. Besides, SF2/ASF also regulates the splicing of RON (a receptor tyrosine kinase). Through binding to the exon 12 of RON, SF2/ASF stimulates the skipping of exon 11 of RON to produce the constitutively active kinase (known as delta-RON) promoting the cell motility and invasion [128]. These studies indicate that SF2/ASF promotes several steps of cancer progression through coordinating changes in splicing.

1.6 Previous studies about Sm proteins

In 1990, Neiswanger *et al*. [129] physically and genetically mapped the *snrpE* gene to human chromosome 1q25-43. Later, Fautsch *et al*. [130] found that both coding region and transcriptional control sequence of *snrpE* gene are conserved between human and mouse. As the basal components of spliceosome, the Sm proteins are generally considered to only function during the control of global gene expression through the mRNA splicing, however, a growing body of evidences [131-134] suggests that Sm proteins have other functions distinct from mRNA processing. In 1999, Seto *et al*. [131] first reported that the telomerase of *Saccharomyces cerevisiae* (*S. cerevisiae*) is a Sm snRNP particle. The amount of

Sm proteins or the integrity of Sm binding site in telomerase RNA (TLC1) affects the activity of telomerase. In addition, Barbee *et al*. found that Sm proteins play the role in controlling the integrity and localization of germ granule (P granule) [132] and also regulate the germ cell specification during the early embryogenesis of *Caenorhabditis elegans* (*C. elegans*) [134]. Later, Bilinski *et al*. [133] showed that Sm proteins bind to mRNAs (such as Xcat2), which are the components of germinal granules, and facilitate the transport of these mRNAs from nucleus to the nuage in *Xenopus* oogenesis.

Recent research has reported that mutations within SmE are identified as the cause of some diseases [135-137], and the Sm proteins also play role in the cancer development [138, 139]. Pasternack *et al*. [137] reported the mutations in SmE are the cause of autosomal-dominant HS with the phenotype of hair loss. They identified two heterozygous mutations. One is located in the first amino acid leading to form the N-terminal truncated protein, whereas the second mutation (G45S) doesn't affect the integration capacity of SmE protein. Weiss *et al*. [140] identified a dominant heterozygous E51D (Aspartic acid to Glutamic acid) mutation in the *snrpE* gene from the mice with hypogonadal phenotype via ENU (N-ethyl-N-nitrosourea) mutagenesis.

In cancer studies, Anchi *et al*. [138] reported that SmE regulates the expression of androgen receptor (AR), which is essential for the cell proliferation and progression of high-grade prostate cancer (PC). Quidville *et al*. [139] reported that the Sm proteins are overexpressed in the lung, breast and ovarian cancers. The depletion of SmE or SmD1 in breast, lung and melanoma cancer cell lines leads to remarkable reduction of cell viability through induction of autophagy, whereas it has little effect on the survival of the nonmalignant cells. These studies suggest that spliceosome might be a novel target for cancer drug development.

1.7 The goal of this study

RNA splicing is an intricate process in humans and higher metazoans. The AS is widely used by higher eukaryotes to expand the transcriptome complexity and proteome diversity without increasing the number of genes. With the increase of complexity along the evolution, the susceptibility of splicing to mis-regulation also increases.

Although lots of human diseases have been reported to link to various mutations within splicing factors so far, only very few mutations in the basal components of spliceosome have been associated with human diseases due to the pivotal functions in shaping the whole transcriptome and lethal effect of their dysfunction.

This thesis aims to identify and functionally characterize a key *de novo* mutation in a basal component of spliceosome that is putatively associated with a human disease. In this study, the whole exome sequencing was first applied and identified a missense heterozygous mutation in the gene *snrpE* from a family with a child suffering from non-syndromal microcephaly and intellectual disability. The traditional biochemical experiment was performed to explore the effect of identified mutation on the function of SmE protein. Then, the genetic modification technology and RNA-seq were further applied to construct the *in vitro* cell model and establish the cellular link between the identified mutation and disease.

2 MATERIALS AND METHODS

2.1 Materials

2.1.1 Cell lines

(1) HEK293 Flp-In T-Rex cell line: Human embryonic kidney cell line that contains a single integrated FRT site and stably expresses high levels of Tet repressor protein (Thermofischer Scientific).

(2) Stable HEK293 Flp-In T-Rex cell line with inducible expression of HA tagged wild-type or mutated SmE protein: exogenous wild-type and mutant SmE protein were integrated into HEK293 Flp-In T-Rex cell line, respectively.

(3) Stable HEK293 Flp-In T-Rex cell line with inducible co-expression of HA-or 2A tagged wild-type or mutated SmE: wild-type and mutated SmE proteins were co-expressed with 2A or HA tag, respectively.

(4) Fibroblast cell lines: derived from normal controls and one patient carrying the specific point-mutation site (T>C) in *snrpE* gene. These cell lines are kindly provided by Kaindle Lab in Charite.

(5) HEK293 Flp-In T-Rex cell line with the introduction of the heterozygous point-mutation (T>C) into the endogenous *snrpE* gene: the heterozygous point-mutation was introduced by the combinatorial application of CRISPR/Cas9 and *piggyBac* transposon system.

2.1.2 Vectors

pENTR-D-TOPO vector (Thermofischer Scientific)

pOG44 vector (Part of the Flp-In Core System from Thermofischer Scientific, courtesy of Landthaler Lab at MDC)

pcDNA5/FRT/TO vector (Thermofischer Scientific)

pcDNA3.1/Hygro^{+} vector (Thermofischer Scientific)

pGEM-T easy vector (Promega)

2.1.3 Cell culture mediums

2.1.3.1 Maintainance medium for HEK293 Flp-In T-Rex cell line

High glucose DMEM (Gibco) supplemented with 10% final concentration of FBS (heat inactivated, Gibco), 100 U/ml of penicillin and 100 µg/ml streptomycin (Gibco), 15 µg/ml blasticidin (InvivoGen) and 100 µg/ml zeocin (InvivoGen).

2.1.3.2 Antibiotic-free medium for transfection

DMEM with 10% FBS.

2.1.3.3 Stable cell line selection medium

For maintaining the HEK293 cell line with stable overexpression of interested genes, the medium composition is DMEM with 10% final concentration of FBS, 100 U/ml of penicillin and 100 µg/ml streptomycin, 100 µg/ml hygromycin, 15 µg/ml blasticidin and 100 µg/ml zeocin.

For maintaining the point-mutation introduced HEK293 cell lines, the first round selection medium is composed by DMEM with 10% FBS, 100 U/ml of penicillin and 100 µg/ml streptomycin and 3 µg/ml puromycin whereas the second round selection medium contains DMEM with 10% FBS, 100 U/ml of penicillin, 100 µg/ml streptomycin and 0.5µM fialuridine (FIAU).

2.1.3.4 Freezing medium

90% FBS and 10% DMSO.

2.1.4 Enzymes

GoTaq DNA Polymerase (Promega)

Phusion High-Fidelity DNA polymerase (Finnzymes)

Proteinase K (lyophilizate) (Roche Diagnostics)

Restriction enzymes (NEB)

T4 DNA ligase (NEB)

SuperScript III reverse transcriptase (Thermofischer Scientific)

Gibson Assembly master mix (NEB)

T7 endonuclease I (NEB)

2.1.5 Kits

AgencourtAMPure XP beads (Beckman Coulter)

Agilent RNA 6000 Nano and Pico Kit (Agilent Technologies)

Agilent DNA 1000, 7500 Kit (Agilent Technologies)

TURBO DNA free Kit (Ambion)

Dynabeads Protein G (Thermofischer Scientific)

GeneJET Plasmid Miniprep Kit (Fermentas)

Lipofectamine 2000 reagent (Thermofischer Scientific)

Lipofectamine RNAiMAX reagent (Thermofischer Scientific)

pENTR/D-TOPO Cloning Kit (Thermofischer Scientific)

QIAquick PCR Purification and Gel Extraction Kit (Qiagen)

Qubit dsDNA HS Assay (Thermofischer Scientific)

Qubit RNA HS Assay (Thermofischer Scientific)

Immobilon Western Chemiluminescent HRP Substrate (Millipore)

mMESSAGE mMACHINE T7 ULTRA Transcription Kit (Ambion)

SYBR Green I Master (Roche)

Trueseq DNA Library Preparation Kit (Illumina)

Illumina Sequencing Kit (Illumina)

2.1.6 Chemicals

The chemicals used in this project were purchased from the following companies: Roth (Germany), Sigma (Germany), Roche (Germany), Invitrogen (Germany), Millipore (Germany) and QIAGEN (Germany).

2.1.7 Other reagents

Reagents not listed in the above list are specified in the experimental methods, including siRNA, antbodies, solutions, buffers and primer sequences etc.

2.1.8 Major Equipments

Agarose gel electrophoresis equipment (Bio-Rad)

Agilent 2100 Bioanalyzer (Agilent Technologies)

Bio-Rad Tetrad 2 Peltier Thermo Cycler (Bio-Rad)

Cell incubator (Hereaus Instruments)

Centrifuges: 5417R, 5804R (Eppendorf)

Illumina HiSeq 2000 (Illumina)

MilliQ Biocel water purification system (Millipore)

NanoDrop ND-1000 (NanoDrop Technologies)

PAGE electrophoresis equipment (Thermofischer Scientific)

pH Meter 537 (WTW)

Qubit Fluorometer (Thermofischer Scientific)

Roche LightCycler 480 Real-Time PCR System (Roche)

Thermomixer (Eppendorf)

Trans-Blot SD Semi-Dry Transfer Cell (Bio-Rad)

2.2 EXPERIMENTAL METHODS

2.2.1 Cloning of exogenous SmE into mammalian expression Vector

2.2.1.1 Total RNA extraction

Following the manufacturer's protocol, total RNA was extracted from control and patient fibroblast cells by TRIzol reagent (Thermofischer Scientific). The concentration of total RNA was quantified by NanoDrop (NanoDrop Technologies) and the quality was assessed by Agilent 2100 Bioanalyzer (Agilent Technologies).

2.2.1.2 Reverse transcription (RT) and PCR

The extracted total RNA was first treated with TURBO DNA free kit (Ambion) following the manufacturer's protocol. Reverse transcription was performed

using 500 ng of DNase treated total RNA, oligo $(dT)_{20}$ and Superscript III reverse transcriptase (Thermofischer Scientific), according to manufacture's instruction. After the reverse transcription, the RT product was used as template to amplify the full length coding sequence (CDS) of SmE by Phusion High-Fidelity DNA Polymerase (Finnzymes) in the following 50 µl reaction system: 32.5 µl H2O, 10µl 5 x Phusion HF buffer, 1 µl dNTPs (10 mM), 2 µl *snrpE*_cDNA forward primer (10 µM), 2 µl *snrpE*_cDNA reverse primer (10 µM), 0.5 µl Phusion Polymerase (2 U/µl) and 2 µl RT product. The PCR program is: 98 °C 30 sec, (98 °C 10 sec, 55°C 30 sec, 72 °C 30 sec) x 35 cycles, 72 °C 10 min, 4 °C hold. The specificity of PCR product was first checked on argarose gel and then purified by QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's protocol and quantified the concentration by Nanodrop.

Sequences of PCR primers:

*snrpE*_cDNA_Forward 5'-CACCATGGCGTACCGTGGCCAGGGTCAGA-3'
(CACC is the protruding sequence in pENTR-D-TOPO vector, which entails the directional insertion of PCR product into the vector)

*snrpE*_cDNA-Reverse 5'-CTAGTTGGAGACACTTTGTAGCAGA-3'

2.2.1.3 Cloning of *snrpE* cDNA into pENTR-D-TOPO vector

The purified full-length CDS of *snrpE* was ligated into pENTR-D-TOPO vector. The ligated product was transformed into One Shot TOP10 chemically competent *E. coli* cells (Thermofischer Scientific) following the manual protocol. Colonies were screened by colony PCR using GoTaq DNA polymerase (Promega) together with M13_For and *snrpE*_cDNA_Reverse primers. The colonies with right insertion were expanded in the LB medium supplemented with 50 µg/ml Kanamycin (Sigma-Aldrich) for plasmid preparation. The plasmid was extracted using GeneJET Plasmid Miniprep Kit (Fermentas) and quantified by Nanodrop. The composition of inserted fragment was verified by Sanger sequencing using M13_F primer (LGC Genomics, Germany).

Sequences of primers:

M13_For: 5'-GTAAAACGACGGCCAGTGAAT-3'

*snrpE*_cDNA-Reverse 5'-CTAGTTGGAGACACTTTGTAGCAGA-3'

2.2.1.4 Construction of HA-tagged wild-type or mutant SmE expression pcDNA5/FRT/TO vector

According to the Sanger sequencing result, we picked out the right colonies containing the wild-type or mutant SmE cDNA, and amplified the *snrpE* cDNA from the pENTR-D-TOPO vector by primers (*snrpE*_Single_Forward/Reverse) to add the HA-tag at the C-terminal and restriction enzyme sites. The fragments were first digested by restriction enzyme, and then ligated into the linearized pcDNA5/FRT/TO vectors to get the expression vector, pcDNA5/FRT/TO_SmE_HA (Figure 2.1).

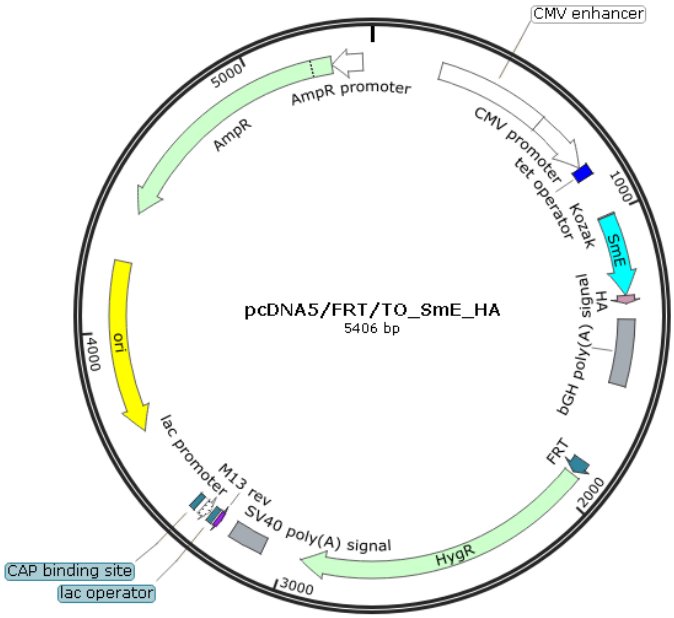

Figure 2.1: The vector map of pcDNA5/FRT/TO_SmE_HA. Vector properties are shown, including replication origins, antibiotic resistance and recombination sites.

2.2.1.5 Construction of wild-type and mutant SmE co-expression pcDNA5/FRT/TO vector

To construct the vector co-expressing the wild-type and mutant SmE proteins, the self-cleaving 2A peptide was applied to separate the wild-type from mutant. The first protein is tagged by 2A whereas the second one is tagged by HA. The 2A sequence is artificially synthesized (MWG, Germany). The construction of SmE (wild-type)-2A-SmE (mutant)-HA co-expression vector is shown as an example here (Figure 2.2). First, the wild-type and mutant SmE CDS was amplified from the corresponding pENTR-D-TOPO vector by primers Co-expression-Up-Forward/Reverse and Co-expression-Down-Forward/Reverse, respectively. For the wild-type part, it contains the restriction enzyme site at N-terminal and partial of 2A sequence at C-terminal whereas the mutant part has partial 2A sequence at N-terminal and enzyme site at C-terminal. Next, the amplified wild-type and mutant SmE CDS fragment and artificially synthesized 2A sequence were fused together by overlap PCR with primers Co-expression-Up-Forward/Co-expression-Down-Reverse. Finally, the fused fragment was digested and ligated into the linearized pcDNA5/FRT/TO vectors. The ligated product was transformed into One Shot TOP10 chemically competent *E. coli* cells (Thermofischer Scientific) following the manufacturer's protocol. The colonies were checked by PCR and Sanger sequencing. For SmE (mutant)-2A-SmE (wild-type)-HA co-expression vector construction, all of the steps are the same, excepting the first step.

Sequences of primers:

Co-expression-Up-Forward:

5'CGCGGATCCGCCACCATGGCGTACCGTGGCCAGG3'

Co-expression-Up-Reverse:

5'ACTTCCTCTGCCCTCGTTGGAGACACTTTGTAGCAGAGTA3'

Co-expression-Down-Forward:

5’GAGAATCCCGGCCCTATGGCGTACCGTGGCCAGGGTCAGA3’

Co-expression-Down-Reverse:

5’ATTTGCGGCCGCCTAAGCGTAATCTGGAACATCGTATGGGTAGTTGG AGACACTTTGTAGCA3’

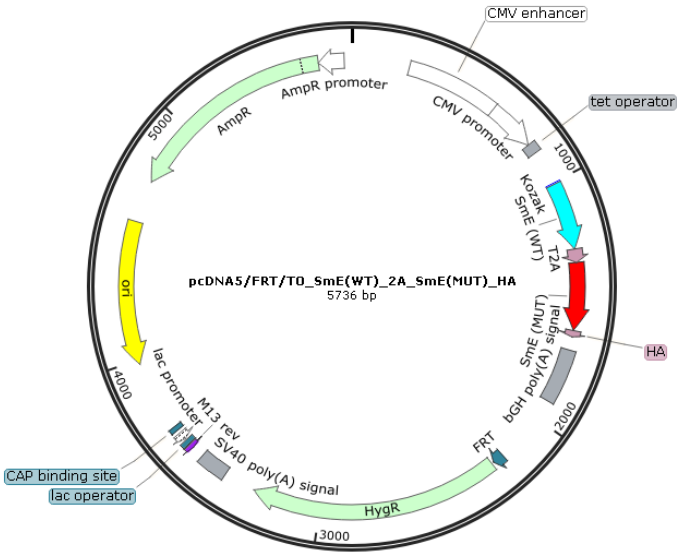

Figure 2.2: The vector map of pcDNA5/FRT/TO_SmE (WT)_2A_SmE (MUT)_HA. Vector properties are shown, including replication origins, antibiotic resistance and recombination sites.

2.2.2 Construction of single guide RNA (sgRNA) expression vector

To screen a suitable position for Cas9, we applied the E-CRISP developed by Heigwer *et al.* [141] to scan the possible positions of sgRNAs within the *snrpE*. After scanning, we selected two sgRNAs located in the 3’ end of first intron and third exon, respectively. These two sgRNAs are named as sgRNA-8-0 and sgRNA-0-1278, respectively. The fragment was artificially synthesized. We first anneal the two single strands together and then perform the Gibson Assembly reaction to fuse each fragment into the linearized vector (Addgene 41824). The reaction product was transformed into *E. coli* competent cells and spread onto

agar plate with antibiotics. The colonies were checked by colony PCR and Sanger sequencing. The sequence of synthesized fragments is listed below:

sgRNA8-0-Sense:

5’TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTCTGC ACCACAGTGGCTAGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGG CTAGTC 3’

sgRNA8-0-anti-Sense:

5’GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACTAGCCACT GTGGTGCAGAACGGTGTTTCGTCCTTTCCACAAGATATATAAAGCCAA GAAA 3’

sgRNA0-1278-Sense:

5’TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATGATT ATTTCAGAGATCGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGG CTAGTC 3’

sgRNA0-1278-anti-Sense:

5’GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACCGATCTCT GAAATAATCATCGGTGTTTCGTCCTTTCCACAAGATATATAAAGCCAA GAAA 3’

2.2.3 Construction of homologous recombination donor vector

To introduce the point mutation into the genome via homologous recombination, the donor vector containing the homologous arm and selection cassette is needed. Using the human genomic DNA as template, the 5’ and 3’ homologous arm was amplified by primer sets 5’HR_F/R and 3’HR_F/R, respectively. The 5’and 3’ ITR regions of *piggyBac* were amplified by primer sets 5’ITR_F/R and 3’ITR_F/R from the vector pUC19XLneo (a gift from Izsvak Lab). Using the vector pT2RMCE-OSKM-puDTK (a gift from Izsvak Lab) as the template, the pGK promoter, selection cassette (containing pGK promoter, positive/negative selection marker and polyA signal) and polyA signal region were amplified by

pGK_F/R, pGK_F/polA_R and polA_F/R, respectively. Next, the overlap PCR was performed with these purified fragments. The first round of two independent overlap PCR was performed to get two fragments (5'HR-5'ITR-pGK and polA-3'ITR-3'HR). Then the second round overlap PCR put the 5'HR-5'ITR-pGK and selection cassette together to form the long fragment (5'HR-5'ITR-pGK-puDTK-polA). After purification, the long fragment was inserted into the linearized pcDNA 3.1 Hygro^{+} vector (Thermofischer Scientific) by Gibson Assembly reaction (NEB). The reaction product was transformed into the *E. coli* competent cells. The plasmid was extracted from correct colony and linearized by BbsI enzyme. Then, the fragment polA-3'ITR-3'HR is inserted into the linearized vector to get the final plasmid, pcDNA 3.1 Hygro^{+} 5HR-5ITR-puDTK-3ITR-3HR (Figure 2.3).

5'HR_F:

5'AGCTCGGATCCACTAGTCCAGTGTGGTGGAATTCTGCAGATATCCAG CACAGTGGCGGCCCGGCACTCCTCCTACCACGGCCACTAT 3'

5'HR_R:

5'- GATTATCTTTCTAGGGTTAACTCCCGTTTTCACTCCGCA-3'

5'ITR_F:

5'- GAAAACGGGAGTTAACCCTAGAAAGATAATCATATTGTG-3'

5'ITR_R:

5'- GCTGAAGTTCCTATACCCCCCGGGCTGCAGGAATTCGATAA-3'

pGK_F:

5'- TCCTGCAGCCCGGGGGGTATAGGAACTTCAGCTTGATATCG-3'

pGK_R:

5'- GCGCTTTTGAAGCGTGCAGAATGCC-3'

polA_F:

5'- CCGTCCCATGCACGTCTTTATCCTGG-3'

polA_R:

5’- TCGATACCGTCGACCTGCCCATAGAGCCCACCGCATCCCCA-3’

3’ITR_F:

5’- GGTGGGCTCTATGGGCAGGTCGACGGTATCGATAAGCTTGAT-3’

3’ITR_R:

5’- GTTTCCCAAAACCGTTAACCCTAGAAAGATAGTCTGCGTA-3’

3’HR_F:

5’- CTTTCTAGGGTTAACGGTTTTGGGAAACTCCAGGGGGAT-3’

3’HR_R:

5’GGGGATACCCCCTAGAGCCCCAGCTGGTTCTTTCCGCCTCAGAAGCC ATAGAGCCCACCGCATCCCCAGCATGCCTGCTATTTCATCCTAAAGGG GAGAGGTTCC-3’

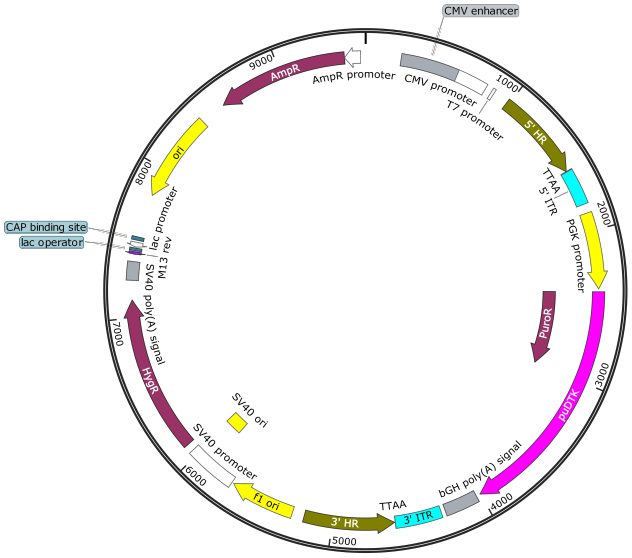

Figure 2.3: The vector map of pcDNA 3.1 Hygro^{+} 5HR-5ITR-puDTK-3ITR-3HR. Vector properties are shown, including homologous recombination arm (5’ and 3’ HR), selection cassette (puDTK) and *piggyBac* transposase recognition specific region (5’ and 3’ ITR).

2.2.4 Establishing exogenous SmE stable expression HEK293 cell lines

0.9 µg pOG44 and 0.1 µg pcDNA5/FRT/TO_SmE-HA or pcDNA5/FRT/TO_SmE (WT)_2A_SmE (MUT)_HA were mixed and transfected

into HEK293 Flp-In T-REx cells using Lipofectamine 2000 (Thermofischer Scientific) according to the manufacturer's recommendations. A mixture of pOG44 and pmaxGFP was transfected as control. Transfection efficiency was assessed by GFP expression. After 48 hours, the cells were digested and seeded into a 10-cm cell culture plate, and incubated with antibiotic-free medium for 12 hours until they were well attached. Then, selection medium with appropriate antibiotics was added and changed every 2-3 days until the distinct cell colony can be identified. Several single colonies were carefully picked out and expanded sequentially with selection medium. During the cell expansion process, a small fraction of cells from each picked colony were seeded into 6-well plate. A serial of doxycycline (Doxy) concentrations (0, 5, 10, 100, 500, 1000 ng/ml) were tested to obtain optimal protein expression levels for future experiments. Established stable cell lines were used for future experiments, or kept in freezing medium and frozen in liquid nitrogen. All the cell culture was performed in a 37 °C incubator with 5 % CO_2.

2.2.5 T7 endonuclease assay

The T7 endonuclease assay was applied to evaluate the gene targeting efficiency of sgRNA. The HEK293 cells were first seeded into a 6-well plate and then transfected with different ratio of sgRNA and Cas9 expression vector (ratio of sgRNA/Cas9: 0.5:1, 1:1, 2:1) by lipofectamine 2000. The amount of Cas9 expression vector was kept constant (500 ng). The group transfected with only transfection reagent was used as control. After 48 hours, the cells were collected and extracted the genomic DNA. Using the genomic DNA as template, the PCR reaction was performed to amplify the corresponding region with specific primer sets. The PCR product was purified and measured the concentration by Nanodrop. 300 ng of purified PCR product was mixed with 2 µl 10× NEBuffer 2 and water was added to 19 µl in total. The mixture was putted into the thermocycler to perform the denaturation and annealing as following conditions

(Table 1). Then, 1 µl of T7 endonuclease I was directly added into the tube and incubated at 37 ^{o}C for 1 h. After the incubation, the mixture was checked on 1% agarose gel and the cleavage efficiency was calculated based on the formula below:

$$\% \text{ gene modification} = 100 \times (1 - (1 - \text{fraction cleaved})^{1/2})$$

The primer sets used for amplification of fragment are listed below:

sg8-0-T7-F: 5' TTCCACAGCGCAGACGCCTCAGTTT 3'

Sg8-0-T7-R: 5' GTACTCCTTCCCCTCAATTACCCTCC 3'

sg1278-T7-F: 5' TTTGGCATCCCAAAGTATGGGAACG 3'

Sg1278-T7-R: 5' CCTCTCCATACTAACATTGTCGCTC 3'

Table 1. The conditions for DNA denaturation and annealing

Step	Temperature	Ramp Rate	Time
Initial Denaturation	95 ^{o}C		5 min
Annealing	95-85 ^{o}C	-2 ^{o}C /s	
	85-25 ^{o}C	-0.1 ^{o}C /s	
Hold	4 ^{o}C		Hold Forever

2.2.6 Establishing HEK293 cell line with point-mutation in *snrpE*

To precisely introduce the point mutation into genome, the combination of CRISPR/Cas9 and *piggyBac* transposon system is applied just the same as Xie *et al*. [142] described. The procedure of establishing this cell line was classified into two steps.

The first step was the insertion of template containing selection cassette, into genome via homologous recombination. A combination of 500 ng sgRNA-8-0 vector, 500 ng Cas9 vector and 3 µg pcDNA3.1 Hygro^{+} 5HR-5ITR-puDTK-

3ITR-3HR vector was co-transfected into a 6-well plate by lipofectamine 2000 whereas the combination of 500ng sgRNA-8-0 vector and 500ng Cas9 vector, was used as control. After 48 h, the cells were transfer into 10 cm plate and incubated with antibiotics free medium for 12 h until they were well attached. Then, selective medium containing 3 μg/ml puromycin was added and changed every 2-3 days until distinct cell colonies could be identified. Single colonies were carefully picked out and further expanded sequentially with selective medium. The successful homologous recombination was checked by genomic PCR and Sanger sequencing.

For the second step, the selection cassette was removed by *piggyBac* transposase. The cell colony with successful insertion was expanded and seeded into 6-well plate. 3 μg *piggyBac* transposase expression vector (a gift from Izsvak Lab) was transfected by lipofectamine 2000. After 48 h, the cells were seeded into 10 cm plate and incubated with antibiotics free medium for 12 h until they were well attached. Then, negative selection medium containing 0.5 μM FIAU was added and changed every 2-3 days until the distinct cell colony could be identified. Single colonies were carefully picked out and further sequentially expanded. The successful remove of selection cassette was check by genomic PCR, and the mutation introduction was checked by Sanger sequencing.

2.2.7 Immunoprecipitation (IP)

Cell pellet was lysed in 3 volumes of lysis buffer (50 mM HEPES-NaOH pH 7.5, 150 mM NaCl, 1mM EDTA, 1% NP40, complete protease inhibitor (Roche)) and incubated on ice for 15 min. To fully lyse cells, the cell suspension was further go through the syringe with 26G needles. The cell lysate was centrifuged at 18,000 g for 15 min at 4^{o}C to get rid of cell debris and the supernatant was further filtered through a 5 μm membrane syringe filter (Pall). The protein complex was immunoprecipitated from filtered cell lysate using magnetic Protein G Dynabeads (Thermofischer Scientific) conjugated with corresponding antibody (such as: anti-

SmY12, anti-HA and anti-SMN), and incubated for 4 hour at 4^{o}C on the rotation wheel. Then, beads were washed three times with wash buffer (50 mM HEPES-NaOH pH 7.5, 300 mM NaCl, 0.05% NP-40). After washing, the beads were washed one more time with PBS and re-suspended in 50 µl PBS with 10 µl 6 x SDS-PAGE loading buffer. The beads were further denatured at 95^{o}C for 10min and the supernatant was transferred into new tube for gel running or stored at -20^{o}C.

2.2.8 Western blot (WB)

Cells were lysed in RIPA lysis buffer (50 mM Tris pH 8.0, 150 mM NaCl, 1.0% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, complete protease inhibitor (Roche)) and sonicated by using Biorupter (UCD-300, Diagenode). The cell lysate was centrifuged at 18,000 g for 10 min at 4°C. The supernatant in 1 x SDS-PAGE loading buffer was denatured at 95°C for 10 minutes and resolved by 10-15% SDS- PAGE in 1 x SDS running buffer (25 mM Tris base, 190 mM glycine, 0.1% SDS) at 200 V for ~1 hour until the dye ran to the bottom of the gel chamber. Proteins were transferred to PVDF membrane (Millipore) using semi-dry blotting apparatus (BioRad) at 20 V for ~1 hour. The membrane was blocked in 5% non-fat milk, washed three times with TBST (50 mM Tris.HCl, pH 7.4, 150 mM NaCl, 0.1% Tween 20) and incubated with primary antibody for 2 h at room temperature or overnight at 4°C with gentle shaking. After primary antibody incubation, the membrane was washed three times with 1 x TBST and incubated with HRP-conjugated secondary antibody at room temperature for 1 hour. The secondary antibody incubated membrane was washed three times with TBST, developed with SuperSignal Kit (Thermo) and detected by LAS-4000 imaging system (Fujifilm) following the manufacturer's instructions. The following antibodies were used: rabbit polyclonal anti-*snrpE* (Millipore, ABN313, 1:1000), rabbit polyclonal anti-2A (Millipore, ABS31, 1:1000), mouse monoclonal anti-HA (Covance, MMS-101P, 1:1000), mouse monoclonal anti-

Smith Antigen antibody (anti-SmY12, Abcam, ab3138, 1:1000), and secondary HRP-conjugated goat anti-mouse or Rabbit IgG (Santa Cruz, 1:2000). The SMN (7B10) and pICln antibody were the gift from Fischer Lab.

2.2.9 Immunofluorescence

For immunostaining, HEK293 cells were seeded onto the coverslips and transfected with pcDNA5/FRT/TO_SmE (WT)-HA or pcDNA5/FRT/TO_SmE (MUT)-HA at 70% confluency by lipofectamine 2000. After 48 h, cells were washed and fixed with 4% paraformaldehyde for 10 min. The cells were washed for three times by PBS and permeabilized for 20 min using 0.2% Triton X-100 in PBS. Then, the cells were washed again for three times by PBS and blocked with 5% BSA in PBS for 1h. The blocked cells were transferred to humid chamber, incubated with 200 µl primary antibody freshly diluted in antibody dilution solution (0.5% BSA in PBS) for 2 h at room temperature or 4 °C overnight. Subsequently, the cells were washed 3 times in PBS. Then, the cells were incubated with 200 µl fluorochrome-conjugated secondary antibody freshly diluted in antibody dilution solution for 40 min at room temperature. After washing away the excess secondary antibody solution, the cells were stained with 1:5000 DAPI (Thermofischer Scientific) diluted in PBS for 5 min at room temperature, then the cells were washed with PBS for 3 times. Finally, cells were mounted on the microscopic glass slide with Prolong Gold Antifade Reagent (Thermofischer Scientific) and observed under Leica DM IRB microscope (Leica microsystems). The images were processed by Fuji software. The following antibodies were used: rabbit polyclonal anti-Coilin (Santa Cruz Biotechology, sc-3260, 1:200), mouse monoclonal anti-HA (Covance, MMS-101P, 1:200).

2.3 COMPUTATIONAL METHODS

2.3.1 Exome sequencing and bioinformatics

All of family members were subjected to exome sequencing. The DNA was extracted from the blood samples. According to the manufacture's protocol, the genomic DNA was enriched by Agilent Human All Exon V4 Kit (Agilent Technologies, Santa Clara, CA, USA). The whole exome libraries were subjected to Illumina HiSeq2000 system for 100 cycles single end sequencing. After sequencing, the data analysis for exome sequencing was performed as described before by Fröhler *et al*. [143].

2.3.2 Analysis of splicing by RNA-seq

Total RNA from each cell lines and conditions was prepared by TRIzol reagent (Thermofischer Scientific) according to the manual instruction. mRNA libraries were prepared and subjected to 100 cycles of single end sequencing on Illumina Hiseq 2000 system. Before starting the mapping process, a database containing exon-exon junction was first constructed based on the gene annotation (gencode V17). If the length of exon is less than 6 nt, it will be discarded. Based on the annotation, the alternative splicing events were classified into the following categories: exon skipping, alternative 5' or 3' splice sites, and intron retention. The mRNA-seq reads were first mapped to the human reference genome (gencode GRCh37.p12) using bowtie2 with default parameters, meanwhile they were aligned to the constructed junction database with the requirement that the reads should overlap with either side exon for at least 6 nt. For the read could be aligned to both genome reference and junction reference, it will be assigned to genome or junction according to the alignment score. If the alignment score is equal, it will be defined as aligning to genome. The method described by McManus *et al*. [144] was applied to estimate the PSI (percentage of splicing in) for each exon and PIR (percentage of intron retention) for each intron.

In brief, for exon skipping, PSI value was calculated using the junction reads as follows: avg(C1:A, A:C2)/[C1C2+avg(C1:A+A:C2)]. The A represents the alternative exon whereas the C1 and C2 are the flanking exons. The C1:A and A:C2 are the number of junction reads supporting the exon inclusion whereas the C1C2 represents the number of junction reads for exon skipping. If the length of alternative exon is less than 100 nt, the C1:A:C2 reads can also be identified. In such situation, the C1:A:C2 will be included to calculate the PSI.

For 5' alternative splice site, the PSI value was calculated as follows: D1:A/(D1:A+D2:A). The D1:A and D2:A mean the number of junction reads supporting Donor1-Acceptor and Donor2-Acceptor, respectively. For 3' alternative splice site, the PSI was: D:A1/(D:A1+D:A2). The D:A1 and D:A2 represent the number of junction reads supporting Donor-Acceptor1 and Donor-Acceptor2, respectively.

For intron retention, PIR was calculated as follows: avg(C1:I+I:C2))/[C1C2+avg(C1:I+I:C2)]. The I represents the alternative exon whereas the C1 and C2 are the flanking exons. The C1:I and I:C2 are the number of exon-intron junction reads supporting the intron retention whereas the C1C2 represents the number of exon-exon junction reads for intron exclusion.

After calculating the PSI or PIR for each type of splicing events, each splicing event is compared between different conditions in the same cell line, such as the patient fibroblast versus healthy control, SmE siRNA mediated SmE knock down HEK 293 versus mock siRNA treated HEK293, mutation introduced HEK293 versus original host HEK293 cell. We applied the rank product test as described by Wang *et al*. [145] to detect the splicing events showing significant difference between two conditions (FDR < 0.01, delta PSI or PIR > 0.2).

2.3.3 Gene expression quantification

The reads mapped to the exon or exon-exon junction were used to quantify gene expressions. Simply for each gene, the reads per kilobase of exon per million mapped reads (RPKM) [146] value was calculated using uniquely mappable reads.

3 RESULTS

3.1 Identification of a *de novo* mutation from the patient with non-syndromal microcephaly and intellectual disability by whole exome sequencing

Given that the family pedigree (Figure 3.1A) in our study is too small, the linkage analysis is impossible to perform. To circumvent this problem, we took advantage of the whole exome sequencing, as a powerful and cost effective tool, to identify the potential causative mutations in genes associate with such clinical phenotype. The exome sequencing result from the parents and the affected child was listed in Table 2. For each family member, at least 180 M single end 101 bp reads were obtained. After filtering out those redundant reads, 42-48 M unique reads were mapped to the targeted protein coding regions, which covers 91% of exon region of genome. The average base coverage for coding region was at least 84 times.

After excluding those known and unlikely variants, we identified a *de novo* heterozygous missense point mutation (T>C) in the gene of *snrpE*. The missense point mutation was located in the 2nd exon of *snrpE* gene. The identified heterozygous missense mutation was further validated by traditional Sanger sequencing (Figure 3.1B). The gene *snrpE* encodes the protein of SmE, the basal component of spliceosome complex. It has been well studied that the SmE and other six Sm proteins co-operate and form the heptameric ring around the Sm site of snRNAs to form the snRNP complex as the basal building blocks of splicing machinery [7].

The conservation of SmE protein indicates that this protein is deeply conserved between different species, even down to yeast (Figure 3.1C). More importantly, we found that the mutation site is also deeply conserved. It is known that the SmE protein can interact with Gemin2 protein, a component of SMN complex that plays important role in the assembly of snRNPs. Based on the currently available

structure data [34], we predicted that the mutated site in the SmE was the exact position that interacts with Gemin2: Phe22 of SmE could interact with Pro49 and Tyr52 of Gemin2 (Figure 3.1D). The missense mutation has changed the polarity of amino acid from hydrophobic to hydrophilic amino acid. Therefore, the protein-protein interaction between SmE and Gemin2 would likely be disrupted by this missense mutation. Therefore, we hypothesized that the integration of mutant SmE into the snRNP complex might be affected, considering that the SMN complex plays an important role in the biogenesis of snRNPs.

Table 2: The exome sequencing result

	Total reads	Non-redundant reads	Reads in exon	Mean coverage	Exome coverage %
Child	186.74 M	72.28 M	43.99 M	87.89	91.97
Father	200.19 M	77.66 M	47.53 M	94.81	92.18
Mother	179.61 M	66.58 M	41.95 M	84	91.25

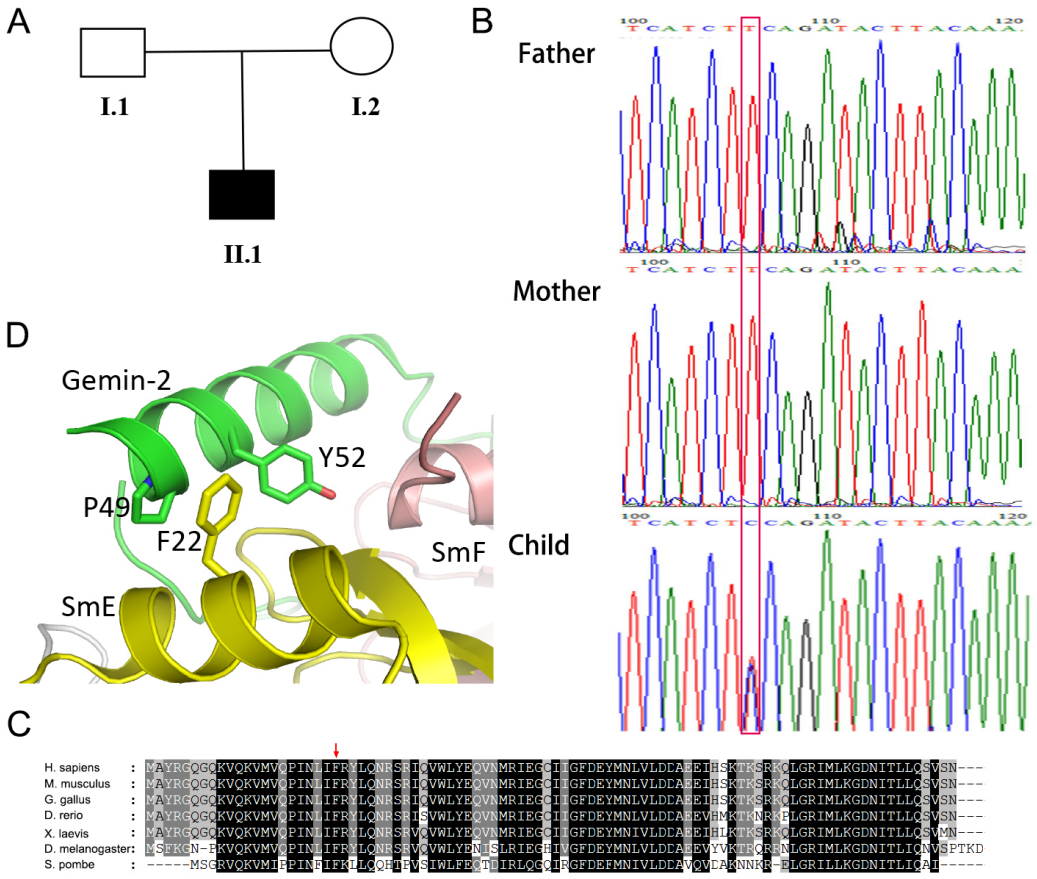

Figure 3.1: Identification of potential causative mutation by whole exome sequencing. A, Family pedigree. Filled symbol indicates individual suffering from non-syndromal microcephaly and intellectual disability. B, The traditional Sanger sequencing validated the identified mutation. The red box labels the *de novo* heterozygous mutation identified in the affected child. C, The conservation of SmE proteins in different species. The SmE protein is conserved in different species and the mutated site is also deeply conserved, even down to yeast. The red arrow indicates the site of mutated residue. D, The predicted protein-protein interaction between Gemin2 and SmE protein. The identified missense mutation in SmE might disrupt the protein-protein interaction between SmE and Gemin2 by changing the polarity of amino acid. Gemin2 and SmE are colored in green and lemon, respectively.

3.2 The integration capacity and subcellular distribution of mutant SmE protein is affected

To investigate whether the integration capacity of mutated SmE will be affected as we hypothesized, the SmY12-IP experiment was carried out to pull down the snRNP complex (see Methods). Since there are no any available antibodies that can be used to distinguish the wild-type from mutant proteins, we first constructed a HEK293 cell line that co-express the wild-type and mutant SmE protein with different tags (2A and HA tag), respectively. The co-expression of differentially tagged mutant and wild-type SmE was then induced by adding a series concentration of Doxy (0, 5, 10, 100, 500 and 1000 ng/ml). The expression level of the exogenous SmE was checked by WB (Figure 3.2A and B). As shown in Figure 3.2A and B, the exogenous SmE protein was successfully induced to express by addition of Doxy whereas no exogenous SmE was detected without adding the Doxy. The induction with 100 ng/ml of Doxy was considered as optimal, and was used for further experiments.

To compare the integration capacity of wild-type and mutant SmE protein, we performed the anti-SmY12 IP experiment (Figure 3.2C and D) to pull down the snRNP complex containing the SmB/B' protein (SmY12 antibody recognizes the epitopes located in the proteins SmB/B', SmD1 and SmD3). After IP, the IP product was checked by WB with corresponding antibodies (anti-SmY12, anti-Vinculin, anti-2A and anti-HA). To avoid the effect of imbalance expression of wild-type and mutant SmE proteins due to their position cloned in the exogenous constructs, we used the abundance of each protein in the cell lysates for normalization. As a result, in the cell line expressing the wild-type SmE protein with HA tag and mutant SmE protein with 2A tag, the amount of wild-type SmE protein in the IP product was almost three times higher than that of mutant SmE protein after normalization (Figure 3.2C).

To further avoid the potential artifact due to the cloned position and the different tag used to label the wild-type and mutant protein, we switched the position and tag of the wild-type and mutant SmE protein in the construct, and then repeated the same IP experiment in the cell line expressing the wild-type SmE protein with 2A tag and mutant SmE with HA tag. As shown in Figure 3.2D, we observed the same pattern: compared to wild-type SmE, the mutant SmE protein showed decreased integration capacity. Together, these results indicated that the integration capacity of mutant SmE protein is dramatically decreased comparing to the wild-type one.

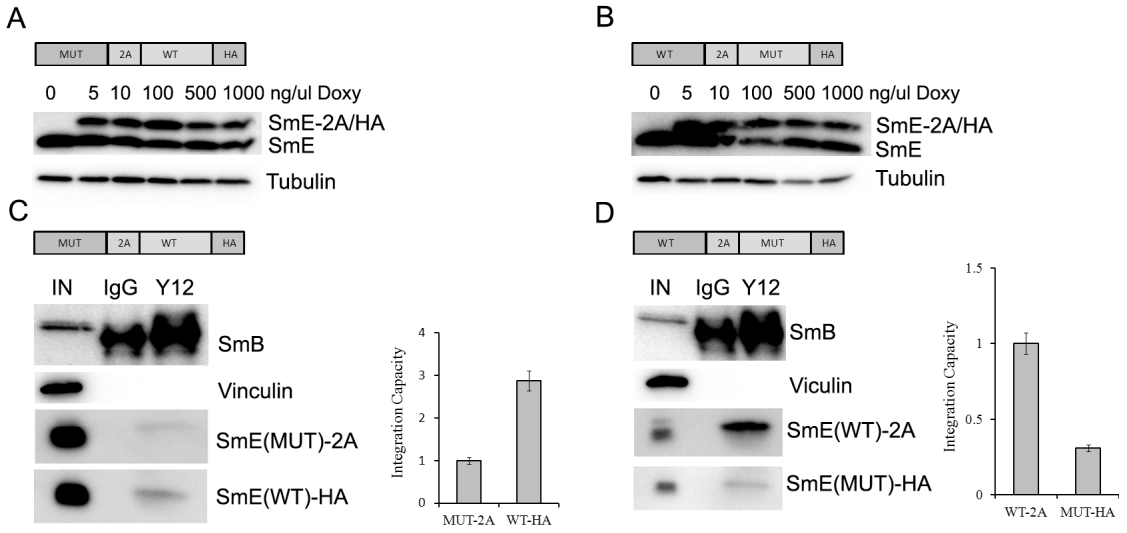

Figure 3.2: The integration capacity of SmE protein was affected by this point mutation. A, B, The induction concentration of Doxy was optimized in the two co-expression HEK293 cell lines. A series concentration of Doxy was applied and the abundance of exogenous SmE protein was detected by WB. The β-tubulin was used as the loading control. C, D, The integration capacity of wild-type and mutant SmE protein was evaluated by the SmY12 IP experiment. The SmY12 IP products were blotted with different antibodies to check the integration capacity (left panel). The integration capacity for both types of SmE protein was quantified (right panel). IN, input; IgG, mock IP; Y12, SmY12 IP.

The assembly process of snRNPs is majorly classified into two steps: cytoplasmic assembly and nucleus maturation [24], and the Sm proteins are mainly distributed in the nucleus. Based on the SmY12 IP results shown above, we postulated that the subcellular distribution of SmE might be affected by this mutation. To prove this hypothesis, we performed the immunostaining experiment to check the cellular distribution of wild-type and mutant SmE proteins. For this purpose, we transiently transfected the same amount of single overexpression vector into the HEK293 cells to over-express the wild-type or mutant SmE protein with HA-tag at the C-terminal. According to a previous study [42], the immature snRNPs are considered to concentrate in the CBs to finish the maturation after re-import into the nucleus. As shown in Figure 3.3, the wild-type SmE protein was mainly located in the nucleus that is consistent with the previous study, whereas the mutant SmE protein showed totally different pattern. It was mainly distributed in the cytoplasm of HEK293 cells. Therefore, it seemed likely that the cellular distribution of mutant SmE has been affected. Taken together, the results indicated that the integration capacity of mutant SmE protein is reduced and the subcellular distribution of mutant SmE protein is also shifted from nucleus to cytoplasm.

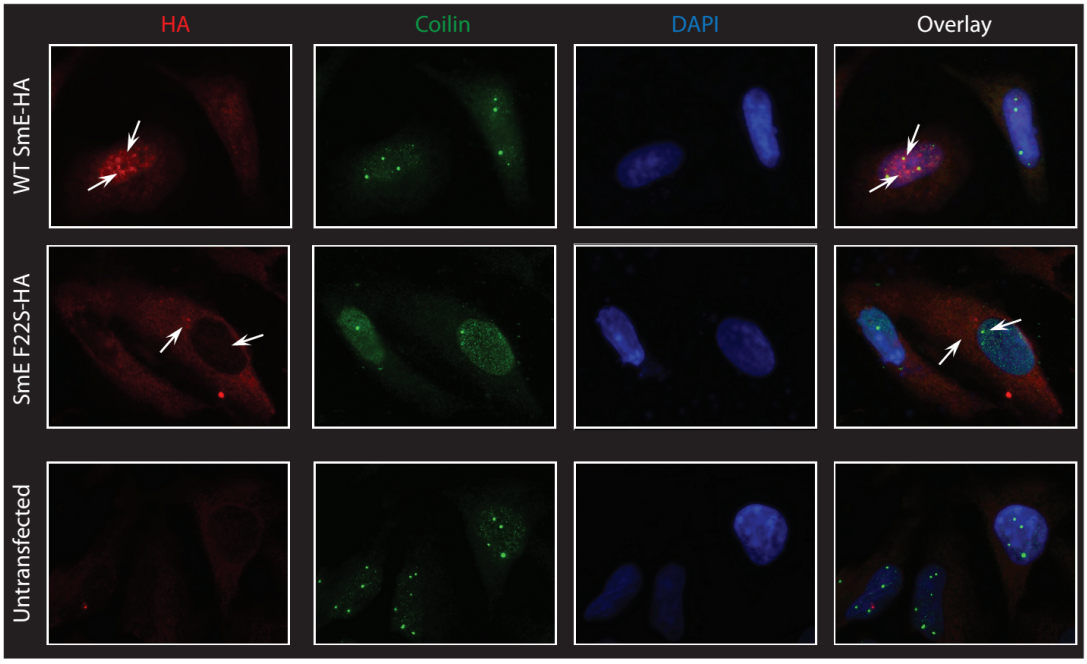

Figure 3.3: The subcellular distribution of SmE protein is affected by this mutation. The immunostaining is applied to check the subcellular distribution of exogenous SmE protein. The cells are stained by HA and coilin antibody to check the subcellular distribution of HA-tagged SmE protein and CBs, respectively. The nucleus of cell is stained by DAPI.

3.3 The early assembly phase of snRNPs is affected by this mutation

Given the assembly of snRNPs includes cytoplasmic assembly [29] and nuclear maturation [24], we speculated that the abnormal cytoplasmic distribution of mutant SmE protein might be caused by abnormal cytoplasmic assembly. To test this hypothesis, we established a stable HEK293 cell line to overexpress the exogenous wild-type or mutant SmE proteins with HA tag at the C-terminal. We first induced the cell lines with a series concentration of Doxy (0, 5, 10, 100, 500 and 1000 ng/μl). As shown in Figure 3.4A and B, the induction with 100 ng/ml of Doxy was considered as optimal, and was used for further experiments.

We then performed the HA-tag IP to pull down the complexes containing the

exogenous SmE protein and then the WB was carried out to assess the abundance of the IPed proteins using corresponding antibodies (such as: HA, SMN, pICln and Vinculin antibodies). As shown in Figure 3.4C, the amount of SMN protein in the IP product of mutant SmE group was much less than that of wild-type SmE group whereas almost no difference can be found for the amount of pICln protein between the wild-type and mutant group.

Based on the WB results of the HA-tag IP, we speculated that the association between subcomplex containing SmE protein and SMN complex was affected by this mutation whereas the formation of subcomplex containing SmE and pICln protein was not influenced.

To further support this, we further performed the SMN-IP to pull down the complex containing the SMN protein and then WB to detect the corresponding proteins. The blot for HA-tag antibody detecting the exogenous SmE protein indicated that the amount of HA-tagged wild-type SmE protein in IP product is much higher than that of mutant (Figure 3.4D). However, we couldn't detect any band for pICln protein in the IP product (Figure 3.4D). This was consistent with previous study that the pICln protein is dissociated from the SMN complex once the Sm proteins are loaded onto the SMN complex. All of these IP results indicate that the interaction between Sm proteins and SMN complex is impacted by this mutation whereas the formation of subcomplex between Sm proteins and pICln protein is not. Therefore, the early assembly phase of snRNPs is partially disrupted by this mutation.

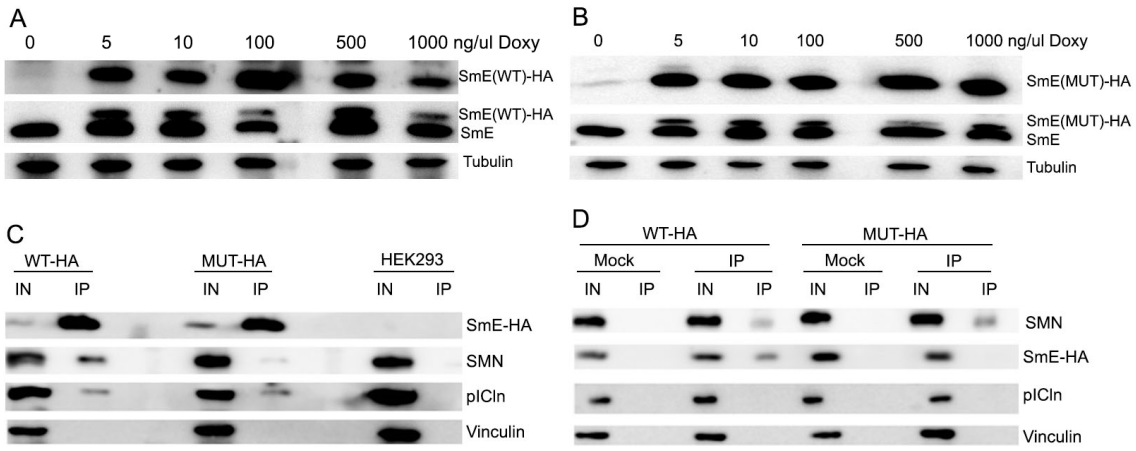

Figure 3.4: The early phase of cytoplasmic assembly is affected by the point mutation. A, B, Optimization of the induction concentration of Doxy. A series concentration of Doxy was used to induce the expression of exogenous SmE protein. The expression level of exogenous SmE protein was detected by WB. The β-tubulin was used as the loading control. C, D, The early cytoplasmic assembly phase was impacted by the identified point mutation. The HA-tag (C) and SMN (D) IP are performed in the HEK293 cell line stably overexpressing tagged SmE and the IP products are blotted for corresponding antibodies to check the deficiency of cytoplasmic assembly. The Vinculin is used as the loading control.

3.4 The general mRNA splicing in the patient is affected

Given that the early cytoplasmic assembly of snRNPs is affected by this mutation and the snRNPs are the major building blocks of spliceosome, we assume that the general mRNA splicing in the patient might be affected by this mutation. To test this, we extracted the RNA from the fibroblast cells derived from the patient and healthy people for transcriptome wide mRNA splicing analysis using RNA-seq. After mapping the reads back to the genome and transcriptome, we classified the reads into the exonic and intronic region. We found that, more reads were located in the intronic region from the patient fibroblast cells, and the percentage of reads located in the intron region was almost 1.7 fold than that of control (Table 3). This result indicated that the general mRNA splicing in the patient is affected.

Table 3, Reads distribution for each experimental sample

Cell type	Group	Exon %	Intron %
Fibroblast	Control	87.8	12.2
	Patient	79	21
HEK293	Control	85.2	14.8
	siRNA KD	80.6	19.4
	point-mutation	85	15

Since there are several types of alternative splicing events, we further analyzed the splicing pattern in the patient and control cells to check which type of splicing events was mostly influenced. Our bioinformatics analysis showed that the IR was the major affected type of splicing events. In the patient, 451 introns are significantly more retained (Figure 3.5A). The 451 retained introns derived from 357 genes. Since the IR might affect the host gene's expression level through introducing the PTC, we further checked the expression level of genes with IR and found that their expression was significantly decreased (Figure 3.5B).

In summary, these results (Figure 3.5) indicate that the general mRNA splicing in the patient is abnormal, and the IR as the major type of splicing events leads to the decreased gene expression.

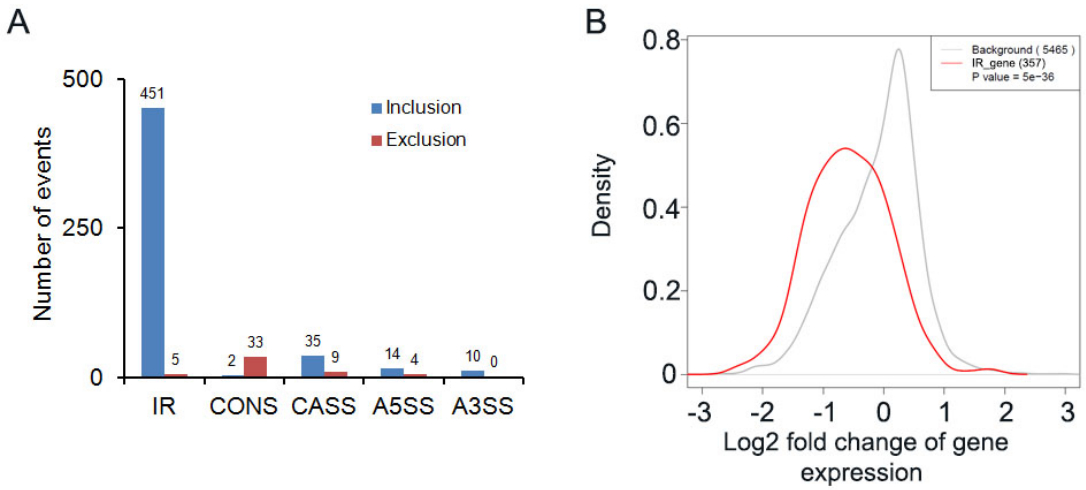

Figure 3.5: The general mRNA splicing in the patient is affected. A, The IR is the major type of splicing events affected in the patient sample. The splicing events are compared between the patient and healthy control. Number of changed events for each type is indicated above bars. IR, intron retention; CONS, constitutive exon splicing; CASS, cassette exon splicing; A5SS, alternative 5' splice site; A3SS, alternative 3' splice site. B, The expression of genes with IR is significantly decreased.

3.5 The deficiency of SmE protein in HEK293 cells causes abnormal splicing

Although the abnormal splicing was observed in the fibroblast of patient, we still couldn't conclude that the observed abnormal splicing in the patient fibroblasts is caused by the identified mutation of *snrpE* since it is not the only mutation existed in the patient. However, due to the mutation leads to changed subcellular distribution and decreased integration capacity of SmE protein detected in the HEK293 cells, it is likely that this mutation leads to partial loss of function of SmE protein contributing to the observed aberrant splicing.

To test the assumption, we decreased the SmE expression level in HEK293 cells by siRNA and checked whether we could observe similar abnormal splicing pattern as that in the patient fibroblasts. For this purpose, we transfected 40nM SmE siRNA into the HEK293 cells by lipofectamine RNAiMAX. The cells were

collected for total RNA and protein extraction after 48 h post transfection. The RT-qPCR and WB were performed to detect the transcript and protein level of SmE after siRNA mediated knocking down. The result of qRT-PCR and WB indicated that the endogenous SmE protein has been successfully decreased by *snrpE* specific siRNA (Figure 3.6A). Compared to control, the residual SmE mRNA level decreased to less than 5%.

The extracted total RNA was further analyzed by RNA-seq. After mapping reads back to the genome and transcriptome, we observed that, the percent of intron reads in SmE deficient HEK293 cells increased compared to the control group (19.4% vs 15%) (Table 3). In addition to the IR as the major type of splicing events, it showed that the exon skipping was also observed as the prominent events in comparison to the control, especially the skipping of constitutive exons (Figure 3.6B). In total, 2317 introns were retained, 724 constitutive exons and 360 cassette exons were with increased skipping, respectively. We then checked the expression level of genes with IR and found that the expression of the 1235 gens with IR was significantly decreased (Figure 3.6C).

To sum up, these results indicate that the depletion of SmE in HEK293 cells leads to the abnormal splicing, especially IR and constitutive exon skipping, and the IR also results in decreased gene expression. Most importantly, it suggests that knocking down SmE in HEK293 cells could recapitulate the splicing defect in the patient fibroblasts.

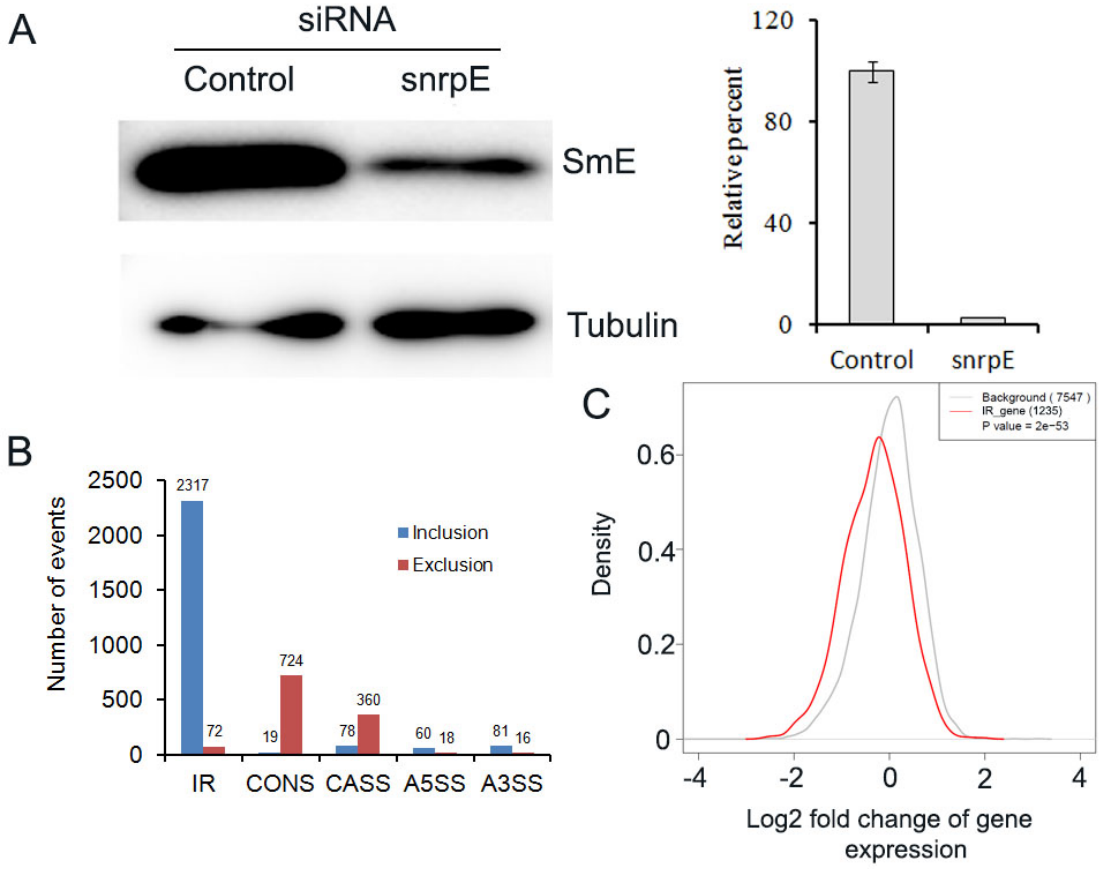

Figure 3.6: The siRNA mediated SmE deficiency in HEK293 leads to aberrant mRNA splicing. A, The endogenous SmE can be successfully decreased by SmE specific siRNA. The residual expression level of SmE protein and mRNA was measured by WB (left panel) and RT-qPCR (right panel). The β-tubulin was used as the loading control of WB. B, The mRNA splicing in SmE deficienct HEK293 cells is abnormal. The IR is the major type splicing events. In addition, a lot of exon skipping events was observed for CONS and CASS. Number of changed events for each type is indicated above the bars. C, The expression of genes with IR is significantly decreased.

3.6 The mutation site can be successfully introduced into HEK293 cells by the combinational application of CRISPR/Cas9 and *piggyBac* transposon system

To clearly pinpoint the function of this point mutation in *snrpE*, we constructed the mutation introduced HEK293 cell line to mimic the situation by applying the

techniques of CRISPR/Cas9 and *piggyBac* transposon system (Figure 3.7A).

To introduce the mutation, we first screened the suitable sgRNAs to guide the Cas9 nuclease mediated DNA cleavage. Using the E-CRISP [141] to scan the possible positions of sgRNA in the snrpE gene, we selected two sgRNAs which located in the 3' end of its first intron (sgRNA8-0)and 3rd exon (sgRNA0-1278) (Figure 3.7B), respectively. Then, we transfected the combination of sgRNA and Cas9 vector into HEK293 cells with different ratio. The cells were collected and the genomic DNA was extracted after 48 h post transfection. The fragment containing each sgRNA targeting site was amplified, and T7 endonuclease assay was performed to evaluate the cleavage efficiency of each sgRNA. The result indicated that both sgRNAs could effectively mediate the cleavage. The insertion or deletion (indel) frequency for both sgRNAs was almost 30%. The different combination ratio of sgRNA and Cas9 didn't induce significant difference in cleavage efficiency. We chose the sgRNA8-0 and the ratio between sgRNA and Cas9 as 1:1 for further experiment.

Next, we proceeded to introduce the mutation into the HEK 293 cells (Figure 3.7A). After two rounds of selection, including positive and negative selection sequentially, the sequence composition of cell colonies was checked by Sanger sequencing (Figure 3.7D). To check whether the potential off target sites were also modified, two off target sites of highly potential were also checked by Sanger sequencing (Figure 3.7E). The Sanger sequencing results (Figure 3.7D and E) demonstrated that the point mutation has been successfully introduced into the HEK293 cells as heterozygous and the potential off target sites were kept intact.

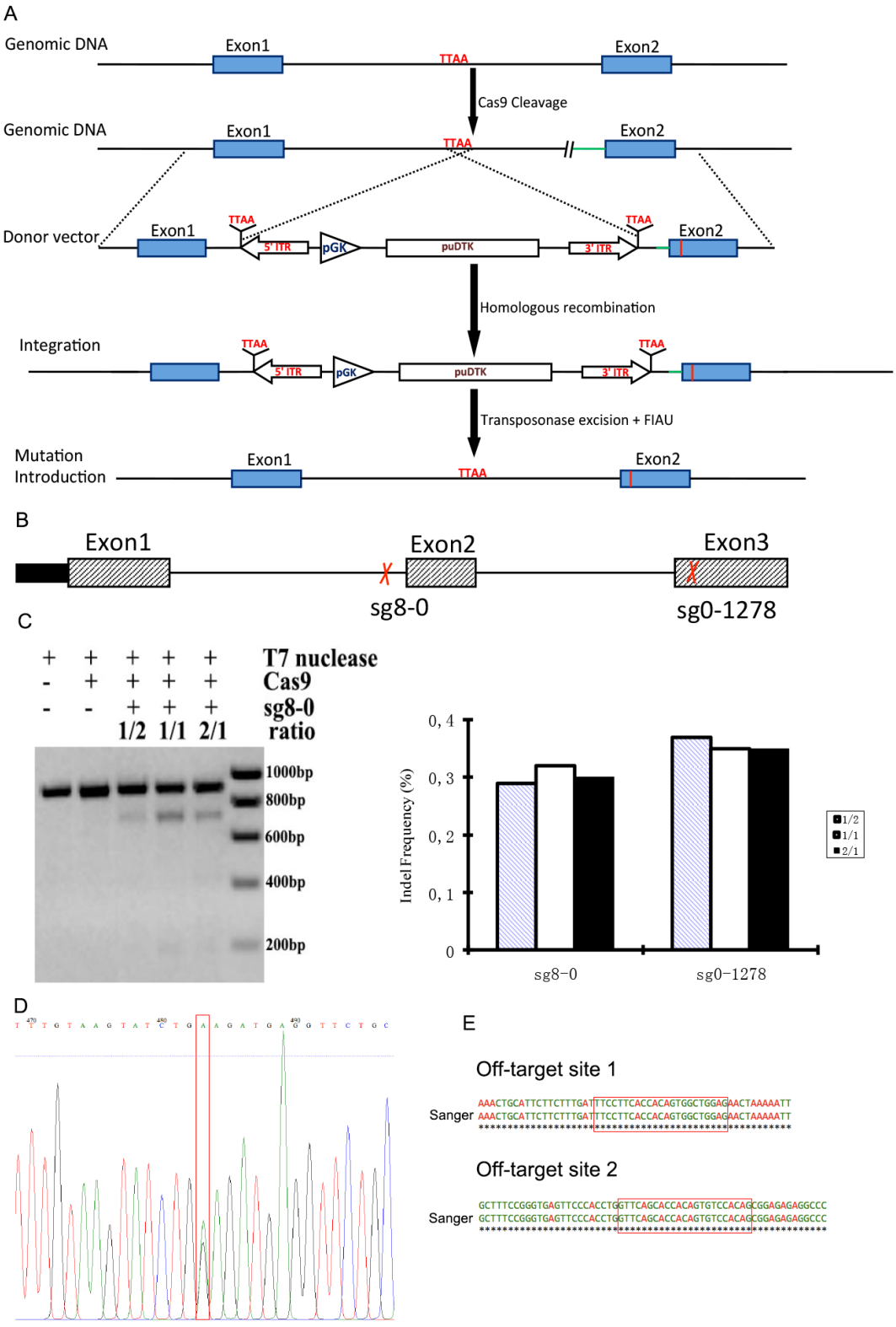

Figure 3.7: Introduction of point mutation into the genome by combination of

CRISPR/Cas9 and *piggyBac* system. A, Illustration of the strategy used for CRISPR/Cas9 mediated introduction of the point mutation into the genome. The vertical red line represents the point mutation. TTAA is the *piggyBac* recognition site. The double slash represents the cleavage site of Cas9 nuclease. B, Illustration of the relative position of sgRNAs selected. The red cross represents the sgRNA site. C, The sgRNA guided cleavage efficiency of Cas9 nuclease was evaluated by T7 endonuclease assay. The cleaved products were checked by agarose gel. The sgRNA8-0 was used as an example (left panel). The cleavage efficiency was quantified for both sgRNAs (right panel). D, Sequencing analysis of excised clone shows the precise introduction of point mutation into the genome as heterozygous. The red box marks the substitution site. E, Sequencing analysis of two highly potential off target sites to check whether these sites were also unwantedly modified.

3.7 The HEK293 cells with point mutation showed the abnormal splicing pattern as observed in the patient fibroblasts

To evaluate whether the aberrant splicing pattern observed in the patient fibroblasts is resulted from the identified missense mutation of *snrpE* gene, the transcriptome profiling of HEK293 cell line introduced with identified point mutation was investigated. Based on the reads distribution, although we could not observe the reads accumulated in the intronic region (Table 3), the in-depth mRNA splicing analysis indicated that the mRNA splicing pattern in this cell line is affected and the IR was still the major type of affected splicing events. In the mutation introduced HEK293 cells, 584 introns were retained. In addition, 110 constitutive exons and 93 cassette exons were observed with increased skipping (Figure 3.8A). The RNA-seq result of mutation introduced HEK293 cells indicated that the splicing pattern of this HEK293 cell line shows some similarity as that observed in patient fibroblasts (Figure 3.5A and 3.8A). Since the IR was still the major type of splicing events in the mutation introduced HEK293 cells, we further checked the expression status of genes with IR. We found that these genes also showed the decreased expression level (Figure 3.8B).

Then, we further compared the splicing change between SmE deficient HEK293 cells and mutation introduced HEK293 cells. Compared to the siRNA mediated SmE deficiency, the extent of aberrant splicing induced by the heterozygous mutation was much milder, consistent with residue function retained by the other wild-type allele (Figure 3.6B and 3.8A).

Given that the IR was the major splicing defect in both SmE deficient HEK293 cells and mutation introduced HEK293 cells, we further checked the overlap of genes with IR between these two conditions and found that almost 93% of genes with IR in mutation introduced HEK293 cells were also identified in the SmE deficient HEK293 cells, suggesting that this mutation disrupts the function of SmE protein.

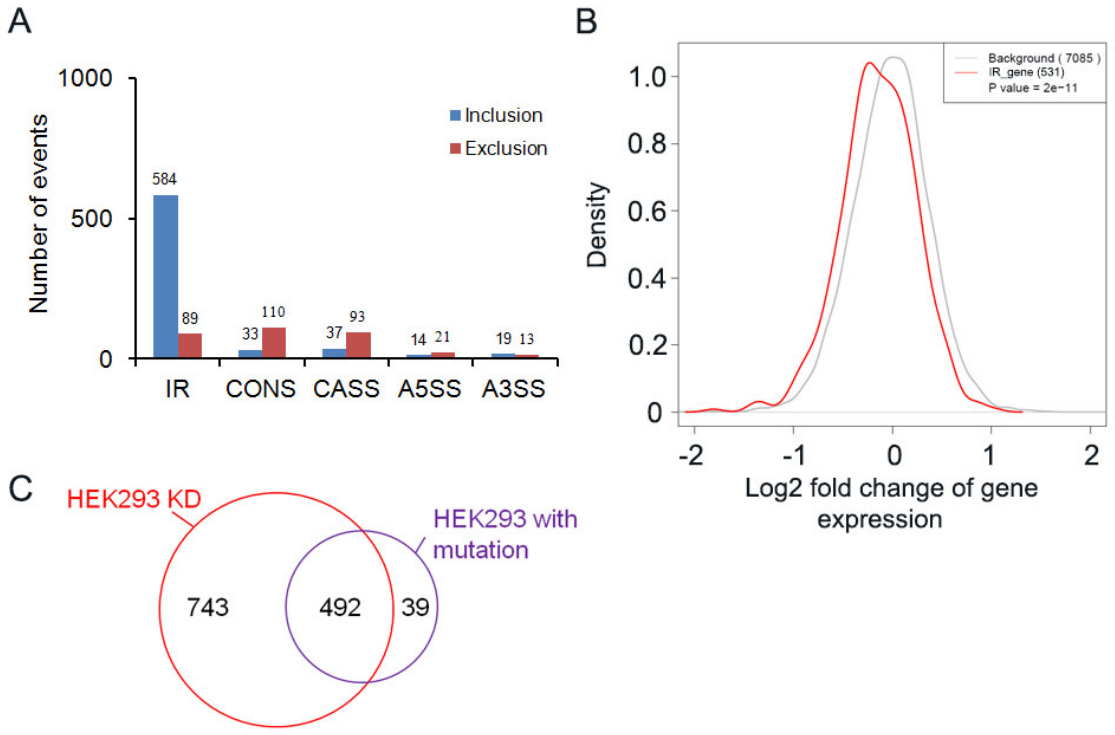

Figure 3.8: The aberrant mRNA splicing can be observed in the missense mutation introduced HEK293 cells. A, The mRNA splicing in missense mutation introduced HEK293 cells is abnormal. Although the IR is the major type splicing events, the extent

decreases compared to SmE deficient HEK293 cells. Number of changed events for each type is indicated above the bars. B, The expression of genes with IR decreases significantly. C, The genes with IR are highly overlapped between SmE deficient HEK293 cells and mutation introduced HEK293 cells.

These results indicate that the mRNA splicing in the mutation introduced HEK293 cells is disturbed. The genetic modified HEK293 cell line with the heterozygous introduction of the point mutation in *snrpE* gene could serve as a cell model to study the functional relevance of the missense mutation.

4 DISCUSSION

In this study, we first took advantage of powerful exome sequencing to identify the potential causative mutations from the patient afflicted with non-syndromal microcephaly and intellectual disability. A heterozygous missense mutation in the gene of *snrpE* was identified as the most potential candidate. To investigate the effect of this mutation on the function of this protein, a series of biochemical experiments were performed to carefully characterize the impact of this mutation. The biochemical results revealed that the function of SmE protein was severely affected by this missense mutation and this mutation interfered the biogenesis of snRNPs complex leading to the changed subcellular distribution. In addition to the biochemical assays, the high throughput RNA-seq was applied for different cell lines under different conditions to check the transcriptome wide gene expression and mRNA splicing. The RNA-seq data demonstrated that the transcriptome wide splicing and gene expression were affected in the patient fibroblasts and siRNA-mediated SmE deficient HEK293 cells. Although we could both observe the splicing change in the patient fibroblasts and siRNA-mediated SmE deficient HEK293 cells, the splicing pattern was different. The IR was the only major type of splicing events in the patient fibroblast whereas the constitutive exon skipping was also the prominent type in the siRNA-mediated SmE deficient HEK293 cells.

To precisely check the contribution of identified mutation to the observed splicing change, the newly developed genomic modification system, CRISPR/Cas9, was combined with *piggyBac* transposon system to introduce the identified mutation into the genomic loci of *snrpE* in HEK293 cells. The transcriptome wide mRNA splicing and gene expression status of mutation introduced HEK293 cells were also checked by RNA-seq. The result showed that the splicing and gene expression were also affected in this mutation introduced

HEK293 cells, and although IR was the major type splicing events, the aberrant splicing in mutation introduced HEK293 cells decreased compared to SmE deficient HEK293 cells. For all of cell lines and conditions, the expression of transcripts with IR was significantly decreased. Taken together, all of these results demonstrate that the identified missense mutation impacts the biogenesis of snRNP complex through interfering the protein-protein interaction between the SmE and Gemin2 protein, and the compromised assembly of snRNP complexes, the major building blocks of spliceosome, leads to the abnormal mRNA splicing and gene expression.

4.1 Exome sequencing is a powerful tool for identifying the potential causative mutations or genes

The exome sequencing as one of applications of 2nd generation of high-throughput sequencing technologies, is a powerful and cost-effective tool to dissect the genetic basis of diseases. Several strategies have been developed to find the disease causing rare variants by exome sequencing, including sequencing and filtering across multiple unrelated, affected individuals to identify variants in the same gene or sequencing parent-child trio for identifying *de novo* mutations [147]. Especially, sequencing of parent-child trio is a highly effective way to identify the *de novo* disease causing mutations, since it is very unlikely by chance to find multiple *de novo* events in patients [148]. The sequencing of trio has been successfully applied to find the *de novo* mutations responsible for genetically heterogeneous disorders, such as intellectual disabilities [149], autism [150].

In this study, we applied the whole exome sequencing for a family (Figure 3.1A) with child suffered from non-syndromal microcephaly and intellectual disability. For each individual, 42-48 M unique reads could map to the targeted protein coding regions covering 91% of exon region of genome and the average base coverage for coding region is at least 84 times (Table 2). Finally, 13 *de novo*

changes, including substitutions, insertions or deletions, were identified. However, most of the *de novo* changes were located in the intronic region, only one heterozygous missense *de novo* mutation located in the coding region of *snrpE* gene (Figure 3.1 B). This mutation is in the 2nd exon of *snrpE* gene encoding the protein of SmE, a member of Sm family proteins and the general component of snRNP complex that are the building blocks of spliceosome.

The Sm proteins are a family of conserved RNA binding proteins present in all three domains of life [151]. In bacteria and archea, Sm homologs form either homohexameric or homoheptameric ring-shaped complexes [152] to regulate the stability and translation of mRNAs by facilitating base pairing interactions between small RNAs and mRNAs [153, 154]. In eukaryotes, more than 20 Sm protein homologs exist and assemble into several distinct heteroheptameric rings. The eukaryotic Sm proteins are classified into two classes: canonical Sm proteins and Sm-like (Lsm) proteins. The canonical Sm proteins, including SmD1, D2, D3, B/B', E, F and G, are the basal constituent of snRNP complexes carrying out the important cellular function of pre-mRNA splicing and 3' end processing. The most striking thing is that the mutation site we found is also deeply conserved (Figure 3.1 C), and therefore, we assume that the mutation could be functionally relevant.

4.2 The mutation affects the function of SmE protein by interfering the early assembly phase

snRNPs as the major building blocks of the major and minor spliceosome, constitute the essential factors in the splicing of all cellular pre-mRNAs. snRNPs contain a common core, composed of seven Sm proteins bound to snRNA, which forms in a stepwise and factor-mediated reaction. The formation of Sm core has been investigated under *in vitro* and *in vivo* conditions. *In vitro* studies showed that the formation of Sm core can happen spontaneously [25] whereas its *in vivo*

formation depends on a surprisingly large number of assisting factors [155]. These factors act in the context of biogenesis cycle of snRNPs, including nuclear exportation of pre-snRNA, cytoplasmic assembly and subsequent re-import into the nucleus [24].

Within this cycle, the PRMT5 complex containing PRMT5, pICln and WD45, is first applied to sequester the Sm proteins in the cytoplasm and to symmetrically dimethylate the distinct arginine residues in the C-terminal tails of SmB/B', SmD1 and SmD3 to increase the interaction affinity [30]. In addition, the pICln subunit functions as an assembly chaperone for Sm proteins to form two higher order sub-complex, a ring shaped 6S complex and SmB/B'-SmD3-pICln sub-complex [29], to avoid the unspecific binding. To continue the assembly process, additional *trans*-acting factors are needed to release the Sm proteins from the pICln-imposed kinetically trapped state. All of these required factors are united in the SMN complex including SMN protein and eight other proteins referred to as Gemin2-8 and unrip [156, 157]. The SMN complex functions as the catalyst of snRNP assembly through releasing the Sm protein in 6S complex from pICln-imposed kinetic trap and loading the Sm proteins onto snRNAs [29]. It is believed that the core machinery mediating snRNP assembly is evolutionarily and functionally conserved in metazoans [158].

Recent years, more and more structural data relating Sm protein has shed light on the architecture of snRNP cores and the mechanism of assembly, for instance, the structural investigation of Sm protein hetero-oligomers and even assembled snRNPs [159-161], the structure of a complex composed of the Sm proteins D1, D2, E, F and G bound to Gemin2 and partial SMN [34], 6S complex and 8S complex composed of 6S complex and SMN complex subunits SMN and Gemin2 [162]. The structure data demonstrates that the Gemin2 is the protein of SMN complex that binds the pentamer of Sm proteins comprised of SmD1/D2/E/F/G

[34].

Given the current structural data, the mutation site we identified in our patient is the interaction site between the SmE and Gemin2 proteins (Figure 3.1D) and has changed the polarity of amino acid from the hydrophobic to hydrophilic. Therefore, this structural prediction result gives us the confidence that the mutation we identified might interfere the function of SmE protein.

To investigate whether the mutation we identified will affect the function of SmE protein, the integration capacity is first checked by IP experiment in the co-expression HEK293 cell line. The data indicates that the integration capacity of mutated SmE protein is severely affected, compared to the wild-type SmE protein (Figure 3.2C and D). Further checking the subcellular localization found that the subcellular distribution of SmE protein is also affected by this mutation （Figure 3.3). TThe seven Sm proteins first form the Sm core onto the snRNA in the cytoplasm, the complex is then imported into the nucleus for further maturation to fulfill the splicing function. Therefore, we speculated that the cytoplasmic assembly step might be interfered by this mutation.

Given that the SMN complex plays important role in the assembly of Sm core and Gemin2 is the protein of SMN complex that binds the pentamer of Sm proteins comprised of SmD1/D2/E/F/G [34], we further checked the cytoplasmic assembly process by IP experiment in the HEK293 overexpression cell lines to uncover which step of early assembly phase is affected. The biochemical data shows that the interaction between the Sm proteins and SMN complex is affected whereas the interaction between the Sm proteins and pICln protein is still normal (Figure 3.3C and D), these results are consistent with our hypothesis based on structural data. Taken together, all of the above results demonstrates that the missense mutation we identified in our patient affects the biogenesis of snRNPs through interfering the early assembly phase of snRNPs.

4.3 The heterozygous missense mutation causes the aberrant mRNA splicing

Previous researches have shown that mutation, deletion or knock down of core spliceosomal and spliceosome assembly factors can lead to the altered splicing patterns in yeast [163], fly [164] and mammalian cells [119, 165]. The components of snRNPs are differentially expressed in mammalian cells and tissues [166, 167]. It is believed that the SS selection decisions can be affected by the relative concentration or activity of general splicing factors. For instance, the PRMT5 as the splicing regulator mediates the symmetrical dimethylation of SmB/B', SmD1 and SmD3 to increase their affinity for the Tudor domain of SMN1 [30]. Bezzi et al. [168] investigated the role of PRMT5 in mammalian development by selective deletion of PRMT5 in neural stem/progenitor cells (NPCs). They found that the absence of PRMT5 leads to postnatal death in mice and causes aberrant splicing of mRNAs with weak 5' SS encoding the proteins belonging to specific biological pathways, such as post-transcriptional RNA processing, membrane organization and cell cycle progression. The downstream gene Mdm4 is considered as the sensor to sense the defects in spliceosomal machinery and transduce the signal to p53 pathway leading to the observed phenotype.

In addition, the gene *snrpB* encoding the protein SmB/B' as the basal component of spliceosome, is found to contain a PTC-introducing alternative exon [169]. The expression of SmB is autoregulated by itself through the inclusion of highly conserved PTC-introducing alternative exon [169, 170]. Saltzman et al. [170] further explored the role of SmB/B' in regulating AS by depletion of SmB/B' in human cells. The deficiency of SmB/B' results in reduced level of snRNPs and a striking increasing of alternative exon skipping enriched in genes encoding RNA processing and other RNA-binding proteins whereas only few effects on constitutive exon splicing can be observed.

Since our biochemical data showed that the function and subcellular distribution of SmE protein are affected by this misense mutation through interfering the early cytoplasmic assembly, the transcriptome-wide mRNA splicing and gene expression were examined in patient fibroblast cell and HEK293 cells with SmE deficiency mediated by siRNA or by introducing the heterozygous missense mutation into the endogenous snrpE loci. The mRNA splicing is affected in all of cell lines and IR is the major type of splicing events under all analyzed conditions (Figure 3.5A, 3.6B and 3.8A). This is in contrast to the study of Saltzman et al. [170]. In that study, although few constitutive exon skipping can be found, the alternative exon skipping is the major splicing events. They did not find significant IR. It is however unclear whether this difference is caused by different function of different Sm proteins or cell specificity or the researcher just did not investigate the effect of SmB/B' deficient on IR. To address this, we reanalyzed their data and observed that the defect of SmB/B' in HeLa cells also leads to the IR and the IR is the major type of affected splicing events.

Due to the IR as the major splicing events under all conditions, the expression of genes with IR was checked and found that compared to the control, the expression of intron retained genes is significantly decreased (Figure 3.5B, 3.6C and 3.8B).

4.4 Aberrant splicing and human diseases

The higher eukaryotes apply the AS to expand the transcriptome complexity and proteome diversity without increasing the number of genes. Due to the important role of AS in physiology and cellular function, the AS is strictly regulated in a tissue, developmental stage or gender specific manner, and also modulated in response to the external stimuli or intracellular signals. Compared to other organisms, the humans utilize the most complex AS events. The correct complement of RAN and proteins in the right cell at the right time is required to

keep the homeostatic of RNA processing. With the increase of complexity, the susceptibility of splicing to mis-regulation also increases. Mutations in the *cis*-elements or *trans*-acting factors and over/under expression of *trans*-acting factors can potentially impact the formation of functional spliceosome resulting in deleterious consequences to cells and leading to a variety of human diseases. Analyses of abnormal splicing code in human diseases not only uncover the underlying maladies of splicing regulation in pathological conditions, but also allow us to obtain an insight into splicing controls in physiological conditions. Understanding physiological and pathological splicing mechanisms also paves the way for development of therapeutic strategies.

It has been reported that mutations in the consensus SS or in the auxiliary splicing *cis*-elements contribute to 15 to 20% of human genetic diseases [171]. In addition to the mutations affecting the consensus splice sites, mutations affecting auxiliary *cis*-elements (ISEs/ESEs and ISSs/ESSs) [172, 173] and RNA secondary structure [174] also have profound effects on aberrant splicing. Once the SS is affected, the aberrantly spliced transcript may lack the essential *cis*-element for mRNA degradation or encode the protein lacking essential domains or signals or form the truncated protein with abnormal functions. In addition, if the alternative SS is affected, the homeostatic balance between the isoforms may also be perturbed resulting in compromised cellular process.

Compared to *cis*-acting splicing defects, the consequence caused by *trans*-acting defects will be more detrimental, since even a single *tans*-factor deficiency can affect multiple target genes. The affected *trans*-factors might be the essential constituents of splicing machinery or the auxiliary factors regulating the AS. Recent years, mutations in the core component of spliceosome have been linked with many human diseases [118, 175]. For instance, in a study of whole-exome sequencing of 29 myelodysplasia specimens, the highly potential causative mutations were identified in the SF3B1, SRSF2, U2AF35 and ZRSR2 genes

indicating that the compromised assembly of spliceosome E/A complex might be the potential cause of myelodysplasia [175]. The RP characterized by the loss of retinal rod photoreceptor cells, is another interesting example, in which the mutations located in the PRPF3, PRPF8, PRPF31 and RP9, are identified as the cause of this disease [118]. The evidence for splicing defects in these diseases is sound, however, it bears mentioning that a largely unanswered question is how mutations in the core machinery, that are expected to impair splicing of most if not all introns, lead to very specific disease phenotypes that are often limited to specific organs or cell types.

In addition to the mutations in the core component of spliceosome machinery, the deficiency of auxiliary splicing *trans*-factors also contributes to many human diseases [176-179]. The auxiliary *trans*-factors whose defect is linked with human disease, including TDP-43 [176], FUS [176], RBFOX1 [177], hnRNPsA2B1 [180], NOVA-1 [178], hnRNP L [179], SRSF1 [123], SRSF3 [181] and QKI [182]. A remarkable example is TDP-43, an RNA/DNA-binding protein with structural resemblance to hnRNP family. It is reported that a significant number of genetic mutations have been identified in TDP-43 in sporadic and familial amyotrophic lateral sclerosis (ALS) [183].

Although a lot of mutations from core machinery components or auxiliary *trans*-acting factors have been reported and linked with human diseases, there are only a few events reported that the mutations from basal component of spliceosome are associated with human diseases. For instance, the mutations within the U4atac snRNA are identified as the cause of MOPD I (also known as Taybi-Linder syndrome (TALS)) [121]. Lynch *et al.* [135] reported that the heterozygous mutations within the alternative exon of *snrpB* are responsible for the CCMS. These mutations cause increased inclusion of the alternative exon introducing the PTC into the transcript, which triggers the nonsense-mediated mRNA decay and

results in the decreased overall expression of SmB. Bacrot *et al*. [136] also identified five heterozygous mutations within the alternative exon of *snrpB* from five unrelated CCMS patients as the cause. In addition, the heterozygous mutations from the *snrpE* gene are considered as the cause of HS characterized by a diffuse and progressive loss of hair [137]. In this case, two mutations are identified, c. 1A>G and c. 133G>A. For c. 1A>G mutation, it results in loss of the start codon of the transcript and use of an alternative in-frame downstream start codon to form the N-terminally truncated protein. The c. 133G>A missense mutation is predicted to disrupt the structure of SmE itself whereas the Sm core assembly is not affected by this mutation. Although the mutations identified from these genes are considered as the cause of these diseases, the target exon/introns responsible for specific phenotypes of these diseases remain unknown.

Since these proteins play really important cellular function, all of the known mutations identified in the basal machinery components are the heterozygous and the normal allele might still be able to keep residue level of functional proteins to maintain the basic cellular function. The requirements of different tissue for the same protein is not the same, therefore, the different tissues might have different sensitivity to the same mutation leading to the tissue specific phenotype.

In this study, we identified a heterozygous missense mutation in *snrpE* gene from one patient with non-syndromal microcephaly and intellectual disability. This heterozygous missense mutation is the first mutation identified within the basal component of spliceosome that is located at the protein interaction site and directly affects the function of basal spliceosomal component. This missense mutation disturbs the function and subcellular distribution of SmE protein by interfering the early cytoplasmic assembly phase. Using the transcriptome wide splicing analysis, we detected that the mRNA splicing in the patient fibroblast cells is affected. Taking advantage of genetic modification technology to

introduce the mutation into the endogenous loci, we found that the splicing pattern observed in the patient fibroblasts can be recapitulated in the HEK293 cells with the heterozygous point mutation. Therefore, we established an *in vitro* cell model to explore the effect of the mutation on host protein function, and also build the cellular link between this mutation and human disease.

4.5 Perspective

Although we have already established the cellular function of this heterozygous missense mutation, the direct link between the genotype and phenotype is still missing. It is still unknown which downstream target genes contribute to the observed phenotype.

The zebrafish (*Danio rerio*) as the animal model has many advantages, such as: transparency of embryos and rapid development, and so on. SmE protein is deeply conserved in zebrafish: only three amino acids are different between the zebrafish and human. Currently, the work using zebrafish to model the effect of SmE defect on brain development, and thereafter investigate the responsible downstream genes, is ongoing.

5 REFERENCES

1. Green, M.R., *PRE-mRNA Splicing.* Ann Rev Genet, 1986. **20**(1): p. 671-708.
2. Patel, A.A. and J.A. Steitz, *Splicing double: insights from the second spliceosome.* Nat Rev Mol Cell Biol, 2003. **4**(12): p. 960-970.
3. Burge, C.B., T. Tuschl, and P.A. Sharp, *Splicing of Precursors to mRNAs by the Spliceosomes.* In The RNA World Second edition, 1999: p. 525-560.
4. Moore, M.J., C.C. Query, and P.A. Sharp, *Splicing of Precursors to mRNA by the Spliceosome*. 1993.
5. Will, C.L. and R. Lührmann, *Spliceosome Structure and Function.* Cold Spring Harb Perspect Biol, 2011. **3**(7).
6. Yong, J., *et al*., *snRNAs Contain Specific SMN-Binding Domains That Are Essential for snRNP Assembly.* Mol Cell Biol, 2004. **24**(7): p. 2747-2756.
7. Will, C.L. and R. Lührmann, *Spliceosome Structure and Function*. The RNA World, 3rd Ed, 2006.
8. Matlin, A.J. and M.J. Moore, *Spliceosome assembly and composition.* Adv Exp Med Biol, 2007. **623**: p. 14-35.
9. Staley, J.P. and J.L. Woolford Jr, *Assembly of ribosomes and spliceosomes: complex ribonucleoprotein machines.* Curr Opin Cell Biol, 2009. **21**(1): p. 109-118.
10. Fox-Walsh, K.L., *et al*., *The architecture of pre-mRNAs affects mechanisms of splice-site pairing.* PNAS, 2005. **102**(45): p. 16176-16181.
11. Sun, J.S. and J.L. Manley, *A novel U2-U6 snRNA structure is necessary for mammalian mRNA splicing.* Genes Dev, 1995. **9**(7): p. 843-854.

12. Raghunathan, P.L. and C. Guthrie, *RNA unwinding in U4/U6 snRNPs requires ATP hydrolysis and the DEIH-box splicing factor Brr2.* Curr Biol, 1998. **8**(15): p. 847-855.
13. Ilagan, J.O., *et al*., *Rearrangements within human spliceosomes captured after exon ligation.* RNA, 2013. **19**(3): p. 400-412.
14. Schwer, B. and C.H. Gross, *Prp22, a DExH-box RNA helicase, plays two distinct roles in yeast pre-mRNA splicing.* EMBO J, 1998. **17**(7): p. 2086-2094.
15. Fourmann, J.-B., *et al*., *Dissection of the factor requirements for spliceosome disassembly and the elucidation of its dissociation products using a purified splicing system.* Genes Dev, 2013. **27**(4): p. 413-428.
16. Hoffman, B.E. and P.J. Grabowski, *U1 snRNP targets an essential splicing factor, U2AF65, to the 3' splice site by a network of interactions spanning the exon.* Genes Dev, 1992. **6**(12b): p. 2554-2568.
17. Reed, R., *Mechanisms of fidelity in pre-mRNA splicing.* Curr Opin Cell Biol, 2000. **12**(3): p. 340-345.
18. Reddy, R., *et al*., *The capped U6 small nuclear RNA is transcribed by RNA polymerase III.* J Biol Chem, 1987. **262**(1): p. 75-81.
19. Huang, Y. and R.J. Maraia, *Comparison of the RNA polymerase III transcription machinery in Schizosaccharomyces pombe, Saccharomyces cerevisiae and human.* Nucleic Acids Res, 2001. **29**(13): p. 2675-2690.
20. Izaurralde, E., *et al*., *A cap-binding protein complex mediating U snRNA export.* Nature, 1995. **376**(6542): p. 709-712.
21. Ohno, M., *et al*., *PHAX, a Mediator of U snRNA Nuclear Export Whose Activity Is Regulated by Phosphorylation.* Cell, 2000. **101**(2): p. 187-198.
22. Kohler, A. and E. Hurt, *Exporting RNA from the nucleus to the cytoplasm.* Nat Rev Mol Cell Biol, 2007. **8**(10): p. 761-773.

23. Askjaer, P., *et al*., *RanGTP-Regulated Interactions of CRM1 with Nucleoporins and a Shuttling DEAD-Box Helicase.* Mol Cell Biol, 1999. **19**(9): p. 6276-6285.
24. Fischer, U., C. Englbrecht, and A. Chari, *Biogenesis of spliceosomal small nuclear ribonucleoproteins.* Wiley Interdiscip Rev RNA, 2011. **2**(5): p. 718-731.
25. Raker, V.A., G. Plessel, and R. Lührmann, *The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle in vitro.* EMBO J, 1996. **15**(9): p. 2256-2269.
26. Kambach, C., *et al*., *Crystal Structures of Two Sm Protein Complexes and Their Implications for the Assembly of the Spliceosomal snRNPs.* Cell, 1999. **96**(3): p. 375-387.
27. Pellizzoni, L., J. Yong, and G. Dreyfuss, *Essential Role for the SMN Complex in the Specificity of snRNP Assembly.* Science, 2002. **298**(5599): p. 1775-1779.
28. Yong, J., *et al*., *Gemin5 Delivers snRNA Precursors to the SMN Complex for snRNP Biogenesis.* Mol Cell, 2010. **38**(4): p. 551-562.
29. Chari, A., *et al*., *An Assembly Chaperone Collaborates with the SMN Complex to Generate Spliceosomal SnRNPs.* Cell, 2008. **135**(3): p. 497-509.
30. Meister, G., *et al*., *Methylation of Sm proteins by a complex containing PRMT5 and the putative U snRNP assembly factor pICln.* Curr Biol, 2001. **11**(24): p. 1990-1994.
31. Friesen, W.J., *et al*., *The Methylosome, a 20S Complex Containing JBP1 and pICln, Produces Dimethylarginine-Modified Sm Proteins.* Mol Cell Biol, 2001. **21**(24): p. 8289-8300.

32. Gonsalvez, G.B., *et al*., *Two distinct arginine methyltransferases are required for biogenesis of Sm-class ribonucleoproteins.* J Cell Biol, 2007. **178**(5): p. 733-740.
33. Pu, W.T., *et al*., *pICln Inhibits snRNP Biogenesis by Binding Core Spliceosomal Proteins.* Mol Cell Biol, 1999. **19**(6): p. 4113-4120.
34. Zhang, R., *et al*., *Structure of a Key Intermediate of the SMN Complex Reveals Gemin2's Crucial Function in snRNP Assembly.* Cell, 2011. **146**(3): p. 384-395.
35. Battle, D.J., *et al*., *The Gemin5 Protein of the SMN Complex Identifies snRNAs.* Mol Cell, 2006. **23**(2): p. 273-279.
36. Mouaikel, J., *et al*., *Hypermethylation of the Cap Structure of Both Yeast snRNAs and snoRNAs Requires a Conserved Methyltransferase that Is Localized to the Nucleolus.* Mol Cell, 2002. **9**(4): p. 891-901.
37. Mattaj, I.W., *Cap trimethylation of U snRNA is cytoplasmic and dependent on U snRNP protein binding.* Cell, 1986. **46**(6): p. 905-911.
38. Houseley, J., J. LaCava, and D. Tollervey, *RNA-quality control by the exosome.* Nat Rev Mol Cell Biol, 2006. **7**(7): p. 529-539.
39. Hamm, J., *et al*., *The trimethylguanosine cap structure of U1 snRNA is a component of a bipartite nuclear targeting signal.* Cell, 1990. **62**(3): p. 569-577.
40. Palacios, I., *et al*., *Nuclear import of U snRNPs requires importin β.* EMBO J, 1997. **16**(22): p. 6783-6792.
41. Narayanan, U., *et al*., *Coupled In Vitro Import of U snRNPs and SMN, the Spinal Muscular Atrophy Protein.* Mol Cell, 2004. **16**(2): p. 223-234.
42. Sleeman, J.E. and A.I. Lamond, *Newly assembled snRNPs associate with coiled bodies before speckles, suggesting a nuclear snRNP maturation pathway.* Curr Biol, 1999. **9**(19): p. 1065-1074.

43. Jády, B.E., *et al*., *Modification of Sm small nuclear RNAs occurs in the nucleoplasmic Cajal body following import from the cytoplasm.* EMBO J, 2003. **22**(8): p. 1878-1888.
44. Nesic, D., G. Tanackovic, and A. Krämer, *A role for Cajal bodies in the final steps of U2 snRNP biogenesis.* J Cell Sci, 2004. **117**(19): p. 4423-4433.
45. Novotný, I., *et al*., *In vivo kinetics of U4/U6·U5 tri-snRNP formation in Cajal bodies.* Mol Biol Cell, 2011. **22**(4): p. 513-523.
46. Staněk, D. and K.M. Neugebauer, *Detection of snRNP assembly intermediates in Cajal bodies by fluorescence resonance energy transfer.* J Cell Biol, 2004. **166**(7): p. 1015-1025.
47. R E Breitbart, a. A Andreadis, and B. Nadal-Ginard, *Alternative Splicing: A Ubiquitous Mechanism for the Generation of Multiple Protein Isoforms from Single Genes.* Ann Rev Biochem, 1987. **56**(1): p. 467-495.
48. Blencowe, B.J., *Alternative Splicing: New Insights from Global Analyses.* Cell, 2006. **126**(1): p. 37-47.
49. Black, D.L., *Mechanisms of Alternative Pre-Messenger RNA Splicing.* Ann Rev Biochem, 2003. **72**(1): p. 291-336.
50. Wahl, M.C., C.L. Will, and R. Lührmann, *The Spliceosome: Design Principles of a Dynamic RNP Machine.* Cell, 2009. **136**(4): p. 701-718.
51. Boutz, P.L., et al., *A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons.* Genes Dev, 2007. **21**(13): p. 1636-1652.
52. Makeyev, E.V., *et al*., *The MicroRNA miR-124 Promotes Neuronal Differentiation by Triggering Brain-Specific Alternative Pre-mRNA Splicing.* Mol Cell, 2007. **27**(3): p. 435-448.

53. Pan, Q., *et al*., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-1415.
54. Wang, E.T., *et al*., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-476.
55. Gerstein, M.B., *et al*., *Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project.* Science, 2010. **330**(6012): p. 1775-1787.
56. Graveley, B.R., *et al*., *The developmental transcriptome of Drosophila melanogaster.* Nature, 2011. **471**(7339): p. 473-479.
57. Baker, B.S., *Sex in flies: the splice of life.* Nature, 1989. **340**(6234): p. 521-524.
58. Nagoshi, R.N. and B.S. Baker, *Regulation of sex-specific RNA splicing at the Drosophila doublesex gene: cis-acting mutations in exon sequences alter sex-specific RNA splicing patterns.* Genes Dev, 1990. **4**(1): p. 89-97.
59. Shin, C. and J.L. Manley, *Cell signalling and the control of pre-mRNA splicing.* Nat Rev Mol Cell Biol, 2004. **5**(9): p. 727-738.
60. Xie, J. and D.L. Black, *A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels.* Nature, 2001. **410**(6831): p. 936-939.
61. Cieply, B. and R.P. Carstens, *Functional roles of alternative splicing factors in human disease.* Wiley Interdiscip Rev RNA, 2015. **6**(3): p. 311-326.
62. Modrek, B. and C. Lee, *A genomic view of alternative splicing.* Nat Genet, 2002. **30**(1): p. 13-19.
63. Bassell, G.J. and S. Kelic, *Binding proteins for mRNA localization and local translation, and their dysfunction in genetic neurological disease.* Curr Opin Neurobiol, 2004. **14**(5): p. 574-581.

64. Keene, J.D., *RNA regulons: coordination of post-transcriptional events.* Nat Rev Genet, 2007. **8**(7): p. 533-543.

65. Green, R.E., *et al*., *Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes.* Bioinformatics, 2003. **19**(suppl 1): p. i118-i121.

66. Graveley, B.R., *Sorting out the complexity of SR protein functions.* RNA, 2000. **6**(9): p. 1197-1211.

67. Tacke, R. and J.L. Manley, *Determinants of SR protein specificity.* Curr Opin Cell Biol, 1999. **11**(3): p. 358-362.

68. Long, Jennifer C. and Javier F. Caceres, *The SR protein family of splicing factors: master regulators of gene expression.* Biochem J, 2009. **417**(1): p. 15-27.

69. Smith, C.W.J. and J. Valcárcel, *Alternative pre-mRNA splicing: the logic of combinatorial control.* Trends Biochem Sci, 2000. **25**(8): p. 381-388.

70. Dreyfuss, G., V.N. Kim, and N. Kataoka, *Messenger-RNA-binding proteins and the messages they carry.* Nat Rev Mol Cell Biol, 2002. **3**(3): p. 195-205.

71. Ule, J., *et al*., *An RNA map predicting Nova-dependent splicing regulation.* Nature, 2006. **444**(7119): p. 580-586.

72. Hui, J., *et al*., *Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing.* EMBO J, 2005. **24**(11): p. 1988-1998.

73. Yeo, G.W., *et al*., *An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells.* Nat Struct Mol Biol, 2009. **16**(2): p. 130-137.

74. Mauger, D.M., C. Lin, and M.A. Garcia-Blanco, *hnRNP H and hnRNP F Complex with Fox2 To Silence Fibroblast Growth Factor Receptor 2 Exon IIIc.* Mol Cell Biol, 2008. **28**(17): p. 5403-5419.

75. House, A.E. and K.W. Lynch, *An exonic splicing silencer represses spliceosome assembly after ATP-dependent exon recognition.* Nat Struct Mol Biol, 2006. **13**(10): p. 937-944.

76. Sharma, S., *et al*., *Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome.* Nat Struct Mol Biol, 2008. **15**(2): p. 183-191.

77. Lallena, M.a.J., *et al*., *Splicing Regulation at the Second Catalytic Step by Sex-lethal Involves 3′ Splice Site Recognition by SPF45.* Cell, 2002. **109**(3): p. 285-296.

78. Bourgeois, C.F., *et al*., *Identification of a Bidirectional Splicing Enhancer: Differential Involvement of SR Proteins in 5′ or 3′ Splice Site Activation.* Mol Cell Biol, 1999. **19**(11): p. 7347-7356.

79. Zuo, P. and T. Maniatis, *The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing.* Genes Dev, 1996. **10**(11): p. 1356-1368.

80. Feng, Y., M. Chen, and J.L. Manley, *Phosphorylation switches the general splicing repressor SRp38 to a sequence-specific activator.* Nat Struct Mol Biol, 2008. **15**(10): p. 1040-1048.

81. Spellman, R. and C.W.J. Smith, *Novel modes of splicing repression by PTB.* Trends Biochem Sci, 2006. **31**(2): p. 73-76.

82. Saulière, J., *et al*., *The Polypyrimidine Tract Binding Protein (PTB) Represses Splicing of Exon 6B from the β-Tropomyosin Pre-mRNA by Directly Interfering with the Binding of the U2AF65 Subunit.* Mol Cell Biol, 2006. **26**(23): p. 8755-8769.

83. Zhou, H.-L. and H. Lou, *Repression of Prespliceosome Complex Formation at Two Distinct Steps by Fox-1/Fox-2 Proteins.* Mol Cell Biol, 2008. **28**(17): p. 5507-5516.

84. Kashima, T. and J.L. Manley, *A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy.* Nat Genet, 2003. **34**(4): p. 460-463.

85. Martins de Araújo, M., *et al*., *Differential 3' splice site recognition of SMN1 and SMN2 transcripts by U2AF and U2 snRNP.* RNA, 2009. **15**(4): p. 515-523.

86. Zahler, A.M., *et al*., *SC35 and Heterogeneous Nuclear Ribonucleoprotein A/B Proteins Bind to a Juxtaposed Exonic Splicing Enhancer/Exonic Splicing Silencer Element to Regulate HIV-1 tat Exon 2 Splicing.* J Biol Chem, 2004. **279**(11): p. 10077-10084.

87. Zhu, J. and A.R. Krainer, *Pre-mRNA splicing in the absence of an SR protein RS domain.* Genes Dev, 2000. **14**(24): p. 3166-3178.

88. Crawford, J.B. and J.G. Patton, *Activation of α-Tropomyosin Exon 2 Is Regulated by the SR Protein 9G8 and Heterogeneous Nuclear Ribonucleoproteins H and F.* Mol Cell Biol, 2006. **26**(23): p. 8791-8802.

89. Licatalosi, D.D., *et al*., *HITS-CLIP yields genome-wide insights into brain alternative RNA processing.* Nature, 2008. **456**(7221): p. 464-469.

90. Martinez-Contreras, R., *et al*., *Intronic Binding Sites for hnRNP A/B and hnRNP F/H Proteins Stimulate Pre-mRNA Splicing.* PLoS Biol, 2006. **4**(2): p. e21.

91. Wang, Z. and C.B. Burge, *Splicing regulation: From a parts list of regulatory elements to an integrated splicing code.* RNA, 2008. **14**(5): p. 802-813.

92. Chen, M. and J.L. Manley, *Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches.* Nat Rev Mol Cell Biol, 2009. **10**(11): p. 741-754.

93. Bonnal, S., *et al*., *RBM5/Luca-15/H37 Regulates Fas Alternative Splice Site Pairing after Exon Definition.* Mol Cell, 2008. **32**(1): p. 81-95.

94. Batsche, E., M. Yaniv, and C. Muchardt, *The human SWI/SNF subunit Brm is a regulator of alternative splicing.* Nat Struct Mol Biol, 2006. **13**(1): p. 22-29.

95. de la Mata, M. and A.R. Kornblihtt, *RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20.* Nat Struct Mol Biol, 2006. **13**(11): p. 973-980.

96. Sims Iii, R.J., *et al*., *Recognition of Trimethylated Histone H3 Lysine 4 Facilitates the Recruitment of Transcription Postinitiation Factors and Pre-mRNA Splicing.* Mol Cell, 2007. **28**(4): p. 665-676.

97. Moldon, A., *et al*., *Promoter-driven splicing regulation in fission yeast.* Nature, 2008. **455**(7215): p. 997-1000.

98. Luco, R.F., *et al*., *Epigenetics in Alternative Pre-mRNA Splicing.* Cell, 2011. **144**(1): p. 16-26.

99. Gómez Acuña, L.I., et al., *Connections between chromatin signatures and splicing.* Wiley Interdiscip Rev RNA, 2013. **4**(1): p. 77-91.

100. de Almeida, S.F. and M. Carmo-Fonseca, *Design principles of interconnections between chromatin and pre-mRNA splicing.* Trends Biochem Sci, 2012. **37**(6): p. 248-253.

101. Iannone, C. and J. Valcárcel, *Chromatin's thread to alternative splicing regulation.* Chromosoma, 2013. **122**(6): p. 465-474.

102. Das, R., *et al*., *SR Proteins Function in Coupling RNAP II Transcription to Pre-mRNA Splicing.* Mol Cell, 2007. **26**(6): p. 867-881.

103. Auboeuf, D., *et al*., *Differential recruitment of nuclear receptor coactivators may determine alternative RNA splice site choice in target genes.* PNAS, 2004. **101**(8): p. 2270-2274.

104. Kornblihtt, A.R., *et al*., *Alternative splicing: a pivotal step between eukaryotic transcription and translation.* Nat Rev Mol Cell Biol, 2013. **14**(3): p. 153-165.

105. Cartegni, L., S.L. Chew, and A.R. Krainer, *Listening to silence and understanding nonsense: exonic mutations that affect splicing.* Nat Rev Genet, 2002. **3**(4): p. 285-298.

106. Cooper, T.A., L. Wan, and G. Dreyfuss, *RNA and Disease.* Cell, 2009. **136**(4): p. 777-793.

107. Gallo, J.M., W. Noble, and T.R. Martin, *RNA and protein-dependent mechanisms in tauopathies: consequences for therapeutic strategies.* Cell Mol Life Sci, 2007. **64**(13): p. 1701-1714.

108. Hasegawa, M., *et al*., *FTDP-17 mutations N279K and S305N in tau produce increased splicing of exon 10.* FEBS Lett, 1999. **443**(2): p. 93-96.

109. Dawson, H.N., *et al*., *The Tau N279K Exon 10 Splicing Mutation Recapitulates Frontotemporal Dementia and Parkinsonism Linked to Chromosome 17 Tauopathy in a Mouse Model.* J Neurosci, 2007. **27**(34): p. 9155-9168.

110. CARMEL, I., *et al*., *Comparative analysis detects dependencies among the 5′ splice-site positions.* RNA, 2004. **10**(5): p. 828-840.

111. Cheishvili, D., *et al*., *IKAP/hELP1 deficiency in the cerebrum of familial dysautonomia patients results in down regulation of genes involved in oligodendrocyte differentiation and in myelination.* Hum Mol Genet, 2007. **16**(17): p. 2097-2104.

112. Buratti, E., *et al*., *Nuclear factor TDP-43 and SR proteins promote in vitro and in vivo CFTR exon 9 skipping.* EMBO J, 2001. **20**(7): p. 1774-1784.

113. Neumann, M., *et al*., *Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis.* Science, 2006. **314**(5796): p. 130-133.

114. Lynch, K.W. and T. Maniatis, *Assembly of specific SR protein complexes on distinct regulatory elements of the Drosophila doublesex splicing enhancer.* Genes Dev, 1996. **10**(16): p. 2089-2101.

115. Huh, G.S. and R.O. Hynes, *Regulation of alternative pre-mRNA splicing by a novel repeated hexanucleotide element.* Genes Dev, 1994. **8**(13): p. 1561-1574.

116. Laule, M., *et al*., *Interaction of CA repeat polymorphism of the endothelial nitric oxide synthase and hyperhomocysteinemia in acute coronary syndromes: evidence of gender-specific differences.* J Mol Med, 2003. **81**(5): p. 305-309.

117. Ranum, L.P.W. and T.A. Cooper, *RNA-MEDIATED NEUROMUSCULAR DISORDERS.* Ann Rev Neurosci, 2006. **29**(1): p. 259-277.

118. Mordes, D., *et al*., *Pre-mRNA splicing and retinitis pigmentosa.* Mol Vis, 2006. **12**: p. 1259-1271.

119. Zhang, Z., *et al*., *SMN Deficiency Causes Tissue-Specific Perturbations in the Repertoire of snRNAs and Widespread Defects in Splicing.* Cell, 2008. **133**(4): p. 585-600.

120. Edery, P., *et al*., *Association of TALS Developmental Disorder with Defect in Minor Splicing Component U4atac snRNA.* Science, 2011. **332**(6026): p. 240-243.

121. He, H., *et al*., *Mutations in U4atac snRNA, a Component of the Minor Spliceosome, in the Developmental Disorder MOPD I.* Science, 2011. **332**(6026): p. 238-240.

122. Venables, J.P., *Unbalanced alternative splicing and its significance in cancer.* Bioessays, 2006. **28**(4): p. 378-386.

123. Karni, R., *et al*., *The gene encoding the splicing factor SF2/ASF is a proto-oncogene.* Nat Struct Mol Biol, 2007. **14**(3): p. 185-193.

124. Anczuków, O., *et al*., *The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation.* Nat Struct Mol Biol, 2012. **19**(2): p. 220-228.

125. David, C.J. and J.L. Manley, *Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged.* Genes Dev, 2010. **24**(21): p. 2343-2364.

126. Zhang, J. and J.L. Manley, *Misregulation of Pre-mRNA Alternative Splicing in Cancer.* Cancer Discov, 2013. **3**(11): p. 1228-1237.

127. Biamonti, G., *et al*., *The alternative splicing side of cancer.* Semin Cell Dev Biol, 2014. **32**: p. 30-36.

128. Collesi, C., *et al*., *A splicing variant of the RON transcript induces constitutive tyrosine kinase activity and an invasive phenotype.* Mol Cell Biol, 1996. **16**(10): p. 5518-26.

129. Neiswanger, K., *et al*., *Assignment of the gene for the small nuclear ribonucleoprotein E (SNRPE) to human chromosome 1q25-q43.* Genomics, 1990. **7**(4): p. 503-508.

130. Fautsch, M.P., *et al*., *Conservation of coding and transcriptional control sequences within the snRNP E protein gene.* Genomics, 1992. **14**(4): p. 883-890.

131. Seto, A.G., *et al*., *Saccharomyces cerevisiae telomerase is an Sm small nuclear ribonucleoprotein particle.* Nature, 1999. **401**(6749): p. 177-180.

132. Barbee, S.A., A.L. Lublin, and T.C. Evans, *A Novel Function for the Sm Proteins in Germ Granule Localization during C. elegans Embryogenesis.* Curr Biol, 2002. **12**(17): p. 1502-1506.

133. Bilinski, S.M., *et al*., *Sm proteins, the constituents of the spliceosome, are components of nuage and mitochondrial cement in Xenopus oocytes.* Exp Cell Res, 2004. **299**(1): p. 171-178.

134. Barbee, S.A. and T.C. Evans, *The Sm proteins regulate germ cell specification during early C. elegans embryogenesis.* Dev Biol, 2006. **291**(1): p. 132-143.

135. Lynch, D.C., *et al*., *Disrupted auto-regulation of the spliceosomal gene SNRPB causes cerebro–costo–mandibular syndrome.* Nat Commun, 2014. **5**.

136. Bacrot, S., *et al*., *Mutations in SNRPB, Encoding Components of the Core Splicing Machinery, Cause Cerebro-Costo-Mandibular Syndrome.* Hum Mutat, 2015. **36**(2): p. 187-190.

137. Pasternack, Sandra M., *et al*., *Mutations in SNRPE, which Encodes a Core Protein of the Spliceosome, Cause Autosomal-Dominant Hypotrichosis Simplex.* Am J Hum Genet, 2013. **92**(1): p. 81-87.

138. Anchi, T., *et al*., *SNRPE is involved in cell proliferation and progression of high-grade prostate cancer through the regulation of androgen receptor expression.* Oncol Lett, 2012. **3**(2): p. 264-268.

139. Quidville, V., *et al*., *Targeting the Deregulated Spliceosome Core Machinery in Cancer Cells Triggers mTOR Blockade and Autophagy.* Cancer Res, 2013. **73**(7): p. 2247-2258.

140. Weiss, J., *et al*., *ENU mutagenesis in mice identifies candidate genes for hypogonadism.* Mam Genome, 2012. **23**(5-6): p. 346-355.

141. Heigwer, F., G. Kerr, and M. Boutros, *E-CRISP: fast CRISPR target site identification.* Nat Meth, 2014. **11**(2): p. 122-123.

142. Xie, F., *et al*., *Seamless gene correction of β-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac.* Genome Res, 2014.

143. Fröhler, S., *et al*., *Exome sequencing helped the fine diagnosis of two siblings afflicted with atypical Timothy syndrome (TS2).* BMC Med Genet, 2014. **15**(1): p. 1-6.

144. McManus, C.J., *et al*., *Evolution of splicing regulatory networks in Drosophila.* Genome Res, 2014.

145. Wang, Y., *et al*., *Integrative analysis revealed the molecular mechanism underlying RBM10-mediated splicing regulation.* EMBO Mol Med, 2013. **5**(9): p. 1431-1442.

146. Mortazavi, A., *et al*., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Meth, 2008. **5**(7): p. 621-628.

147. Bamshad, M.J., *et al*., *Exome sequencing as a tool for Mendelian disease gene discovery.* Nat Rev Genet, 2011. **12**(11): p. 745-755.

148. Nachman, M.W. and S.L. Crowell, *Estimate of the Mutation Rate per Nucleotide in Humans.* Genetics, 2000. **156**(1): p. 297-304.

149. Vissers, L.E.L.M., *et al*., *A de novo paradigm for mental retardation.* Nat Genet, 2010. **42**(12): p. 1109-1112.

150. O'Roak, B.J., *et al*., *Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations.* Nat Genet, 2011. **43**(6): p. 585-589.

151. Salgado‐Garrido, J., *et al*., *Sm and Sm‐like proteins assemble in two related complexes of deep evolutionary origin.* EMBO J, 1999. **18**(12): p. 3451-3462.

152. Törö, I., *et al*., *Archaeal Sm Proteins form Heptameric and Hexameric Complexes: Crystal Structures of the Sm1 and Sm2 Proteins from the Hyperthermophile Archaeoglobus fulgidus.* J Mol Biol, 2002. **320**(1): p. 129-142.

153. Sledjeski, D.D., C. Whitman, and A. Zhang, *Hfq Is Necessary for Regulation by the Untranslated RNA DsrA.* J Bacteriol, 2001. **183**(6): p. 1997-2005.

154. Zhang, A., *et al*., *The Sm-like Hfq Protein Increases OxyS RNA Interaction with Target mRNAs.* Mol Cell, 2002. **9**(1): p. 11-22.

155. Chari, A., E. Paknia, and U. Fischer, *The role of RNP biogenesis in spinal muscular atrophy.* Curr Opin Cell Biol, 2009. **21**(3): p. 387-393.

156. Gubitz, A.K., W. Feng, and G. Dreyfuss, *The SMN complex.* Exp Cell Res, 2004. **296**(1): p. 51-56.

157. Meister, G., *et al*., *A multiprotein complex mediates the ATP-dependent assembly of spliceosomal U snRNPs.* Nat Cell Biol, 2001. **3**(11): p. 945-949.

158. Kroiss, M., *et al*., *Evolution of an RNP assembly system: A minimal SMN complex facilitates formation of UsnRNPs in Drosophila melanogaster.* PNAS, 2008. **105**(29): p. 10045-10050.

159. Leung, A.K.W., K. Nagai, and J. Li, *Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis.* Nature, 2011. **473**(7348): p. 536-539.

160. Pomeranz Krummel, D.A., *et al*., *Crystal structure of human spliceosomal U1 snRNP at 5.5[thinsp]A resolution.* Nature, 2009. **458**(7237): p. 475-480.

161. Weber, G., *et al*., *Functional organization of the Sm core in the crystal structure of human U1 snRNP.* EMBO J, 2010. **29**(24): p. 4172-4184.

162. Grimm, C., *et al*., *Structural Basis of Assembly Chaperone- Mediated snRNP Formation.* Mol Cell, 2013. **49**(4): p. 692-703.

163. Campion, Y., *et al*., *Specific splicing defects in S. pombe carrying a degron allele of the Survival of Motor Neuron gene.* EMBO J, 2010. **29**(11): p. 1817-1829.

164. Park, J.W., *et al*., *Identification of alternative splicing regulators by RNA interference in Drosophila.* PNAS, 2004. **101**(45): p. 15974-15979.

165. Bäumer, D., *et al*., *Alternative Splicing Events Are a Late Feature of Pathology in a Mouse Model of Spinal Muscular Atrophy.* PLoS Genet, 2009. **5**(12): p. e1000773.

166. Grosso, A.R., *et al*., *Tissue-specific splicing factor gene expression signatures.* Nucleic Acids Res, 2008. **36**(15): p. 4823-4832.

167. Castle, J.C., *et al*., *Digital Genome-Wide ncRNA Expression, Including SnoRNAs, across 11 Human Tissues Using PolyA-Neutral Amplification.* PLoS ONE, 2010. **5**(7): p. e11779.

168. Bezzi, M., *et al*., *Regulation of constitutive and alternative splicing by PRMT5 reveals a role for Mdm4 pre-mRNA in sensing defects in the spliceosomal machinery.* Genes Dev, 2013. **27**(17): p. 1903-1916.

169. Saltzman, A.L., *et al*., *Regulation of Multiple Core Spliceosomal Proteins by Alternative Splicing-Coupled Nonsense-Mediated mRNA Decay.* Mol Cell Biol, 2008. **28**(13): p. 4320-4330.

170. Saltzman, A.L., Q. Pan, and B.J. Blencowe, *Regulation of alternative splicing by the core spliceosomal machinery.* Genes Dev, 2011. **25**(4): p. 373-384.

171. Matlin, A.J., F. Clark, and C.W.J. Smith, *Understanding alternative splicing: towards a cellular code.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 386-398.

172. Masuda, A., *et al*., *hnRNP H enhances skipping of a nonfunctional exon P3A in CHRNA1 and a mutation disrupting its binding causes congenital myasthenic syndrome.* Hum Mol Genet, 2008. **17**(24): p. 4022-4035.

173. Cartegni, L. and A.R. Krainer, *Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1.* Nat Genet, 2002. **30**(4): p. 377-384.

174. Graveley, B.R., *Mutually Exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures.* Cell, 2005. **123**(1): p. 65-73.

175. Yoshida, K., *et al*., *Frequent pathway mutations of splicing machinery in myelodysplasia.* Nature, 2011. **478**(7367): p. 64-69.

176. Vance, C., *et al*., *Mutations in FUS, an RNA Processing Protein, Cause Familial Amyotrophic Lateral Sclerosis Type 6.* Science, 2009. **323**(5918): p. 1208-1211.

177. Voineagu, I., *et al*., *Transcriptomic analysis of autistic brain reveals convergent molecular pathology.* Nature, 2011. **474**(7351): p. 380-384.

178. Ule, J., *et al*., *CLIP Identifies Nova-Regulated RNA Networks in the Brain.* Science, 2003. **302**(5648): p. 1212-1215.

179. Goehe, R.W., *et al*., *hnRNP L regulates the tumorigenic capacity of lung cancer xenografts in mice via caspase-9 pre-mRNA processing.* J Clin Invest, 2010. **120**(11): p. 3923-3939.

180. Jensen, K.B., *et al*., *Nova-1 Regulates Neuron-Specific Alternative Splicing and Is Essential for Neuronal Viability.* Neuron, 2000. **25**(2): p. 359-371.

181. Tang, Y., *et al*., *Downregulation of splicing factor SRSF3 induces p53[beta], an alternatively spliced isoform of p53 that promotes cellular senescence.* Oncogene, 2013. **32**(22): p. 2792-2798.

182. Zong, F.-Y., *et al*., *The RNA-Binding Protein QKI Suppresses Cancer-Associated Aberrant Splicing.* PLoS Genet, 2014. **10**(4): p. e1004289.

183. Kabashi, E., *et al*., *TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis.* Nat Genet, 2008. **40**(5): p. 572-574.

PUBLICATION

1. *Li, N.★, You, X.★, **Chen, T.★**, Mackowiak, S.D., Friedlander, M.R., Weigt, M., Du H., Gogol-Döring , A., Chang, Z., Dieterich, C., Hu, Y., Chen, W. Global profiling of miRNAs and the hairpin precursors: insights into miRNA processing and novel miRNA discovery. Nucleic Acids Res. 2013, 41(6): 3619-3634. (★ equal contribution)*
2. *Sun, W., You, X., Gogol-Döring, A., He, H., Kise, Y., Sohn, M., **Chen, T.**, Klebes, A., Schmucker, D., Chen, W. Ultra-deep profiling of alternative splicing Drosophila Dscam isoforms by circularization-assisted multi-segment sequencing. EMBO J. 2013, 32: 2029-2038.*
3. *You, X., Vlatkovic, I., Babic, A., Will, T., Epstein, I., Tushev, G., Akbalik, G., Wang, M., Glock, C., Quedenau, C., Wang, X., Hou, J., Liu, H., Sun, W., Sambandan, S., **Chen, T.**, Schuman, E.M., Chen, W. Neural circular RNAs are derived from synaptic genes, regulated by development, plasticity. Nat Neurosci. 2015, 18: 603-610.*

CURRICULUM VITAE

For reasons of data protection, the curriculum vitae is not published in the electronic version.

Appendix

RNP: ribonuceloprotein complex

SS: splice site

BS: branch point site

PPT: polypyrimidine tract

snRNPs: small nuclear ribonucleoproteins

CBC: cap binding complex

PHAX: phosphorylated adaptor of export protein

NES: nuclear export signal

PRMT5: protein arginine methyltransferase 5

TMG: 2,2,7-trimethylguanosine

CBs: cajal bodies

AS: alternative splicing

IR: intron retention

UTR: untranslated region

PTC: premature termination codon

NMD: nonsense mediated decay

ESEs: exonic splicing enhancers

ESSs: exonic splicing silencers

ISEs: intronic splicing enhancers

ISSs: intronic splicing silencers

hnRNPs: heterogeneous nuclear ribonucleoproteins

NOVA1: neuro-oncological ventral antigen 1

FTDP-17: Front temporal dementia with parkinsonism linked to chromosome 17

FD: Familial Dysautonomia

IKBKAP: i-kappa-B kinase complex associated protein

RP: rentinitis pigmentosa

SMA: spinal muscular dystrophy

TALS/MOPD1: Taybi-Linder Syndrome/Microcephalic osteodysplastic primordial dwarfism type 1

TDP43: TAR DNA binding protein 43 kDa

FTD: Fronto temporal dementia

DM: Myotonic dystrophy

CCMS: Cerebro-Costo-Mandibular Syndrome

HS: Hypotrichosis Simplex

AR: androgen receptor

PC: prostate cancer

IP: immunoprecipitation

WB: western blot

PSI: percentage of splicing in

PIR: percentage of intron retention

KD: knock down