Gene expression regulation by well-known non-coding RNAs

Tan, Jiazi

2014

Tan, J. (2014). Gene expression regulation by well-known non-coding RNAs. Doctoral thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/64260

https://doi.org/10.32657/10356/64260

GENE EXPRESSION REGULATION BY WELL-KNOWN NON-CODING RNAS

TAN JIAZI

School Of Biological Sciences

A thesis submitted to the Nanyang Technological University in partial fulfillment of the requirement for the degree of Doctor of Philosophy

2014

ACKNOWLEDGEMENTS

In the course of my Ph.D. studies, I have had the utmost good fortune to work with some of the brightest minds in their fields. First and foremost, I would like to express my gratitude towards Asst. Professor Francesc Xavier Roca Castella, for being quite possibly the most patient person I have known. His invaluable supervision, guidance, and unwavering support in our investigation of RNA splicing has been, and still is, absolutely appreciated. In working on the X-chromosome inactivation projects, I would like to thank Asst. Professor Zhang Li Feng for his patient guidance and kind tutelage throughout my time in his laboratory. As events transpired, I could not complete my thesis there, thus transferring to Asst. Professor Roca's lab. The willingness of my supervisors to take me, an inexperienced researcher, as their very first Ph.D. student, is something that I will always be grateful for.

I would also like to show my undying gratitude to all my fellow lab members, past and present, for their unwavering support and assistance. Particular thanks go to Mdm Luo Shufang, for her help in constructing and transfecting some of the minigenes used in the study of 5' splice site recognition, Ms Malini Bhadra, for her help in proofreading this very thesis, Ms Tang Sze Jing, for her instructive help and advice, especially in regards to CD46, Mr. Shao Yu, for his assistance during DNA-FISH genotyping of the CH29-76M9-m*Tsix* ES cells, and Mdm. Chelliah Rosi, for her work on the pEZ-Frt-loxP-DT-zeo-RFP plasmid. Special thanks go out to my FYP student Ms Khoo Bee Luan, for without her dedication and hard work, the second XCI project would be still in its infancy.

In addition, I would like to express thanks to all my friends, and everyone who has rendered assistance to me in the course of my Ph.D. journey. Special thanks go out to Dr. Lawrence Ho Chun Loong and Dr. Chen Ming Wei, whose indefatigable friendship has helped me through both smooth and rough periods.

I would like to thank my parents, whom I adore immensely.

Last but not least, thank you, dear reader, for your time and patience.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Abbreviation	**Expanded form**
3'ss	3' splice site
5'ss	5' splice site
A	Adenosine
ABCC12	ATP-Binding Cassette, Sub-Family C (CFTR/MRP), Member 12
ARHGAP12	Rho GTPase activating protein 12
ATP	Adenosine Triphosphate
BAC	Bacterial artificial chromosome
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
BPRC	BACPAC Resources Center
BPS	Branch point sequence
BSA	Bovine serum albumin
C	Cytidine
CCDC132	Coiled-Coil Domain Containing 132
CD46	CD46 Molecule, Complement Regulatory Protein
CHORI	Children’s Hospital Oakland Research Institute
CRISPR	Clustered regularly interspaced short palindromic repeat
DAPI	4',6-diamidino-2-phenylindole
DC	dyskeratosis congenita
DHODH	Dihydroorotate Dehydrogenase (Quinone)
DMEM	Dulbecco’s Modified Eagle’s medium
DNA	Deoxyribonucleic acid
DNAI1	Dynein, Axonemal, Intermediate Chain 1
dNTP	Deoxyribonucleotide triphosphate
dUTP	2´-Deoxyuridine, 5´-Triphosphate
EB	Embryonic body
EDTA	Ethylenediaminetetraacetic acid
ES	Embryonic stem
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer

FACS	Fluorescence-activated cell sorting
FBS	Fetal bovine serum
FBXL13	F-Box And Leucine-Rich Repeat Protein 13
FISH	Fluorescence *in situ* hybridization
G	Guanosine
hnRNP	Heterogeneous nuclear ribonucleoprotein
Hprt	Hypoxanthine guanine phosphoribosyl transferase
HPS4	Hermansky-Pudlak Syndrome 4
ISE	Intronic splicing enhancer
ISS	Intronic splicing silencer
Itga4	Integrin alpha 4
kb	Kilobase
LB	Luria-Bertani
LIF	Leukemia inhibitory factor
MCAD	Acyl-coenzyme A dehydrogenase, C-4 to C-12 straight chain (*ACADM*)
MCS	Multiple cloning site
MEF	Mouse embryonic fibroblasts
miRNA	MicroRNA
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
ncRNA	Non-coding RNA
nt	Nucleotide
ori	Origin of replication
PAGE	Polyacrylamide gel electrophoresis
PAK3	P21 Protein (Cdc42/Rac)-Activated Kinase 3
PARP14	Poly (ADP-Ribose) Polymerase Family, Member 14
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PFA	Paraformaldehyde
PGK	Phosphoglycerate kinase
Phf6	Plant homeodomain finger gene 6
PIK3R4	Phosphoinositide-3-kinase, regulatory subunit 4

PNK	Polynucleotide kinase
POLQ	Polymerase (DNA directed), theta
PPT	Polypyrimidine tract
PRC	Polycomb repressive complex
pre-mRNA	Precursor mRNA
RA	Retinoic acid
RARS2	Arginyl-tRNA synthetase 2, mitochondrial
RFP	Red fluorescent protein
RNA	Ribonucleic acid
RNAi	RNA interference
Rnf12	Ring finger protein 12
RNF170	Ring Finger Protein 170
RPS6KC1	Ribosomal Protein S6 Kinase, 52kDa, Polypeptide 1
rRNA	Ribosomal RNA
RT-PCR	Reverse transcription PCR
SEMA3A	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3A
SGCE	Sarcoglycan, epsilon
shRNA	Short hairpin RNA
siRNA	Short interfering RNA
SLC5A8	Solute Carrier Family 5 (Sodium/Monocarboxylate Cotransporter), Member 8
snoRNA	Small nucleolar RNA
snoRNP	Small nucleolar ribonucleoprotein
SNP	Single nucleotide polymorphism
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
SR	serine/arginine
SSC	Saline sodium citrate
T	Thymidine
TAE	Tris/acetate/EDTA
TALEN	Transcription activator-like effector nuclease
TBE	Tris/borate/EDTA

Tm	Melting temperature
tRNA	Transfer RNA
Tsix	X (inactive)-specific transcript, antisense
U	Uridine
U2AF	U2 auxiliary factor
UMV	Universal Minigene Vector
UTR	Untranslated region
VWA3B	von Willebrand factor A domain containing 3B
Xa	Active X chromosome
XCI	X chromosome inactivation
Xi	Inactive X chromosome
XIC	X inactivation centre
Xist	X (inactive)-specific transcript
Zcchc5	Zinc-finger, CCHC domain containing 5
Zeo	Zeocin resistance gene
ZFN	Zinc-finger nuclease
ΔG	Minimum free energy
Ψ	Pseudouridine

ABSTRACT

In this thesis, I studied the functions of two well-known non-coding ribonucleic acids (ncRNAs), namely U1 small nuclear ribonucleic acid (snRNA) and *Xist*, which exert their influence over gene expression via different pathways.

5’ splice site (5’ss) recognition by U1 snRNA binding is one of the first steps in pre-mRNA splicing, which is critical for gene expression in eukaryotes. U1 classically base pairs to the 5’ss in a specific ‘canonical’ register, yet there is proof that other non-canonical registers exist. In this thesis, we verify the existence of non-canonical 1-nucleotide asymmetric loop registers, as well as present proof for the usage of non-canonical registers with 2 bulged nucleotides. We also show evidence which implies that bulge registers longer than 2 may not be tolerated. We also demonstrate that if the fifth intronic nucleotide of the 5’ss is a guanine, U1 always base pairs with it in the canonical register, despite thermodynamic predictions to the contrary. In addition, we report that a uridine residue at position +4 of the 5’ss can establish a non-canonical base pair with a pseudouridine in U1, thus contributing to 5’ss recognition. Our results extend our knowledge on the flexibility of the 5’ss/U1 RNA duplex structure that leads to productive splicing.

The *Xist* long ncRNA is essential for random X chromosome inactivation (XCI). XCI acts upon one of the two X chromosomes in female mammalian cells during differentiation, silencing most of the genes on that chromosome. In one project, we attempted to insert a second *Xist* gene into the single X chromosome of male murine embryonic stem (ES) cells to cause ectopic XCI upon differentiation, leading to cell lethality as important X-linked genes were inactivated. From there, we could screen for genes important for XCI by rescuing the differentiated transgenic ES cells by gene silencing. Although ectopic XCI was achieved, the expected 100% lethality did not materialize, preventing us from establishing the screen. In another project, we screened for activators of XCI. By transfecting male ES cells with sequences derived from a region of the X-chromosome known to be important for XCI, and screening for induction of ectopic XCI, our lab had previously identified four novel sequences that may contribute to XCI activation. Using the same strategy, we may have located yet another genomic sequence that fuels XCI activation.

1. INTRODUCTION

In this thesis, two immensely fascinating non-coding ribonucleic acids (ncRNAs) that regulate gene expression are examined. This touches on some very basic and very important themes of RNA biology, which are briefly reviewed in this introduction.

1.1. Nucleic acids

Nucleic acids are polymeric biomolecules that serve critical roles in all organisms. There are two major classes of nucleic acids, namely deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) (Soukup, 2001). They are made from nucleotide monomers, and exist as chains of nucleoside residues linked by phosphates. Each nucleotide consists of three components: a 5-carbon sugar, a nitrogenous base attached to the 1’ carbon, and a phosphate group attached to the 5’ carbon of the sugar (Figure 1.1A).

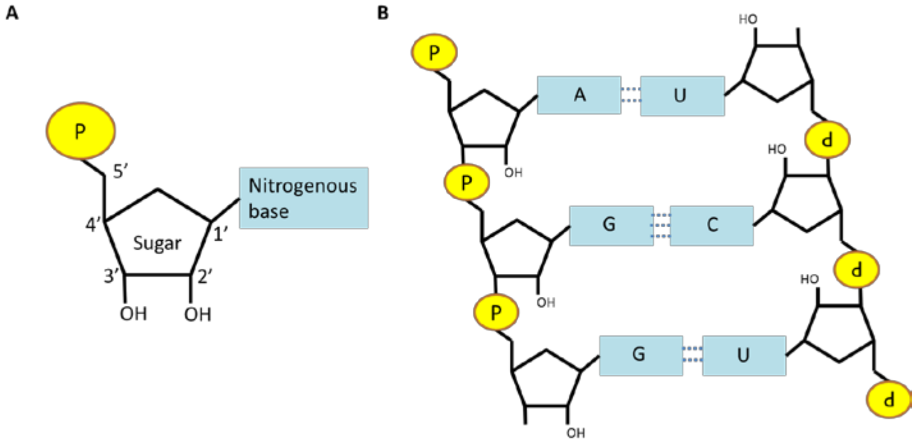


Figure 1.1 RNA nucleotides and chains
The schematic structure of: **(A)** a RNA nucleotide and **(B)** a pair of base paired RNA chains. The P indicates the phosphate group. The positions of carbons in the sugar ring are indicated by numbers. Dotted lines indicate hydrogen bonds.

1.2. RNA structure

In RNA, the 5-carbon sugar is ribose, which has a hydroxyl group attached to its 2’ carbon. This differentiates it from DNA, which instead contains deoxyribose, which lacks such a group (Soukup, 2001). During RNA assembly, the phosphate group is linked to the 3’ hydroxyl group of the preceding nucleotide by RNA polymerase(s), forming a chain that proceeds to elongate solely in the 5’ to 3’ direction. This grants RNA molecules directionality.

The nucleosides found in RNA are typically adenosine (A), guanosine (G), cytidine (C), and uridine (U), which have their respective adenine, guanine, cytosine, and uracil nitrogenous bases attached to a ribose. Further post-transcriptional modifications to these residues can and do occur (Soukup, 2001). One example would be the isomerization of uridine to pseudouridine (Ψ) by Ψ-synthase (Hamma and Ferré-D'Amaré, 2006).

The nitrogenous bases are capable of forming hydrogen bonds with a compatible partner base, a process known as base pairing. Watson-Crick base pairing occurs when hydrogen bonds are formed between opposing A-U/Ψ and G-C residues: two bonds between A-U/Ψ, and three bonds between G-C (Donohue and Trueblood, 1960; Traub and Elson, 1966). Other base pairs are possible; wobble base pairing allows the formation of G-U/Ψ base pairs with two hydrogen bonds (Chan et al., 1972), in addition to various other combinations (Lee and Gutell, 2004).

RNA is also capable of base pairing with DNA, and DNA with DNA, according to the same rules, although in DNA, uracil is replaced with thymine, forming the thymidine nucleoside (T). When base pairing occurs, the two nucleic acid strands will inevitably be oriented antiparallel to each other (Figure 1.1B). These base pairing interactions allow RNA to form duplexes with complementary nucleic acids. The more base pairs established, the higher the stability (minimum free energy, ΔG) of the duplex. Duplex stability can be estimated by measurement of the melting temperature (Tm), which is the temperature at which half of the nucleic acid strands are unpaired (Owczarzy et al., 1997). The higher the Tm, the more stable the duplex.

The sequence of nucleotides, each bearing their respective bases, determines the role of a particular RNA. This is because the formation of hydrogen bonds and other electrostatic interactions as well as hydrophobic stacking interactions between RNA bases and other molecules is essential for RNA function (Marz and Stadler, 2011). Since such interactions can be highly specific, and are dependent on the molecular geometries of the RNA base(s) present, RNAs can selectively bind to and recognize target regions of other molecules, especially other nucleic acids. Also, many RNAs contain self-complementary sequences, which are regions that can base pair with each other within the same RNA strand. This encourages folding, looping, and the formation of highly structured double helices, in a manner reminiscent of proteins (Tinoco Jr and Bustamante, 1999). This allows certain RNAs to perform catalytic activities (Lilley, 2011). Such secondary structures are also important for RNA interaction with other macromolecules. In particular, it allows specific proteins to precisely recognize and bind to the RNA, forming RNA-protein complexes that can act synergistically to perform important functions.

1.3. RNA classes and their functions

Biologically available RNAs can be divided into two groups: coding and non-coding RNAs. Coding RNAs, also known as messenger RNA (mRNA), serve as a template for protein expression. They act as a medium for the transfer of genetic information from DNA and its subsequent translation into proteins by ribosomes as per the Central Dogma of biology (Crick, 1970; Watson and Crick, 1953). The sequence of the mRNA determines the sequence of the resulting protein. In eukaryotes, they are transcribed from a DNA template in the nucleus. These nascent transcripts, also known as precursor mRNA (pre-mRNA), are then modified by RNA editing, by the addition of a 5’ cap, by splicing to remove intervening sequences (called introns) and join the protein-coding and/or untranslated regions (known as exons), as well as by polyadenylation at the 3’ end. Such modifications are essential to ensure proper protein expression and mRNA stability. The mature mRNA is then exported out of the nucleus to a ribosome to be translated into protein.

On the other hand, non-coding RNAs (ncRNA) consist of any RNA that is not translated into a protein. They are a vast and diverse group of RNA molecules, many of which participate in translation, RNA splicing, gene expression regulation, genomic defense, and even in regulation of chromatin structure (Mattick and Makunin, 2006).

The best characterized ncRNAs are ribosomal RNA (rRNA) and transfer RNA (tRNA), both of which participate in protein biosynthesis. rRNA is the RNA component of the ribosome, comprising roughly 50% of its total mass (in eukaryotes), the rest is taken up by protein (Ben-Shem et al., 2011). Ribosomes consist of two major subunits, namely the large and small subunits. The small subunit reads an mRNA sequence while the large subunit catalyzes the linking of amino acids to form a polypeptide chain. Both these functions are dependent upon the rRNA present in both subunits (Nissen et al., 2000). On the other hand, tRNA serves as the link between the genetic information encoded in the mRNA and the amino acid sequence of proteins. It determines which genetic sequence corresponds to which amino acid.

Small nuclear RNAs (snRNA) are ncRNAs that play an indispensable role in RNA processing, in particular RNA splicing (Matera and Wang, 2014). Each snRNA is an average of 150 nucleotides long, and found primarily in the nucleus of eukaryotic cells, hence the name. When complexed with their attendant proteins, they form small nuclear ribonucleoproteins (snRNPs). These RNA-protein complexes can assemble upon primary RNA transcripts, along with various other accessory proteins, to form the spliceosome (Wahl et al., 2009). This massive multi-molecular complex is responsible for pre-mRNA splicing, which removes the (non-coding) introns from the mRNA sequence and joins the (coding) exons together. The snRNA is responsible for recognizing the sequences within the RNA that act as splicing signals by binding to them, thus delineating exon-intron boundaries; some are also involved in the catalysis of the transesterification reactions that take place during splicing (Fica et al., 2013). One such snRNA, the U1 snRNA, recognizes 5’ splice site (5’ss) sequences on the RNA transcript. The binding of U1 snRNP to the 5’ss is one of the first steps in RNA splicing.

Mammalian cells contain an assortment of small ncRNAs, including small nucleolar RNAs (snoRNAs), microRNAs (miRNAs), short interfering RNAs (siRNAs), and small double-stranded RNAs (Huang et al., 2013). These RNAs regulate gene expression at many levels, and are processed by complex pathways from longer primary RNA transcripts. Most display distinctive temporal- and tissue-specific expression patterns; and some are imprinted. Small RNAs are known to control a wide range of developmental and physiological pathways in animals.

Long ncRNA includes any ncRNA longer than 200 nucleotides, distinguishing them from the small regulatory RNAs. They regulate gene-specific expression, control post-transcriptional mRNA processing, stability, and translation, and are involved in chromatin modification pathways (Kung et al., 2013). X-chromosome inactivation (XCI) is one of the best-characterized processes regulated by long ncRNA (Huang et al., 2013; Kung et al., 2013; Payer and Lee, 2008), in particular by the *Xist* ncRNA (Clemson et al., 1996). XCI occurs in female mammals in order to equalize gene dosage. *Xist* RNA is expressed from the future inactive X-chromosome during differentiation. The *Xist* transcripts coat the inactive X-chromosome, leading to irreversible chromatin modifications that involve the loss of active chromatin markers and the recruitment of repressive chromatin modifications, silencing the inactive X-chromosome. This inactive X-chromosome will henceforth be clonally maintained in the daughter cells.

1.4. Pre-mRNA splicing

1.4.1. The role and mechanisms of action of pre-mRNA splicing

A preponderance of eukaryotic genes encoding for proteins are transcribed into precursor-messenger RNAs (pre-mRNAs). In order to produce a mature, translatable mRNA from pre-mRNA, the introns must be excised and the exons ligated in an RNA processing step known as pre-mRNA splicing. Pre-mRNA splicing was discovered in 1977 when adenoviral transcripts in mammalian cells were found to contain sequences from noncontiguous sites in the viral genome (Berget et al., 1977; Chow et al., 1977).

A human gene, on average, contains 7.8 introns and 8.8 exons (Sakharkar et al., 2004). Approximately 80% of human exons are less than 200 nucleotides long. Human intron size is much more variable, and characteristically significantly larger, with an average length of about 3,000 nucleotides. About 10% of human introns extend to a length of more than 11,000 nucleotides.

The biochemical mechanism of splicing has been studied in various contexts. Specific sequences located at the exon-intron boundaries and within an intron determine its recognition and removal by the spliceosomal components (Sheth et al., 2006). Such sequences are known as *cis*-acting elements. The three essential reactive *cis*-acting elements on the pre-mRNA are the 5' splice site (5'ss), the 3' splice site (3'ss), which include a conserved polypyrimidine tract (PPT) in metazoans, and the branch point sequence (BPS) (Figure 1.3A) (Sheth et al., 2006).

Introns are removed from primary transcripts by cleavage at conserved sequences called splice sites. These sites are found at the 5' and 3' ends of introns, and are therefore termed the 5'ss and the 3'ss respectively. The majority of metazoan introns are of the major class or U2-type; they possess canonical GU – AG intron boundaries. This means the excised RNA sequence typically starts with a GU dinucleotide at its 5' end, and ends with AG at its 3' end. These consensus sequences are known to be critical, as altering one of the conserved nucleotides results in inhibition of proper splicing. Other splice site sequences are found that begin with the dinucleotide AU and end with AC;

these are spliced through a distinct but similar mechanism (Turunen et al., 2013).

Another essential *cis*-acting element is called the branch point sequence, located 18 to 40 nucleotides upstream from the 3' end of an intron. The branch point always contains an adenine, but it is otherwise loosely conserved (Gao et al., 2008).

Splicing consists of 2 sequential transesterification reactions (Figure 1.2B) (Matera and Wang, 2014). The first reaction involves the 2' hydroxyl group of a conserved intronic adenosine nucleotide in the BPS performing a nucleophilic attack upon the 5'ss phosphodiester bond at the upstream exon-intron junction. This forms a lariat intermediate and paves the way for the 3' hydroxyl group of the upstream exon to perform the second nucleophilic attack upon the 3'ss phosphate group, joining the two exons and releasing the intron lariat.

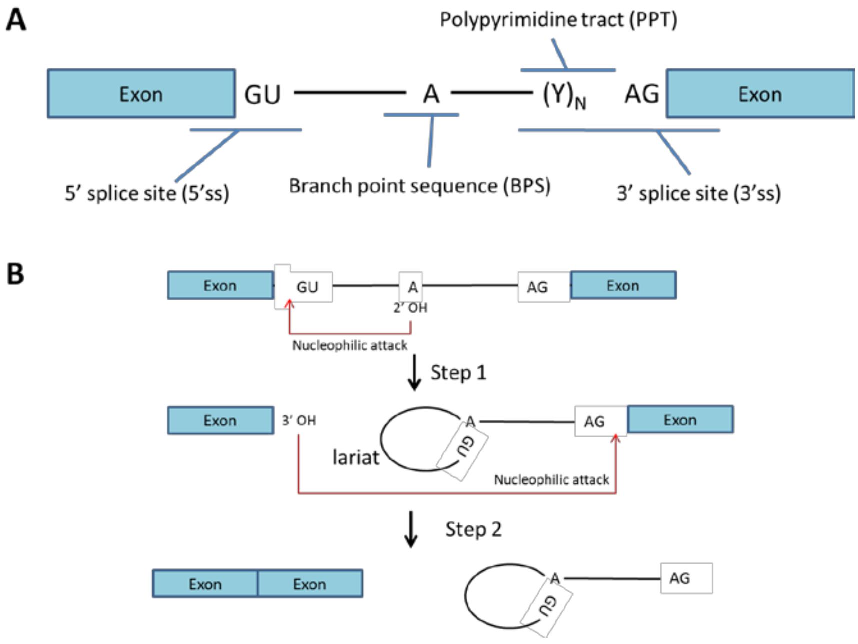


Figure 1.2 RNA splicing
(A) A representative schematic of a precursor RNA, displaying essential *cis*-acting elements for RNA splicing to occur. The region flanked by exons is the intron. **(B)** Transesterification steps in RNA splicing.

1.4.2. The spliceosome plays an essential role in RNA splicing

Splicing is a multi-step process catalyzed by a large multimeric complex known as the spliceosome (Konarska et al., 2006; Matera and Wang, 2014; Staley and Guthrie, 1998; Wahl et al., 2009). The spliceosome encompasses a core of 5 small nuclear ribonucleoproteins (snRNPs), namely U1, U2, U4, U5 and U6, as well as numerous other accessory proteins. Every snRNP particle itself comprises a small ribonucleic acid (snRNA) component, several Sm or LSm proteins that are shared across snRNPs, as well as a variable number of snRNP-specific proteins.

Each snRNP contains a single strand of snRNA about 150 nucleotides long, capable of base pairing with either the pre-mRNA substrate and/or interacting with other snRNAs (Matera and Wang, 2014; Wahl et al., 2009). The Sm or LSm proteins form cores that act as scaffolds or chaperones for the snRNA, thereby ensuring the snRNA adopts and retains the correct 3D confirmation essential for snRNP functionality. Other snRNP-specific proteins help establish protein-protein interactions between the snRNPs and the other *trans*-acting factors.

The spliceosome assembles stepwise upon the pre-mRNA substrate (Figure 1.3), forming a number of different complexes that position the reactive sites for productive splicing (Jamison et al., 1992; Matera and Wang, 2014; Staley and Guthrie, 1998; Wahl et al., 2009). Initially, the U1 snRNP binds to the 5’ss, while the non-snRNP proteins SF1 and U2 auxiliary factor (U2AF) bind to the BPS and the 3’ss respectively, forming the E complex in an ATP-independent fashion. Several rearrangements of the snRNPs occur in the presence of ATP: U2AF recruits the U2 snRNP, which displaces SF1 and binds to the BPS, thereby forming the A complex. After that, the preassembled U4/U6.U5 tri-snRNP is incorporated to form the B complex. Subsequently, another series of rearrangements cause the U1 and U4 snRNPs to be released, followed by the association of the U6 snRNP with the 5’ss and with the U2 snRNA. This rearrangement catalyzes the first transesterification reaction of splicing, and leads to C complex formation. Consequently, the second transesterification reaction occurs, releasing the intron lariat and joining the two exons. Both these transesterification reactions are catalyzed by the U6 snRNA, which positions divalent metal ligands that stabilize the leaving groups during each reaction (Fica et al., 2013).

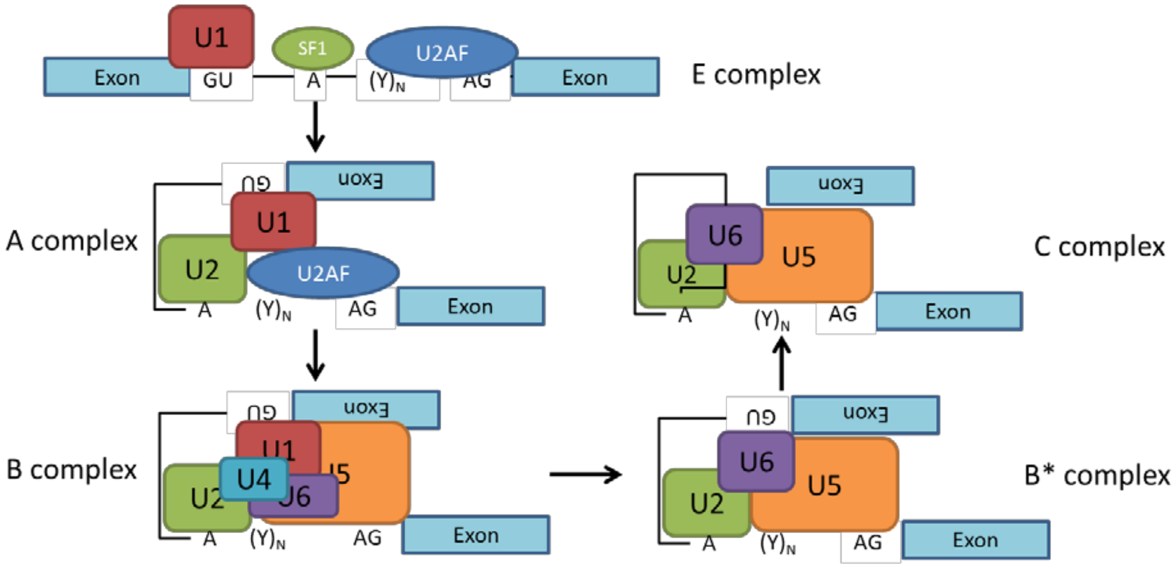


Figure 1.3 Spliceosomal assembly
This cartoon diagram (not to scale) displays the various stages of spliceosomal assembly upon the pre-mRNA substrate. U1, U2, U4, U5 and U6 are snRNPs, whereas SF1 and U2AF are spliceosomal accessory proteins.

1.4.3. Cis-acting elements and trans-acting factors

In simpler eukaryotes, the essential *cis*-acting elements (5’ss, 3’ss, BPS) are sufficient for productive splicing. These sites are conserved in simpler eukaryotes, but are highly degenerate in higher eukaryotes like humans. Therefore, more information is required for accurate splicing in more complex eukaryotes. This is provided by the presence of other *cis*-acting elements, like exonic splicing enhancer (ESE) and silencer (ESS) or intronic splicing enhancer (ISE) and silencer (ISS) sequences, which contribute to the correct demarcation of the intron/exon boundaries (Cartegni et al., 2002; Sheth et al., 2006). Such sequences are recognized by a wide variety of molecules known as *trans*-acting factors.

The binding of *trans*-acting factors to *cis*-acting elements drives RNA splicing (Wahl et al., 2009; Wang and Burge, 2008). SnRNPs and their accessory proteins bind to the 3 essential *cis*-acting elements. Splicing activators like serine/arginine (SR)-rich proteins usually recognize enhancer elements (Busch and Hertel, 2012). They help to recruit snRNPs and other spliceosomal proteins, thus enhancing the use of the splice sites. On the other hand, splicing repressors, like some of the heterogeneous nuclear ribonucleoproteins (hnRNPs) for example, typically bind to silencer elements (Cartegni et al., 2002). They act to restrict snRNP and spliceosomal protein binding, hence blocking the use of splice sites. If an enhancer and a silencer element are sufficiently close to each other, a *trans*-acting factor that is bound to one of the elements can physically block or limit the binding of a factor to the other element, consequently neutralizing its effect.

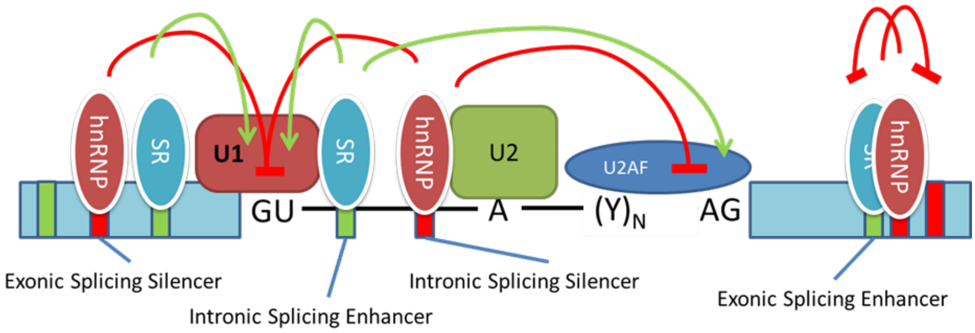


Figure 1.4 *Cis*-acting elements and *trans*-acting factors
A schematic showing the interplay of *cis*-acting elements and *trans*-acting factors. Red barred lines indicate inhibitory action. Green arrows indicate activatory action.

1.4.4. Alternative splicing

Alternative splicing (or differential splicing) is a process by which the exons of the RNA produced by transcription of a gene (a primary gene transcript or pre-mRNA) are reconnected in multiple ways during RNA splicing (Nilsen and Graveley, 2010). The complex interplay of *cis*-acting elements and *trans*-acting factors can lead to a wide variety of alternative splicing events (Figure 1.5). These then give rise to different isoforms from the same pre-mRNA transcript, generating a diversity of products with possibly varied functions. Alternative splicing allows a single gene to code for multiple proteins, enhancing the complexity of the proteome of eukaryotes.

There are several types of alternative splicing events, including cassette exons, mutually exclusive exons, alternate 5’ss, alternate 3’ss, intron retention, mutually exclusive 5’ untranslated regions (UTRs), and mutually exclusive 3’ UTRs. Cassette exon events result from the skipping versus inclusion of the alternatively spliced exon in the mature transcript. Mutually exclusive alternative splicing occurs when a pre-mRNA which contains contiguous exons includes only one or the other exon, but not both exons in the same mRNA. Intron retention, or lack of splicing, is also possible, and is typically observed with short introns. Alternative 5’ss arise when competing 5’ss are available. If an upstream 5' splice site is selected, the exon is truncated at its 3' end as compared to the downstream 5’ss. A parallel situation occurs when different 3’ splice sites are available. Alternative promoters that modify transcription start sites cause mutually exclusive first exons. Similarly, mutually exclusive last exons occur due to regulation of alternative polyadenylation sites.

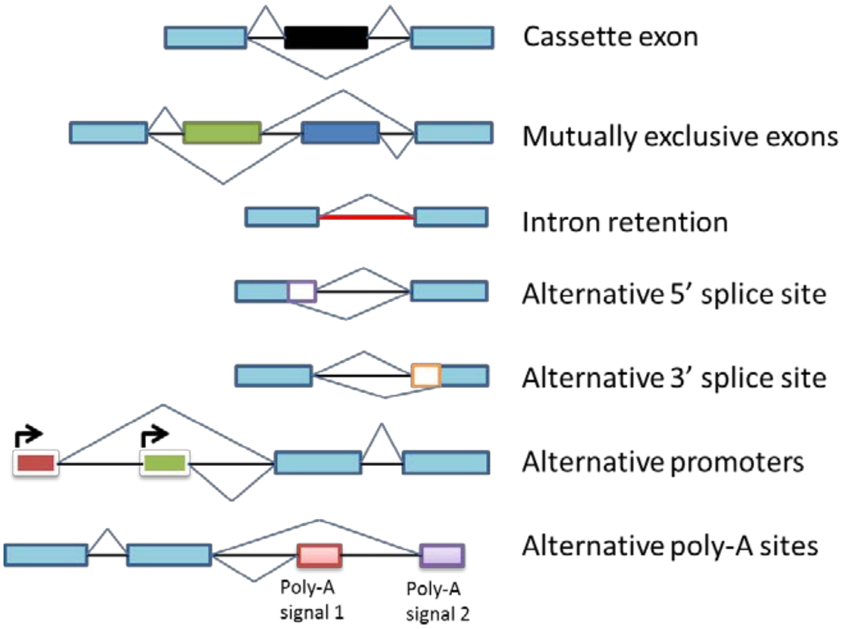


Figure 1.5 Alternative splicing
Examples of how different splicing patterns can give rise to different RNA isoforms.

1.4.5. U1 and 5’ splice site recognition

5’ss recognition is the first step in pre-mRNA splicing, in which the 5’ss is bound by the U1 snRNP particle, in particular by the 5’ end of the U1 snRNA (Hall and McLaughlin, 1991; Siliciano and Guthrie, 1988).

The U1 snRNA is 164 nucleotides long in humans. The 5’ tail of U1 snRNA was found to be complementary to the 5’ss consensus sequence, which is comprised of the most common nucleotide at each position in the 5’ss (Figure 1.6A, see for nucleotide position numbering convention). This allows them to establish up to 11 base pairs spanning 3 exonic nucleotides and 8 intronic nucleotides, forming a double helix (Roca et al., 2013). Therefore, this suggested that U1 snRNA base pairs with the 5’ss consensus sequence in a constant register (Lerner et al., 1980; Rogers and Wall, 1980), hereby called the “canonical register”.

Initially, the presence of a 5’ss sequence alone was assumed to be sufficient for 5’ss recognition and usage, with the consensus sequence being the optimal 5’ss. Additionally, the base pairing between the U1 and the 5’ss was expected to be via the canonical register across all 5’ss (Rogers and Wall, 1980).

However, this assumption was challenged by further research. There are about 200,000 known 5’ss sequences in humans, and most of them deviate from the consensus sequence (Figure 1.6B). A recent study revealed more than 9000 sequence variants in the -3 to +6 region of the 5’ss (Roca et al., 2012). Moreover, most pre-mRNAs contain multiple pseudo-5’ss: sequences that matched the 5’ss consensus sequence equally or better than the actual 5’ss, but were not used in splicing (Sun and Chasin, 2000). Furthermore, studies of β-globin pre-mRNA splicing detected cryptic 5’ss, which are sequences that are used as 5’ss when the natural 5’ss is inactivated (Roca et al., 2003; Treisman et al., 1983). In addition, it was discovered that the two mutually exclusive alternative 5’ss in the adenovirus E1A gene were used at different ratios relative to each other in a sequence-dependent fashion, implying competition between such 5’ss.

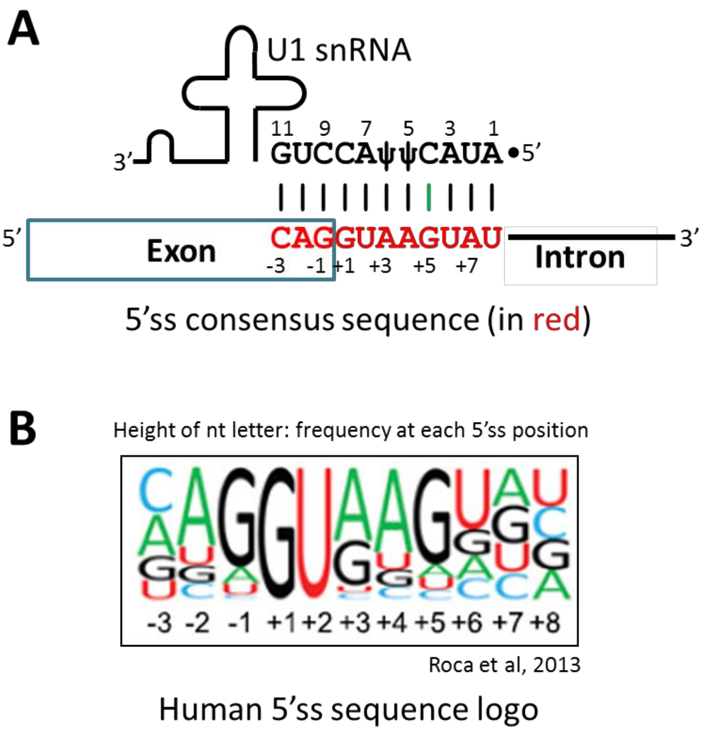


Figure 1.6 5'ss recognition and conservation
(A) 5'ss consensus sequence (nucleotides shown in red), canonical base pairing register shown. The nucleotide position numbering convention for 5'ss is as follows: exonic nucleotide numbers start at -1 and become increasingly negative while intronic nucleotide numbers start at +1 and become increasingly positive relative to the exon-intron junction. For U1, the nucleotide position numbers simply increase linearly in a 5'-to-3' fashion. Base pairs between the 5'ss and the U1 5' end are written with the 5'ss nucleotide first, then the corresponding U1 nucleotide. As an example, the base pair "+5G-C4" refers to the base pair in green. **(B)** Human 5'ss sequence logo as derived from >200,000 human 5'ss (Roca et al., 2013). The height of each nucleotide letter corresponds to its conservation at a particular 5'ss position.

The discovery of competitive alternative splicing made possible genetic tests of the role of U1 snRNA. Transfecting cells with "suppressor" U1 snRNA genes containing mutations in the 5' tail complementary to one of the two alternative 5'ss in adenovirus E1A transcripts caused the relative usage of the 5'ss to shift (Zhuang and Weiner, 1986). This reaffirmed the role of the 5' tail of U1 snRNA in 5'ss recognition, and also indicated that the strength of the base pairing interaction between U1 and the 5'ss has an effect on competition. Using mutant U1 snRNA genes to suppress mutations in 5'ss, thereby rescuing correct splicing in yeast (Siliciano and Guthrie, 1988) reinforced the importance of U1 snRNA. It also resolved the initial confusion regarding the role of U1 snRNA in yeast, as the 5' end of U1 snRNA is fully conserved but mismatched with the yeast consensus 5'ss (Siliciano et al., 1987).

5'ss competition experiments also allowed examination of the relationship between 5'ss sequences, their strength, and their U1 base pairing potential. 5'ss sequences could be tested by introducing them into plasmid constructs (as minigenes) in competition with an alternative 5'ss as a reference site, allowing relative comparison of their respective strengths. The first experiments demonstrated that in human (HeLa) cells, the consensus sequence was the most potent, being capable of silencing the reference 5'ss (Eperon et al., 1986). Several 5'ss were ranked based on their splicing efficiency, and these ranks correlated effectively with thermodynamic estimates of their base pairing strength.

1.4.6. 5'ss/U1 snRNA base pairing

As mentioned earlier, the maximum length of the double helix formed during 5'ss/U1 interaction is 11 base pairs, because the 12th nucleotide of U1 forms an internal base pair in stem I. The contribution of each 5'ss/U1 base pair to productive splicing is correlated to the conservation of the 5'ss positions (Figure 1.6B). Mismatches can be tolerated in 5'ss/U1 base pairing during splice site recognition, but only at certain positions.

The most highly conserved 5'ss positions are the first two intronic nucleotides, +1G and +2U, which display Watson-Crick base pairing with nucleotides A7 and C8 in U1 snRNA. Mutations in these positions typically completely abolish use of the 5'ss. The last exonic nucleotide, -1G, and the intronic +5G 5'ss positions are also strongly conserved in humans, forming G-C base pairs with U1. 5'ss nucleotide positions -2A, +3A, +4A, and +6U also contribute to U1 base pairing and thereby to 5'ss strength by forming relatively weaker A-U base pairs, accounting for their lower levels of conservation. -3C in the 5'ss forms a strong base pair with U1 but is also less conserved, possibly due to its proximity to the adjacent U1 stem I weakening the 5'ss/U1 interaction. Although positions +7 and +8 are poorly conserved in humans, they are still capable of base pairing to U1 and assisting in splicing (Hartmann et al., 2008);

base pairs at these positions enhance splicing kinetics in human cells and extracts (Freund et al., 2005).

The two Ψ in the U1 5’ end are indicated to play an important role in 5’ss recognition. They are highly conserved in across species. Only the 5’ end of U1, the region which interacts with 5’ss, contains Ψ (Wu et al., 2011). The presence of Ψ leads to an increase in base-stacking and extra hydrogen bonds between base and own phosphate backbone, contributes to stable intermolecular interactions. Since Ψ-A base pairs are stronger than U-A base pairs (Hudson et al., 2013), they may play a role in 5’ss/U1 helix stability during recognition.

1.4.7. Analyses of 5’ss strength

5’ss strength-scoring algorithms rely on either large-scale collections of genomic 5’ss sequences or computational estimates of 5’ss/U1 base pairing stability. Approaches that use the former criterion assume that the most commonly conserved 5’ss positions are the most efficient for productive splicing. 5’ss alignments were used to derive position-weight matrices (PWMs) which account for the frequency of each nucleotide at each position (Shapiro and Senapathy, 1987). Such a technique assumes 5’ss position independence, but evidence has been found for complex interdependencies between 5’ss positions (Roca et al., 2008). Other algorithms do take these links into consideration, like the maximum entropy models, first-order Markov models, and decision trees (Yeo and Burge, 2004). Also, overall 5’ss sequence patterns can be adapted by machine-learning approaches based on neural networks to infer 5’ss strength. Another effective process involves analyzing the frequency of the entire test 5’ss sequence across the pool of normal human 5’ss (Sahashi et al., 2007). The other group of techniques assume that U1 binding is the only force influencing 5’ss selection, with the most common method calculating the minimum free energy of each 5’ss/U1 helix using experimentally-derived thermodynamic parameters known as nearest-neighbor “Turner” rules (Mathews et al., 1999). Although all these algorithms can provide comparable rankings, and effective estimates of 5’ss strength, they are often inadequate versus the experimental data on 5’ss strength.

One possible explanation is that many such methods ignore the contribution of 5’ss positions +7 and +8. Also, most estimates assume base pairing in the canonical base pairing register, which is defined as U1 C8 base pairing to 5’ss +1G without any bulged nucleotides (Figure 1.6A).

Relatively recent mutational analyses and suppressor U1 experiments indicate that certain classes of 5’ss are recognized by alternate base pairing registers. A handful of ostensibly weak 5’ss were found to be efficiently selected due to U1 base pairing in a register shifted by 1 nucleotide upstream on the 5’ss, such that U1 C9 instead base pairs with 5’ss position +1G (Figure 1.7A) (Roca and Krainer, 2009). Further investigation (Roca et al., 2012) revealed that numerous other 5’ss could establish more stable base pairing interactions when a nucleotide is bulged on either the 5’ss (at positions +2 to +5) or the 5’ tail of U1 (primarily the pseudouridines, Ψ at positions 5 and 6) .

These base pairing schemes are collectively termed bulge/asymmetric loop registers. A bulge in a RNA (or DNA) duplex is defined as one or more nucleotides that are not opposed by any nucleotide on the other strand. These nucleotides bulge out and cause a kink in the helix. Similarly, asymmetric loops occur when an uneven number of unpaired nucleotides that are flanked by base pairs are present on both strands of the helix, forming a lopsided kink.

The shifted register is predicted to affect only a small number of 5’ss, 59 in humans. On the other hand, bulge/asymmetric loop registers occur far more regularly, with an estimated 5% of all human 5’ss (present in 40% of human genes) recognized via this format (Roca et al., 2012). It explains the efficient recognition of many authentic 5’ss otherwise predicted to be weak. These additional registers also increase the number of possible pseudo-5’ss present in the transcriptome.

These registers highlight the flexibility of the 5’ss/U1 interaction, which allows varied base pairing interactions to initiate 5’ss recognition and thus productive splicing. Another consequence of these registers is that relevant 5’ss positions may vary according to the type of register used; for example, the 5’ss +9 position might be base paired in the shifted +1 as well as certain bulge registers. More accurate scoring methods could be developed if these new registers were taken into account.

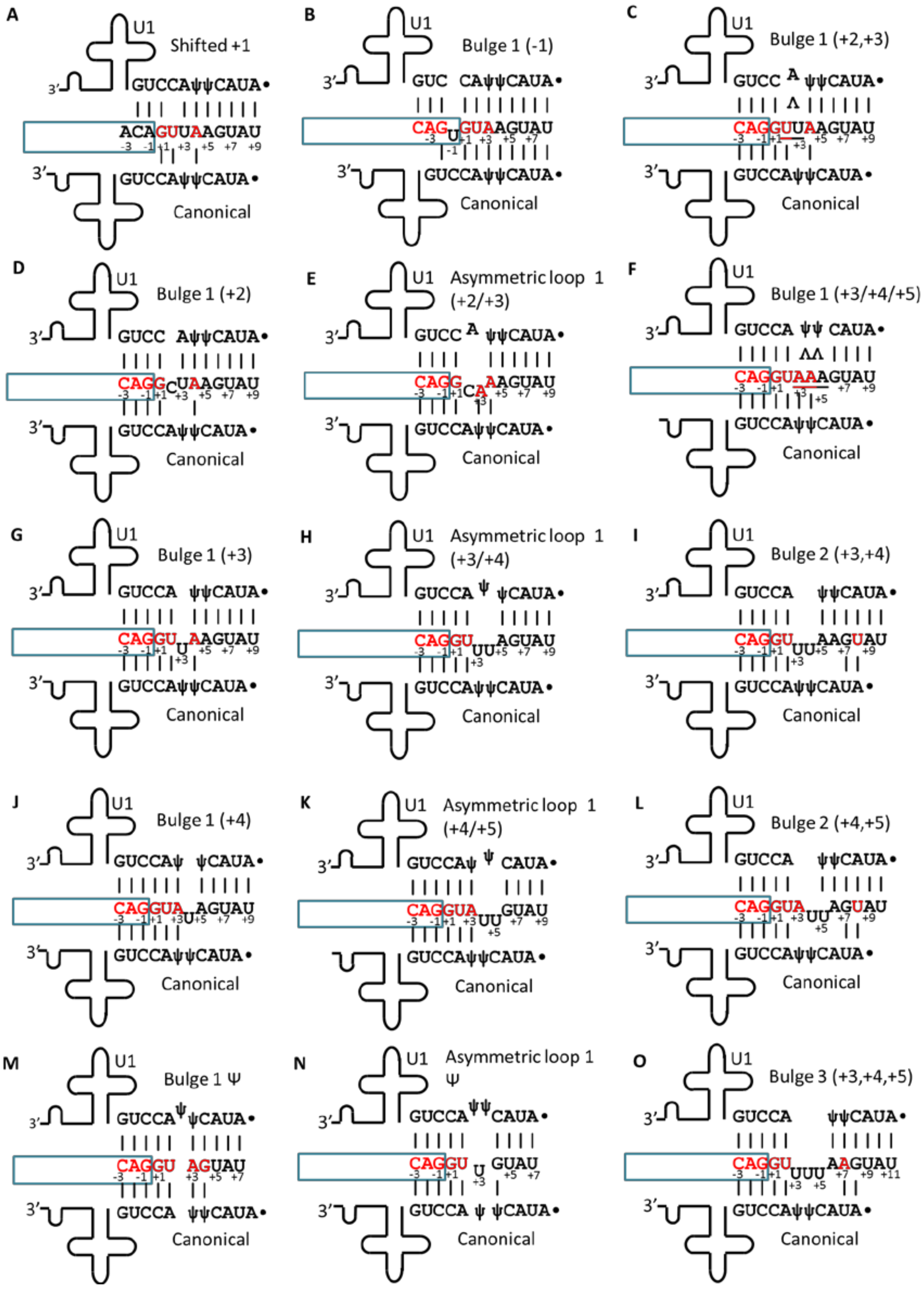


Figure 1.7 Examples of alternate 5'ss/U1 registers

Diagrams illustrating certain 5'ss/U1 base pairings and their naming convention. The blue box indicates the extent of the exon. The alternate register is displayed above the 5'ss sequence while the canonical register is displayed below for comparison of base pairing. These representative 5'ss sequences can establish the maximum possible number of base pairs with U1 in the alternate register. Red nucleotides in the 5'ss match those in the consensus 5'ss. Underlined nucleotides in the 5'ss indicate that they will bulge out if they do not base pair with a U1 5' end nucleotide, while

(Continued from page 18, **Figure 1.7**) an inverted “V” indicates that the U1 nucleotide at its apex can potentially base pair with both the nucleotides at its base. The bulge 1 (+5) register is omitted for space. It has an ideal 5’ss sequence of “CAG/GUAA**U**GUAU”, with the underlined nucleotide being bulged and the slash mark indicating the exon-intron junction.

1.4.8. Splicing disorders and disease implications

Elucidating the mechanisms that dictate 5’ss recognition and selection is important for our understanding of human genetics, in particular genetic diseases (Cooper et al., 2009). Around 10% of all disease-causing mutations affect either one of the two splice sites (Krawczak et al., 2007), and roughly half of such mutations affect 5’ss.

The two most important parameters of a splice site mutation are the severity and the molecular consequence of that mutation. Severity refers to the extent of reduction of correct splicing. The molecular consequences denote the effect on the final spliced product, which in order of frequency in humans, are: exon skipping, cryptic splice site activation, and intron retention. These two factors often correlate with disease severity. *Ab initio* predictions of mutation severity by 5’ss scoring methods comparing the wild-type 5’ss with the mutant one are usually accurate, as the larger the difference in strength, the more severe the effect (Roca et al., 2013). Normally, the higher the conservation of a particular 5’ss position, the more severe the disruption caused by a mutation and henceforth the disease.

Not all cases conform to the *in silico* predictions; a 5’ss +5 A-to-G transition in the *RARS2* gene causes pontocerebellar hypoplasia due to improper splicing. However, such a mutation would lead to an extra G-C base pair formation in the canonical 5’ss/U1 register, theoretically increasing the 5’ss strength. This phenomenon can be explained by taking the shifted +1 5’ss/U1 base pairing register into account (Roca and Krainer, 2009).

Therapies to rescue splicing defects, in particular 5’ss mutations that do not affect 5’ss positions +1 and +2, are being developed. Such approaches include the use of antisense oligonucleotides or larger RNA molecules capable of influencing splicing. By better elucidating the mechanisms of 5’ss recognition and selection, 5’ss mutations can be better diagnosed and effective treatments developed more efficiently.

Single-nucleotide polymorphisms (SNPs) can affect splicing signals. More than 1000 SNPs in the human genome map to natural human 5’ss (Roca et al., 2008). Usually, these deviations do not significantly alter 5’ss strength and use, but some can influence splicing (Lu et al., 2012). Improving our understanding of 5’ss selection will therefore also aid in the identification of SNPs that might alter splicing patterns with phenotypic consequences.

1.4.9. Testing new non-canonical 5’ss recognition registers

As mentioned earlier, prior work has authenticated that 5’ss positions +2 to +5 and the Ψ at U1 position 5 or 6 can be bulged in certain 5’ss/U1 RNA helices to form alternate registers for 5’ss recognition (Roca et al., 2012). In this research, a data set of 201,541 well-annotated human 5’ss sequences was generated, each sequence encompassing 15 nucleotides on either side of the exon-intron junction for a total length of 30 nucleotides. The base pairing register and minimum free energy (ΔG1) for each sequence and the 5’ end of U1 was estimated using a predictive algorithm known as UNAFold hybrid (Markham and Zuker, 2008). A second run of UNAFold calculated the free energies for these 5’ss by forcing canonical base pairing (ΔG2). If ΔG1<ΔG2, the 5’ss was predicted to base pair to U1 via a bulge register, comprising a total of 10,248 5’ss, or 5.1% of all analyzed 5’ss. These 5’ss were designated as “bulge 5’ss”.

Bulge 5’ss occurred in 6577 genes, amounting to 41% of the 15,894 genes covered by the data set. The energetic advantage of the bulge over the canonical register was calculated as the difference between ΔG1 and ΔG2 (ΔΔG). Results ranging from −0.1 to −4.9 kcal/mol were obtained. Bulge 5’ss in which the bulge register confers a substantial energetic advantage were defined as cases with a ΔΔG ≤ −1 kcal/mol.

The bulge 5’ss set with ΔΔG ≤ −1 kcal/mol comprised 6940 5’ss (3.4% of all 5’ss) that use a base pairing register with one bulged nucleotide. Of the registers with one bulged nucleotide, they experimentally validated the bulge 1

(+2,+3), bulge 1 (+3), bulge 1 (+3,+4,+5), bulge 1 (+4), bulge 1 (+5), and the bulge 1 Ψ registers (See Figure 1.7).

The rest of the untested predicted 1-nucleotide registers consisted of: a bulge at 5’ss position −1, also known as the bulge 1 (-1) register; a bulge at either position +3/+4 or +4/+5, which are now designated as the asymmetric loop 1 (+3/+4) or asymmetric loop 1 (+4/+5) registers respectively; and a bulge at GC 5’ss including the C at position +2, which are known as the bulge 1 (+2) or asymmetric loop 1 (+2/+3) registers (See Figure 1.7). In addition to single-nucleotide bulges, UNAFold also predicted many registers involving longer bulges at the 5’ss, ranging from 2 to 8 nt. These registers were not experimentally tested yet, but they would account for the recognition of 3294 5’ss (1.6% of total 5’ss). The number of candidates and the ΔΔG became smaller as the bulge length increased.

In order to study 5’ss recognition, we transfect human cell lines with splicing minigenes, which essentially consist of three exons and their intervening introns cloned into a suitable expression vector. These minigenes carry the test 5’ss at the junction of the second exon and the second intron. Transcription of this minigene generates a short pre-mRNA transcript which is then spliced. If the 5’ss is successfully recognized by U1, the exon is included in the final product, which can then be revealed via vector-specific RT-PCR and gel electrophoresis. However, if the 5’ss is not recognized, these other two outcomes may occur, leading to bands of different sizes appearing in the gel. Combinations of each possible outcome are possible. Once we know the typical splicing pattern of a particular minigene, mutational analysis can be performed, whereby point mutations are introduced. Mutations are made that affect both canonical and test registers, as well as mutations that only affect the test register, and changes in the splicing pattern.

If there is evidence of a shift in the splicing pattern due to the mutations, plasmids that code for U1 suppressors are co-transfected together with the mutant test minigenes. In this context, U1 suppressors are actually U1 snRNA with compensatory mutations that restore base pairing at the mutant 5’ss in either the canonical or test register. We can then see which U1 suppressor can rescue splicing, and from there infer the type of register used. A pictorial description of this testing methodology is shown below (Figure 1.8).

1.4.10. Non-canonical 5’ss recognition register testing objectives

As explained earlier, not all the bulge/asymmetric loop registers were experimentally verified. Therefore, the aim in this project was to characterize the other 1-nucleotide bulge/asymmetric loop registers, as well as to test whether the longer bulge registers predicted *in silico* by UNAfold hybrid are tolerated in 5’ss recognition by U1. This would permit the refinement of the dataset of candidate 5’ss.

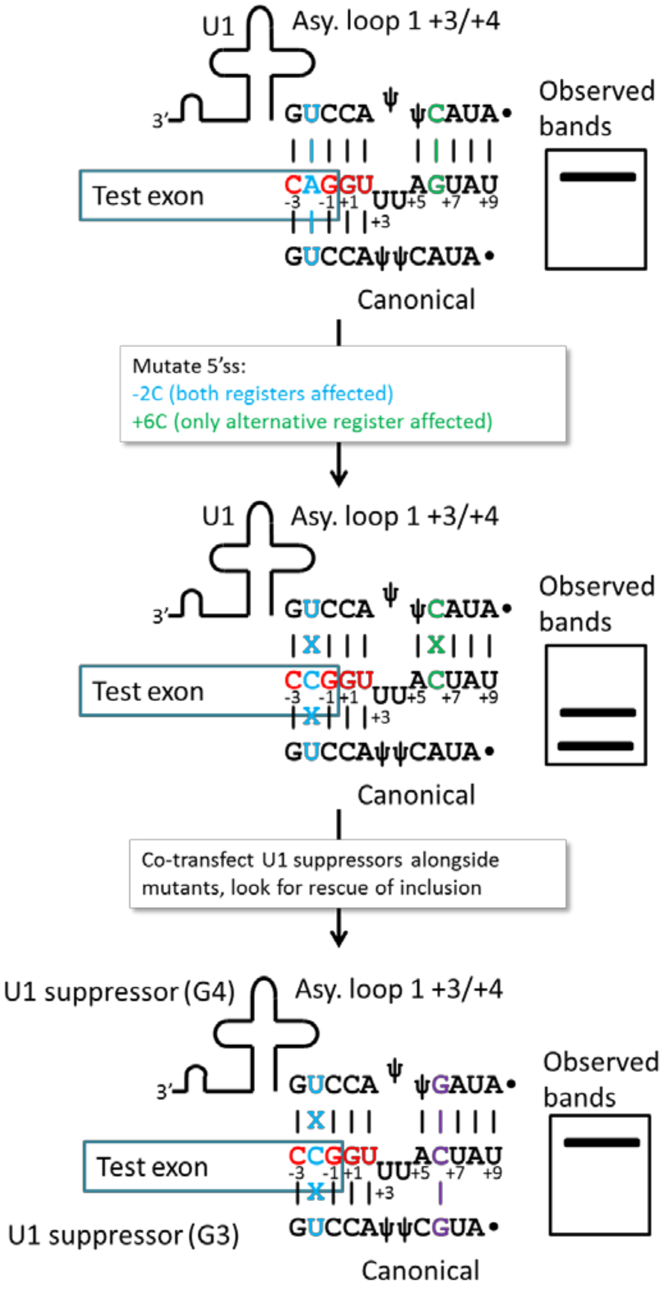


Figure 1.8 Workflow of test 5'ss mutational analysis and U1 suppressor usage

1.5. X-chromosome Inactivation (XCI): a crucial sex-specific epigenetic process

XCI is an epigenetic phenomenon that occurs specifically in female mammalian somatic cells (Lyon, 1961). In this elaborate multi-factorial process, one of the two X chromosomes in the female cell is rendered transcriptionally silent, eventually condensing into a discrete heterochromatic form known as a Barr body (Barr and Bertram, 1949). Instead of irreversible genomic sequence alterations, XCI efficiently produces stable, heritable chromatin structures that permit genes on the silenced chromosome to function in the next generation (Riggs and Porter, 1996).

Female mammalian cells contain two X chromosomes, as compared to the single X chromosome in males. This difference in X-linked gene copy number might lead to gene dosage problems if unregulated. Numerical aberrations in chromosome number, or aneuploidy, usually result in abortion, developmental abnormality, and mental retardation in humans (Hassold and Hunt, 2001). Human examples include trisomy 21, whereby one extra copy of chromosome 21 is present, leading to Down’s Syndrome (Patterson, 2009). XCI is thought to have evolved as a mechanism of gene dosage compensation (Payer and Lee, 2008) in order to prevent such a situation. It acts to balance X-linked gene expression levels between the sexes by inactivating one of the two copies of the X-chromosomes in female cells.

During early embryonic development, or in embryonic stem (ES) cells, both X chromosomes are transcriptionally active (Xa) (Figure 1.9). However, when these cells differentiate, each cell independently, randomly, and irreversibly inactivates one copy of the X chromosome (Xi), a process known as random XCI (Payer and Lee, 2008). From then on, all descendants of the cell will continue to inactivate the same X chromosome, making XCI clonally maintained. XCI and cell differentiation are interdependent, as disrupting one process will disturb the other (Lee, 2005; Silva et al., 2008).

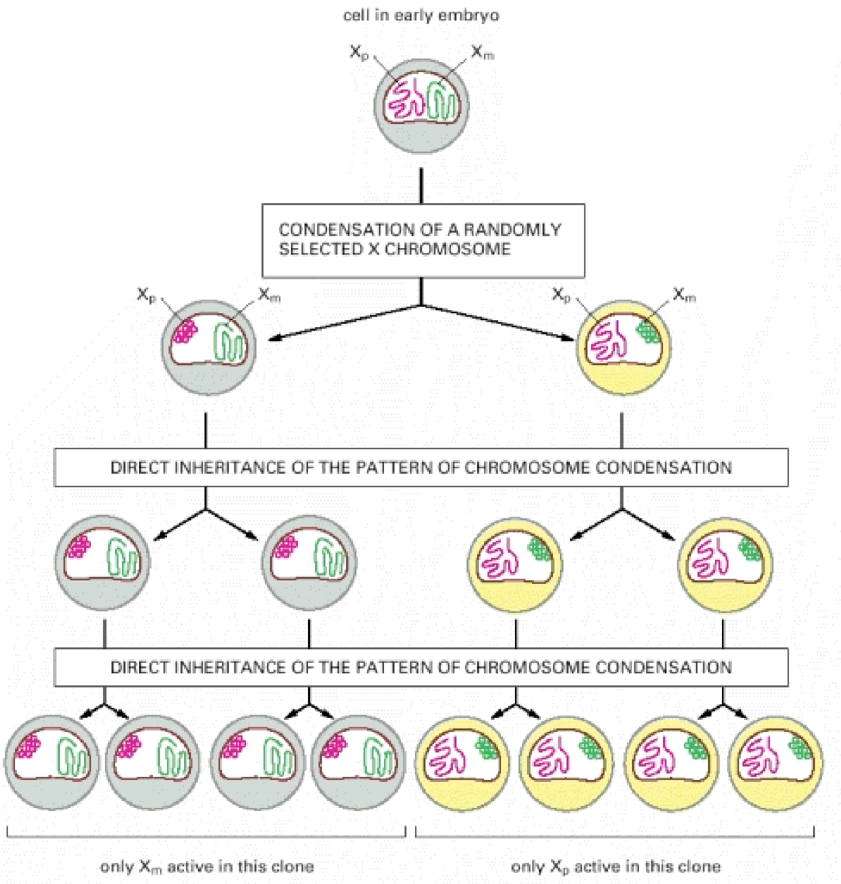

cell in early embryo
X_p
X_m
CONDENSATION OF A RANDOMLY SELECTED X CHROMOSOME
X_p
X_m
X_p
X_m
DIRECT INHERITANCE OF THE PATTERN OF CHROMOSOME CONDENSATION
DIRECT INHERITANCE OF THE PATTERN OF CHROMOSOME CONDENSATION
only X_m active in this clone
only X_p active in this clone

Figure 1.9 X-inactivation inheritance pattern
Adapted from Molecular Biology of the Cell, Fifth Edition (Boyle, 2008). Note how the same X chromosome is inactivated across all descendants of a particular cell.

1.5.1. Stages of random XCI

Random XCI can be separated into four major steps, namely counting, choice, silencing, and maintenance (Payer and Lee, 2008). This process has been studied in mice at length (Wutz, 2011).

1.5.1.1. Counting: Determination of X chromosome copy number

The cell decides whether to initiate XCI progression by determining the X-to-autosome ratio. Essentially, the cell "counts" the number of X chromosomes present in the cell (Payer and Lee, 2008; Wutz, 2011). Pluripotency factors like Oct4 (octamer-binding transcription factor 4) negatively regulate XCI counting in *trans* by influencing X-encoded dose-dependent activators of XCI (Donohoe et al., 2009). During differentiation, levels of the pluripotency factors decrease, allowing XCI activator levels to rise (Navarro et al., 2011). Each additional X chromosome in the cell nucleus further boosts the level of XCI activators, and thereby the chances of XCI occurring on any X chromosome (Monkhorst et al., 2008). Since male cells only have one X chromosome, they are usually unable to attain the threshold of XCI activators necessary to initiate XCI. Known *cis*-acting activators of XCI include the *RNF12* protein-coding gene (Jonkers et al., 2009) and the long ncRNA *Jpx12* (Tian et al., 2010).

1.5.1.2. Choice: Selection of X chromosomes to be inactivated

The female cell randomly selects the X chromosome that will remain active and the X chromosome(s) that will be inactivated. XCI initiation on any X chromosome within the nucleus is an independent probability event, according to the stochastic model of XCI (Monkhorst et al., 2008). The “blocking factor” model hypothesizes that a limited amount of autosomally-encoded dose-dependent blocking factor(s) binds to the Xa to-be, breaking the symmetry between X chromosomes and inhibiting XCI on a single X chromosome (Payer and Lee, 2008).

1.5.1.3. Silencing: the roles of Xist and Tsix

The X inactivation centre (XIC) is a 100 to 500kb-long region on the X chromosome (Lee et al., 1999b; Lee et al., 1996) that is essential for XCI to occur. Numerous sequences located in the XIC play an important role in X-inactivation (Payer and Lee, 2008; Rastan and Brown, 1990). Many such sequences contain genes that express long ncRNA. These include the X (inactive)-specific transcript (*Xist*) gene (Brown et al., 1991) and its antisense counterpart, *Tsix* (Lee et al., 1999a).

Xist codes for a spliced and polyadenylated ~18kb long ncRNA product in mice while its human counterpart produces a ~19kb long ncRNA (Flicek et al., 2014). *Xist* RNA acts in *cis* to elicit XCI, and is necessary for XCI to occur (Penny et al., 1996; Wutz and Jaenisch, 2000). *Xist* transcripts are exclusively confined to the cell nucleus, in particular to the nuclear territory of the Xi (Jonkers et al., 2008). *Xist* RNA accumulates on and propagates across the Xi-to-be (Clemson et al., 1996) in a two-step manner which targets gene-rich regions at the outset before spreading to other gene-poor domains (Simon et al., 2013), recruiting heterochromatic factors as they do.

The *Tsix* long ncRNA product negatively regulates *Xist* XCI induction (Luikenhuis et al., 2001) by affecting *Xist* chromatin configuration (Sado et al., 2005), masking important domains for silencing (Shibata and Lee, 2003), inducing DNA methylation of *Xist* (Navarro et al., 2006), engaging RNA interference (RNAi) pathways (Ogawa et al., 2008), as well as inhibiting the ability of *Xist* to target Polycomb group proteins to the X chromosome (Payer and Lee, 2008). Deletion of *Tsix* on one X chromosome in female cells skews XCI to that X chromosome (Lee and Lu, 1999).

In undifferentiated female cells, both *Xist* and *Tsix* ncRNA are expressed at low levels on the X chromosomes, forming pin-point signals upon RNA fluorescence in-situ hybridization (RNA-FISH) (Figure 1.10A). During differentiation, as levels of pluripotency factors (like Nanog, Sox2, Oct2/4) decline, transcription of *Xist* ncRNA is up-regulated on the putative Xi (Navarro et al., 2008). At the same time, *Tsix* expression declines on the Xi-to-be. *Xist* transcripts spread outwards from the XIC, accumulating within the chromosome territory of the Xi, coating it (Clemson et al., 1996). The recruitment of repressive chromatin modification factors and complexes, like polycomb repressive complexes PRC1 and PRC2 (Plath et al., 2003; Silva et al., 2003), by the *Xist* RNA causes most of the genes on the Xi to become transcriptionally inert (Payer and Lee, 2008). At this time point, *Xist* RNA can be visualized as a cloud signal within the cell nucleus by RNA-FISH (Figure 1.10B), which can be used as a marker of XCI. In contrast, *Tsix* transcription levels persist on the Xa, continuing to repress *Xist* transcription and spreading (Payer and Lee, 2008).

1.5.1.4. Maintenance of XCI

Once the cell is fully differentiated, and XCI successfully established, *Xist* RNA levels are clonally maintained on the Xi (Figure 1.10C), which is now heavily methylated with repressive marks on both the DNA and the histones (Beard et al., 1995; Panning and Jaenisch, 1996). Concurrently, neither *Xist* nor *Tsix* is expressed on the Xa (Payer and Lee, 2008). However, once the inactive state is established, *Xist* is not necessary for XCI maintenance (Brown and Willard, 1994; Csankovszki et al., 1999), as multiple repressive pathways synergize to prevent loss of silencing (Csankovszki et al., 2001).

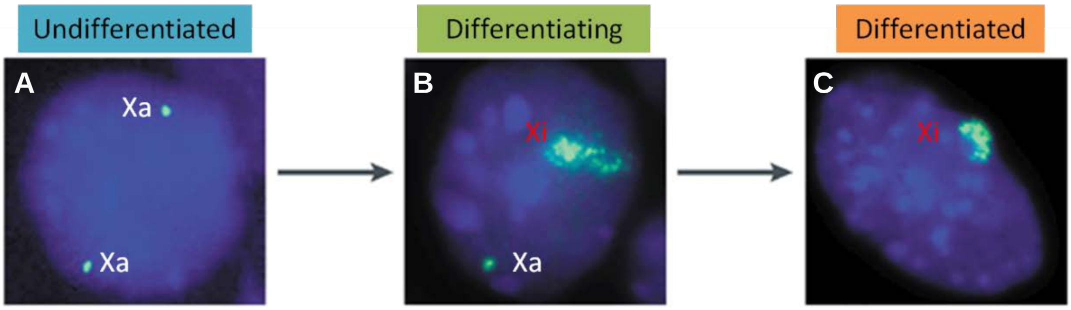


Figure 1.10 *Xist* RNA-FISH in different stages of ES cell differentiation
Xist RNA is labelled with a green FITC probe. ES cell nucleus is stained with DAPI (blue). Adapted from: "X-chromosome inactivation: counting, choice and initiation" (Avner and Heard, 2001).

1.5.2. XCI as a model for the study of epigenetics and related diseases

XCI is an epigenetic process that influences gene expression. The study of XCI enhances our understanding of the various epigenetic pathways that exist in cells, as well as how such pathways can control the flow of genetic information. However, many of the processes supporting XCI are poorly-characterized. In particular, counting and choice of X chromosomes for XCI needs to be further investigated (Payer and Lee, 2008) – although some progress has been made in this regard.

Epigenetic errors can and do lead to a wide variety of disorders, including cancer, autoimmune diseases, imprinting disorders, as well as developmental and behavioral problems (Agrelo and Wutz, 2010). Using XCI as a model for the study of epigenetics may lead to new breakthroughs in treatment of such diseases.

Recent work has also revealed the possibility of leveraging the chromosome-wide silencing functionality of XCI to rescue polyploidy-related syndromes. Inducible *XIST* transgenes have been used to effectively silence the extra chromosome in human induced pluripotent stem cells with trisomy 21, successfully relieving Down Syndrome phenotypes (Jiang et al., 2013). Therefore, mechanisms involved in XCI can be exploited in the generation of new therapeutic tools to cure or mitigate genetic/epigenetic disorders.

1.5.3. XCI projects: Objectives

In order to advance our understanding of the complex pathways at play in XCI, it becomes necessary to identify the genes that are involved in XCI, and thereby realize their function and purpose in the XCI process. Thus, two screening projects were conceived.

1.5.3.1. Project 1: Establishing a genetic screen for the identification of genes involved in XCI

In this project, we aimed to implement a genetic screen to identify genes involved in XCI. This would be executed by exploiting the ability of *Xist* RNA expression to cause chromosome-wide silencing by initiating XCI pathways.

We intended to insert an extra copy of the XIC or the *Xist* gene into the lone X chromosome of male murine ES cells. When differentiated, these cells were expected to trigger XCI due to miscounting, inducing cell death as X-linked genes would be inactivated. Screens could then be carried out by rescuing such cell mortality, silencing genes responsible for XCI in the transgenic ES cells with use of a lentiviral shRNA library, and identifying them from the lentiviral tags.

1.5.3.2. Project 2: Screen for activators of XCI

Rnf12 was the first *trans*-acting activator of XCI discovered (Jonkers et al., 2009). They presented evidence in the same paper that suggested the existence of other activators. The aim of this project was to search for these other XCI activators. The study was focused towards the HD2-HD3 breakpoint region on the X chromosome, which is essential for XCI (Rastan and Robertson, 1985) and encompasses the *Rnf12* locus.

The screen can be performed by integrating sequences from this region into the genome of male murine stem cells, and observing the resultant transgenic cell lines for any ectopic XCI triggered by increased XCI activator dosage.

2. MATERIALS AND METHODS

2.1. Cloning procedures

2.1.1. Universal Minigene Vector (UMV) construction

The Universal Minigene Vector was designed to facilitate the cloning and testing of many 5’ss in their native exonic context. A *MCAD* (officially defined as acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain, *ACADM*) minigene was used as the template (Figure 2.1). This minigene consists of exons 8 to 10 of the *MCAD* gene with internally-deleted intervening introns, each retaining 250 nt of their native 5’ and 3’ ends, resulting in 500 nt-long introns. This insert was cloned into a pcDNA +3.1 plasmid using HindIII and XhoI.

PCR mutagenesis was performed on the template plasmid to remove the middle exon and introduce KpnI and EcoRI restriction sites, using the forward primer *MCAD* B-Fwd (5’-gaatctga**ggtacctagcattagaattc**ctttgcaggccatatctctgtc-3’) and the reverse primer *MCAD* A-Rvs (5’-ctgcaaag**gaattctaatgctaggtacc**tcagattcaaacagagcacgca-3’); see section 0 for detailed steps. This caused exon 9 and its 250 nt flanking intronic sequences to be replaced by a 20 nt-long sequence consisting of KpnI and EcoRI restriction sites separated by a 8 nt spacer (see bold sequence in primers, underlined sequences are restriction sites). Bacterial cells were transformed with the PCR product and incubated overnight. Colonies were cultured overnight, minipreps were performed, and the plasmids were sequenced with the pcDNA-F primer to confirm the presence of the correct sequence. The resulting construct, now termed “UM”, was digested for 2 h at 37°C with KpnI and EcoRI (New England Biolabs, USA). The digested DNA fragment was resolved on 1% agarose gel in TAE buffer and gel-extraction was performed. The extracted DNA was then ethanol-precipitated.

In order to introduce a multiple cloning site (MCS) into the UM plasmid, first the short oligonucleotides *MCAD* MCS Fwd (5’-CCCGCGGGGATCCATCGATGCGGCCGCTTAATTAAG-3’) and *MCAD* MCS Rvs (5’-AATTCTTAATTAAGCGGCCGCATCGATGGATCCCCGCGGGGTAC-

3') were annealed by combining them in equimolar amounts (100 µM, 2 µl each) in 1x T4 Ligase Buffer (New England Biolabs). The mixture was heated at 95 °C for 5 min and then allowed to cool to room temperature for 1 h, permitting them to form an uneven-ended double-stranded oligonucleotide able to base pair with the sticky ends on the KpnI / EcoRI-digested UM plasmid. To phosphorylate the oligonucleotide ends, polynucleotide kinase (PNK) (New England Biolabs, USA) was added to the mixture, along with additional 10x T4 Ligase buffer and ddH$_2$O to ensure correct buffer dilution. This was subsequently incubated at 37 °C for 30 min and heat-inactivated at 65°C for 20 min.

To form the complete vector via ligation, the digested UM DNA was combined with the oligonucleotide mixture, together with T4 Ligase (New England Biolabs, USA) and supplementary 10x T4 Ligase buffer plus ddH$_2$O to ensure correct buffer dilution. The ligation product was incubated at 16 °C overnight. Bacterial cells were transformed with 4 µl of the ligation product, plated and incubated overnight on LB-ampicillin agar plates. Colonies were cultured overnight in liquid LB-ampicillin media. Plasmids were extracted and then double-digested with EcoRI and KpnI. Plasmids that displayed the correct band pattern on 1% agarose gel were sent for sequencing with pcDNA-F primer (Table 2.2). Midipreps were made of the plasmid with the correct sequence. This final construct was designated as the "Universal Minigene Vector" (UMV).

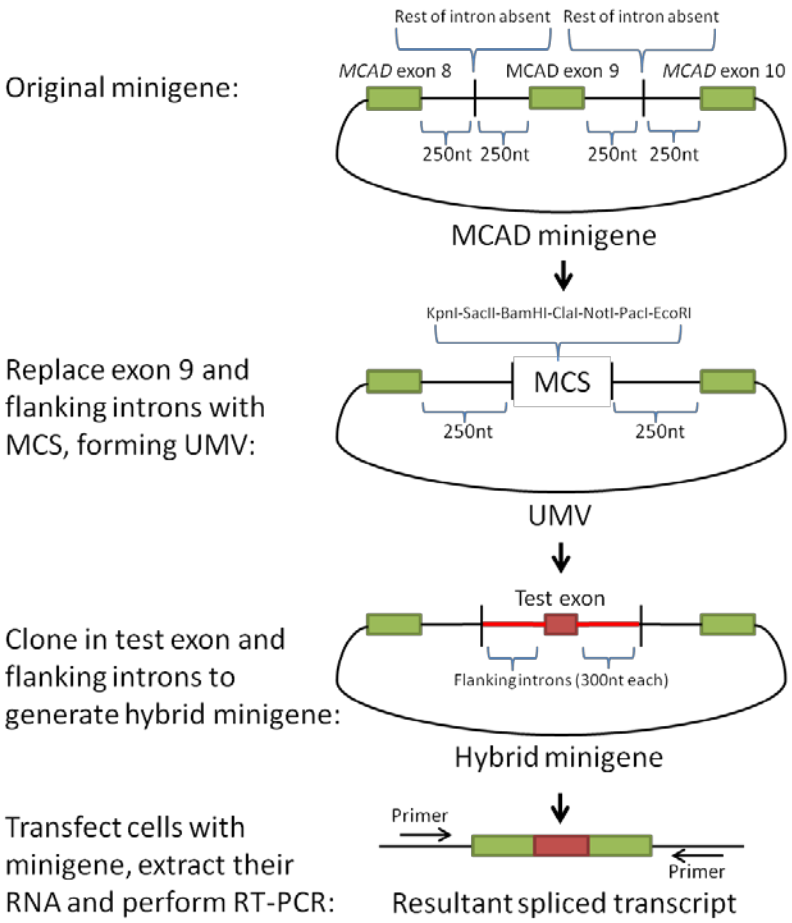

Original minigene:
Rest of intron absent
Rest of intron absent
MCAD exon 8
MCAD exon 9
MCAD exon 10
250nt 250nt
250nt 250nt
MCAD minigene
KpnI-SacII-BamHI-ClaI-NotI-PacI-EcoRI
Replace exon 9 and flanking introns with MCS, forming UMV:
MCS
250nt
250nt
UMV
Test exon
Clone in test exon and flanking introns to generate hybrid minigene:
Flanking introns (300nt each)
Hybrid minigene
Transfect cells with minigene, extract their RNA and perform RT-PCR:
Primer
Primer
Resultant spliced transcript

Figure 2.1 Universal Minigene Vector (UMV) construction and usage
The first two steps detail the construction of the UMV. The multiple cloning site (MCS) consists of the restriction sites listed in the diagram, in that order. The last two steps illustrate the cloning of the test exon into UMV to form the hybrid minigene, and the resultant mRNA product.

2.1.2. Strategy for test 5’ splice site identification and test exon selection

The list of naturally-occurring 5’ss predicted by the UNAFold hybrid tool (Markham and Zuker, 2008) with an energetic advantage in the bulge register versus the canonical register (ΔΔG) in the human genome has been described (Roca and Krainer, 2009). In order to identify potential test 5’ss, the list was sorted according to predicted bulge positions and then by energetic advantage, and only sequences with the highest ΔΔG were selected. These selected sequences, together with their corresponding exon and flanking introns, were checked in Ensembl (Flicek et al, 2013) for the following: correct positioning of the 5’ss sequence at the annotated exon-intron junction, suitable exon size (50-200 nt) and position (avoiding the first or last exons), as well as appropriate flanking intron sizes (>600 nt) for easy cloning. Alternative and/or infrequently used 5’ss on exons, as annotated on Ensembl, were discarded in order to increase the probability of test exon inclusion, which would facilitate further mutational analysis.

2.1.3. Cloning of test exons with flanking intronic sequences into UMV

High fidelity PCR, using primers bearing restriction sites on their 5’ ends (see Table 2.1) was employed to acquire DNA fragments consisting of test exons together with 300 nt of flanking intronic sequences on both ends (Figure 2.1). In a single reaction, the primers were adjusted to a final concentration of 200 nM each, together with 40 ng of human genomic DNA (Promega, USA) as the template, 25 μl of PrimeSTAR® Max DNA Polymerase Premix (Takara Bio, Japan), and additional ddH_2O to a final reaction volume of 50 μl. Reactions were run using the following thermocycler program: 35 cycles of 95 °C for 10 s, 55 °C for 5 s, and 72 °C for 5 s; then 4 °C forever. Reaction products were purified with Qiaquick® PCR Purification Kit (Qiagen, Germany). Both the purified product and UMV plasmids were digested for 2 h at 37 °C with their respective pair of restriction enzymes (New England Biolabs, USA), see list in Table 2.1. The resultant DNA fragments were separated by size on 1% agarose gel in 1x TAE buffer and gel-extracted.

To ligate the insert and vector, the digested PCR product and UMV were combined in a molar ratio of 8:1, along with T4 DNA Ligase (New England Biolabs, USA), and 10x T4 DNA Ligase Buffer plus ddH_2O to a final volume of 10 μl. The ligation was incubated at 16 °C overnight. DH5α cells were transformed with the ligation product. Minipreps were performed and extracted plasmids were double-digested with the same pair of restriction enzymes used earlier. Plasmids with the correct digestion pattern resolved in 1x TAE on 1% agarose gels were sequenced using UMV-seq-R primer (Table 2.2). Subsequently, plasmids with the correct sequence were midiprepped and used for transfection.

Table 2.1 List of primers used for cloning.

Gene with test 5’ss	Primer name	Primer sequence[1]	Restriction enzyme
ABCC12	*ABCC12*-BamHI-F	agtctctt**ggatcc**tgctgcttctaggagaaatacaccaagac	BamHI
	ABCC12-EcoRI-R	tatggaag**gaattc**aatggactggccactgctgtac	EcoRI
PARP14	*PARP14*-BamHI-F	caacattt**ggatcc**cttttccctgttgaattttaactgtttttctct	BamHI
	PARP14-EcoRI-R	gatttggt**gaattc**tgtctctcaaatttgtctattttggaaaatacatgt	EcoRI
SLC5A8	*SLC5A8*-BamHI-F	acaagaag**ggatcc**agtagtataactaggtacctttataaagctaacaat	BamHI
	SLC5A8-EcoRI-R	taaaaaac**gaattc**acaaacaacaacaacaacaaaaaactgtg	EcoRI
DNAI1	*DNAI1*-KpnI-F	cctgatgt**ggtacc**atttgttctgcctgctgggg	KpnI
	DNAI1-EcoRI-R	agcacact**gaattc**ctcttccagaaggcatatagtgctttc	EcoRI
CCDC132	*CCDC132*-KpnI-F	atggtctt**ggtacc**ggcagtattgagaacctggtctctg	KpnI
	CCDC132-EcoRI-R	ataaacag**gaattc**tcactgagaaagcatacatttctctatgatagtgg	EcoRI
FBXL13	*FBXL13*-KpnI-F	tatgaaaa**ggtacc**gaagttctggttttacaaagaacaaccttgtttaaa	KpnI
	FBXL13-EcoRI-R	actacaaa**gaattc**agctgggtgtggtggcag	EcoRI
RNF170	*RNF170*-BamHI-F	ttttgaag**ggatcc**caccaaagaaacaaatgttctttacattacttg	BamHI
	RNF170-EcoRI-R	tcagaaaa**gaattc**ttttctactacatttcattaaggtcaattagacag	EcoRI
HPS4	*HPS4*-BamHI-F	aaggtcag**ggatcc**tggctcgaggtaatagaacagacctcatggagatac	BamHI
	HPS4-EcoRI-R	actgaccg**gaattc**aggacctgttttaaacctatcccatcacagatccac	EcoRI
PIK3R4	*PIK3R4*-BamHI-F	actgttta**ggatcc**aagtccctgcagagaattaggc	BamHI
	PIK3R4-EcoRI-R	ttctttag**gaattc**tgtttctaatttttcttttatgag	EcoRI
POLQ	*POLQ*-BamHI-F	tgttatgc**ggatcc**taagagtagaagcagaaacatcttaggagaaatac	BamHI
	POLQ-NotI-R	tcatgcaa**gcggccg**caccaatcatttcttcaacaaatacttagtgagtc	NotI

1: Underlined primer sequences in bold indicate the restriction site sequence.

Table 2.2 List of primers used for sequencing and RT-PCR

Primer function	Primer name	Primer sequence
Reverse transcription	Oligo-dT	TTTTTTTTTTTTTTTTTT
Sequencing *SMN1/2* minigenes and their products, and radio-labeled PCR; forward	pCI-FwB	GACTCACTATAGGCTAGCCTCG
Radio-labeled PCR for *SMN1/2* minigenes; reverse	pCI-Rv	GTATCTTATCATGTCTGCTCG
Radio-labeled PCR for UMV minigenes; forward	pcDNA-F	GAGACCCAAGCTGGCTAGCGTT
Radio-labeled PCR for UMV minigenes; reverse	pcDNA-R	GAGGCTGATCAGCGGGTTTAAAC
Sequencing UMV minigenes; reverse	UMV-seq-R	cttgctacaatggcagaactg

Table 2.3 List of primers used for PCR mutagenesis of UMV minigenes.

Minigene	Primer function	Primer name	Primer sequence[1]
ABCC12	-2C, forward	*ABCC12* -2C F	CGTACATTAAGGCTTCTGG**C**Ggttcagtataaaacaacaagtttcttg
	+6C, forward	*ABCC12* 6C F	CGTACATTAAGGCTTCTGGAGgttca**c**tataaaacaacaagtttcttg
	+7C, forward	*ABCC12* 7C F	CGTACATTAAGGCTTCTGGAGgttcag**c**ataaaacaacaagtttcttg
	Common reverse	*ABCC12* cmn R	CAGAAGCCTTAATGTACGTGTGATATGTTTTCCAGGTCACGG
PARP14	-2C, forward	*PARP14* -2C F	CGAAGGCTAAAGATACACA**C**Ggttcagtaaagcttctaaattgagaagtg
	+6C, forward	*PARP14* 6C F	CGAAGGCTAAAGATACACAAGgttca**c**taaagcttctaaattgagaagtg
	+7C, forward	*PARP14* 7C F	CGAAGGCTAAAGATACACAAGgttcag**c**aaagcttctaaattgagaagtg
	Common reverse	*PARP14* cmn R	GTGTATCTTTAGCCTTCGGAATTTTGTCACTGACGAGATTTCC
SLC5A8	-2C, forward	*SLC5A8* -2C F	GAAAGTGTCTGCACCAGACC**C**Ggttcagtaccatgtctttcttacaggtg
	+6C, forward	*SLC5A8* 6C F	GAAAGTGTCTGCACCAGACCAGgttca**c**taccatgtctttcttacaggtg
	+7C, forward	*SLC5A8* +7C F	GAAAGTGTCTGCACCAGACCAGgttcag**c**accatgtctttcttacaggtg
	Common reverse	*SLC5A8* cmn R	GTCTGGTGCAGACACTTTCTTGGCTGTCCAAGGATCACAGTC
DNAI1	-2C, forward	*DNAI1* -2C F	GCTGACATCTATGGAGTCTC**C**Ggtttggtgttagttcctacagctctgcc
	+4C, forward	*DNAI1* 4C F	GCTGACATCTATGGAGTCTCAGgtt**c**ggtgttagttcctacagctctgcc
	+5C, forward	*DNAI1* 5C F	GCTGACATCTATGGAGTCTCAGgttt**c**gtgttagttcctacagctctgcc
	+4C+5C, forward	*DNAI1* 4C5C F	GCTGACATCTATGGAGTCTCAGgtt**cc**gtgttagttcctacagctctgcc
	+6C, forward	*DNAI1* 6C F	GCTGACATCTATGGAGTCTCAGgtttg**c**tgttagttcctacagctctgcc
	-2C+6C, forward	*DNAI1* -2C6C F	GCTGACATCTATGGAGTCTC**C**Ggtttg**c**tgttagttcctacagctctgcc
	Common reverse	*DNAI1* cmn R	AGACTCCATAGATGTCAGCTTCCTCATGGCCATCTTCCCTG
CCDC132	-2C, forward	*CCDC132* -2C F	TGGACTTACACGAATATGGC**C**Ggtttggtttttttaaaattatttttttc
	+4C, forward	*CCDC132* 4C F	TGGACTTACACGAATATGGCAGgtt**c**ggtttttttaaaattatttttttc
	+5C, forward	*CCDC132* 5C F	TGGACTTACACGAATATGGCAGgttt**c**gtttttttaaaattatttttttc
	+4C+5C, forward	*CCDC132* 4C5C F	TGGACTTACACGAATATGGCAGgtt**cc**gtttttttaaaattatttttttc
	+6C, forward	*CCDC132* 6C F	TGGACTTACACGAATATGGCAGgtttg**c**tttttttaaaattatttttttc
	-2C+6C, forward	*CCDC132* -2C6C F	TGGACTTACACGAATATGGC**C**Ggtttg**c**tttttttaaaattatttttttc
	Common reverse	*CCDC132* cmn R	CCATATTCGTGTAAGTCCATGTTCTAATTTCTTTTTTATGTAGCCACGAT
FBXL13	-1C, forward	*FBXL13* -1C F	AAAAAGAAAGAAGATGAGCT**C**gtattgtatattgaaacaatttttttaag
	+6C, forward	*FBXL13* 6C F	AAAAAGAAAGAAGATGAGCTGgtatt**c**tatattgaaacaatttttttaag
	+7C, forward	*FBXL13* 7C F	AAAAAGAAAGAAGATGAGCTGgtattg**c**atattgaaacaatttttttaag
	+8C, forward	*FBXL13* 8C F	AAAAAGAAAGAAGATGAGCTGgtattgt**c**tattgaaacaatttttttaag
	Common reverse	*FBXL13* cmn R	CTCATCTTCTTTCTTTTTACTCTTATGTCTTGCTGTATTCCGCC
RNF170	-1C, forward	*RNF170* -1C-F	GAACAGCTTCAAACAGAACA**C**gtattgtatatgtatttatttgaggag
	+6C, forward	*RNF170* +6C-F	GAACAGCTTCAAACAGAACAGgtatt**c**tatatgtatttatttgaggag

	+7C, forward	*RNF170* +7C-F	GAACAGCTTCAAACAGAACAGgtattgcatatgtatttatttgaggag
	+8C, forward	*RNF170* +8C-F	GAACAGCTTCAAACAGAACAGgtattgtctatgtatttatttgaggag
	Common reverse	*RNF170* cmn-R	TTCTGTTTGAAGCTGTTCTCGAAGTACCCTTACTAG
HPS4	-2C, forward	*HPS4* -2C F	CCTGTTTCCCTAGCTTATGCGgtacaagtatggggttgaggagtcttac
	+7C, forward	*HPS4* 7C F	CCTGTTTCCCTAGCTTATGAGgtacaactatggggttgaggagtcttac
	+8C, forward	*HPS4* 8C F	CCTGTTTCCCTAGCTTATGAGgtacaagcatggggttgaggagtcttac
	+9C, forward	*HPS4* 9C F	CCTGTTTCCCTAGCTTATGAGgtacaagtctggggttgaggagtcttac
	-2C +7C, forward	*HPS4* -2C7C F	CCTGTTTCCCTAGCTTATGCGgtacaactatggggttgaggagtcttac
	-2C +8C, forward	*HPS4* -2C8C F	CCTGTTTCCCTAGCTTATGCGgtacaagcatggggttgaggagtcttac
	-2C +9C, forward	*HPS4* -2C9C F	CCTGTTTCCCTAGCTTATGCGgtacaagtctggggttgaggagtcttac
PIK3R4	-2C, forward	*PIK3R4* -2C F	CAAATGGAAATTATGACACCGgttgtataattttctcttccagatttc
	+4C, forward	*PIK3R4* 4C F	CAAATGGAAATTATGACACAGgttctataattttctcttccagatttc
	+5C, forward	*PIK3R4* 5C F	CAAATGGAAATTATGACACAGgttgcataattttctcttccagatttc
	Common reverse	*PIK3R4* cmn R	TGTCATAATTTCCATTTGGATGTGTAACCTCATCTATTTCTTC
POLQ	-2C, forward	*POLQ* -2C F	AGTTCAGATGACATCGCTGCGgttgtatcatggggctagggatatag
	+4C, forward	*POLQ* 4C F	AGTTCAGATGACATCGCTGAGgttctatcatggggctagggatatag
	+5C, forward	*POLQ* 5C F	AGTTCAGATGACATCGCTGAGgttgcatcatggggctagggatatag
	+6C, forward	*POLQ* 6C F	AGTTCAGATGACATCGCTGAGgttgtctcatggggctagggatatag
	-2C +6C, forward	*POLQ* -2C6C F	AGTTCAGATGACATCGCTGCGgttgtctcatggggctagggatatag
	Common reverse	*POLQ* cmn R	AGCGATGTCATCTGAACTGGTGAAAGAAAAACTGTGGCC

1: Mutated nucleotide in red, 5’ss underlined, exonic nucleotides capitalized

Table 2.4 List of primers used for PCR mutagenesis of *SMN1/2* minigenes

Minigene	Primer function	Primer name	Primer sequence[1]
SMN1/2	Common reverse	R-SMN-mutag	TTAATTTAAGGAATGTGAGCACCTTCC
	Mutate 5'ss to bulge 2 +3,+4 register, forward	SMN_bulge2nt_3_4_F	CTCACATTCCTTAAATTAACAGgtttaagtatgccagcattatgaaagtg
	Mutate 5'ss to bulge 3 +3,+4,+5 register, forward	SMN_bulge3nt_3_4_5F	CTCACATTCCTTAAATTAACAGgttttaagtatgccagcattatgaaagt
	Mutate 5'ss to bulge 2 +3,+4 –ISS, forward	SMN-ISS-Bulge2-F	TAACAGgtttaagtatgcccgcattatgaacgtgaatcttacttttgtaa
	Mutate 5'ss to bulge 2 +3,+4 –ISS, reverse	SMN-ISS-Bulge2-R	ttacaaaagtaagattcacgttcataatgcgggcatacttaaacCTGTTA
	Mutate 5'ss to bulge 3 +3,+4,+5 -ISS, forward	SMN-ISS-Bulge3-F	AACAGgttttaagtatgcccgcattatgaacgtgaatcttacttttgtaa
	Mutate 5'ss to bulge 3 +3,+4,+5 -ISS, reverse	SMN-ISS-Bulge3-R	ttacaaaagtaagattcacgttcataatgcgggcatacttaaaacCTGTT
	Mutate bulge 2 +3/+4 –ISS to -1C, forward	SMN_b2nt_3_4_-1C-ISS	CTCACATTCCTTAAATTAACA**C**gtttaagtatgcccgcattatgaacgtg
	Mutate bulge 2 +3/+4 –ISS to +5C, forward	SMN_b2nt_3_4_5C-ISS	CTCACATTCCTTAAATTAACAGgttt**c**agtatgcccgcattatgaacgtg
	Mutate bulge 2 +3/+4 –ISS to +6C, forward	SMN_b2nt_3_4_6C-ISS	CTCACATTCCTTAAATTAACAGgttta**c**gtatgcccgcattatgaacgtg
	Mutate bulge 2 +3/+4 –ISS to +7C, forward	SMN_b2nt_3_4_7C-ISS	CTCACATTCCTTAAATTAACAGgtttaa**c**tatgcccgcattatgaacgtg
	Mutate bulge 2 +3/+4 –ISS to +8C, forward	SMN_b2nt_3_4_8C-ISS	CTCACATTCCTTAAATTAACAGgtttaag**c**atgcccgcattatgaacgtg
	Mutate bulge 2 +3/+4 –ISS to +9C, forward	SMN_b2nt_3_4_9C-ISS	CTCACATTCCTTAAATTAACAGgtttaagt**c**tgcccgcattatgaacgtg
	Mutate 5'ss to asymmetric loop 1 +3/+4, forward	SMN asyloop1nt_3_4 F	CTCACATTCCTTAAATTAACAGgtttagtatgccagcattatgaaagtg
	Mutate asymmetric loop 1 +3/+4 to -2C, forward	SMN aslp1nt_3_4 -2C	CTCACATTCCTTAAATTAAC**C**Ggtttagtatgccagcattatgaaagtg
	Mutate asymmetric loop 1 +3/+4 to +6C, forward	SMN aslp1nt_3_4 6C	CTCACATTCCTTAAATTAACAGgttta**c**tatgccagcattatgaaagtg
	Mutate asymmetric loop 1 +3/+4 to +7C, forward	SMN aslp1nt_3_4 7C	CTCACATTCCTTAAATTAACAGgtttag**c**atgccagcattatgaaagtg
	Mutate asymmetric loop 1 +3/+4 to -8C, forward	SMN aslp1nt_3_4 8C	CTCACATTCCTTAAATTAACAGgtttagt**c**tgccagcattatgaaagtg
	Mutate asymmetric loop 1 +3/+4 to +9C, forward	SMN aslp1nt_3_4 9C	CTCACATTCCTTAAATTAACAGgtttagta**c**gccagcattatgaaagtg
	Mutate 5'ss to asymmetric loop 1 +4/+5, forward	SMN asyloop1nt_4_5 F	CTCACATTCCTTAAATTAACAGgtattgtatgccagcattatgaaagtg
	Mutate asymmetric loop 1 +4/+5 to -2C, forward	SMN aslp1nt_4_5 -2C	CTCACATTCCTTAAATTAAC**C**Ggtattgtatgccagcattatgaaagtg

	Mutate asymmetric loop 1 +4/+5 to +6C, forward	SMN aslp1nt_4_5 6C	CTCACATTCCTTAAATTAACAGgtatt**c**tatgccagcattatgaaagtg
	Mutate asymmetric loop 1 +4/+5 to +7C, forward	SMN aslp1nt_4_5 7C	CTCACATTCCTTAAATTAACAGgtattg**c**atgccagcattatgaaagtg
	Mutate asymmetric loop 1 +4/+5 to +8C, forward	SMN aslp1nt_4_5 8C	CTCACATTCCTTAAATTAACAGgtattgt**c**tgccagcattatgaaagtg
	Mutate asymmetric loop 1 +4/+5 to -2C +7C, forward	SMN aslp1nt_4_5 -2C7C	CTCACATTCCTTAAATTAAC**C**Ggtattg**c**atgccagcattatgaaagtg
	Mutate asymmetric loop 1 +4/+5 to -2C +8C, forward	SMN aslp1nt_4_5 -2C8C	CTCACATTCCTTAAATTAAC**C**Ggtattgt**c**tgccagcattatgaaagtg
	Mutate 5’ss to asymmetric loop 1 Ψ, forward	SMN_Asyloop_psi_F	CTCACATTCCTTAAATTAACAGgttgtatgccagcattatgaaagtg
	Mutate asymmetric loop 1 Ψ to -2C, forward	SMN_Asyloop_psi -2C	CTCACATTCCTTAAATTAAC**C**Ggttgtatgccagcattatgaaagtg
	Mutate asymmetric loop 1 Ψ to +4C, forward	SMN_Asyloop_psi 4C	CTCACATTCCTTAAATTAACAGgtt**c**tatgccagcattatgaaagtg
	Mutate asymmetric loop 1 Ψ to +5C, forward	SMN_Asyloop_psi 5C	CTCACATTCCTTAAATTAACAGgttg**c**atgccagcattatgaaagtg
	Mutate asymmetric loop 1 Ψ to +6C, forward	SMN_Asyloop_psi 6C	CTCACATTCCTTAAATTAACAGgttgt**c**tgccagcattatgaaagtg

1: Mutated nucleotide in red, 5’ss underlined, exonic nucleotides capitalized

2.1.4. Testing 5’ss in heterologous context, SMN1/2

The *SMN1/2* minigenes were used extensively in previous work (Cartegni et al., 2006; Roca and Krainer, 2009; Roca et al., 2012). In brief, the *SMN1/2* minigenes consist of exons 6 to 8 as well as a truncated intron 6 (retaining only 62 nt of the 5’ end and 139 nt of the 3’ end) and a full-length intron 7 of the *SMN1*/2 genes cloned into the pCI vector. Although the *SMN1* and *SMN2* pre-mRNAs are paralogous, *SMN1* tends to exhibit high levels of exon 7 inclusion, but exon 7 is predominantly skipped in *SMN2*. This is due to a point mutation (C6T) in exon 7 of *SMN2* which disrupts an exonic splicing enhancer critical for exon 7 inclusion (Monani et al., 1999). Therefore, by replacing the native 5’ss of exon 7 in the *SMN1*/2 minigene with ideal test 5’ss sequences, which can establish the maximum possible number of base pairs with the U1 5’ end via the predicted non-canonical register (via PCR mutagenesis), it becomes possible to examine the efficiency of the test 5’ss in two different heterologous contexts. Primers used to introduce test 5’ss and point mutations are listed in Table 2.4. See Figure 2.2A for illustrated diagram.

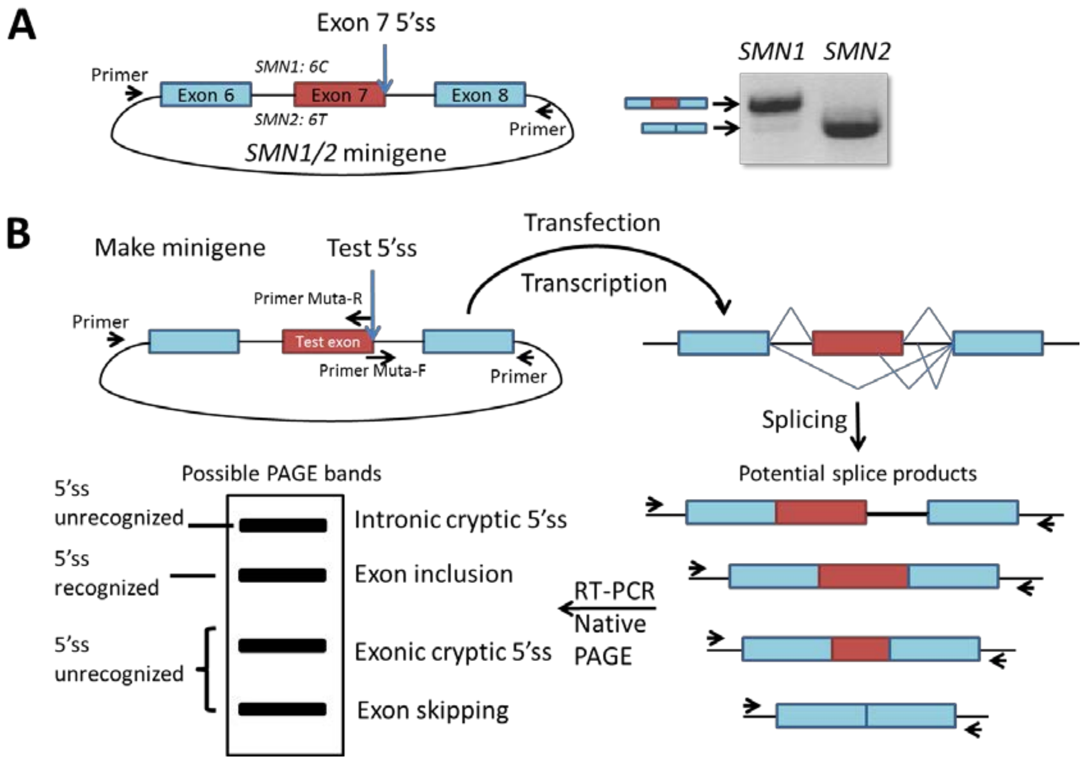


Figure 2.2 Minigene system for testing 5'ss
(A) A schematic of the *SMN1/2* minigene design on the left, and the typical splicing pattern of wild-type *SMN1* and *SMN2* minigenes on the right. **(B)** The workflow for a typical minigene experiment, with an explanation of potential results.

2.1.5. U1 suppressors

U1 suppressor plasmids (Zhuang and Weiner, 1986) code for a full-length version of U1 snRNA with the respective mutations introduced via PCR mutagenesis in the 5' tail. This allows the U1 suppressors to establish more base pair(s) with the mutant 5'ss, thereby promoting splicing by acting as native U1 snRNA, "suppressing" the effect of the mutation. By co-transfecting cells with U1 suppressor plasmids together with the test minigene plasmids, the effects of the suppressors on splicing of the test minigene transcript can be evaluated. See section 2.2 for transfection protocols, and Figure 1.8 for an example of U1 suppressor usage.

2.1.6. PCR mutagenesis

In order to incorporate the test 5’ss or introduce point mutations into the minigenes, PCR mutagenesis was performed on template plasmids using HiFi PCR (Kapa Biosystems, USA) kits. The sequences of the primers used can be found in Table 2.3 and Table 2.4.

Common reverse primers and specific forward primers bearing the mutation are used for each particular 5’ss. The 3’ ends of the forward primers and the reverse primer are complementary to each other. Each reaction consisted of 20 ng of template, primers at final concentration of 300 nM each, dNTP mix at a final concentration of 300 µM each, 5 µl of 5x HiFi Buffer, and 0.5 U of polymerase, with ddH$_2$O added to achieve a total reaction volume of 25 µl. Reactions were run using the following thermocycler program: an initial denaturation at 95 °C for 2 min; then 18 cycles of denaturation at 98 °C for 30 s, annealing at 50 °C for 1 min, and extension at 72 °C for 5 min; followed by a final extension at 72 °C for another 5 min; and finally 4 °C forever for storage. The PCR products were incubated with DpnI (New England Biolabs) at 37 °C for 2 h to digest template DNA, and then at 80 °C for 20 min to inactivate the DpnI. DH5α cells were transformed with the mixture; colonies were picked and cultured in liquid LB-ampicillin overnight. Minipreps were made and all mutants were verified by DNA sequencing. Midipreps were made from the confirmed clones.

2.1.7. Protocols for cloning, bacterial cell culture, and plasmid extraction

All bacterial transformations were performed as follows: 4 µl of the plasmid construct was mixed with 50 µl of chemically-competent DH5α *E. coli* cells. The mixture was incubated on ice for 30 min and then heat-shocked at 42 °C for 45 s. After 3-5 min recovery on ice, 946 µl of liquid LB media was added to the mixture. This was followed by incubation at 37 °C for 1 h with shaking. Cells were spun down at 10,000 g for 1 min and 900 µl of the supernatant was discarded. The cells were re-suspended in the remaining supernatant by gentle pipetting, before being transferred to LB-ampicillin agar plates (LB supplemented with 70 µg/ml of ampicillin) (Merck, USA) and incubated at 37 °C overnight.

All colonies picked from the plates were cultured in 3 ml of liquid LB media supplemented with 70 µg/ml of ampicillin (Merck, USA), at 37 °C overnight. All minipreps were performed on such cultures with the E.Z.N.A.® Plasmid Mini Kit (Omega Bio-tek, USA). All midipreps were made using the PureLink® HiPure Plasmid Midiprep Kit (Invitrogen, USA). All agarose gel extractions were performed with the Qiaquick Gel Extraction Kit (Qiagen, Germany). All sequencing reactions were performed by 1st BASE (Singapore) to verify the construct/DNA fragment sequences.

2.2. Cell transfection

HEK293T cells were cultured in Hyclone Dulbecco's Modified Eagle's medium (DMEM) (Thermo Scientific, USA) with 10% (v/v) FBS and antibiotics (100 U ml^{-1} penicillin and 100 mg ml^{-1} streptomycin). For each experiment, ~50% confluent HEK293T cells in 12-well plates were transfected with 1 µg of DNA per well, using 3 µl of X-tremeGENE 9 DNA Transfection Reagent (Roche, Switzerland) diluted in 100 µl of Hyclone Opti-MEM (Thermo Scientific, USA). Typically, test constructs (UMV or *SMN1/2* minigenes) were mixed with control plasmids in a 1:11 ratio. For suppressor experiments, test constructs were mixed with U1 suppressor plasmids and control plasmids in a ratio of 1:10:1.

2.3. RNA extraction, reverse transcription, and PCR

Cells were harvested 48 h after transfection and the total RNA was extracted with PureLink® RNA Mini Kit (Life Technologies, USA). Residual DNA was eliminated by RQ1 RNase-Free DNaseI (Promega, USA) digestion, and the RNA was ethanol-precipitated. A total of 1 µg of RNA was used for reverse transcription with Moloney Murine Leukemia Virus Reverse Transcriptase (New England Biolabs, USA) according to the manufacturer's instructions, with oligo-dT (18 T) as a primer.

Amplification of cDNAs derived from expression of the UMV (derived from the pcDNA3.1+ vector) or the *SMN1*/2 (derived from the pCI vector) constructs were performed via semi-quantitative (radioactive) PCR, using primer pairs pcDNA F and pcDNA R, or pCI-FwB and pCI-Rv (Roca and Krainer, 2009) respectively (see Table 2.2 for description and sequences). These primers anneal to the transcribed portion of the plasmids upstream of the 5’ exon and downstream of the 3’ exon in the minigene. The 5’ end of the forward primer (10 pmol) was radio-labeled using 10 U of T4 PNK (New England Biolabs, USA) in 1x PNK buffer, and γ-^{32}P-ATP (Perkin-Elmer, USA) at a final concentration of 90 µCi/µl. The labeled primer was purified via MicroSpin G-25 columns (GE Healthcare, USA) and mixed with 90 pmol of unlabeled forward primer as well as 100 pmol of the reverse primer, creating the primer mix. Each PCR consists of primers at a final concentration of 200 nM each, dNTP mix at a final concentration of 200 µM each, $MgCl_2$ adjusted to a final concentration of 1 mM, 2.5 µl of 5x Colourless GoTaq reaction buffer (which was provided without added $MgCl_2$), and 0.625 U of GoTaq DNA polymerase (Promega, USA), with ddH_2O added to a total volume of 12.5 µl. Reactions were done using the following thermocycler program: an initial denaturation at 95 °C for 5 min; then 23 cycles of denaturation at 95 °C for 30 s, annealing at 58 °C (for pCI-FwB and pCI-Rv) or 54 °C (for pcDNA F and pcDNA R) for 40 s, and extension at 72 °C for 50 s; followed by a final extension at 72 °C for another 5 min; and finally 4 °C forever. With only 23 cycles, the PCR amplification remains within the exponential phase, ensuring that amplimer abundances correspond to the abundances of their templates.

PCR products were separated by 6% native PAGE at 10 V/cm for 6 h in 1x TBE buffer. The gels were vacuum-dried with a Model 583 gel-dryer (Bio-Rad, USA), then exposed to a storage phosphor screen (GE Healthcare Life Sciences, USA). The screen was then scanned with a Typhoon Trio variable mode imager (GE Healthcare Life Sciences, USA), and band intensity was quantified by 1D gel analysis using ImageQuant TL software (GE Healthcare Life Sciences, USA).

Data from three experimental replicas (RT-PCRs of total RNA acquired from three independent transfections) allowed us to derive the mean percentage of inclusion for each experiment. If the standard deviations did not exceed 5%, it indicates that the exon-inclusion percentages are highly reproducible between experiments. If the mean percentages of inclusion between two experiments are distinct enough so that the standard deviations do not overlap, these values can be deemed ‘different’.

Figures were generated by exposing Medical X-ray Film General Purpose Green (Kodak, USA) to the radioactive gels at -80 °C and developing them with a Kodak Model 2000 X-Ray Film Processor. Developed films were scanned at the highest possible resolution with a GS-800 Calibrated Densitometer (Bio-Rad, USA).

PCR products were identified by agarose gel-extraction with Qiaquick Gel Extraction Kit (Qiagen, Germany) followed by sequencing utilizing one of the primer pairs used in the reaction (1st BASE, Singapore).

2.4. Large DNA constructs for the XCI projects

All BAC and fosmid bacterial clones were ordered from BACPAC Resources Center (BPRC) at Children's Hospital Oakland Research Institute (CHORI), and delivered as LB stabs. Clones were streaked out on selective LB agar plates containing 12.5 µg/ml of chloramphenicol. Single colonies on the plates were picked and grown overnight in 3 ml of liquid LB media containing 12.5 µg/ml of chloramphenicol. Construct DNA was extracted using generally available miniprep methods, which was then purified by isopropanol and ethanol precipitation. The purified DNA was digested with EcoRI (New England Biolabs, USA) and separated on a 1% agarose gel at 100 V for 45 min. Clones with the highest DNA yield and the correct restriction pattern were made into glycerol stocks, with one part liquid LB bacterial culture to one part sterile 80% glycerol. When required, maxipreps were performed using the Large Construct Kit (Qiagen, Germany) or the NucleoBond® BAC 100 Kit (Macherey-Nagel, Germany), to acquire sufficient amounts of construct DNA (30-70 µg) for downstream applications.

2.5. Generation of irradiated mouse embryonic fibroblast (MEF) feeder cells

Mouse embryonic fibroblasts (MEF) were extracted from day 13.5 DR4 mouse embryos. Fibroblasts were cultured and expanded in EF medium, which is composed of DMEM, $NaHCO_3$, HEPES (GIBCO, USA), non-essential amino acids (GIBCO, USA), glutamine (GIBCO, USA), penicillin-streptomycin (GIBCO, USA), 2-mercaptoethanol (Sigma-Aldrich, USA), and Hyclone characterized Fetal Bovine Serum (FBS) (Life Technologies, USA). MEF cells were harvested by trypsinization, irradiated at 30 Gy, and slow-frozen to -80°C after the addition of an equal volume of freezing medium (ES medium with 5% DMSO, see section 2.6 for details), with 4 x 10^6 cells per cryovial.

2.6. ES cell culture

Male murine ES cell lines J1 or Ainv15 of low passage number were cultured with ES+LIF medium on a layer of irradiated MEF feeder cells as per usual practice. ES+LIF medium contains DMEM, $NaHCO_3$, HEPES (GIBCO, USA), non-essential amino acids (GIBCO, USA), glutamine (GIBCO, USA), penicillin-streptomycin (GIBCO, USA), 2-mercaptoethanol (Sigma-Aldrich, USA), Hyclone characterized Fetal Bovine Serum (FBS) (Life Technologies, USA), plus leukemia inhibitory factor (LIF) (Millipore, USA). ES cells were allowed to grow to about 70-80% confluency before being trypsinized (treated with 0.05% trypsin-EDTA for 5 min) and split in a 1:6 or 1:8 ratio. All ES cells were slow-frozen and placed into liquid nitrogen after mixing in freezing media (ES medium with 5% DMSO) with an equal volume of ES cell suspension.

2.7. ES cell electroporation and generation of transgenic ES cell lines

ES cells were cultured in Falcon T25 flasks, split into Falcon T75 flasks, and allowed to reach log phase growth one or two days before electroporation. Sufficient 10 cm dishes of irradiated feeder cells were prepared one day prior to the electroporation (conditioned 10 cm feeder dishes). The ES+LIF medium of all the T75 ES cell flasks and the 10 cm feeder dishes were changed at least 2 h before electroporation. The ES cells were trypsinized and neutralized in ES+LIF medium, and counted. They were then washed with 1x PBS, and resuspended in ice-cold 1x PBS to 5 x 10^6 cells/ml. 800 µl of this cell suspension was mixed with the DNA to be electroporated inside a Bio-Rad 4 mm electroporation cuvette and incubated on ice for 5 min. Electroporation was performed in a Bio-Rad Gene Pulser II at 500 µF, 0.24 kV, and ∞ Ω. After electroporation, ES cells were allowed to recover on ice for 5 min, and then transferred into the conditioned 10 cm feeder dishes. Transfected ES cells were allowed to recover for 24 h before selection with appropriate antibiotic, G418 (400 µg/ml) for 7 days or hygromycin (300 µg/ml) throughout the experiment. Sufficiently mature/large survivor ES cell colonies were picked from the selection dishes and allowed to proliferate in 6-well plates before being harvested and frozen down in 1 well: 2 cryovial ratio.

2.8. ES cell differentiation and slide preparation

ES cells were cultured in T25 flasks with feeders to about 70-80% confluency. Cells were trypsinized, neutralized in ES+LIF medium, then incubated in gelatin-coated T75 flasks for 15 min at 37 °C to allow feeder attachment. ES cell suspension was removed from the flasks and the number of ES cells counted. ES cells were resuspended in ES medium (ES+LIF medium without LIF), and 400,000 ES cells per cell line placed into a Petri dish. 24 h later, the embryoid bodies (EB) formed were transferred into fresh ES+RA medium (ES medium with added retinoic acid) in a new Petri dish and cultured for a further 48 h, before being transferred into gelatin-coated T25 flasks, allowing the EBs to attach. After 24-48 h (or even later, if required), differentiated cells were harvested by partial trypsinization with 0.1 ml of trypsin-EDTA, and neutralized in ES medium. The cells were counted and resuspended in ES medium to a concentration of 7x10^5 cells/ml. 100 µl of this cell suspension was cytospun at 1,000 rpm for 10 min onto polysine-coated glass slides. These slides were then washed with 1x PBS for 5 min, fixed in 4% paraformaldehyde (PFA) for 10 min at room temperature, and stored in 70% ethanol at 4 °C.

2.9. Nick translation

Probes for RNA- and DNA-FISH (Fluorescence In Situ Hybridization) were prepared by labeling 2.5 µg of the respective large DNA constructs with Cy3-12-dUTP using the Roche Nick Translation Kit. Reactions were performed at 15 °C for 2 h before heat-inactivation at 65 °C for 10 min. Probes were ethanol-precipitated together with 25 µg mouse Cot-1 DNA and dissolved in hybridization buffer composed of 50% formamide (Merck, USA), 2x saline-sodium citrate (SSC) (pH 7.4, Sigma-Aldrich, USA), 2mg/ml BSA (Roche, Switzerland), and 10% dextran sulphate (Sigma-Aldrich, USA).

2.10. RNA-FISH

Slides were dehydrated sequentially in 80%, 90% and 100% ethanol for 2 min each, dried, and warmed to 42 °C. Cy3-labeled DNA probes were denatured at 80 °C for 10 min, pre-hybridized at 42 °C for 20 min, and applied to the slides. Cover-slips were placed on the slides and sealed with rubber cement. The slides were then incubated at 42 °C for at least 3 h in dark and humid conditions. Slides were subsequently washed at 45 °C under shaking: three times with 50% formamide in 2x SSC, and three times with 2x SSC. Slides were then cleaned in 1x PBS + 0.2% Tween 20, counterstained with Vectorshield anti-fade medium (Vector Laboratories, USA) containing 0.2 μg/ml of 4',6-diamidino-2-phenylindole (DAPI), and sealed with 50 mm x 50 mm cover-slips (Fisher Scientific, USA). Slides were examined by fluorescent microscopy.

2.11. Chromosome spreads

Transgenic ES cells were cultivated in T25 flasks till about 80% confluency and treated with colcemid (GIBCO, USA) at a final concentration of 0.2 μg/ml for 2 h. Cells were harvested by trypsinization, neutralized in ES+LIF medium, then resuspended and allowed to swell in 75 mM KCl for 15 min at 37 °C. After this, a few drops of methanol/acetic acid (3:1 v/v) fixative was added to the cell suspension and the mixture was spun down at 200 g for 5 min. The supernatant was removed and the cells were resuspended in methanol/acetic acid fixative. This suspension was spun at 900 g for 5 min and then the fixative was refreshed – both steps were performed twice. Cells were dropped onto polysine-coated glass slides, allowed to air-dry, and fixed with 4% PFA in 1x PBS for 10 min.

2.12. DNA-FISH genotyping of cell lines

Chromosome spreads made from the cell lines to be genotyped were dehydrated in 80%, 90% and 100% ethanol, 2 min each. Slides were treated with 400 μg/ml RNAse H (New England Biolabs) in 1x PBS for 40 min at 37°C, washed with 1x PBS + 0.2% Tween-20 three times at 3 min each, and denatured with 70% formamide in 2x SSC buffer for 10 min at 80°C. Slides were once more dehydrated sequentially in 80%, 90% and 100% ethanol, for 2 min each, dried, and warmed up to 42 °C. The corresponding Cy3-labeled DNA probes (made by nick translation) were denatured at 80 °C for 10 min, prehybridized at 42 °C for 20 min, and applied to the slides. Cover-slips were placed on the slides and sealed with rubber cement. The slides were then incubated at 42 °C overnight in dark and humid conditions. Slides were washed at 45 °C under shaking: three times with 50% formamide in 2x SSC, and three times with 2x SSC. Slides were then cleaned in 1x PBS + 0.2% Tween 20, counterstained with Vectorshield anti-fade medium (Vector Laboratories, USA) containing 0.2 µg/ml of DAPI, and sealed with 50mm x 50mm cover-slips (Fisher Scientific, USA). Slides were examined by fluorescence microscopy.

2.13. Fluorescent microscopy

Fluorescence images were obtained by a Nikon Eclipse Ti-E inverted microscope with Nikon software (NES-Elements AR 3.1). The images were processed with Adobe Photoshop and Adobe Illustrator.

2.14. Fluorescence-activated cell sorting (FACS)

Fluorescence-expressing ES cells were cultured on feeder cells in T25 flasks until at least 4 x 10^6 cells were available. Cells were trypsinized and neutralized in ES+LIF medium, then incubated in gelatin-coated T75 flasks for 15 min at 37°C to allow feeder adhesion. ES cell suspension was removed from the flasks and the number of ES cells counted. The cells were washed and re-suspended in 1x PBS to a concentration of $1x10^6$ cells/ml, then sorted by FACSAria (BD Biosciences). 2 x 10^5 cells per fraction were placed back into T25 conditioned feeder flasks if further culturing was needed.

2.15. DNA extraction from transgenic ES cells

Harvested transgenic ES cells were resuspended in 1x PBS. Cell suspensions were incubated at 42°C overnight with an equal volume of lysis buffer (100 mM NaCl, 10 mM Tris-HCl (pH 7.5), 1 mM EDTA (pH 7.5), 1% SDS, and 500 mg/ml proteinase K). Genomic DNA was phenol-chloroform extracted, ethanol precipitated, and dissolved in ddH_2O.

2.16. Colony PCR for im*Xist*-transfected Ainv15 cells

In order to locate Ainv15 cells with Cre-mediated insertions of im*Xist*, genomic DNA was extracted and colony PCR performed. Primers used: 5'-GGCCACCATGGTGTCGATAAC-3' and 5'-TGGATACTTTCTCGGCAGGAG-3'. Normal Ainv15 cells would not give any product whereas Ainv15 cells with Cre-mediated insertions would generate a fragment of roughly 400bp.

2.17. Strategy for generation of ihXIST plasmid

2.17.1. Long primer design

Starting (*XIST*START) and ending (*XIST*END) homologous arm sequences of human *XIST* were acquired from the NCBI Nucleotide database (http://www.ncbi.nlm.nih.gov/nuccore/NG_016172.1). Primers were designed to be 200 nt long. Each primer contains a 21 nt complementary sequence for annealing to the pEZ-Frt-loxP-DT-zeo-RFP template.

The forward primer consisted of the 179 nt *XIST* ending sequence (*XIST*END) plus a short 21 nt sequence downstream of the pBR322 origin of replication: (5’-CTTGAACTTGTGAACTGATGTGAAATGCAGAATCTCTTTTGAGTCTTTGCTGTTTGGAAGATTGAAAAATATTGTTCAGCATGGGTGACCACCAGAAAGTAATCTTAAGCCATCTAGATGTCACAATTGAAACAAACTGGGGAGTTGGTTGCTATTGTAAAATAAAATATACTGTTTTGCCACAGAATCAGGGGATAACG-3’)

The reverse primer consisted of the reverse complement to the 21 nt sequence at the end of the PGK promoter, an ATG start sequence (in order to complete the neomycin resistance cassette in the Ainv15 cells) plus the modified loxP site specific to Ainv15, and the *XIST* starting sequence (*XIST*START): (5’-CCCAAGTGCAGAGAGATCTTCAGTCAGGAAGCTTCCAGCCCCGAGAGAGTAAGAAATATGGCTGCAGCAGCGAATTGCAGCGCTTTAAGAACTGAAGGATGCAACTTCGTATAATGTATGCTATACGAAGTTATCGACACCATGGTGGCCTCCAGATCCTTCGAGATCTAGATGGATGCAGGTCGAAAGGCCCGGAGATG-3’)

2.17.2. Long PCR

Long PCR was performed with the long primers using the Expand High Fidelity PCR System (Roche, Switzerland) according to the manufacturer's instructions and the program in Table 2.5.

Table 2.5 PCR program for generation of long PCR fragment.

Step	Temperature	Time	Cycles
Initial Denaturation	94°C	2 min	1x
Denaturation	94°C	15 s	10x
Annealing	55°C	30 s	
Elongation	68°C	4 min	
Denaturation	94°C	15 s	15-20x
Annealing	65°C	30 s	
Elongation	68°C	4 min, +10s cumulatively per subsequent cycle	
Final Elongation	68°C	7 min	1x
Cooling	4°C	Forever	

2.17.3. Transformation of BAC-carrying bacterial cell line with pRed/ET plasmid

The BAC-carrying *E. coli* cells were transfected with the pRed/ET plasmid using the BAC Subcloning Kit (Gene Bridges, Germany). The cells were plated on tetracycline + chloramphenicol LB agar at 30 °C, and glycerol stocks were made from colonies cultured in tetracycline + chloramphenicol LB broth the next day.

2.17.4. Red/ET recombination

The entire process was carried out according to the instructions provided in the BAC Subcloning Kit (Gene Bridges). In brief, a culture of bacteria made in the previous stage was induced with L-arabinose to cause expression of recombinases at 37°C for 1 h. The bacteria were then electroporated with the long PCR fragment, and then incubated at 37°C for 70 min or more before being plated onto zeocin LB agar and subsequently the plates were incubated overnight at 37°C.

2.17.5. Colony PCR

To detect the ih*XIST* plasmid, a pair of primers was designed that would produce a ~350bp fragment covering *XIST*END and the region downstream of pBR322: Forward 350bp: 5'-TATGGAAAAACGCCAGCAACG-3' and Reverse: 5'-TAAGTGGCTTCGTCATTGTCC-3'. Also, another primer was designed so that when paired with the Reverse primer, it would detect the un-recombined BAC by generating a ~550bp fragment covering *XIST*END and part of the BAC insert: Forward 550bp: 5'-ATGGCAAAACCCCGTCTCTAC-3'. Colonies produced after the Red/ET recombination step were picked and colony PCR performed, using the following reaction mixture for one colony: all three primers were diluted to a final concentration of 300 nM in the reaction, 4 µl of dNTP mix (2.5 mM), 5 µl of 10x Buffer, 0.5 µl of Taq polymerase and made up with ddH_2O for a total reaction volume of 50 µl; using the following program: an initial denaturation at 95 °C for 5 min; then cycled 30 times: denaturation at 95 °C for 30 s, annealing at 57 °C for 20 s, and extension at 72 °C for 35 s; then a final extension at 72°C for 10 min, and 4 °C forever.

3. RESULTS

3.1. Splicing: Asymmetric loops

Earlier work (Roca et al., 2012) revealed the existence of 1-nucleotide bulge registers for 5’ss/U1 base pairing. In such an arrangement, one nucleotide on one strand remains unpaired and unopposed by any nucleotide on the opposite strand, thereby bulging out of the duplex and forming a kink in the helix (Figure 3.1C, D, and E). The energetic disadvantage generated by this kink is compensated by the additional 5’ss/U1 base pairs that can be established in these registers (up to a maximum of 11 base pairs) versus the canonical register, thereby enhancing the stability of the 5’ss/U1 duplex.

From the same dataset, 1-nucleotide asymmetric loop registers were also predicted. They are remarkably similar to 1-nucleotide bulge registers. While bulge registers are defined as having one or more unpaired nucleotides on only one side of the helix that are flanked by base pairs, asymmetric loop registers have an uneven number of unpaired nucleotides on both sides of the helix which are flanked by base pairs. The difference in number of the unpaired nucleotides on both sides of the loop is used as the loop number. For example, as seen in Figure 3.1F and Figure 3.1G, where the two uridines at the 5’ss and a single pseudouridine at the U1 are unpaired.

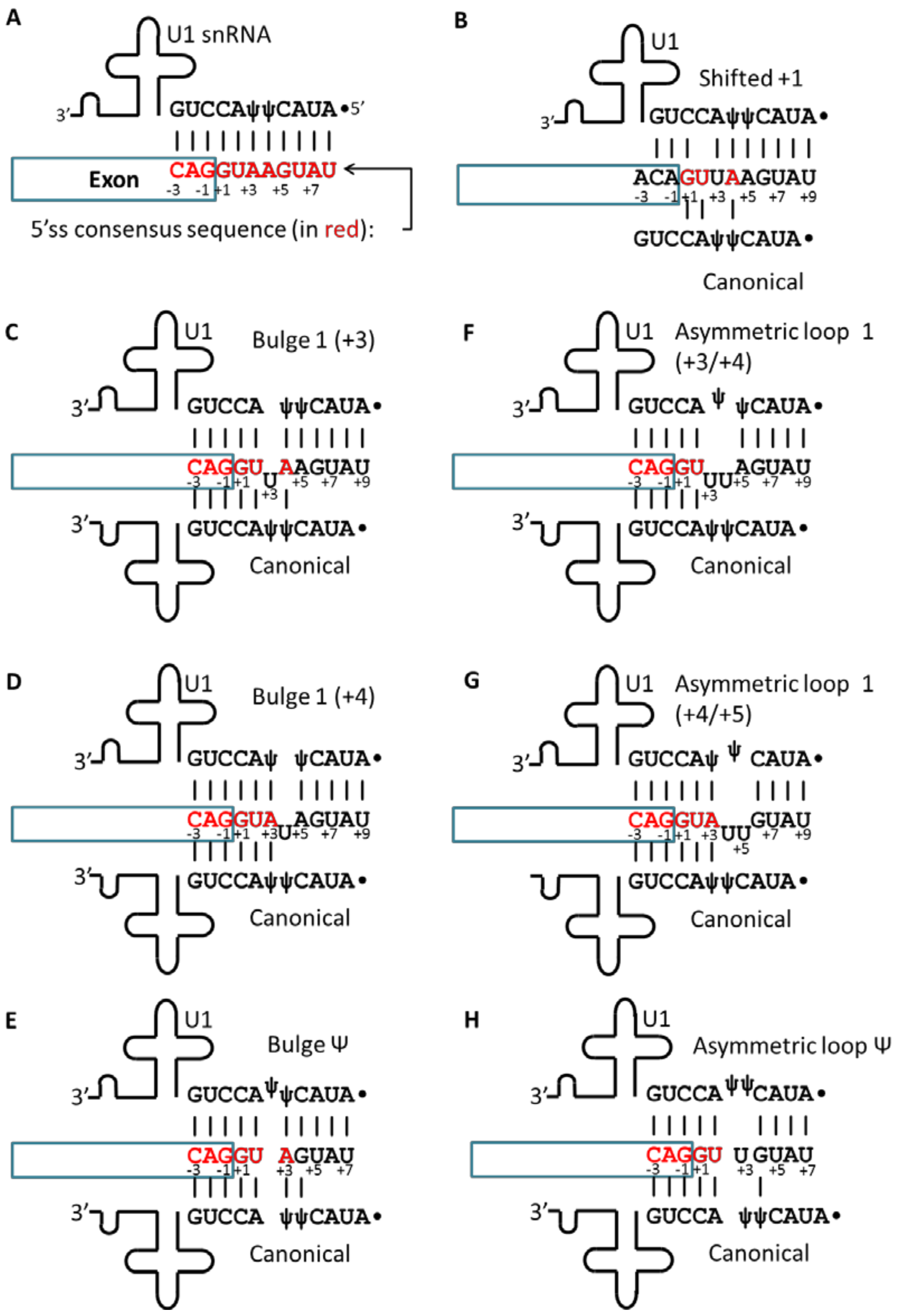


Figure 3.1 Bulge and asymmetric loop registers

Blue boxes represent the test exons. The 5’ss sequence is depicted, with red letters indicating consensus nucleotides in the 5’ss. The sequences above and below represent the U1 snRNA and the registers by which it can interact with the 5’ss. Vertical lines between sequences represent base pairs.

3.1.1. Asymmetric loop 1 (+3/+4) register

The asymmetric loop 1 (+3/+4) register is a register in which 5’ss nucleotides at positions +3 and +4 as well as U1 Ψ6 form the loop in the 5’ss/U1 helix, which leaves one nucleotide on the 5’ss unmatched (Figure 3.1F). A total of 348 human 5’ss sequences were predicted to possess an energetic advantage if base paired to U1 via this register, which can increase the number of maximum possible base pairs in the 5’ss/U1 duplex to 10, versus about 5 in the canonical register.

3.1.1.1. UMV minigene testing reveals evidence for use of the asymmetric loop (+3/+4) register for 5’ss recognition in native context

To test whether U1 recognition via the asymmetric loop 1 (+3/+4) register occurs in certain 5’ss, candidate naturally-occurring 5’ss were selected based on the criteria provided in the Methods (2.1.2). Three candidates were chosen, which were the 5’ss from *ABCC12* exon 17, *PARP14* exon 9, and *SLC5A8* exon 7 (Figure 3.2A, B, and C respectively). Test exons and their flanking intronic sequences from these three genes containing the 5’ss of interest were cloned into UMV to create their respective minigenes. HEK293T cells were transfected with these minigenes, the total RNA extracted after 48 hours, and radioactive RT-PCR performed on the total RNA as detailed in the methods. The level of test exon inclusion, which reflects the recognition of the test 5’ss, was determined.

All three of the *ABCC12*, *PARP14*, and *SLC5A8* minigenes produced transcripts with complete or almost complete test exon inclusion, reflecting use of the test 5’ss (Figure 3.2D, lane 1 for each minigene). Mutational analysis of the test 5’ss was carried out. Mutants that affected both canonical and asymmetric loop 1 +3/+4 register base pairs, the -2C mutants, caused significant loss of exon inclusion as expected (Figure 3.2D, lanes 2; indicated in blue). This indicated that it was possible to alter the splicing pattern by mutating the 5’ss.

Point mutations that affected only the asymmetric loop register, at positions +6 and +7 of the 5’ss (+6C and +7C mutants), also caused significant loss of exon inclusion (Figure 3.2D, indicated in green, lane 3 and 6 respectively for each minigene). This revealed that the nucleotide identity at these positions was important for 5’ss recognition. As these positions can only base pair to U1 in the asymmetric register, these findings suggest that these 5’ss are recognized via the asymmetric loop 1 (+3/+4) register.

Suppressor U1 snRNA experiments were performed to determine whether base pairing interactions between the test 5’ss and the U1 were occurring via the canonical or the asymmetric loop 1 (+3/+4) register. In the +6C mutants, test exon inclusion was partially rescued by suppressor U1 rescuing a base pair in the asymmetric loop 1 +3/+4 register, U1 with the G4 mutation (Figure 3.2D, lane 5 for each minigene). The suppressor U1 rescuing a base pair in the canonical register, U1 with the G3 mutation (Figure 3.2D, lane 4 for each minigene) did not restore correct splicing as effectively as U1 G4. Similarly, for the +7C mutants, test exon inclusion was rescued by suppressor U1 acting via the asymmetric loop 1 (+3/+4) register, U1 with the G3 mutation (Figure 3.2D, lanes 8), while suppressor U1 acting via the canonical register, U1 with the G2 mutation (Figure 3.2D, lane 7 for each minigene) could not restore correct splicing as effectively as U1 G3. Thus, in all cases, the suppressor U1 acting via the canonical register did not restore correct splicing patterns or performed much less effectively versus their asymmetric loop register counterparts. Also, the effect of the same suppressor U1 (U1 with G3 mutation) varied between the different mutations in the 5’ss (compare lane 4 with lane 8 for each minigene in Figure 3.2D), demonstrating that the suppressor U1 was effective and that the suppressor effect corresponds to the register being tested. All this evidence further supported the hypothesis that the test 5’ss were recognized via the asymmetric loop 1 (+3/+4) register.

Interestingly, each transcript exhibited different changes in their splicing pattern when the 5’ss was altered. *ABCC12* experienced high levels of exon skipping. *PARP14* made use of a cryptic intronic 5’ss 62nt downstream of the test 5’ss. *SLC5A8* experienced both exon skipping as well as usage of a cryptic 5’ss 52nt upstream of the test 5’ss. Sequencing of the RT-PCR products confirmed that splicing only occurred at the GU exon-intron boundary.

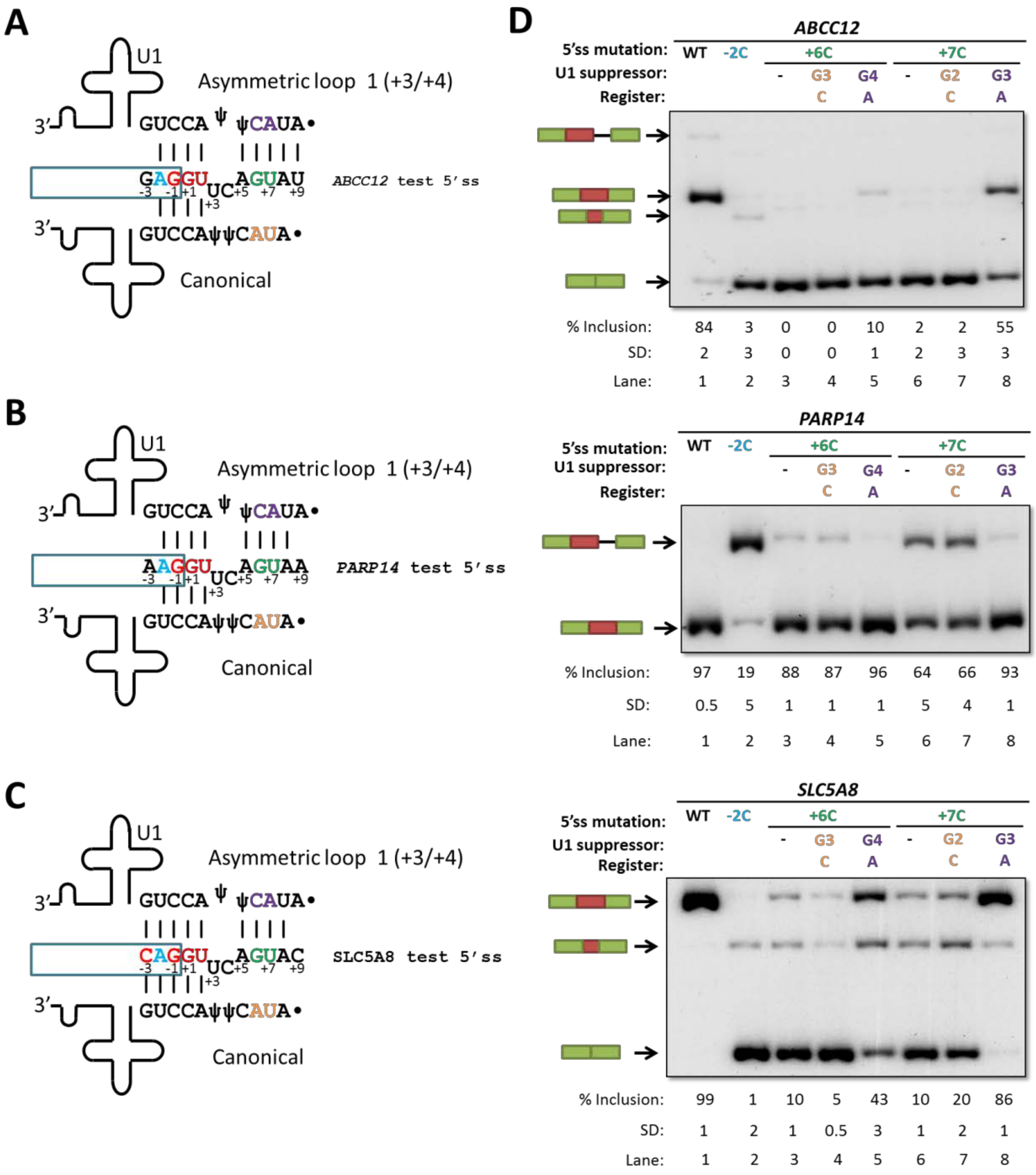


Figure 3.2 *ABCC12*, *PARP14* and *SLC5A8* minigene analysis demonstrates asymmetric loop 1 (+3/+4) register

(A, B, C) Sequences of test 5’ss from these three genes, and representations of both the asymmetric loop 1 (+3/+4) register and the canonical register that the U1 snRNA can adopt to base pair with these 5’ss. **(D)** Native PAGE of radioactive RT-PCR of RNA products of these three minigenes. The red box indicates the test exon, green boxes the flanking *MCAD* exons, and the black line represents intronic sequences. 5’ss mutations in blue represent mutations that affect both registers, while those in green represent mutations that affect only the asymmetric loop register. U1 suppressors in orange represent U1 suppressors that affect the canonical register, also denoted with a C, while those in purple represent U1 suppressors that affect the asymmetric loop register, also indicated with an A. The largest band in *ABCC12* represents utilization of a putative intronic 5’ss downstream of the test 5’ss which has yet to be precisely mapped, the band immediately below that the exon inclusion band, the next band represents use of an exonic cryptic 5’ss 20 nt upstream of the test 5’ss, while the lowest band is the exon skipping band. The higher band in *PARP14* indicates use of an intronic cryptic 5’ss 62 nt downstream of test 5’ss, while the lower band is the exon inclusion band. The highest band in *SLC5A8* is the exon inclusion band, the second band from the top represents use of an exonic cryptic 5’ss 52nt upstream of the test 5’ss, and the lowest band indicates the test exon skipping band.

3.1.1.2. SMN1/2 minigene testing indicates that the asymmetric 1 (+3/+4) register can be used in a heterologous context

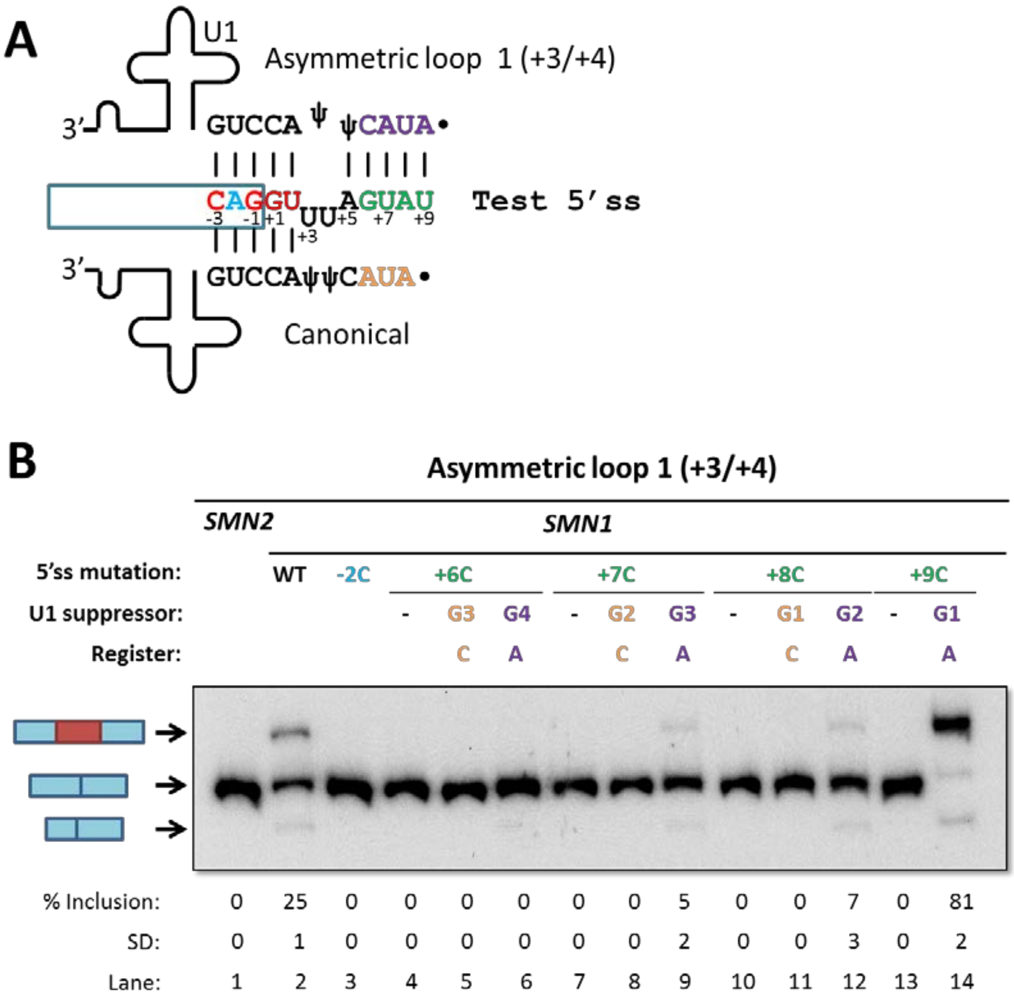
In order to prove the general applicability of the asymmetric loop 1 (+3/+4) register, the natural 5’ss of exon 7 in the *SMN1/2* minigenes was replaced with a representative ideal test 5’ss sequence (Figure 3.3A) which can establish the maximum possible number of base pairs with U1 via the asymmetric loop 1 (+3/+4) register, via PCR mutagenesis. After transfection, RNA extraction, and RT-PCR, *SMN2* transcripts displayed complete exon skipping (Figure 3.3B, lane 1), while *SMN1* retained some exon inclusion (lane 2). Therefore, only the test *SMN1* minigene was used for mutational analysis.

The -2C mutant that affects both the canonical and asymmetric loop 1 (+3/+4) registers resulted in complete exon skipping (Figure 3.3B, lane 3; indicated in blue). This indicated that it was possible to alter the splicing pattern of the test *SMN1* by mutating the 5’ss, and that the test 5’ss was relatively weak in this context.

Point mutations that affected only the asymmetric loop register, at positions +6, +7, +8, and +9 of the test 5’ss (+6C, +7C, +8C, and +9C mutants), all caused complete exon skipping (Figure 3.3B, indicated in green, lanes 4, 7, 10, and 13 respectively). This revealed that the nucleotides at these positions were important for 5’ss recognition. As these positions can only base pair to U1 in the asymmetric register, this strongly suggested that the test 5’ss was recognized via the asymmetric loop 1 (+3/+4) register.

Suppressor U1 snRNA experiments were performed to determine whether base pairing interactions between the test 5’ss and the U1 were occurring via the canonical or the asymmetric loop 1 (+3/+4) register. In the +6C mutant, test exon inclusion could not be rescued by either suppressor U1, perhaps because the mutation proved too strong to be overcome. However, for the +7C mutant, test exon inclusion was weakly rescued by suppressor U1 acting via the asymmetric loop 1 (+3/+4) register, U1 with the G3 mutation (Figure 3.3B, lane 9). Suppressor U1 acting via the canonical register, U1 with the G2 mutation (Figure 3.3B, lane 8) had no observable effect on splicing. A similar situation occurred with the +8C mutant, whereby the test exon inclusion was partially rescued by suppressor U1 acting via the asymmetric loop 1 (+3/+4) register, U1 with the G2 mutation (Figure 3.3B, lane 12), while suppressor U1 acting via the canonical register, U1 with the G1 mutation (Figure 3.3B, lane 11) did not rescue correct splicing. With the +9C mutant, test exon inclusion was strongly rescued by suppressor U1 acting via the asymmetric loop 1 (+3/+4) register, U1 with the G1 mutation (Figure 3.3B, lane 12). As the canonical register only extends to position +8 of the 5’ss, no U1 suppressors in the canonical register can be used for the +9C mutant.

In all cases, the suppressor U1 acting via the canonical register did not affect the splicing pattern. Also, the effect of the same suppressor U1s (U1s with the G3, G2, and G1 mutations) varied between the different mutations in the 5’ss (compare lane 5 with lane 9, lane 8 with lane 12, and lane 11 with lane 13 in Figure 3.3B), indicating that the effect of the suppressor U1 was specific to the register. More convincingly, the +9C mutation was capable of affecting splicing despite not being covered by the canonical register, and it could be rescued via the asymmetric loop 1 (+3/+4) register. All this evidence further supported the hypothesis that the test 5’ss was recognized via the asymmetric loop 1 (+3/+4) register.



% Inclusion:	0	25	0	0	0	0	0	0	5	0	0	7	0	81
SD:	0	1	0	0	0	0	0	0	2	0	0	3	0	2
Lane:	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Figure 3.3 Asymmetric loop 1 (+3/+4) register testing in *SMN1/2* minigenes
(A) Sequence of test 5’ss replacing the natural 5’ss of *SMN1/2* exon 7, and representations of both the asymmetric loop 1 (+3/+4) register and the canonical register that the U1 snRNA can adopt to base pair with the test 5’ss. **(B)** Native PAGE of RT-PCR of RNA products of the *SMN1/2* minigenes. The identity of the various spliced mRNAs is indicated on the left; from large to small, the bands correspond to exon 7 inclusion, exon 7 skipping, and exon 7 skipping with activation of a cryptic 5′ss at position −50 in exon 6. The red box indicates exon 7, while blue boxes represent the flanking exons 6 and 8. 5’ss mutations in blue represent mutations that affect both registers, while those in green represent mutations that affect only the asymmetric loop register. U1 suppressors in orange represent U1 suppressors that affect the canonical register, also denoted with a C, while those in purple represent U1 suppressors that affect the asymmetric loop register, also indicated with an A.

3.1.2. Asymmetric loop 1 (+4/+5) register

The asymmetric loop 1 (+4/+5) register is defined by a loop comprised of the 5’ss nucleotides at positions +4 and +5 together with the U1 Ψ5 during 5’ss/U1 base pairing, which leaves one nucleotide on the 5’ss unmatched (Figure 3.1G). 653 human 5’ss sequences were predicted to possess an energetic advantage if base paired to U1 via this register, which increases the number of maximum possible base pairs in the 5’ss/U1 duplex to 10, versus 6 in the canonical register.

3.1.2.1. UMV minigene testing reveals evidence for use of the asymmetric loop 1 (+4/+5) register for 5’ss in native context

To test whether U1 recognition of the asymmetric loop 1 (+4/+5) register occurs in certain 5’ss, candidate naturally-occurring 5’ss were selected based on the criteria provided in the Methods (2.1.2). Two candidates were found, 5’ss from *FBXL13* exon 5 and *RNF170* exon 3 (Figure 3.4A and B respectively). Test exons and their flanking intronic sequences from these two genes containing the 5’ss of interest were cloned into UMV to create their respective minigenes.

Both *FBXL13* and *RNF170* minigenes produced transcripts with more than 70% test exon inclusion, reflecting relatively efficient use of the test 5’ss (Figure 3.4C, lanes 1). Then, mutational analysis of the test 5’ss was carried out. Mutants that affected both canonical and asymmetric loop 1 (+4/+5) register base pairs, the -1C mutants, caused major loss of exon inclusion (Figure 3.4C, lanes 2; indicated in blue). This indicated that it was possible to alter the splicing pattern by mutating the 5’ss.

Point mutations that affected only the asymmetric loop register, specifically at positions +6 and +7 of the 5’ss (+6C and +7C mutants), also caused substantial loss of exon inclusion (Figure 3.4C, lanes 3 and 6 respectively). This revealed that the nucleotides at these positions were important for 5’ss recognition. As these positions can only base pair to U1 in the asymmetric loop register, this data suggested that these 5’ss are recognized via the asymmetric loop 1 (+4/+5) register.

Suppressor U1 snRNA experiments were performed to determine whether base pairing interactions between the test 5’ss and the U1 occurred via the canonical or the asymmetric loop 1 (+4/+5) register. In the +6C mutants, test exon inclusion could not be rescued by suppressor U1 for both registers (Figure 3.4A, lane 4 and 5 respectively). This was probably because the +6C mutation was too strong to overcome. However, for the +7C mutants, test exon inclusion was strongly rescued by the suppressor U1 acting via the asymmetric loop 1 (+4/+5) register (Figure 3.4C, lane 8). The suppressor U1 acting via the canonical register (Figure 3.4C, lane 7) did not rescue splicing for the +7C mutants. Similarly, the +8C mutants were successfully rescued by suppressor U1 acting via the asymmetric loop 1 (+4/+5) register (Figure 3.4C, lane 11), while the suppressor U1 acting via the canonical register did not rescue splicing (Figure 3.4C, lane 10). Therefore, the suppressor U1 acting via the canonical register did not restore correct splicing patterns versus their asymmetric loop register counterparts for all tested circumstances. Moreover, the effect of the same suppressor U1 (G3 and G2) changed between the different mutations in the 5’ss (compare lane 4 with lane 8, and lane 7 with 11 in Figure 3.4C), indicating that the suppressor U1 effect was specific to the splicing register. All this evidence further supported the idea that the test 5’ss were recognized via the asymmetric loop 1 (+4/+5) register.

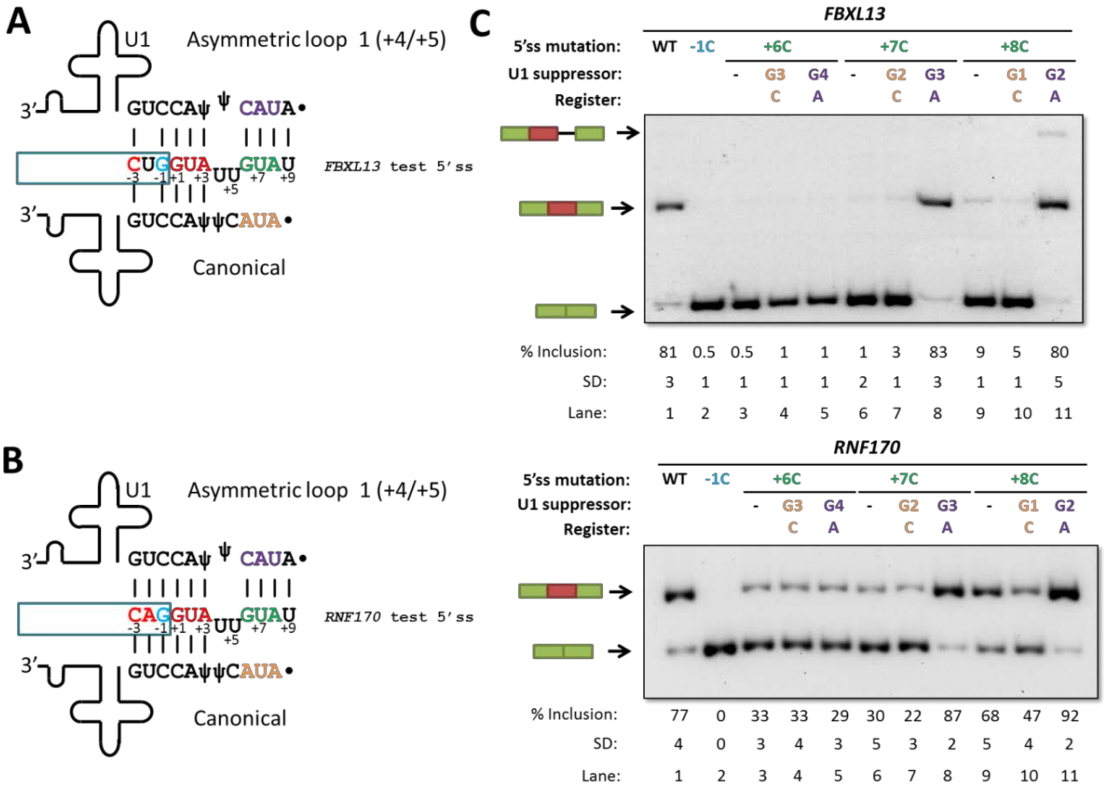

A
U1
Asymmetric loop 1 (+4/+5)
3'
GUCCAψ ψ CAUA•
CUGGUAUUGUAU
-3 -1 +1 +3 +5 +7 +9
FBXL13 test 5' ss
3'
GUCCAψψCAUA•
Canonical
B
U1
Asymmetric loop 1 (+4/+5)
3'
GUCCAψ ψ CAUA•
CAGGUAUUGUAU
-3 -1 +1 +3 +5 +7 +9
RNF170 test 5' ss
3'
GUCCAψψCAUA•
Canonical
C
FBXL13
5'ss mutation: WT -1C +6C +7C +8C
U1 suppressor: - G3 G4 - G2 G3 - G1 G2
Register: C A C A C A
% Inclusion: 81 0.5 0.5 1 1 1 3 83 9 5 80
SD: 3 1 1 1 1 2 1 3 1 1 5
Lane: 1 2 3 4 5 6 7 8 9 10 11
RNF170
5'ss mutation: WT -1C +6C +7C +8C
U1 suppressor: - G3 G4 - G2 G3 - G1 G2
Register: C A C A C A
% Inclusion: 77 0 33 33 29 30 22 87 68 47 92
SD: 4 0 3 4 3 5 3 2 5 4 2
Lane: 1 2 3 4 5 6 7 8 9 10 11

Figure 3.4 *FBXL13* and *RNF170 demonstrate use of asymmetric loop 1 (+4/+5) register*

(A, B) Sequences of test 5'ss from both genes, and representations of both the asymmetric loop 1 (+4/+5) register and the canonical register that the U1 snRNA can adopt to base pair with these 5'ss. **(C)** Native PAGE of radioactive RT-PCR of RNA products of both minigenes. The red box indicates the test exon, green boxes the flanking *MCAD* exons, and the black line represents intronic sequences. The highest band in *FBXL13* represents a putative intronic cryptic 5'ss that has yet to be precisely mapped, the middle band indicates exon inclusion while the lowest band represents exon skipping. The higher band in *RNF170* indicates exon inclusion while the lowest band represents exon skipping. 5'ss mutations in blue represent mutations that affect both registers, while those in green represent mutations that affect only the asymmetric loop register. U1 suppressors in orange represent U1 suppressors that affect the canonical register, also denoted with a C, while those in purple represent U1 suppressors that affect the asymmetric loop register, also indicated with an A.

3.1.2.2. SMN1/2 minigene testing indicates that the asymmetric 1 (+4/+5) register can be used in a heterologous context

In order to prove the general applicability of the asymmetric loop 1 (+4/+5) register, the natural 5’ss of exon 7 in the *SMN1/2* minigenes was replaced, via PCR mutagenesis, with an ideal test 5’ss sequence (Figure 3.5A), which means the 5’ss can form the maximum possible number of base pairs with U1 via the asymmetric loop 1 (+4/+5) register. After transfection, RNA extraction, and RT-PCR, *SMN1/2* transcripts displayed full exon inclusion (Figure 3.5B, lane 1). This indicated strong usage of the test 5’ss. Therefore, both minigenes were used for mutational analysis.

The -2C mutant, that affects both canonical and asymmetric loop 1 (+4/+5) register base pairs, caused low levels of exon skipping (Figure 3.5B, lanes 2; indicated in blue). This indicated that it was possible to alter the splicing pattern of the test *SMN1/2* by mutating the 5’ss, and also reaffirmed that the 5’ss was relatively strong in this context.

Point mutations that affected only the asymmetric loop register at positions +6, +7, and +8 of the test 5’ss (+6C, +7C, and +8C mutants), did not alter the splicing pattern of the *SMN1* minigene (Figure 3.5B, lanes 3, 4, and 8 respectively in the upper panel). In the context of *SMN2*, the +6C and +7C mutations caused a slight increase in exon skipping, while the +8C mutants did not affect exon inclusion (Figure 3.5B, lanes 3, 4, and 8 respectively in the lower panel). The reduced effects of the mutants was perhaps due to the innate strength of the test 5’ss. In order to counteract this, double point mutations were made for each minigene: -2C +7C and -2C +8C (Figure 3.5B, lanes 5 and 9 respectively). The double mutations synergized, triggering a significant loss in exon inclusion that was larger than that of the single mutants combined, in both the *SMN1* and *SMN2* minigene contexts. This revealed that the nucleotides at these positions were important for 5’ss recognition. As the +7 and +8 positions can only base pair to U1 in the asymmetric register, this implied that the test 5’ss was recognized via the asymmetric loop 1 (+4/+5) register.

Suppressor U1 snRNA experiments were performed to determine whether base pairing interactions between the test 5’ss and the U1 were occurring via the canonical or the asymmetric loop 1 (+4/+5) register. For the +7C mutant, test exon inclusion was strongly rescued by suppressor U1 acting via the asymmetric loop 1 (+4/+5) register (Figure 3.5B, lane 7), while suppressor U1 acting via the canonical register was unable to restore correct splicing (Figure 3.5B, lane 6). A similar situation occurred with the +8C mutant, whereby the test exon inclusion was strongly rescued by suppressor U1 acting via the asymmetric loop 1 (+4/+5) register (Figure 3.5B, lane 11), whereas suppressor U1 acting via the canonical register, could not rescue splicing (Figure 3.5B, lane 10).

In all cases, the suppressor U1 acting via the canonical register did not correct the splicing pattern. Also, the effect of the same suppressor U1 (U1 G2) varied between the different mutations in the 5’ss (compare lane 6 with lane 11 in Figure 3.5B), indicating that the effect of the suppressor U1 was specific to the register being tested. All this evidence further supported the hypothesis that the test 5’ss was recognized via the asymmetric loop 1 (+4/+5) register.

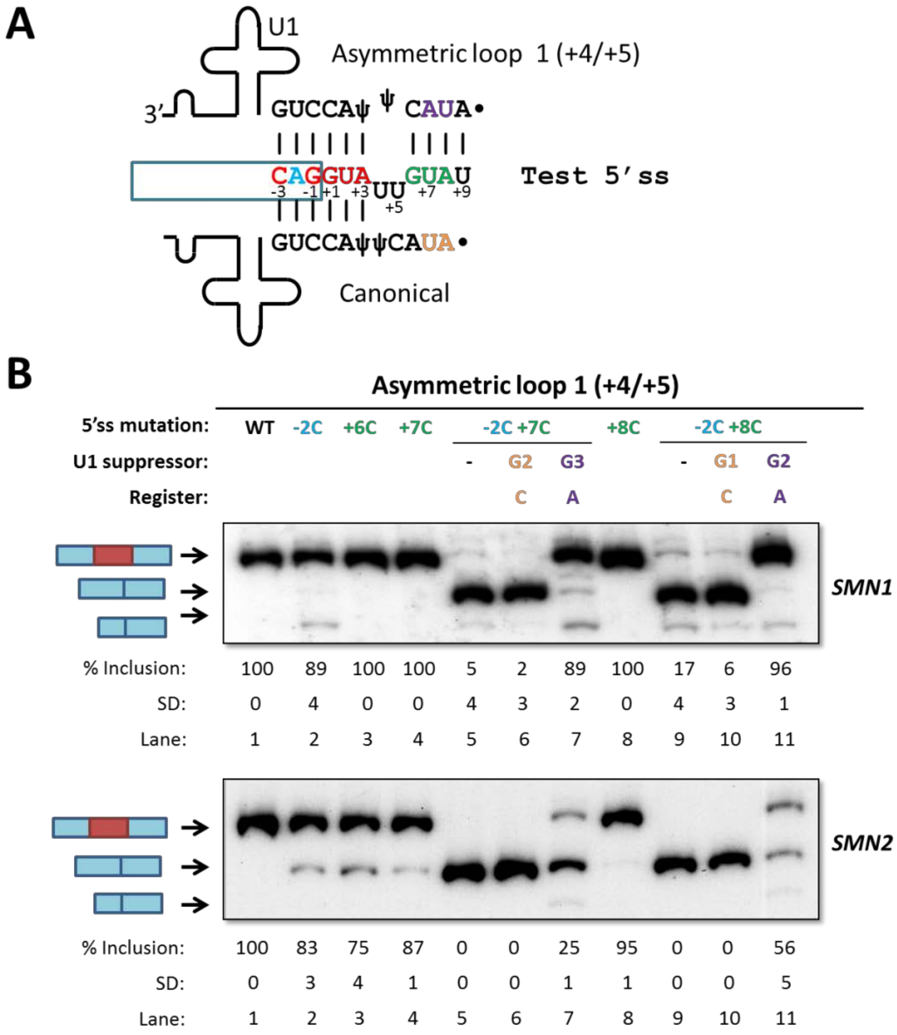


Figure 3.5 Asymmetric loop 1 (+4/+5) register testing in *SMN1/2* minigenes

(A) Sequence of test 5'ss replacing the native 5'ss in the *SMN1* minigene, and representations of both the asymmetric loop 1 (+4/+5) register and the canonical register that the U1 snRNA can adopt to base pair with these 5'ss. **(B)** Native PAGE of RT-PCR of RNA products of the *SMN1/2* minigenes. The identity of the various spliced mRNAs is indicated on the left; from large to small, the bands correspond to exon 7 inclusion, exon 7 skipping, and exon 7 skipping with activation of a cryptic 5′ss at position −50 in exon 6. The red box indicates exon 7, while blue boxes represent the flanking exons 6 and 8. 5'ss mutations in blue represent mutations that affect both registers, while those in green represent mutations that affect only the asymmetric loop register. U1 suppressors in orange represent U1 suppressors that affect the canonical register, also denoted with a C, while those in purple represent U1 suppressors that affect the asymmetric loop register, also indicated with an A.

3.1.3. Asymmetric loop Ψ register

The asymmetric loop 1 Ψ register is defined by a loop formed by nucleotides of the 5’ss at position +3 and the U1 Ψ5 and Ψ6 during 5’ss/U1 base pairing, which leaves one Ψ on the U1 5’ end unmatched (Figure 3.1H). 115 human 5’ss sequences were predicted to possess an energetic advantage if base paired to U1 via this register, which increases the number of maximum possible base pairs in the 5’ss/U1 duplex to 10, versus just 6 in the canonical register.

3.1.3.1. UMV minigene testing demonstrates evidence for usage of asymmetric loop 1 Ψ register for 5’ss recognition in native context

To test whether U1 recognition of the asymmetric loop 1 Ψ register occurs in certain 5’ss, candidate naturally-occurring 5’ss were selected based on the criteria provided in Methods (2.1.2). Two candidates were tested, which were 5’ss from *PIK3R4* exon 5 and *POLQ* exon 20 (Figure 3.6A and B respectively). Test exons and their flanking intronic sequences from these two genes containing the 5’ss of interest were cloned into UMV to create their respective minigenes, and analyzed by transfection and RT-PCR.

Both *PIK3R4* and *POLQ* minigenes produced transcripts with more than 90% test exon inclusion, reflecting use of the test 5’ss (Figure 3.6C, lanes 1). Mutational analysis of the test 5’ss was carried out. Mutants that affected both canonical and asymmetric loop 1 Ψ register base pairs, the -2C and +4C mutants, caused significant loss of exon inclusion for *PIK3R4* (Figure 3.6C, upper panel, lanes 2 and 3 respectively; indicated in blue). The loss of exon inclusion for the same mutations in *POLQ* was much less severe (Figure 3.6C, lower panel, lanes 2 and 3 respectively), especially for the +4C mutant. Nevertheless, this indicated that mutations in the test 5’ss could impact the splicing pattern.

A point mutation that affected only the asymmetric loop register at the +5 position for *PIK3R4* also caused extensive loss of exon inclusion (Figure 3.6C, upper panel, lane 6). However, in the *POLQ* context, the point mutations +5C and +6C only weakly impacted exon inclusion (Figure 3.6C, lower panel, lanes

6 and 9 respectively). To enhance the effect of the +6C mutation, the double mutation -2C +6C was made in the *POLQ* minigene. The double mutation synergized, causing a significant loss of exon inclusion (Figure 3.6C, lower panel, lane 12), more than the sum of the individual mutants. The point mutations revealed that the nucleotides at these positions were important for test 5'ss recognition. As such positions can only base pair to U1 in the asymmetric loop register, this data suggested that these 5'ss are recognized via the asymmetric loop 1 Ψ register.

Suppressor U1 snRNA experiments were performed to determine whether base pairing interactions between the test 5'ss and the U1 occurred via the canonical or the asymmetric loop 1 Ψ register. In the *PIK3R4* context, the +4C mutant was more efficiently rescued by suppressor U1 acting via the asymmetric loop 1 Ψ register, U1 G4 (Figure 3.6C, upper panel, lane 5), versus the canonical register suppressor, U1 G5 (Figure 3.6C, upper panel, lane 4). Similarly, the +5C *PIK3R4* mutant was also more effectively rescued by suppressor U1 acting via the asymmetric loop 1 Ψ register, U1 G4 (Figure 3.6C, upper panel, lane 8), as opposed to the canonical register suppressor, U1 G5 (Figure 3.6C, upper panel, lane 7).

Although the effect of the *POLQ* +4C, +5C, and +6C mutants were small to begin with, U1 suppressors in the asymmetric loop 1 Ψ register indeed rescued exon inclusion (Figure 3.6C, lower panel, lanes 5, 8, and 11 respectively). U1 suppressors in the canonical register did not increase exon inclusion (Figure 3.6C, lower panel, lanes 4, 7, and 10 respectively). U1 suppressors in the asymmetric loop 1 Ψ register also rescued exon inclusion for the -2C +6C double mutant (Figure 3.6A, lower panel, lane 14), while U1 suppressors in the canonical register did not help restore correct splicing in this mutant (Figure 3.6C, lower panel, lane 13)

In all tested situations, the suppressor U1 acting via the canonical register performed far worse in restoring correct splicing patterns versus their asymmetric loop register counterparts. Also, the same suppressor U1 (U1 G4 and U1 G3) affected different mutations in the 5'ss differently (compare lane 5 with lane 7, and lane 8 with 10 in Figure 3.6A), indicating that the effect of the suppressor U1 was specific to the base pairing register. All this evidence further supported the premise that the test 5’ss were recognized via the asymmetric loop 1 Ψ register.

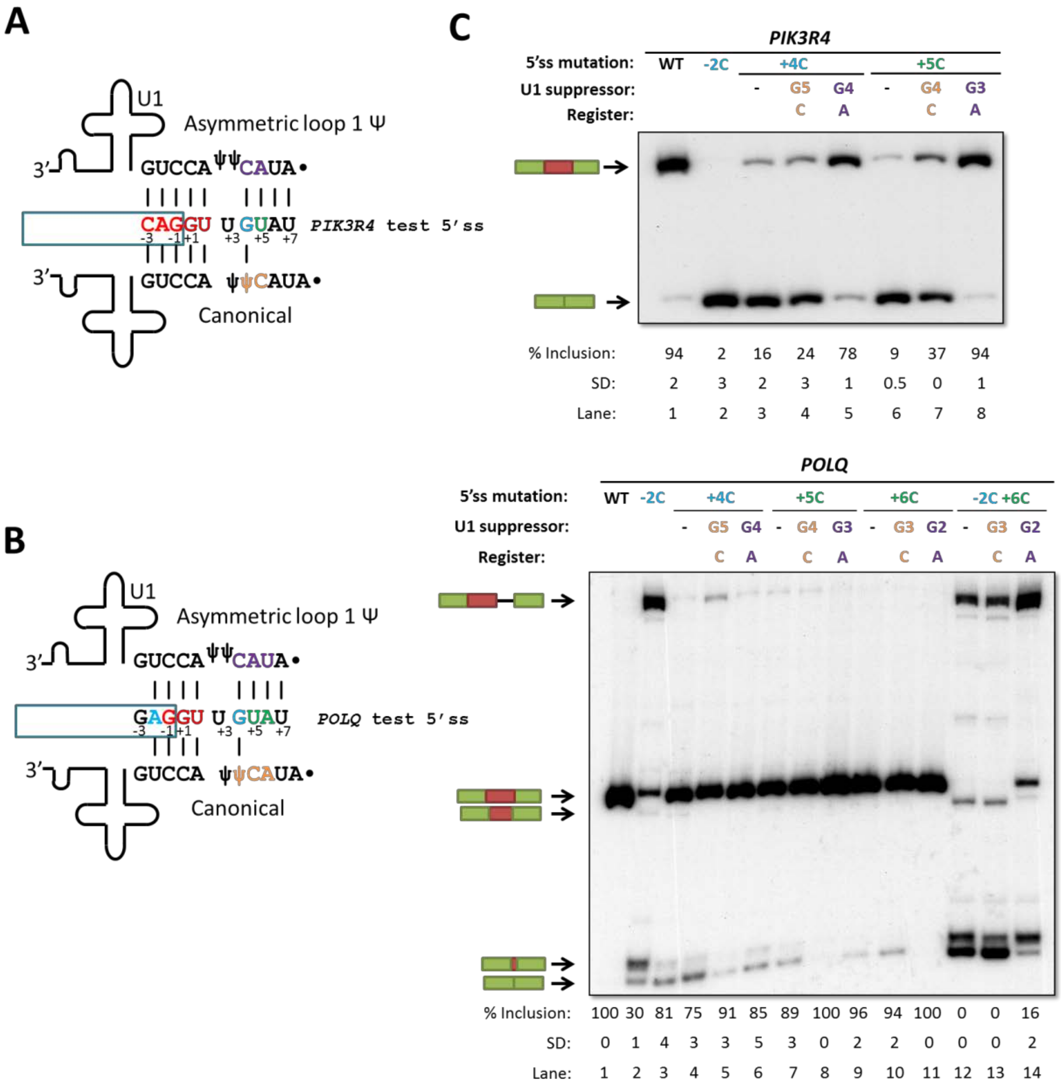

A
U1
Asymmetric loop 1 Ψ
3′
GUCCA ΨΨ CAUA•
CAGGU U GUAU
-3 -1 +1 +3 +5 +7
PIK3R4 test 5′ss
3′
GUCCA ΨΨ CAUA•
Canonical
B
U1
Asymmetric loop 1 Ψ
3′
GUCCA ΨΨ CAUA•
GAGGU U GUAU
-3 -1 +1 +3 +5 +7
POLQ test 5′ss
3′
GUCCA ΨΨ CAUA•
Canonical
C
PIK3R4
5'ss mutation: WT -2C +4C +5C
U1 suppressor: - G5 G4 - G4 G3
Register: C A C A
% Inclusion: 94 2 16 24 78 9 37 94
SD: 2 3 2 3 1 0.5 0 1
Lane: 1 2 3 4 5 6 7 8
POLQ
5'ss mutation: WT -2C +4C +5C +6C -2C +6C
U1 suppressor: - G5 G4 - G4 G3 - G3 G2 - G3 G2
Register: C A C A C A C A
% Inclusion: 100 30 81 75 91 85 89 100 96 94 100 0 0 16
SD: 0 1 4 3 3 5 3 0 2 2 0 0 0 2
Lane: 1 2 3 4 5 6 7 8 9 10 11 12 13 14

Figure 3.6 *PIK3R4* and *POLQ* minigenes exhibit use of the asymmetric loop 1 Ψ register

(A, B) Sequences of test 5’ss from both genes, and representations of both the asymmetric loop 1 Ψ register and the canonical register that the U1 snRNA can adopt to base pair with these 5’ss. **(C)** Native PAGE of radioactive RT-PCR of RNA products of both minigenes. The red box indicates the test exon, green boxes the flanking *MCAD* exons, and the black line represents intronic sequences. The larger band in *PIK3R4* represents the exon inclusion band, while the smaller band represents the exon skipping band. The largest band indicated in *POLQ* represents the use of a cryptic intronic 5’ss, while the second largest band represents exon inclusion, the two bands below that both represent the use of a cryptic exonic 5’ss at two different locations, while the smallest band indicates exon skipping. 5’ss mutations in blue represent mutations that affect both registers, while those in green represent mutations that affect only the asymmetric loop register. U1 suppressors in orange represent U1 suppressors that affect the canonical register, also denoted with a C, while those in purple represent U1 suppressors that affect the asymmetric loop register, also indicated with an A.

3.1.3.2. SMN1/2 minigene mutational analysis and U1 suppressor experiments prove asymmetric 1 Ψ register usage in heterologous context

Again, in order to prove the general applicability of the asymmetric loop 1 Ψ register, the natural 5’ss of exon 7 in the *SMN1/2* minigenes was replaced with a representative optimal test 5’ss sequence (Figure 3.7A), which means the 5’ss can establish the maximum possible number of base pairs with U1 via the asymmetric loop 1 Ψ register, via PCR mutagenesis. After transfection, RNA extraction, and RT-PCR, *SMN1* transcripts displayed 49% exon inclusion while *SMN2* indicated complete exon skipping (Figure 3.5B, lanes 2 and 1 respectively). This indicated usage of the test 5’ss in the *SMN1* context but not *SMN2*, hence only the *SMN1* minigene was used for mutational analysis.

The -2C and +4C mutants, each of which affected both canonical and asymmetric loop 1 Ψ register base pairs, caused complete exon skipping (Figure 3.7B, lanes 3 and 4 respectively; indicated in blue). This indicated that mutating the test 5’ss could alter the splicing pattern of the *SMN1* minigene.

Point mutations that affected only the asymmetric loop register (indicated in green), at positions +5 and +6 of the test 5’ss in the *SMN1* minigene (the +5C and +6C mutants), caused complete exon skipping (Figure 3.7B, lanes 7 and 10 respectively). This revealed that the nucleotides at these positions were important for 5’ss recognition, implying that the test 5’ss was recognized via the asymmetric loop 1 Ψ register.

Suppressor U1 snRNA experiments were performed to determine whether base pairing interactions between the test 5’ss and the U1 were occurring via the canonical or the asymmetric loop 1 Ψ register. For the +4C mutant, neither suppressor U1 acting via the asymmetric loop 1 Ψ register, nor suppressor U1 acting via the canonical register could rescue exon inclusion (Figure 3.7B, lane 6 and 5 respectively); this was perhaps due to the +4C mutation being too strong to rescue. However, with the +5C mutant, suppressor U1 acting via the asymmetric loop 1 Ψ register, U1 G4 (Figure 3.7B, lane 9) weakly rescued exon inclusion, but suppressor U1 acting via the canonical register, U1 G5 (Figure 3.7B, lane 8) could not. For the +6C mutant, exon 7 inclusion was strongly rescued by suppressor U1 acting via the asymmetric loop 1 Ψ register (Figure 3.7B, lane 12), while suppressor U1 acting via the canonical register could not rescue exon 7 inclusion (Figure 3.7B, lane 11).

In all cases, the suppressor U1 acting via the canonical register did not affect the splicing pattern. Only the suppressor U1 acting via the asymmetric loop 1 Ψ register could rescue exon inclusion. Also, the effect of the same U1 G3 suppressor varied between the different mutations in the 5'ss (compare lane 9 with lane 11 in Figure 3.7B), indicating that the effect of the suppressor U1 was specific to each particular register. All this evidence further supported the hypothesis that the test 5'ss was recognized via the asymmetric loop 1 Ψ register.

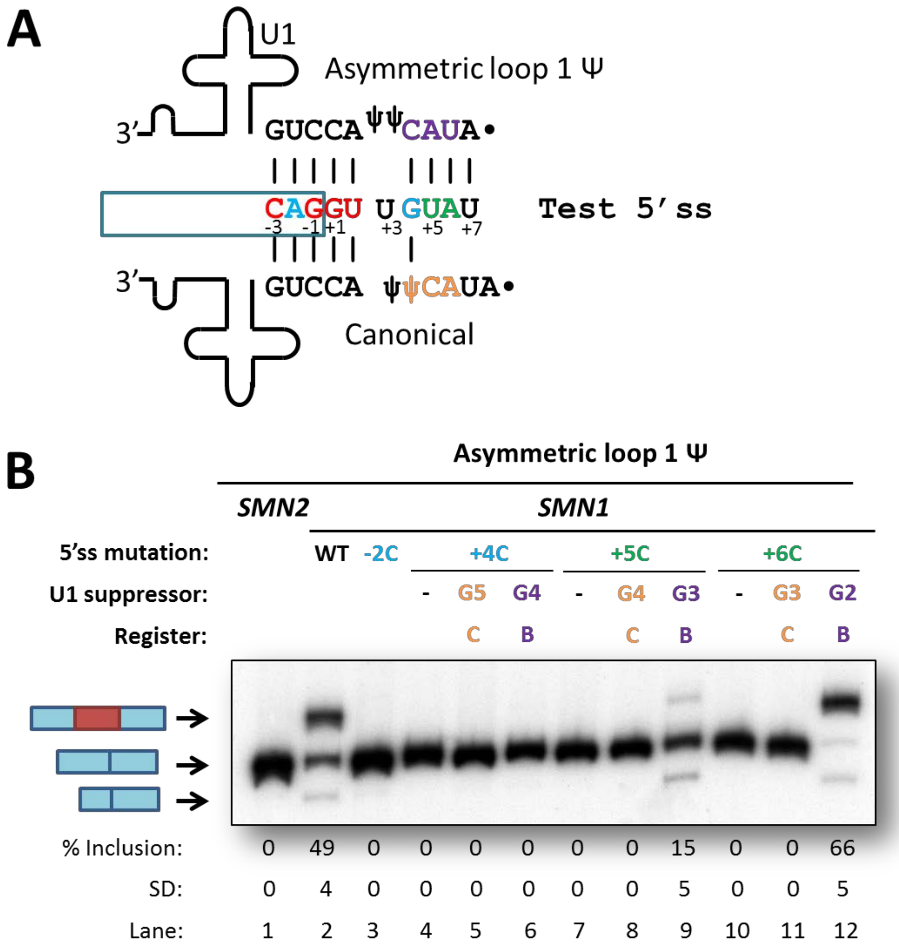


Figure 3.7 *SMN1/2* minigene testing confirms use of the asymmetric loop 1 Ψ register

(A) Sequence of test 5’ss replacing the native 5’ss in the *SMN1/2* minigenes, and representations of both the asymmetric loop 1 Ψ register and the canonical register that the U1 snRNA can adopt to base pair with these 5’ss. **(B)** Native PAGE of radioactive RT-PCR of RNA products of the test *SMN1/2* minigenes. The identity of the various spliced mRNAs is indicated on the left; from large to small, the bands correspond to exon 7 inclusion, exon 7 skipping, and exon 7 skipping with activation of a cryptic 5′ss at position −50 in exon 6. The red box indicates exon 7, blue boxes the flanking exons, and the black line represents intronic sequences. 5’ss mutations in blue represent mutations that affect both registers, while those in green represent mutations that affect only the asymmetric loop register. U1 suppressors in orange represent U1 suppressors that affect the canonical register, also denoted with a C, while those in purple represent U1 suppressors that affect the asymmetric loop register, also indicated with an A.

3.2. Splicing: Registers longer than 1 nucleotide

Roca et al. (2012) also predicted that registers with longer bulges/asymmetric loops, ranging from 2 to 8 nucleotides, might occur in certain 5’ss/U1 helices, for a total of 3,294 cases. Such registers typically involve bulging nucleotides on the 5’ss strand instead of the U1 5’ tail. Although these registers usually confer a smaller energetic advantage (versus the single nucleotide bulge/asymmetric loop registers) over the canonical register, a total of 1,496 such cases were projected to have ΔΔG ≤ −1. These registers were yet to be validated experimentally, and hence testing of some of the shorter predicted bulge registers was carried out in this thesis.

3.2.1. Bulge 2 and bulge 3 registers

Registers that leave two nucleotides unpaired and unopposed on one side of the 5’ss/U1 helix are referred to as bulge 2 registers, while those with three nucleotides unpaired and unopposed are called bulge 3 registers. A total of 1,066 5’ss were predicted to use bulge 2 registers, while another 640 5’ss were predicted to use bulge 3 registers.

3.2.1.1. SMN1/2 testing of bulge 2 (+3,+4) and bulge 3 (+3,+4,+5) registers indicates that bulges larger than 2 might not be energetically stable enough for 5’ss recognition

Bulge 2 (+3,+4) denotes a register whereby positions +3 and +4 in the 5’ss are left unopposed in the 5’ss/U1 helix during 5’ss recognition (Figure 3.8B). They were projected to be used in 320 5’ss. Bulge 3 (+3,+4,+5) designates a register where 5’ss positions +3, +4, and +5 are left unopposed by nucleotides in the U1 5’ tail (Figure 3.8D). 163 5’ss were anticipated to use this register.

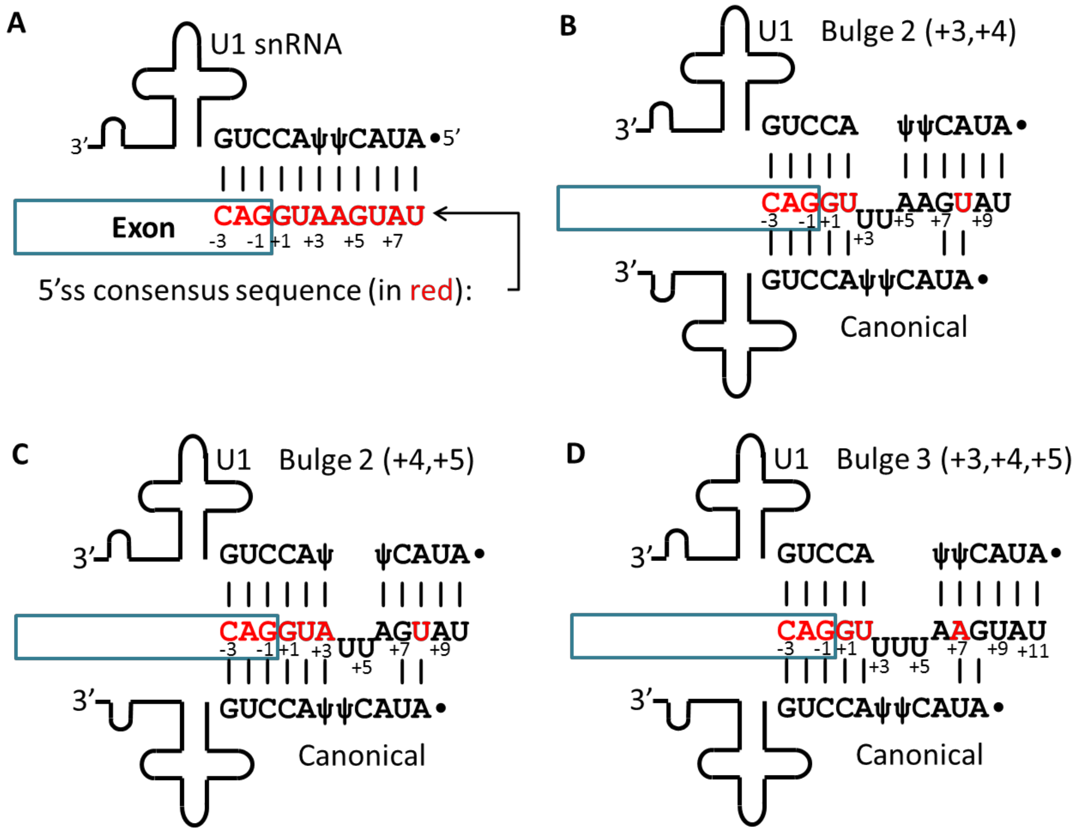


Figure 3.8 Examples of longer bulge registers
Blue boxes represent the test exons. The 5'ss sequence is depicted, with red letters indicating consensus nucleotides in the 5'ss. The sequences above and below represent the U1 snRNA and the registers by which it can interact with the 5'ss. Vertical lines between sequences represent base pairs.

In order to investigate the bulge 2 (+3,+4) and bulge 3 (+3,+4,+5) registers in a heterologous context, the native 5'ss of exon 7 in the *SMN1/2* minigenes (capitalized letters in Figure 3.9A) was replaced with the ideal test 5'ss sequences for the respective bulges, meaning that these 5'ss can establish the maximum possible number of base pairs with U1 via their respective non-canonical registers (Figure 3.9B and C respectively). 'U's were picked to form the bulge in these 5'ss because they were not expected to bind to any other nucleotide in the U1, and because subsequently 'C's would be introduced as mutations. HEK293T cells were transfected with these minigenes, the RNA harvested after 48 hours, and radioactive RT-PCR performed on the total RNA as detailed in the methods. The level of exon 7 inclusion, which reflects the recognition of the test 5'ss, was determined with native PAGE.

Interestingly, all these test 5’ss were found to be very weak in the context of the *SMN1/2* minigenes. In the *SMN1* context, only bulge 2 (+3,+4) retained a low level of exon 7 inclusion (Figure 3.9D, lane 2), while the bulge 3 (+3,+4,+5) experienced complete exon skipping (Figure 3.9D, lane 3). In the *SMN2* context, both bulge 2 and 3 experienced complete exon skipping (Figure 3.9D, lanes 7 and 8 respectively).

In order to enhance the splicing of the bulge 2 and 3 *SMN1/2* minigenes, a known strong intronic splicing silencer (ISS) element in intron 7 just downstream of the test 5’ss (Hua et al., 2008) was mutated (-ISS, see Figure 3.9C and D). The presence of this ISS was shown to decrease exon inclusion of exon 7 via binding of the inhibitory hnRNP A1 or hnRNP A2 protein factors (Hua et al., 2008). As expected, the *SMN1* bulge 2 (+3,+4) -ISS mutant showed improved recognition, with a significant increase in exon 7 inclusion (Figure 3.9A, lane 3). However, its *SMN2* counterpart remained fully skipped (Figure 3.9D, lane 4). This was most instructive, as the -ISS mutation had proved effective in boosting *SMN2* exon 7 inclusion in prior work (Hua et al., 2008), implying that the bulge 2 (+3,+4) test 5’ss was quite weak. Both *SMN1/2* bulge 3 (+3,+4,+5) -ISS mutants also exhibited complete exon 7 skipping (Figure 3.9D, lanes 9 and 10 respectively). This implied that the test 5’ss for bulge 3 (+3,+4,+5) was significantly weaker than that of bulge 2 (+3,+4), which itself was moderately weak. It hinted that bulge registers of 3 nucleotides and longer might not be stable enough for effective 5’ss recognition. Therefore, tests of bulges larger than 2 were not further pursued in our studies.

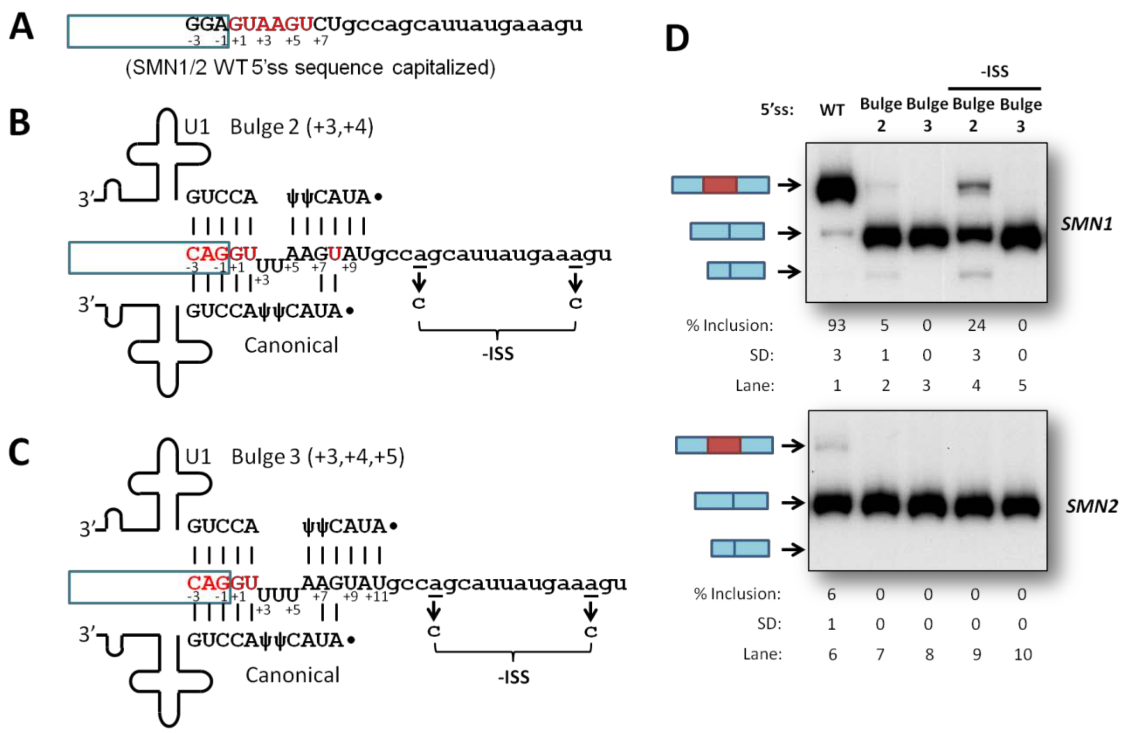


Figure 3.9 *SMN1/2* bulge 2 (+3,+4) and bulge 3 (+3,+4,+5)
(A) *SMN1/2* wild-type sequence, where the capitalized letters represent the natural 5'ss at the 3' end of exon 7. **(B, C)** Sequences of test 5'ss replacing the native 5'ss in the *SMN1/2* minigenes are capitalized, and representations of the bulge 2 (+3,+4) register (in B) and the bulge 3 (+3,+4,+5) register (in C) as well as the canonical register that the U1 snRNA can adopt to base pair with these 5'ss. The -ISS mutation, which eliminates a intronic splicing silencer (ISS) just downstream of the test 5'ss, is also shown. **(D)** Native PAGE of radioactive RT-PCR for the *SMN1*/2 test minigenes. The identity of the various spliced mRNAs is indicated on the left; from large to small, the bands correspond to exon 7 inclusion, exon 7 skipping, and exon 7 skipping with activation of a cryptic 5′ss at position −50 in exon 6. The red box indicates exon 7, blue boxes the flanking exons.

3.2.1.2. SMN1 bulge 2 (+3,+4) -ISS mutational analysis and U1 suppressor experiments show that usage of the bulge 2 (+3,+4) register is possible

Mutational analysis was performed using the *SMN1* bulge 2 (+3,+4) -ISS minigene. As expected, the -1C, +7C, and +8C mutations, which affected both the bulge and canonical registers, led to complete exon 7 skipping (Figure 3.10B, lanes 2, 9, and 12 respectively). This indicated that the splicing pattern could be affected by mutations of the 5'ss. Mutations that affected only the bulge 2 (+3,+4) register, +5C, +6C, and +9C, also resulted in complete exon 7 skipping (Figure 3.10B, lanes 3, 6, and 15 respectively). As these positions can only base pair to U1 in the bulge 2 (+3,+4) register, these findings suggest that these 5’ss are recognized via the bulge 2 (+3,+4) register.

Suppressor U1 snRNA experiments were carried out to define the base pairing register between the test 5’ss and U1, in other words, whether the canonical or the bulge 2 (+3,+4) registers were being used for recognition of the test 5’ss. Thus, co-transfection of HEK293T cells with the appropriate U1 suppressor and the mutant +5C, +6C, +7C, +8C, or +9C *SMN1* bulge 2 (+3,+4) -ISS minigenes was performed. The U1 suppressors for either register did not have any effect on the +6C nor the +7C mutations (Figure 3.10B; lanes 7, 8 and lanes 10, 11 respectively), possibly because the effect of the mutation was too strong to be overcome. For the +5C and +8C mutants, only the U1 suppressor acting in the bulge 2 (+3,+4) register managed to rescue exon 7 inclusion (Figure 3.10B, lanes 5 and 14 respectively), whereas the U1 suppressor for the canonical register had no effect (Figure 3.10B, lanes 4 and 13 respectively). The +9C mutant was also successfully rescued by the U1 suppressor acting in the bulge 2 register, to an even greater extent than that of the unmodified test 5’ss minigene (Figure 3.10B, lane 16). As the canonical register only extends to position +8 of the 5’ss, there were no possible U1 suppressors that can be used for a mutation at +9. Therefore, based on the U1 suppressor data, it demonstrates that the bulge 2 (+3/+4) register is being used for U1 recognition of the artificial 5’ss in the *SMN1* context.

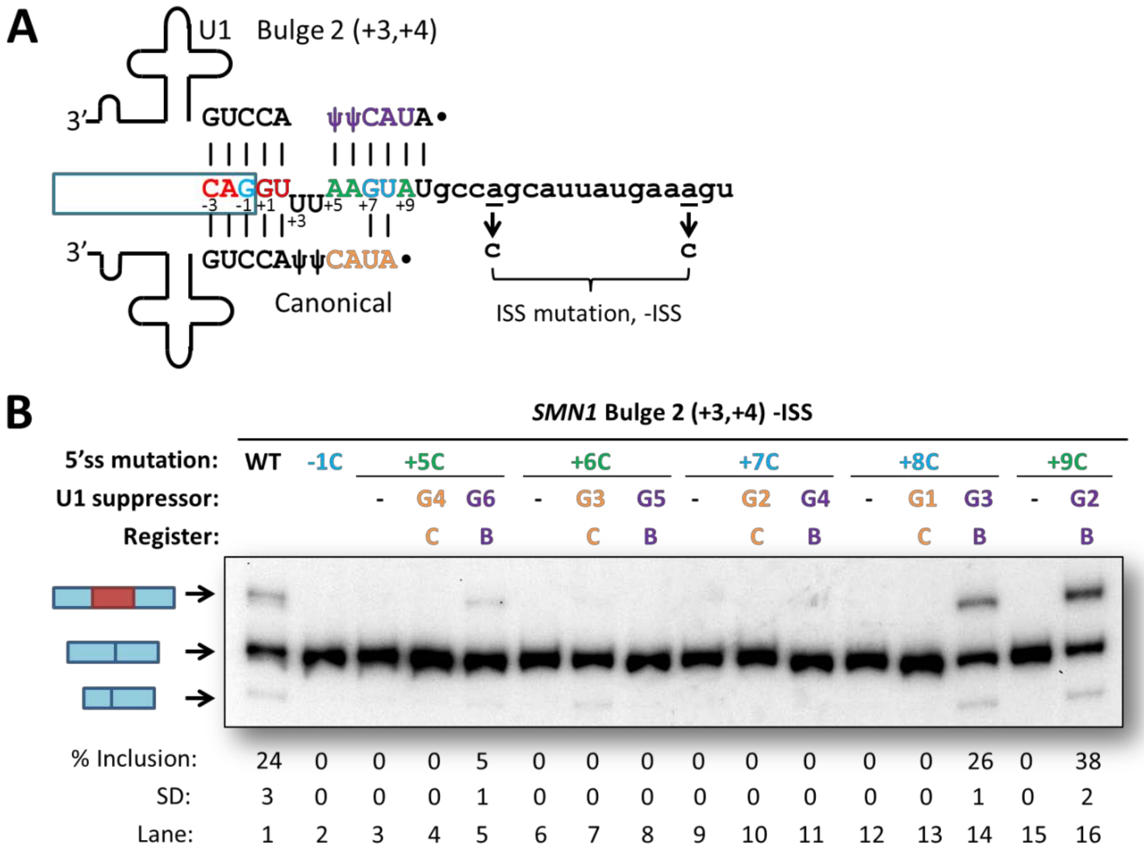

A
U1 Bulge 2 (+3,+4)
3′
GUCCA ψψCAUA•
CAGGUUUAAGUAUgccagcauuaugaaagu
-3 -1 +1 +3 +5 +7 +9
3′
GUCCAψψCAUA•
Canonical
c c
ISS mutation, -ISS
B
SMN1 Bulge 2 (+3,+4) -ISS
5′ss mutation: WT -1C +5C +6C +7C +8C +9C
U1 suppressor: - G4 G6 - G3 G5 - G2 G4 - G1 G3 - G2
Register: C B C B C B C B B
% Inclusion: 24 0 0 0 5 0 0 0 0 0 0 0 0 26 0 38
SD: 3 0 0 0 1 0 0 0 0 0 0 0 0 1 0 2
Lane: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Figure 3.10 *SMN1* bulge 2 (+3,+4) -ISS 5’ss testing
(A) Sequence of test 5’ss replacing the native 5’ss in the *SMN1* minigene, and representations of both the bulge 2 (+3,+4) register and the canonical register that the U1 snRNA can adopt to base pair with these 5’ss. The -ISS mutation is also shown. **(B)** Native PAGE of radioactive RT-PCR of test *SMN1* minigene products. The identity of the various spliced mRNAs is indicated on the left; from large to small, the bands correspond to exon 7 inclusion, exon 7 skipping, and exon 7 skipping with activation of a cryptic 5′ss at position −50 in exon 6. The red box indicates exon 7, blue boxes the flanking exons. 5’ss mutations in blue represent mutations that affect both registers, while those in green represent mutations that affect only the asymmetric loop register. U1 suppressors in orange represent U1 suppressors that affect the canonical register, also indicated with a C, while those in purple represent U1 suppressors that affect the bulge register, also indicated with a B.

3.2.1.3. Predicted natural bulge 2 (+3,+4) 5’ss tested in UMV minigenes showed usage of canonical register

To follow up on the lead that bulge 2 (+3,+4) registers might be viable, candidate naturally-occurring 5’ss were selected based on the criteria provided in the Methods (2.1.2). Many candidates were eliminated due to their location: typically appearing in the middle of an annotated exon in Ensembl (Flicek et al., 2014), indicating that they were uncommonly used (data not shown). Four 5’ss were selected: *PAK3* (exon 7), *SGCE* (exon 10), *DHODH* (exon 6), and *ARHGAP12* (exon 4).

Test exons and their flanking intronic sequences from the genes containing the selected 5’ss of interest were cloned into UMV to create their respective minigenes. HEK293T cells were transfected with these minigenes and radioactive RT-PCR performed on the total RNA as detailed in the methods. The level of test exon inclusion, which reflects the recognition of the test 5’ss, was determined. Here, two candidates were eliminated as the test exons were constitutively skipped; the *PAK3* and the *SGCE* minigenes (data not shown). Mutational analysis on the remaining two minigenes resulted in the conclusion that the 5’ss was being recognized via the canonical register instead (data not shown). Basically, both the tested 5’ss contain +5G, which is a consensus nucleotide that typically base pairs to the U1 in the canonical register. This phenomenon is investigated later on in section 3.3, and is discussed in section 4.1.5.

Therefore, as of this writing, no naturally occurring 5’ss that fit the criteria for bulge 2 (+3,+4) register usage have been shown to actually utilize the register for recognition.

3.2.1.4. Minigene testing of bulge 2 (+4,+5) register indicates possible use of register

Bulge 2 (+4,+5) denotes a register whereby positions +4 and +5 in the 5’ss are left unopposed in the 5’ss/U1 helix during 5’ss recognition (Figure 3.8C). 655

human 5’ss sequences were predicted to possess an energetic advantage if base paired to U1 via this register, which increases the number of maximum possible base pairs in the 5’ss/U1 duplex to 10, versus just 6 in the canonical register.

Exon 5 and its flanking intronic sequences from the *HPS4* gene, which contained the 5’ss predicted to use the bulge 2 (+4,+5) register, were cloned into UMV to create the minigene, and analyzed by transfection and RT-PCR.

The *HPS4* minigene produced transcripts with full test exon inclusion, reflecting use of the test 5’ss (Figure 3.11B, lane 1). Mutational analysis of the test 5’ss was carried out. The -2C mutants caused significant loss of exon inclusion as expected (Figure 3.11B, lane 2). However, not all mutants that affected both canonical and bulge 2 (+4,+5) register base pairs were as effective, as the +7C and +8C mutants did not affect exon inclusion (Figure 3.11B, lanes 3 and 7 respectively). Additionally, the +9C point mutation that affected only the bulge 2 (+4,+5) register in the *HPS4* test 5’ss (indicated in green), did not affect exon inclusion (Figure 3.11B, lane 11).

The reduced effects of the +7C, +8C and +9C mutants was perhaps due to the innate strength of the test 5’ss. In order to counteract this, double point mutations were made. The double mutations -2C +7C, -2C +8C, and -2C +9C synergized weakly, causing a slight but consistent increase in exon skipping beyond the effect of -2C alone (Figure 3.11B, lanes 4, 8, and 12 respectively). This revealed that the nucleotides at these positions played a modest role in 5’ss recognition, although an important one. In fact, as these positions can only base pair to U1 in the bulge register, these findings suggest that these 5’ss are recognized via the bulge 2 (+4,+5) register.

Suppressor U1 snRNA experiments were performed to determine whether base pairing interactions between the *HPS4* test 5’ss and the U1 were occurring via the canonical or the bulge 2 (+4,+5) register. In the -2C +7C mutants, test exon inclusion was not rescued by suppressor U1 rescuing a base pair in the bulge 2 (+4,+5) register (Figure 3.11B, lane 6), nor suppressor U1 rescuing a base pair in the canonical register (Figure 3.11B, lane 5); in fact, the use of suppressors appeared to reduce exon inclusion. Similarly, for the -2C +8C mutants, test exon inclusion could not be rescued by suppressor U1 acting via the bulge 2 (+4,+5) register, U1 G3 (Figure 3.11B, lane 10), nor by suppressor U1 acting via the canonical register, U1 G1 (Figure 3.11B, lane 9). The use of suppressors appeared to reduce exon inclusion in this situation as well. However, with -2C +9C mutants, suppressor U1 acting via the bulge 2 (+4,+5) register managed to restore a certain level of exon inclusion (Figure 3.11B, lane 13). As the canonical register did not reach that far, no U1 suppressors could be used.

In all cases, the suppressor U1 acting via the canonical register did not restore correct splicing patterns. Also, the effect of the same suppressor U1 (U1 G3 and U1 G2) varied between the different mutations in the 5'ss (compare lane 6 with lane 10 and lane 5 with lane 13 respectively in Figure 3.11B), indicating that the effect of the suppressor U1 was specific to the register. This evidence tentatively supported the hypothesis that the test 5'ss were recognized via the bulge 2 (+4,+5) register.

In order to prove the general applicability of the bulge 2 (+4,+5) register, the natural 5'ss of exon 7 in the *SMN1/2* minigenes with the -ISS mutation was replaced with a optimal test 5'ss sequence which can establish the maximum number of base pairs with U1 via the bulge 2 (+4,+5) register (namely the sequence "CAG/GUAUUAGUAU", where the slash mark represents the exon-intron boundary) via PCR mutagenesis. After transfection, RNA extraction, and RT-PCR, both *SMN1* and *SMN2* bulge 2 (+4,+5) -ISS minigene transcripts displayed full exon inclusion (data not shown), showing usage of the test 5'ss. As the only difference between the bulge 2 (+3,+4) and the bulge 2 (+4,+5) *SMN1/2* is the position of the double 'UU' motif that forms the bulge during U1 base pairing in the non-canonical register, this provides a tentative hint that 5'ss using the bulge 2 (+4/+5) register may be stronger than those using the bulge 2 (+3/+4) register, as the corresponding exon inclusion levels are much lower.

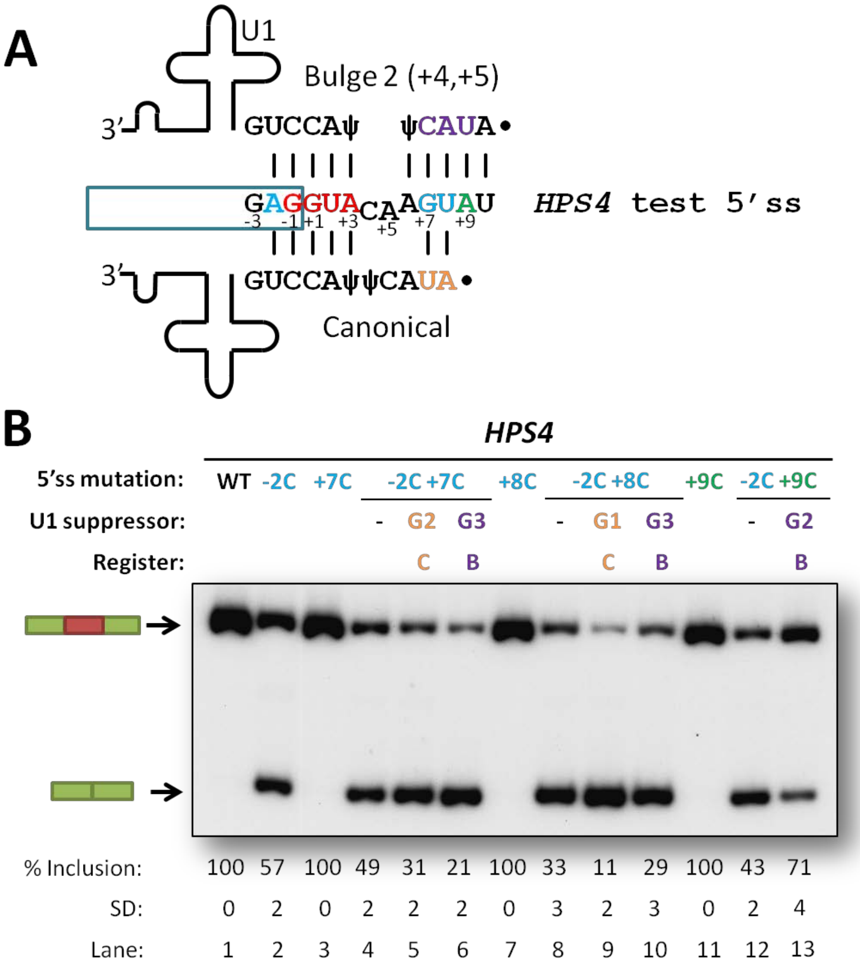

A
U1
Bulge 2 (+4,+5)
3′
GUCCAψ ψCAUA•
GAGGUACAAGUAU
-3 -1 +1 +3 +5 +7 +9
HPS4 test 5′ss
3′
GUCCAψψCAUA•
Canonical
B
HPS4
5′ss mutation: WT -2C +7C -2C +7C +8C -2C +8C +9C -2C +9C
U1 suppressor: - G2 G3 - G1 G3 - G2
Register: C B C B B
% Inclusion: 100 57 100 49 31 21 100 33 11 29 100 43 71
SD: 0 2 0 2 2 2 0 3 2 3 0 2 4
Lane: 1 2 3 4 5 6 7 8 9 10 11 12 13

Figure 3.11 *HPS4* minigene testing of bulge 2 (+4+,5) register
(A) Sequences of *HPS4* test 5’ss, and representations of both the bulge 2 (+4,+5) register and the canonical register that the U1 snRNA can adopt to base pair with these 5’ss. **(B)** Native PAGE of radioactive RT-PCR of RNA products of the *HPS4* minigene. The red box indicates the test exon, and green boxes the flanking *MCAD* exons. 5’ss mutations in blue represent mutations that affect both registers, while those in green represent mutations that affect only the asymmetric loop register. U1 suppressors in orange represent U1 suppressors that affect the canonical register, also indicated with a C, while those in purple represent U1 suppressors that affect the bulge register, also indicated with a B.

3.3. Splicing: 5’ss with +5G always base pair in the canonical register, stabilized by adjacent non-canonical +4U-5Ψ base pair (+5G hypothesis)

While studying the various 5’ss recognition registers, the intriguing behavior of a few test 5’ss warranted further investigation. The *RPS6KC1* exon 4, *DNAI1* exon 9, and *CCDC132* exon 15 test 5’ss were predicted to base pair with U1 in the asymmetric loop 1 (+3/+4) register, while the *DHODH* exon 6 and *ARHGAP12* exon 4 test 5’ss were supposed to base pair with U1 in the bulge 2 (+3/+4) register (Figure 3.12A). However, when mutational analysis was performed, mutations that affected both registers (in blue) impaired 5’ss recognition, but mutations that affected solely the alternative register (in green) did not cause significant loss of exon inclusion. This strongly suggested that these 5’ss base pair in the canonical register, which contradicts the UNAfold predictions based on the free energy of each register. The preliminary characterization of *DHODH* is one such example (Figure 3.12B).

Importantly, it was observed that all these 5’ss have G at the +5 position and U at the +4 position. If the canonical register is used, the base pair between +5G on the 5’ss and C4 is predicted to be thermodynamically unstable (Gutell, 2012), as there are no flanking base pairs to help stabilize it. Nevertheless, our tests showed that +5G makes a significant contribution to 5’ss recognition, because +5C mutations in the 5’ss caused significant loss of exon inclusion (see below). Hence, we hypothesized that 5’ss will base pair with U1 in the canonical register as long as +5G is present in the 5’ss, henceforth called the “+5G hypothesis”. Concurrently, we also postulated that a non-canonical interaction between +4U on the 5’ss and Ψ5 (Figure 3.12A, denoted with a question mark) might help stabilize the ‘lone’ +5G-C4 base pair, as such a role has been proposed before for the same pair of nucleotides in yeast pre-mRNA splicing (Libri et al., 2002).

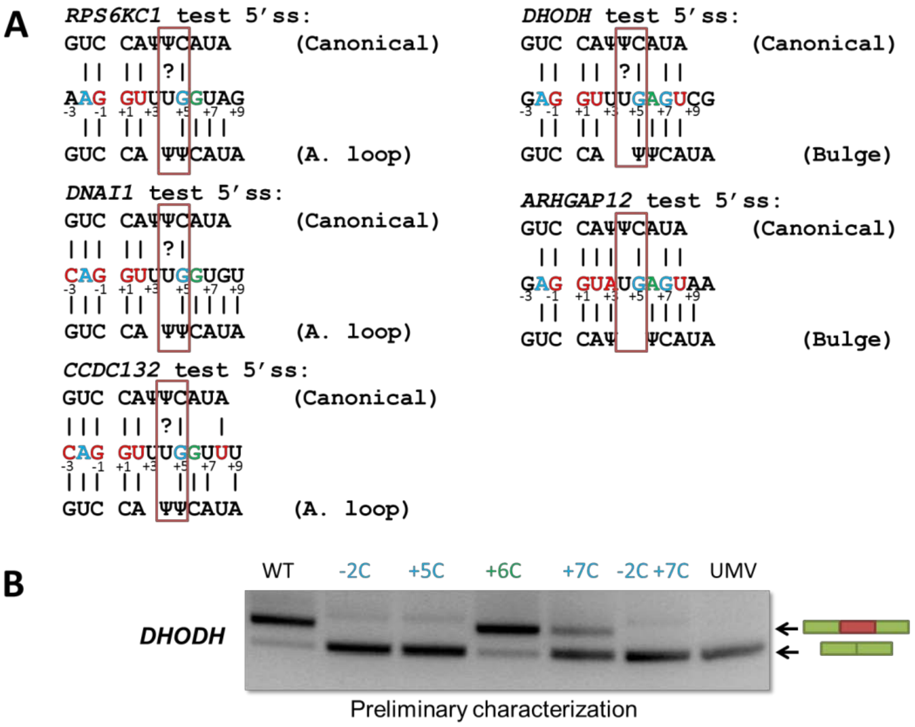


Figure 3.12 5'ss that behaved unexpectedly during characterization (A) Sequences of the 5'ss, with predicted U1 binding registers. Question marks indicate hypothesized U-Ψ interaction. **(B)** An example of test 5'ss characterization, *DHODH* test minigene splice patterns with mutational analysis. UMV splicing pattern is also displayed. Blue denotes mutants that affect both registers, green denotes mutants that affect only the alternative (asymmetric/bulge) registers.

3.3.1. Testing the +5G hypothesis in UMV minigenes reveals canonical register usage in native 5'ss

The *DNAI1* and *CCDC132* minigenes were used to test the +5G hypothesis. Test exons and their flanking intronic sequences from these two genes containing the 5'ss of interest (Figure 3.13A) were cloned into UMV to create their respective minigenes. HEK293T cells were transfected with these minigenes and radioactive RT-PCR performed on the total RNA as detailed in the methods. The level of test exon inclusion, which reflects the recognition of the test 5'ss, was determined.

Both *DNAI1* and *CCDC132* minigenes produced transcripts with more than 90% test exon inclusion, reflecting strong use of the test 5'ss (Figure 3.13B, lanes 1). Mutational analysis of the test 5'ss was carried out. The -2C and the

+5C mutants, each of which affected both canonical and asymmetric loop 1 (+3/+4) register base pairs, caused significant loss of exon inclusion in both the *DNAI1* and *CCDC132* contexts (Figure 3.13B, lanes 2 and 4 respectively; shown in blue). This indicated that mutations in the 5'ss could affect the splicing pattern by disrupting 5'ss/U1 interactions.

Point mutations that specifically affected the asymmetric loop register, at the +6 position for both *DNAI1* and *CCDC132*, had a relatively minor effect on exon inclusion (Figure 3.13B, lanes 6). In order to enhance the effect of the +6C mutation, the double mutation -2C +6C was made in both minigenes. However, the double mutation did not synergize; exon inclusion levels instead seemed to mimic that of the -2C point mutant (Figure 3.13B, lanes 7). These point mutations revealed that the nucleotide at the +6 position was less important for test 5'ss recognition (especially versus position +5). As position +6 can only form a Watson-Crick base pair to U1 in the asymmetric loop register, this data suggested that these 5'ss are probably not recognized via the asymmetric loop 1 (+3/+4) register.

Mutating position +4 of the test 5'ss to C caused a reduction in exon inclusion levels for both *DNAI1* and *CCDC132* (Figure 3.13B, lanes 3). This result was surprising, since the mutation was not projected to affect any 5'ss/U1 base pairing interactions, neither in the canonical nor the asymmetric loop 1 (+3/+4) register. +4C +5C double mutants (Figure 3.13B, lanes 5) synergized to cause further loss of exon inclusion in *DNAI1*, but not in *CCDC132*, where it appeared to mimic the effects of the +5C mutation. This hinted that perhaps the *CCDC132* test 5'ss did not depend on +4U upon loss of +5G, while the *DNAI1* test 5'ss relied on the presence of both +4U and +5G for 5'ss recognition.

Suppressor U1 snRNA experiments were performed to confirm whether base pairing interactions between the test 5'ss and the U1 occurred via the canonical register. Both the +5C mutants were rescued by suppressor U1 acting via the canonical register (Figure 3.13B, lanes 11), and not by the control suppressor (Figure 3.13B, lanes 10). Similarly, the +4C +5C mutants were rescued by suppressor U1 acting via the canonical register (Figure 3.13B, lanes 16), rather than the control suppressor (Figure 3.13B, lanes 15). This confirmed that these 5’ss were recognized via the canonical register. Again, this interaction occurred despite the predicted thermodynamic instabilities of such a configuration, because the +5G-C4 base pair is in principle flanked by mismatches.

To test the contribution of +4U in the test 5’ss to recognition, the +5C and +4C +5C mutants were co-transfected with U1 C5 and U1 G4C5. U1 C5 was anticipated to disrupt the hypothesized +4U-Ψ5 interaction, while U1 G4C5 should simultaneously disrupt the +4U-Ψ5 interaction while restoring base pairing between position +5 in the 5’ss and position 4 in U1.

In the *DNAI1* minigene, U1 C5 suppression caused reduced exon inclusion in both +5C and +4C +5C mutants (Figure 3.13B, lanes 12 and 17 respectively). U1 G4C5 usage also caused a slight increase in the loss of exon inclusion in *DNAI1* +5C and +4C +5C (Figure 3.13B, lanes 13 and 18 respectively), but the exon inclusion levels are still higher than those observed with use of U1 C5. These outcomes show that +4U-Ψ5 interaction plays a role in stabilizing 5’ss/U1 interactions, as well as strengthens the +5G-C4 base pair. However, it also indicates that the base pair between position +5 of the 5’ss and position 4 in U1 can still form even when the +4U-Ψ5 interaction is disrupted, at least in context of the *DNAI1* minigene. Also, based on the +4C +5C results, the +4C-Ψ5 interaction contributes more towards 5’ss/U1 helix stability than that of +4C-C5, which probably does not form a base pair at all.

Strangely, U1 C5 strongly rescued exon inclusion in both *CCDC132* +5C and +4C +5C mutants, even surpassing that of U1 G4. The probable cause for this is the fact that U1 C5 can potentially base pair in a shifted -4 register to the *CCDC132* +5C and +4C +5C mutant 5’ss sequence (Figure 3.13C). We speculate that this might allow a more stable interaction than the canonical register because of an extra strong G-C base pair, allowing mutant 5’ss recognition and encouraging exon inclusion. Paradoxically, U1 G4C5 caused a further loss of exon inclusion in both *CCDC132* +5C and +4C +5C mutants (slightly lower than normal), probably because the G4 mutation breaks another strong G-C base pair at +1 of the shifted -4 register.

Nevertheless, the consistent disruption in correct splicing in the +5C and +4C +5C mutants which can be rescued by U1 suppressors in the canonical register leads us to conclude that the presence of +5G is necessary for recognition of these 5’ss in the canonical register. Additionally, the effect of the +4C mutation, together with the U1 C5 suppression of exon inclusion in the +5C and +4C +5C mutants which can only be partially rescued by U1 G4C5, suggest that the +4U-Ψ5 interaction contributes to 5’ss/U1 helix stability and enhances the strength of the +5G-C4 base pair.

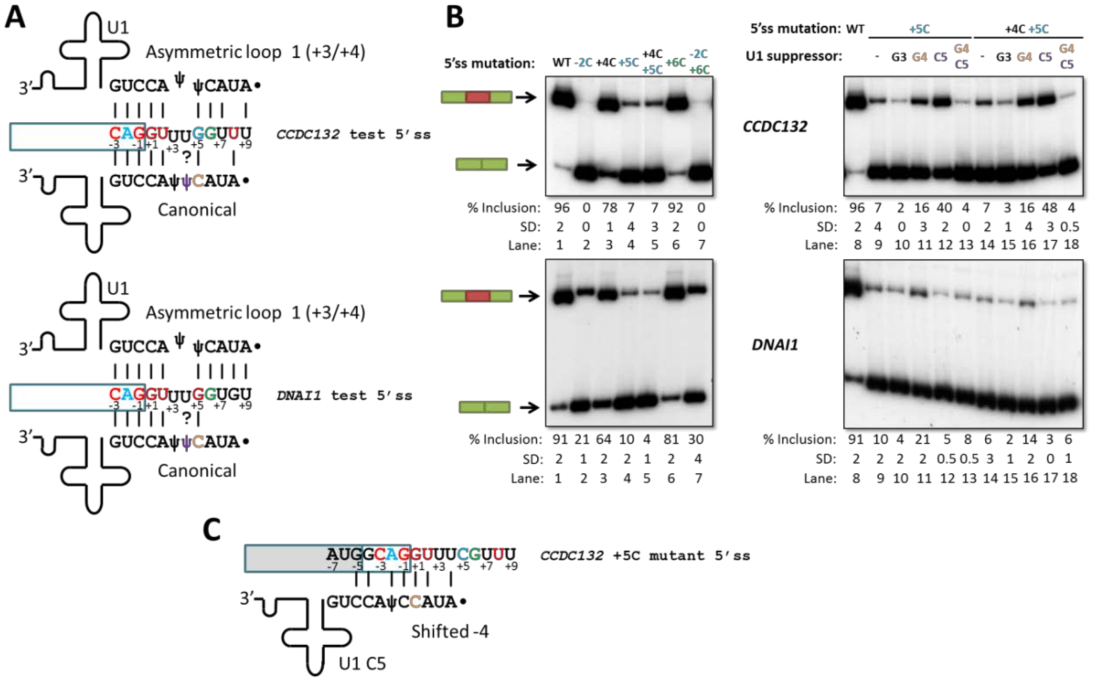

A
U1
Asymmetric loop 1 (+3/+4)
3'
GUCCA ψ ψCAUA•
CAGGUUUGGUUU
-3 -1 +1 +3 +5 +7 +9
CCDC132 test 5'ss
?
3'
GUCCAψψCAUA•
Canonical
U1
Asymmetric loop 1 (+3/+4)
3'
GUCCA ψ ψCAUA•
CAGGUUUGGUGU
-3 -1 +1 +3 +5 +7 +9
DNAI1 test 5'ss
?
3'
GUCCAψψCAUA•
Canonical
B
5'ss mutation: WT -2C +4C +5C +4C +5C +6C -2C +6C
CCDC132
% Inclusion: 96 0 78 7 7 92 0
SD: 2 0 1 4 3 2 0
Lane: 1 2 3 4 5 6 7
5'ss mutation: WT +5C +4C +5C
U1 suppressor: - G3 G4 C5 G4 C5 - G3 G4 C5 G4 C5
% Inclusion: 96 7 2 16 40 4 7 3 16 48 4
SD: 2 4 0 3 2 0 2 1 4 3 0.5
Lane: 8 9 10 11 12 13 14 15 16 17 18
DNAI1
% Inclusion: 91 21 64 10 4 81 30
SD: 2 1 2 2 1 2 4
Lane: 1 2 3 4 5 6 7
% Inclusion: 91 10 4 21 5 8 6 2 14 3 6
SD: 2 2 2 2 0.5 0.5 3 1 2 0 1
Lane: 8 9 10 11 12 13 14 15 16 17 18
C
AUGGCAGGUUUCGUUU
-7 -5 -3 -1 +1 +3 +5 +7 +9
CCDC132 +5C mutant 5'ss
3'
GUCCAψCCAUA•
Shifted -4
U1 C5

Figure 3.13 *CCDC132* and *DNAI1* 5'ss testing of +5G hypothesis
(A) Sequences of test 5'ss from both genes, and representations of both the asymmetric loop register and the canonical register that the U1 snRNA can adopt to base pair with these 5'ss. Question marks represent the potential +4U-Ψ5 interaction. **(B)** Native PAGE of radioactive RT-PCR of RNA products of both minigenes. The red box indicates the test exon, green boxes the flanking *MCAD* exons. 5'ss mutations in blue represent mutations that affect both registers, while those in green represent mutations that affect only the asymmetric loop register. U1 suppressors that target position +5 of the test 5'ss are indicated in orange, while those in purple represent U1 suppressors that affect interactions with position +4 of the test 5'ss. **(C)** Potential *CCDC132* +5C base pairing shifted -4 register with U1 C5. The grey box demarcates the new 5'ss for the shifted register.

3.4. Revised list of 5’ss predicted to use bulge/asymmetric loop registers

In previous work (Roca et al., 2012), it was discovered that 5’ss positions +2 to +5, and Ψ at U1 position 5 or 6, are flexible enough to be bulged in certain 5’ss/U1 RNA helices. In the same study, the base pairing register and minimum free energy for a data set of 201,541 well-annotated human 5’ss sequences and the 5’ end of U1 (ΔG1, in kilocalories per mole) were estimated using UNAFold hybrid (Markham and Zuker, 2008). In a second run, they acquired the free energies for these 5’ss in the canonical register (ΔG2). From this, they predicted that 10,248 5’ss (5.1% of the total) would base pair to U1 using a bulge/asymmetric loop register. Of those 10,248 5’ss, a further 5,877 5’ss were predicted to be significantly more thermodynamically stable using the alternative register as opposed to the canonical (ΔΔG ≤ −1 kcal/mol, where ΔΔG = ΔG1- ΔG2). In the same study, the bulge 1 (+2,+3), bulge 1 (+3), bulge 1 (+3,+4,+5), bulge 1 (+4), bulge 1 (+5), and bulge 1 Ψ registers were experimentally authenticated. They explained that these registers would account for a total of 3,016 5’ss (1.5% of all 5’ss).

Using the experimental results presented earlier, the list of non-canonical 5’ss has been further refined (Table 3.1). If we apply the +5G hypothesis to the entire list of 5’ss, a significant number of the predicted non-canonical 5’ss are eliminated from the pool of potential candidates. Out of the 10,248 predicted non-canonical 5’ss, 4,766 (46.5%) were found to have G at position +5 of the 5’ss (henceforth referred to as +5G-type 5’ss). If the +5G hypothesis is accurate for each case, these 5’ss will employ the canonical register instead of the predicted register. In that scenario, only 5,482 5’ss (2.72% of all 5’ss) can be considered for non-canonical register usage, a significant reduction in potential bulge/asymmetric loop candidate 5’ss. Re-evaluation of the bulge 1 (+2,+3), bulge 1 (+3), bulge 1 (+3,+4,+5), bulge 1 (+4), bulge 1 (+5), and bulge 1 Ψ registers in light of the +5G hypothesis reduces the proposed number of 5’ss using these registers from 3,016 to 2,356, or 1.17% of all 5’ss.

However, certain bulges in the exonic section of the 5’ss, like bulge 1 (-1) and bulge 2 (-1,-2) may not be affected by the +5G hypothesis, which if taken into account in the calculations will give 8,439 total candidates. On the other hand, the bulge 1 (-1) register, which involves bulging the last nucleotide in the exonic section of the 5’ss sequence, did not pass the mutational analysis test in both *SMN1/2* (Roca et al., 2012) and UMV test minigenes (Luo and Roca, unpublished data). In addition, all 5’ss predicted to use the related bulge 2 (-1,-2) register, which involves bulging the last two exonic nucleotides of the 5’ss, have a very low projected thermodynamic stability advantage over the canonical register (ΔΔG = 0.1 kcal/mol). This leads us to postulate that these registers are not feasible for 5’ss recognition anyway, so they were also expunged from the list of curated 5’ss. Doing so reduces the curated number of 5’ss to 5,460.

Analyzing the remaining predicted 5’ss, it was found that the validated asymmetric loop 1 register 5’ss constitute a total of 760 5’ss, which represents 0.38% of all 5’ss. 570 of these validated 5’ss (75%) show a marked increase in predicted 5’ss thermodynamic stability using the predicted register versus the canonical register. In addition, bulge 2 (+3,+4) and bulge 2 (+4,+5) 5’ss comprise a total of 653 curated 5’ss, which account for a further 0.32% of all 5’ss.

From this work, a curated and more reliable list of 5’ss that use bulge/asymmetric loop registers has been created. It is expected that further experiments will further improve the quality and accuracy of this list.

Table 3.1 Numbers and distribution of predicted non-canonical 5’ss

Register	Example 5’ss	Predicted 5’ss number	+5G-type 5’ss	Curated predicted 5’ss number	Curated average ΔΔG, kcal/mol	Number of ΔΔG ≤ -1, kcal/mol
Bulge 1 (-1)	CAGU/GUAUGUAU	2,913	2,891	0	NA	NA
Bulge 1 (+2)	AAG/GCUGAGUAC	1	0	1	-4.90	1
Asy. Loop 1 (+2/+3)	AAG/GCAAAGUUU	2	0	2	-0.95	1
Asy. Loop 1 (+2/+3/+4)	CAG/GCAUAGUUU	6	5	1	-1.10	1
Asy. Loop 1 (+2/+3/+4/+5)	AAG/GCAUCGUAU	1	0	1	-0.10	0
Bulge 1 (+2/+3)	CAG/GUUAAGUAU	68	14	54	-2.21	45
Bulge 1 (+3)	AAG/GUCAAGUAU	51	15	36	-1.65	28
Asy. Loop 1 (+3/+4)	GAG/GUUCAGUAU	348	297	51	-2.31	35
Asy. Loop 1 (+3/+4/+5)	CAG/GUUUUGUAG	52	0	52	-1.73	51
Bulge 1 (+3/+4/+5)	CAG/GUAGAGUAU	579	0	579	-1.16	368
Bulge 1 (+4)	CAG/GUAUAGUAU	1,115	464	651	-1.66	462
Asy. Loop 1 (+4/+5)	CAG/GUAUUGUAU	653	59	594	-1.52	424
Bulge 1 (+5)	CAG/GUAAUGUAU	535	0	535	-1.25	468
Asy. Loop 1 Ψ	CAG/GUUGUAU	115	0	115	-2.27	111
Bulge 1 Ψ	CAG/GUAGUAU	501	0	501	-0.63	75
Bulge 2 (-1,-2)	CAGUU/GUAAGUAU	66	66	0	NA	NA
Bulge 2 (+2,+3)	CAG/GCAUAAGCCA	2	1	1	-0.63	0
Bulge 2 (+3,+4)	CAG/GUUUAAGUGA	320	252	68	-1.09	29
Asy. Loop 2		470	56	414	-1.36	233
Bulge 2 (+4,+5)	CAG/GUACCAGUAU	655	70	585	-1.19	364
Bulge 2 (+5,+6)	CAG/GUGACCGUAU	17	0	17	-0.42	1
Bulge 2 Ψ	CAG/GUGUAU	6	0	6	-0.10	0
Bulge 3 (+3,+4,+5)	GAG/GUUUUAGGUAU	163	62	101	-1.17	54
Bulge 3 (+4,+5,+6)	CAG/GUACCUAGUAU	440	88	352	-0.99	182
Bulge 3 (+5,+6,+7)	CAG/GUAAUUUGUAU	37	0	37	-0.34	0
Asy. Loop 3		285	133	152	-1.06	83
Bulge/Asy. Loop 4-8		839	291	548	-0.61	117
Multiple Bulges		8	2	6	-3.30	6
	Total:	10,248	4,766	5,460		3,139

Registers highlighted in light blue were experimentally validated earlier (Roca et al., 2012), those in dark blue have been validated in this work, and those in pale red are eliminated due to experimental evidence. Boxed registers in blue have experimental evidence indicating their usage, while those in black have experimental evidence indicating otherwise. Registers in grey may be eliminated in the future due to being longer than 2 nucleotides, but are retained as no experimental evidence can yet fully rule them out. “Asy. Loop” is a contraction for “asymmetric loop”. “NA” stands for “non-applicable”. ΔΔG = ΔG1 - ΔG2, where ΔG1 represents the minimum free energy of the 5’ss sequences and the 5’ end of U1 calculated by UNAFold (Markham and Zuker, 2008), and ΔG2 indicates the estimated free energies for these 5’ss in the canonical register.

3.5. XCI PROJECT 1: Establishing a genetic screen for the identification of genes involved in XCI

In this project, the first task was to insert an extra copy of the XIC or the *Xist* gene into the single X chromosome of male ES cells in order to induce ectopic XCI. If this succeeded, the differentiated transgenic cells were expected to experience complete mortality due to the inactivation of X-linked genes critical to survival.

The next step would then be to prepare a lentiviral shRNA screen for XCI-critical genes with these cells. The transgenic cells would be transfected with a lentiviral shRNA library, which would silence a wide variety of genes. If the transfected transgenic ES cells could be rescued, and thus survive differentiation, the genes silenced by the shRNA might be vital for XCI. It would then be possible to identify those genes by high-throughput sequencing of lentiviral tags from the surviving cells. Repeated screens would be able to build up a database of XCI-critical genes and eliminate false positives.

3.5.1. Targeted XIC insertion

Initially, a “shotgun” approach was taken to insert the additional XIC, seeking a random chance insertion in the X chromosome. Electroporation of J1 wild-type male murine ES cells was performed with unaltered and unlinearized RP23-11P22 BAC (containing the XIC), together with BlpI-digested pEZ-Frt-lox-DT plasmid to provide G418 antibiotic resistance. However, this strategy did not yield any usable cell lines, primarily due to low colony viability.

Next, a targeted approach was conceived to achieve XIC insertion. This approach involved the use of a modified BAC known as CH29-76M9-m*Tsix* (Figure 3.14A). Its genomic insert contains a XIC sequence that is ~190kb long, encompassing the entirety of the *Xist* and *Tsix* genes, including their promoter regions. A sequence consisting of a loxP site flanked by hygromycin resistance and red fluorescent protein (RFP) – TomatoRed – selection marker genes is inserted into the *Tsix* promoter, preventing *Tsix* transcription and thus relieving its inhibition of *Xist*. Therefore, targeted insertion of the *Xist* gene should be possible via Cre-mediated recombination of the BAC into the loxP site on the X chromosome of the Ainv15 male murine ES cells (Figure 3.14B).

Ainv15 cells were electroporated with CH29-76M9-m*Tsix*, together with pSALK-CRE plasmid to provide short-lived Cre expression. This strategy generated numerous healthy ES colonies under hygromycin selection. After colony picking and expansion in 6-well plates, all cell lines were checked for RFP expression; cell lines lacking RFP expression were discarded. Consequently, we acquired and froze down a total of 79 RFP-expressing and hygromycin-resistant cell lines. A separate but identical electroporation performed by Dr. Zhang yielded 44 additional cell lines.

3.5.1.1. Transgenic ES cells genotyping identified autosomal insertions

DNA-FISH on metaphase chromosome spreads of all 123 CH29-76M9-m*Tsix* cell lines produced was performed in order to detect the insertion site of the transgenic *Xist* gene. Nick translation was used to generate the FISH probes from the CH29-79M9-m*Tsix* BAC, or from large DNA constructs possessing the *Xist* gene. The integration site was supposed to occur in the X chromosome, especially since the Cre-mediated recombination should target insertions towards the loxP site located downstream of the *Hprt1* gene in Ainv15 cells. However, none of the cell lines generated in the prior experiment had the transgene inserted in the X chromosome, which would appear as two pairs of pinpoint signals on the same chromosome in the DNA-FISH image.

Instead, all the insertions found were autosomal, showing two or more pairs of pinpoint signals on different chromosomes, or only showed up as a single signal pair (Figure 3.15). In some of the single signal cell lines, homologous recombination might have occurred between the transgenic insert and the X-chromosome of the Ainv15 cell, as the sequences are similar. Therefore, the single signal could indicate either the location of the native *Xist* (no insertion) or the exogenous *Xist* (homologous recombination) on the X chromosome. The autosomal-to-single signal genotype occurred roughly in a 1:1 ratio (data not shown).

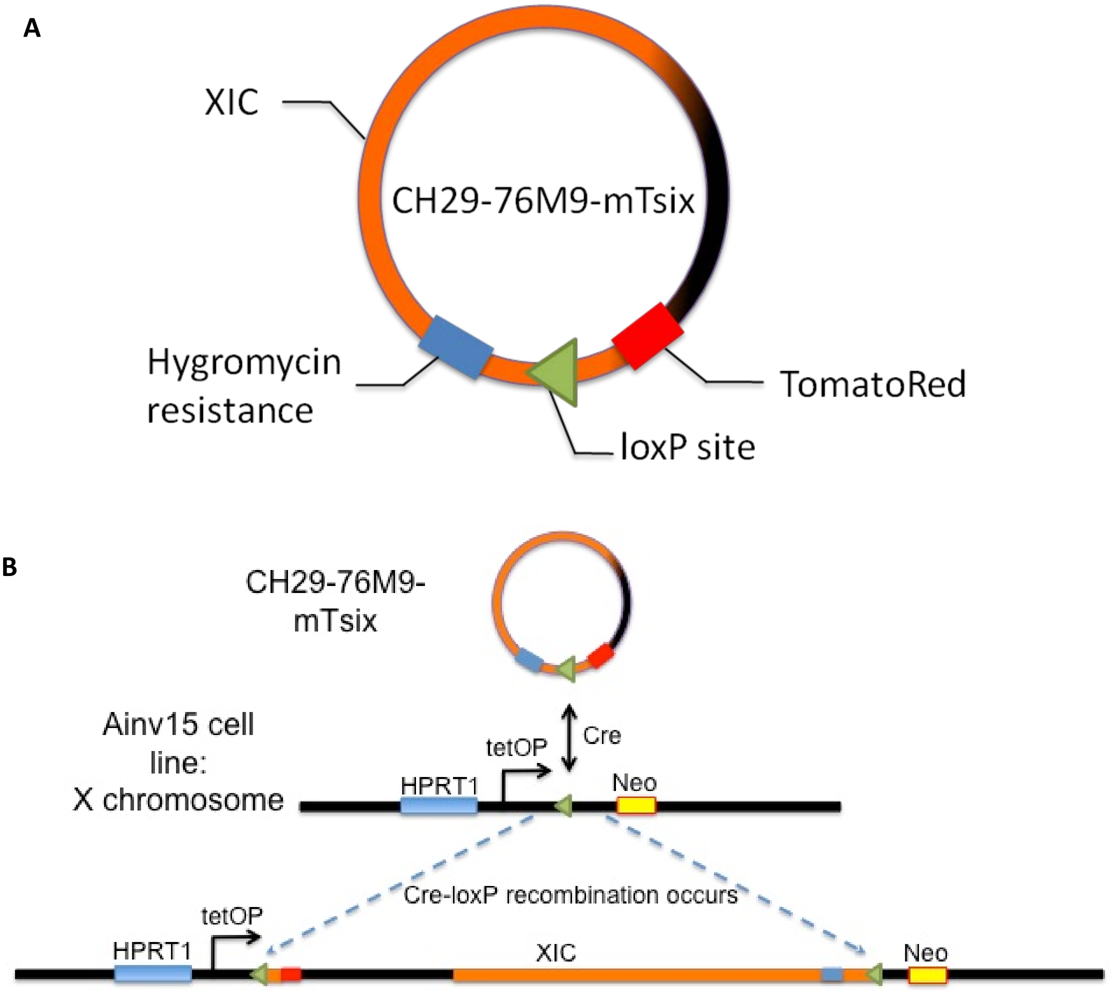


Figure 3.14 Targeted XIC insertion strategy
Diagrams are not to scale. **(A)** Diagram illustrating key features of the CH29-76M9-m*Tsix* BAC, which contains a XIC sequence that is 188,684 base pairs long. The hygromycin resistance cassette (blue), the loxP site (green), and the TomatoRed fluorescent protein gene (red) have been recombined into the promoter region of the *Tsix* gene in the XIC (orange) located within the BAC genomic DNA insert, disrupting the promoter activity and thus *Tsix* expression. The promoter region of *Xist* is present and unaffected. **(B)** Electroporation of CH29-76M9-m*Tsix* into Ainv15 (male) ES cells together with transient Cre recombinase expression should allow Cre-mediated recombination of the entire BAC into the loxP site on the X chromosome of Ainv15 cells. In this case, hygromycin resistance and RFP expression are used to select for the proper recombinants. The tet-ON promoter system (tetOP) and the incomplete neomycin resistance gene (Neo) were not utilized in this design.

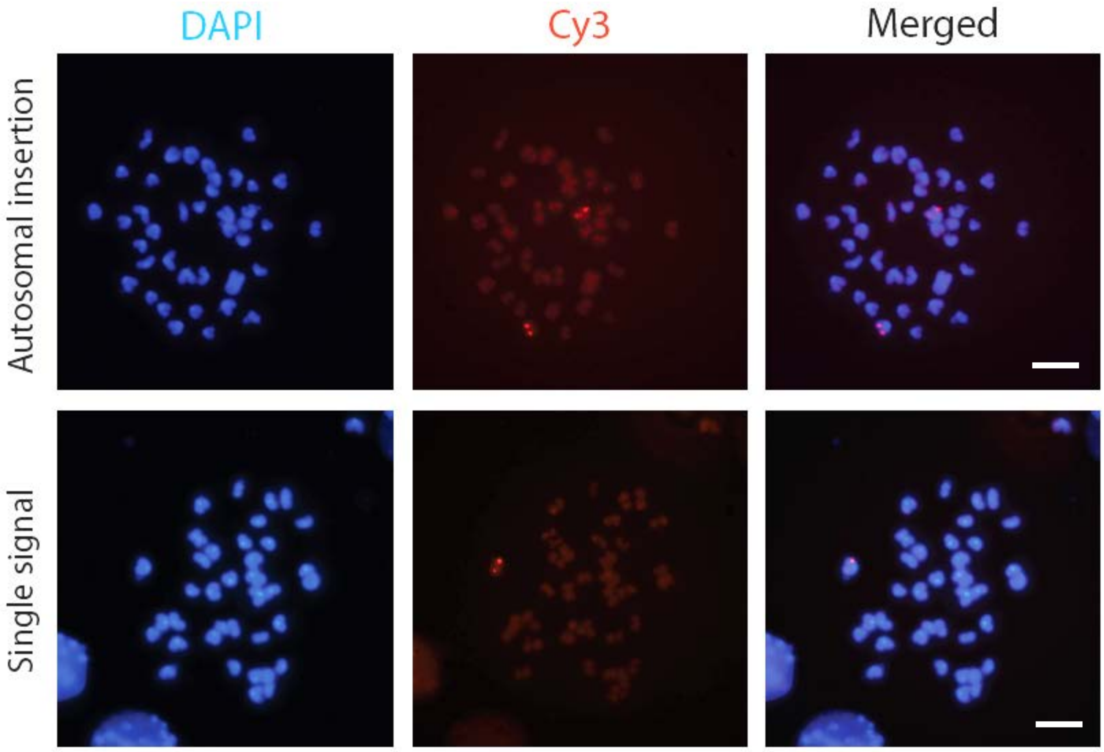


Figure 3.15 DNA-FISH genotyping results of CH29-76M9-m*Tsix*-transfected Ainv15 ES cell lines
Representative examples of autosomal insertions and single signal phenotypes are shown. Cy3 (red) images indicate regions of *Xist* probe binding, with each pinpoint pair representing a signal. Metaphase chromosome spreads cause signals to appear as pairs due to the presence of sister chromatids bearing the same genetic sequence. The white scale bars represent ~4 µm in each row of images.

3.5.1.2. Transgenic ES cell lines showed evidence of transgene inactivation

In the process of cell line selection, it was noticed that CH29-76M9-m*Tsix* colonies predominantly exhibited mosaic patterns of TomatoRed expression (Figure 3.16B), whereby a single colony displays a mix of cells with varying levels of TomatoRed expression. Additionally, numerous colonies did not express any TomatoRed (Figure 3.16A). Although a few colonies expressed TomatoRed in every cell, they were typically small in size (Figure 3.16C).

Strikingly, cells that seemed to have lost their ‘stemness’ typically expressed lower levels of RFP. An example of this can be seen towards the right edge of the colony in Figure 3.16B(iii), where several cells are irregularly shaped and can be clearly distinguished from each other.

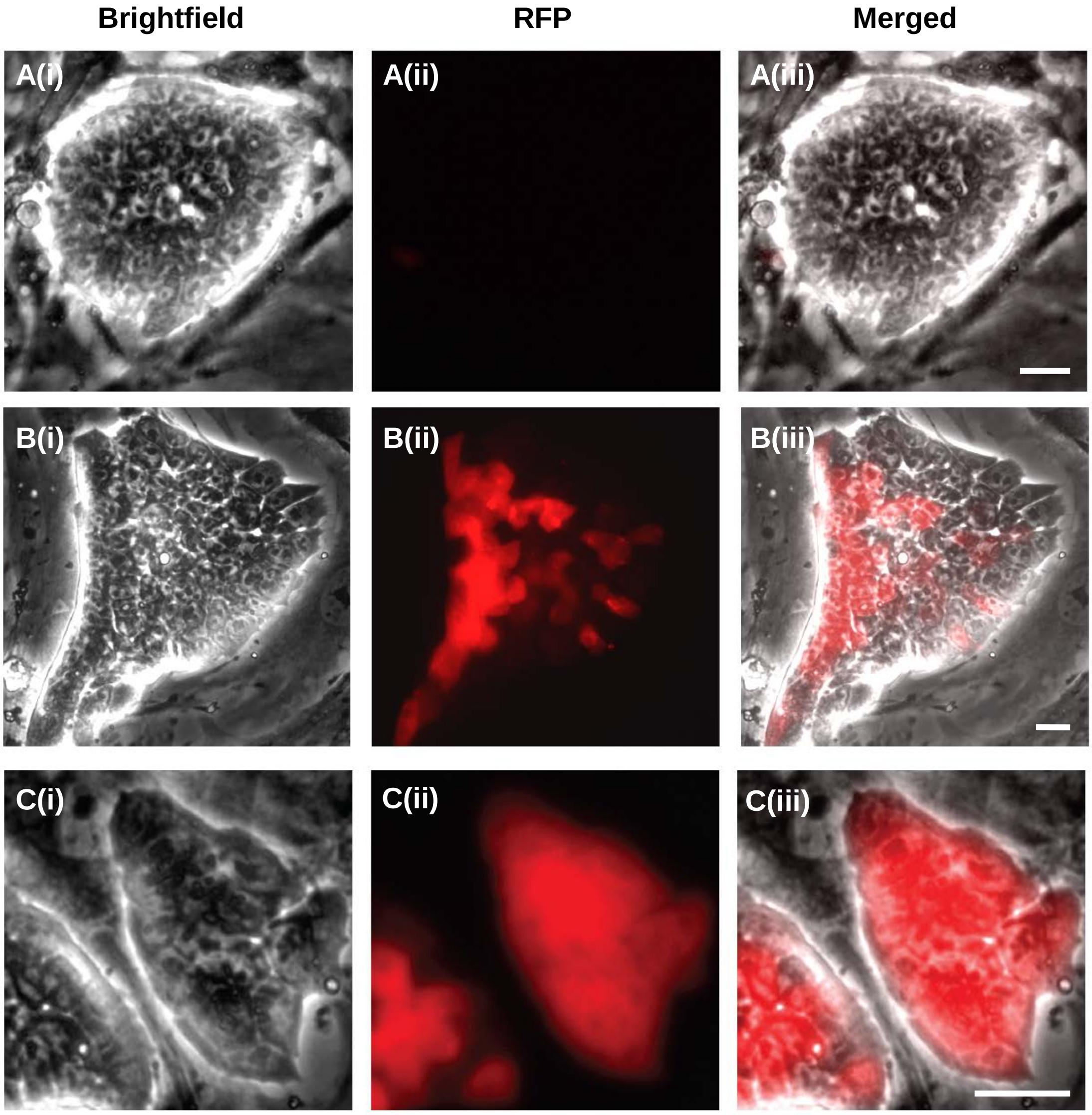


Figure 3.16 Typical CH29-76M9-m*Tsix* ES cell line (ED39) colonies
(A) An example of a colony that did not express RFP (Negative). **(B)** An example of a colony that experienced mosaic expression of RFP (Mosaic). **(C)** An example of a colony that expressed RFP in every cell (Whole). The white scale bars in the Merged column represent ~16 µm across each row of images.

In order to further investigate this phenomenon, FACS was performed on a typical CH29-76M9-m*Tsix* transgenic cell line named ED39. The sorting was based on the relative expression levels of RFP in the cells. Prior to sorting, 100 colonies were counted and their pattern of RFP expression noted (Figure 3.17E). The cells were sorted into two groups (Figure 3.17A): those that exhibited relatively high RFP expression (HIGH) and those with relatively low RFP expression (-VE).

The sorted cells were then cultured for a week. Another 100 colonies from each group were counted and their pattern of RFP expression noted (Figure 3.17E). Both groups of cells were then sorted again according to the same parameters (Figure 3.17B and C). Cells that earlier expressed relatively high RFP levels (HIGH) were found to have segregated once more into two groups: one maintaining relatively high RFP levels, the other with relatively low RFP expression (Figure 3.17B). On the other hand, the population of cells that earlier expressed relatively low RFP levels (-VE) remained homogenously consistent in their low RFP expression levels (Figure 3.17C).

These findings (Figure 3.17D) indicate that the cells experienced down-regulation of RFP expression as they continued to grow and divide. Once the RFP expression was down-regulated, it remained down-regulated. This implied that clonally maintained transgene inactivation was occurring. It was hypothesized that this could be due to low-level up-regulation of the transgenic *Xist* leading to inactivation of the nearby transgenic selection markers. Perhaps this could be exploited to establish the screen.

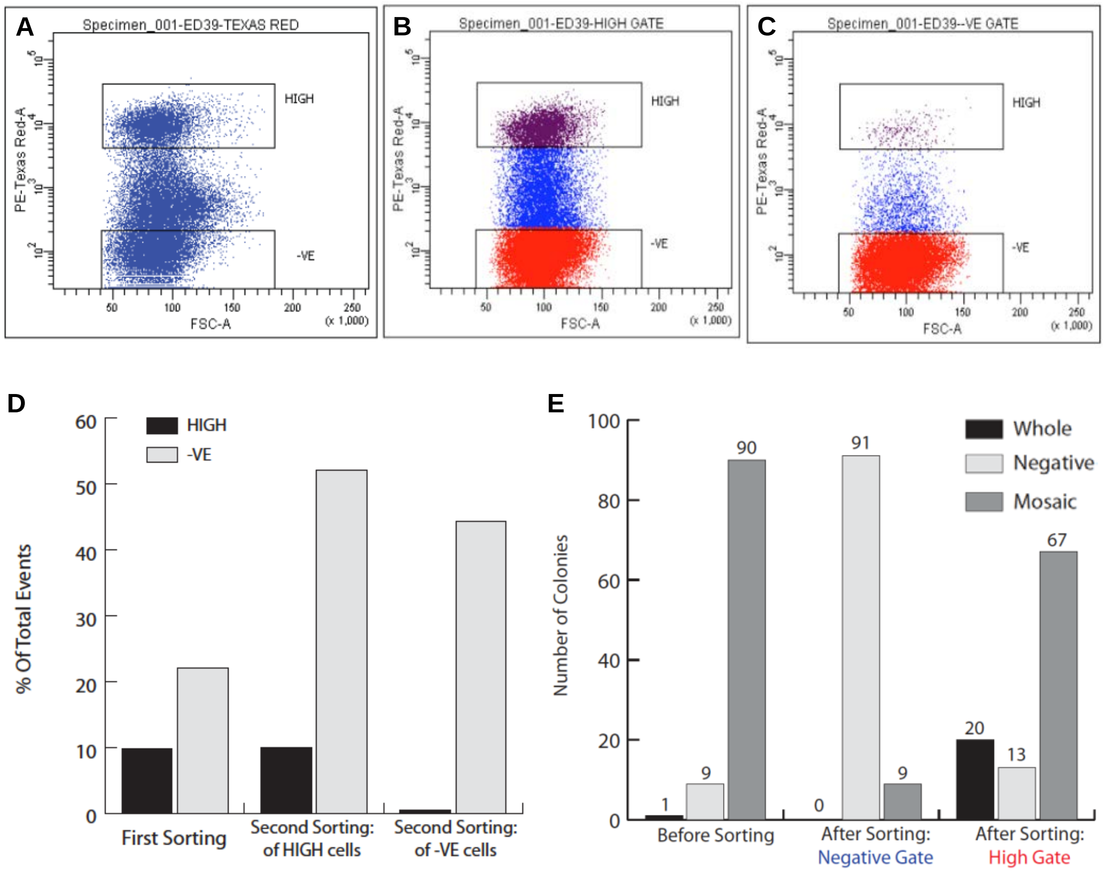


Figure 3.17 RFP expression patterns in CH-76M9-m*Tsix* transgenic ES cells

(A) Cells were sorted by FACS into two groups, HIGH (relatively high expression of RFP) and –VE (relatively low expression of RFP. **(B)** FACS profile of cells that were sorted into the HIGH group, after 7 days of culture. **(C)** FACS profile of cells that were sorted into the –VE group, after 7 days of culture. **(D)** Summary of FACS results. **(E)** RFP expression patterns in ED39 ES cell colonies. 'Whole' indicates the colony expressed RFP in every cell, 'Negative' indicates no visible expression of RFP in each cell in the colony, while 'Mosaic' indicates that individual cells of that colony expressed RFP at different levels to each other.

3.5.1.3. Differentiation of transgenic ES cell lines provided preliminary evidence of XCI

To induce XCI, several CH29-76M9-m*Tsix* transgenic ES cell lines were differentiated by withdrawal of LIF and the addition of retinoic acid (RA) in the absence of hygromycin selection. A few of these cell lines (2.11.2, 2.21.3, 3.12.11, and EE15) were found to experience a larger drop in the percentage of cells expressing RFP relative to their counterparts after differentiation. Those differentiated cells were then cultured under hygromycin selection (300 µg/ml) for 8 days. Cell lines that exhibited relatively high sensitivity to hygromycin, in other words experienced significant cell mortality, were chosen for further testing (Table 3.2).

In these experiments, the 3.12.11, 2.11.2, and EE15 ES cell lines were grown under hygromycin selection for a week, to ensure their resistance to hygromycin while undifferentiated. These resistant cells were then differentiated, while still under hygromycin selection, for another week. All the cell lines tested showed a high degree of sensitivity to hygromycin following differentiation. Roughly 70-90% reductions in cell number for each cell line was observed (data not shown) versus the controls, which were the same cell lines grown in the presence of hygromycin but differentiated in the absence of hygromycin.

Preliminary RNA-FISH results revealed *Xist* RNA cloud formation in roughly 7% of the differentiated cells (Figure 3.18). This phenomenon can be explained by the accumulation of the transgenic *Xist* RNA on its chromosome in *cis* (due to the lack of inhibition by *Tsix*). As a result, the hygromycin resistance gene, which is located near the *Xist* gene, should be silenced due to the repressive effects of *Xist* RNA, thus explaining the increased sensitivity to the antibiotic. However, the expected and/or desired 100% cell mortality rate could not be achieved after 14 days of differentiation, meaning that the hygromycin resistance gene silencing was not total (or the surviving cells were naturally resistant to hygromycin). This limited the usefulness of these cell lines, as any surviving cells following differentiation under hygromycin selection would lead to confusion during screening. This is because the screening process would rely on finding differentiated cells rescued by lentiviral transfection (of a shRNA library silencing a wide range of genes) of those transgenic cell lines. Nevertheless, this was evidence that ES cells could be sensitized to antibiotic selection once the resistance gene was inactivated via ectopic XCI.

Table 3.2 Differentiation of selected CH-29-76M9-m*Tsix* transgenic cell lines

	Transgenic cell line			
	2.11.2	**2.21.3**	**3.12.11**	**EE15**
	Genotyping result:			
	Autosomal	Autosomal	Single signal	Autosomal
	RFP expression when differentiated (% of colonies):			
Day 0	25	80	50	60
Day 6/0H	0	40	0	5
	Survival of differentiated cells in hygromycin (% of EBs):			
Day 7/1H	100	100	100	100
Day 8/2H	95	95	95	95
Day 9/3H	40	45	40	45
Day 10/4H	20	20	20	20
Day 11/5H	5	15	5	15
Day 12/6H	2	12	~0	10
Day 13/7H	~0.2	12	~0	10
Day 14/8H	~0	12	~0	8
	Ranking of relative sensitivity to hygromycin after differentiation			
	2nd	4th	1st	3rd

Day 0 indicates the day LIF was withdrawn and RA added to the growth medium. Day 6 is 24 hours after attachment of the EBs, and also the first day that hygromycin selection was begun (0H).

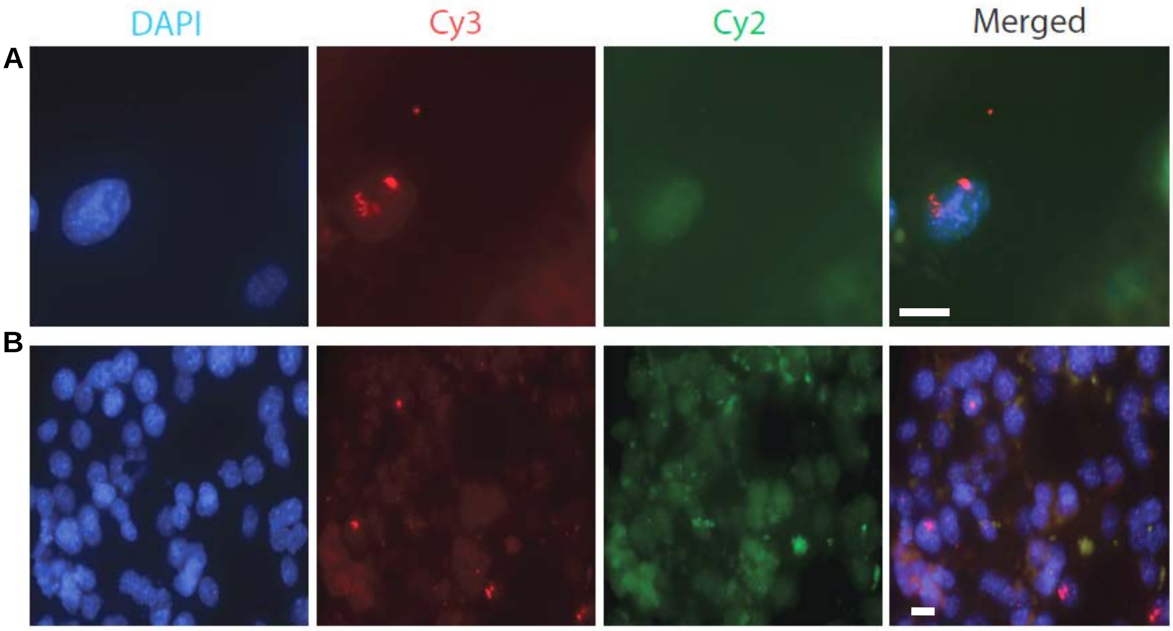


Figure 3.18 *Xist* RNA-FISH of differentiated 2.11.2 transgenic cell line This cell line is indicative of the *Xist* expression seen in the 3.12.11 and EE15 ES cell lines. The DAPI column indicates the DAPI-stained cell nuclei, while the Cy3 column reveals the localization of the Cy3-labeled *Xist* probe – dense red signals indicate *Xist* RNA presence. The Cy2 images are used to detect autofluorescence – when the images are merged, the autofluorescent regions turn yellow while *Xist* signals remain red. The white scale bars in the Merged column represent ~8 µm across each row of images. Row **(A)** indicates a single differentiated cell. Row **(B)** is a representative view of multiple cells.

3.5.2. Targeted Xist insertion

From the previous results, we thought that perhaps the large size of the BAC could be interfering with the insertion. The much smaller im*Xist* plasmid was constructed (Cheliah and Zhang, unpublished data) by means of Red/ET recombination from a construct containing the *Xist* gene. This would have accurately inserted the murine *Xist* gene into the X-chromosome of Ainv15 cells via Cre-mediated recombination. Successful recombination would add the PGK promoter as well as restore the Neo resistance gene start site, returning function to the neomycin resistance cassette. With this, the transgenic *Xist* should be under the control of the tet-On system, allowing induction of ectopic *Xist* transcription with doxycycline whenever required.

Ainv15 cells were electroporated with im*Xist* and pSALK-CRE, optimization of the PCR genotyping done and the first few cell lines processed. All cell lines genotyped did not indicate insert integration (Data not shown). After genomic DNA sequencing, it was found that the loxP site of the Ainv15 cell line was slightly mutated from the wild-type, with two extra nucleotides present (ataacttcgtatagcatacattatacgaagtt**gc**at). This was not reported in the literature (Kyba et al., 2002). This mutation could have prevented Cre-mediated recombination from occurring in the Ainv15 ES cells in all the experiments up to this point. This finding meant that the probability of getting the correct insertion was very low, and any successful recombinants would experience a shift in the reading frame of the transgenic *Xist*. Therefore, this experiment was abandoned.

In response, we designed a similar construct, ih*XIST*, which used human XIST in place of the mouse *Xist* to reduce the probability of homologous recombination, and a loxP site that matched that of the Ainv15 cells. However, Red/ET recombination to subclone the *XIST* gene out of the BAC containing *XIST* to form the ih*XIST* plasmid did not succeed. We hypothesized that perhaps the region to be subcloned – roughly 32kb in size – was too large. This led to the much shorter 4.8kb backbone being unable to bridge the homologous regions to generate ih*XIST*. An alternative explanation would be technical error, but extensive troubleshooting of such issues did not resolve the problem.

3.5.3. Summary of XCI Project 1 results

The goal of inserting the XIC or the *Xist*/XIST gene into the X chromosome of male murine ES cells was not achieved, but some useful information was gathered that will help future experimentation. Avoiding the pitfalls experienced in this project would ensure a more reliable methodology for transgenic insertion of test sequences.

The finding that ectopic XCI could occur in our transgenic ES cells, and that this phenomenon could lead to the cells becoming sensitized to antibiotic selection via inactivation of the resistance gene, hints that further refinement and development of the process might lead to the desired outcome.

3.6. XCI Project 2: Screen for activators of XCI

Rnf12 was the first *trans*-acting activator of XCI discovered (Jonkers et al., 2009). *Rnf12* codes for an E3 ubiquitin ligase that contains zinc finger motifs, and is highly conserved in both mice and humans (Bach et al., 1999). The transgenic over-expression of *Rnf12* in male mouse cells results in the ectopic transcription of endogenous *Xist* on the single X chromosome, giving rise to *Xist* RNA clouds and thus XCI. If one allele of *Rnf12* is deleted in female ES cells, XCI initiation is slowed significantly at the start of differentiation, but manages to recover later on, with cells successfully initiating XCI (Jonkers et al., 2009). This observation hints at the existence of other unknown XCI activators in the cell.

The HD2-HD3 breakpoint region on the X chromosome is critical for XCI activation, with deletion of this ~30Mb region in one of the X chromosomes in diploid female cells precluding XCI (Rastan and Robertson, 1985). The HD2-HD3 breakpoint region was later shown to contain the XIC (Lee et al., 1999b; Lee et al., 1996). As the *Rnf12* gene is located within the HD2-HD3 breakpoint region, we thought that more XCI activators might be concentrated in this region.

In this project, the search for more such XCI activators was expanded after an initial study by Khoo Bee Luan revealed the potential presence of four other genomic regions within the HD2-HD3 breakpoint region that may be important for XCI activation. Four other large DNA constructs, CH29-538N12, RP24-104K20, WI1-667J14, and CH29-484O10, were chosen for this purpose. These constructs contain genomic inserts that cover other sequences within the HD2-HD3 breakpoint region. By integrating these additional sequences into the genome of J1 male murine stem cells, and observing the resultant transgenic cell lines for ectopic XCI triggered by increased XCI activator dosage, it could be possible to identify these genomic regions as potentially being important for XCI activation.

3.6.1. *Khoo Bee Luan's screen for potential XCI activators nets 4 different sequences*

In the beginning, we postulated that the discovery of *Rnf12* as an XCI activator might indicate a role for proteins involved in the ubiquitination pathway and/or zinc finger proteins in XCI activation. Ms. Khoo selected five bacterial artificial chromosome (BAC) clones located within the HD2-HD3 breakpoint region, namely RP23-280L7, RP24-285J22, RP23-282B14, RP24-66B4, and RP24-118E11, each carrying genes that are either part of the ubiquitination pathway, or known zinc finger proteins. The RP24-240J16 BAC containing the *Rnf12* gene was also acquired in order to replicate the findings of Jonkers et al, as well as to act as a set of positive controls denoting XCI activation.

Table 3.3 shows a list of these BACs and the genes that they encompass. Figure 3.19 shows the genomic regions covered by these BACs. J1 (wild-type male murine) ES cells were electroporated with the BACs by Ms. Khoo. At least two transgenic cell lines with BAC insert integration were made for each BAC, with the exception of RP24-66B4, which did not yield any cell lines. Ectopic *Xist* RNA clouds were detected in all transgenic cell lines when they were differentiated (Figure 3.20). As male cells do not typically express *Xist*, this indicated XCI activation. Several of these cell lines even displayed a significantly higher percentage of differentiated cells with *Xist* clouds when compared to differentiated female ES cells (EL16.7). Consequently, it is possible that this screen identified four separate genomic regions that may contribute to XCI activation in *trans*. The genes in these regions can therefore be considered as potential XCI activator candidates.

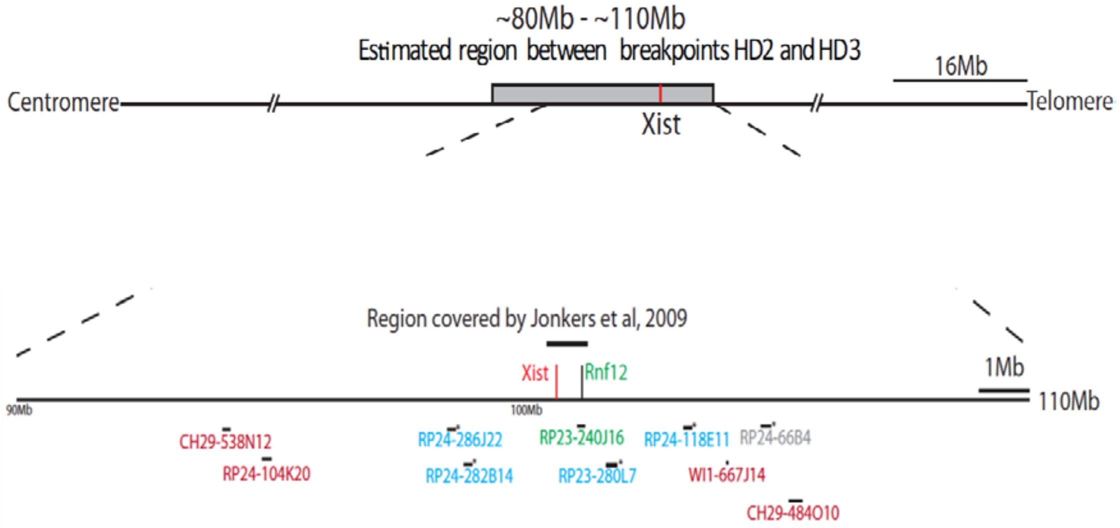
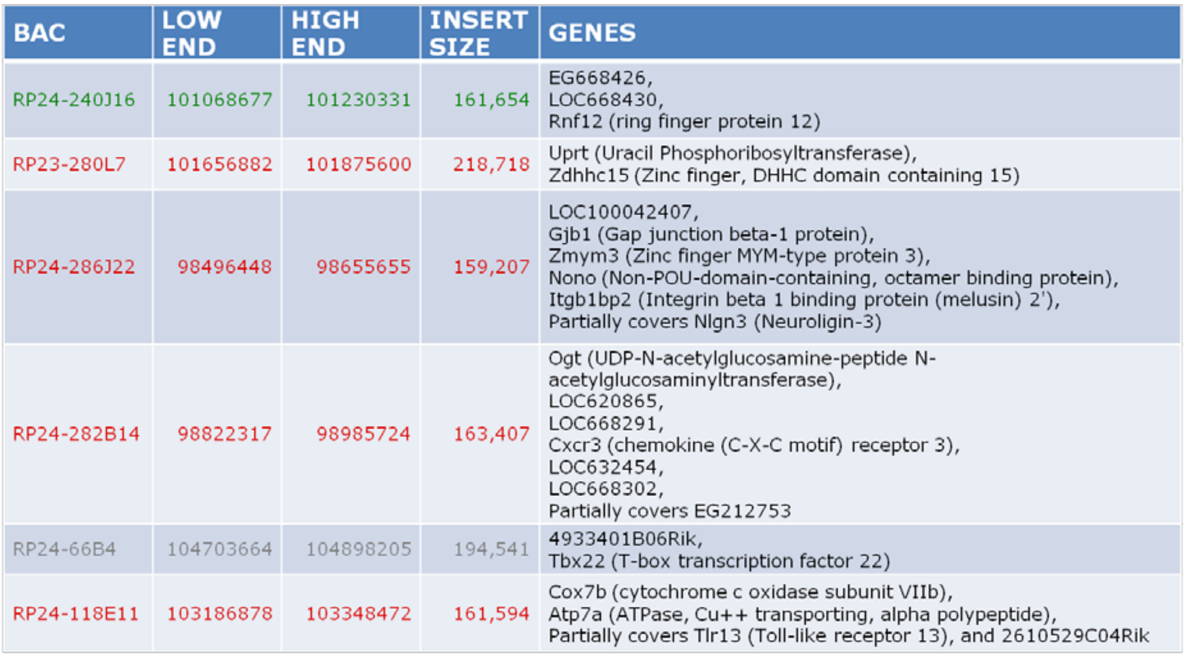


Figure 3.19 Coverage of selected BACs within the HD2-D3 breakpoint region BAC inserts named in light blue are BACs that were shown to be important for XCI. BAC inserts named in red are BACs that were chosen for the phase 2 expansion of the screen. The locations of the *Xist* gene and the *Rnf12* gene have been marked out, as well as the region covered by Jonkers et al (2009) during their search for XCI activators. The coverage of the RP23-240J16 BAC (which contains the *Rnf12* gene) is named in green. The BAC which did not give any transgenic colonies is named in grey.

Table 3.3 A list of Ms. Khoo's BACs and the genes encompassed by their genomic inserts



BAC	LOW END	HIGH END	INSERT SIZE	GENES
RP24-240J16	101068677	101230331	161,654	EG668426, LOC668430, Rnf12 (ring finger protein 12)
RP23-280L7	101656882	101875600	218,718	Uprt (Uracil Phosphoribosyltransferase), Zdhhc15 (Zinc finger, DHHC domain containing 15)
RP24-286J22	98496448	98655655	159,207	LOC100042407, Gjb1 (Gap junction beta-1 protein), Zmym3 (Zinc finger MYM-type protein 3), Nono (Non-POU-domain-containing, octamer binding protein), Itgb1bp2 (Integrin beta 1 binding protein (melusin) 2'), Partially covers Nlgn3 (Neuroligin-3)
RP24-282B14	98822317	98985724	163,407	Ogt (UDP-N-acetylglucosamine-peptide N-acetylglucosaminyltransferase), LOC620865, LOC668291, Cxcr3 (chemokine (C-X-C motif) receptor 3), LOC632454, LOC668302, Partially covers EG212753
RP24-66B4	104703664	104898205	194,541	4933401B06Rik, Tbx22 (T-box transcription factor 22)
RP24-118E11	103186878	103348472	161,594	Cox7b (cytochrome c oxidase subunit VIIb), Atp7a (ATPase, Cu++ transporting, alpha polypeptide), Partially covers Tlr13 (Toll-like receptor 13), and 2610529C04Rik

"Low end" and "high end" indicate the start and stop positions of the BAC genomic insert on the X chromosome (NC_000086.6). The BAC in green was used in identification of the first XCI activator (Jonkers et al., 2009). The BACs in red were shown to be important for XCI. The BAC in grey did not give any transgenic colonies.

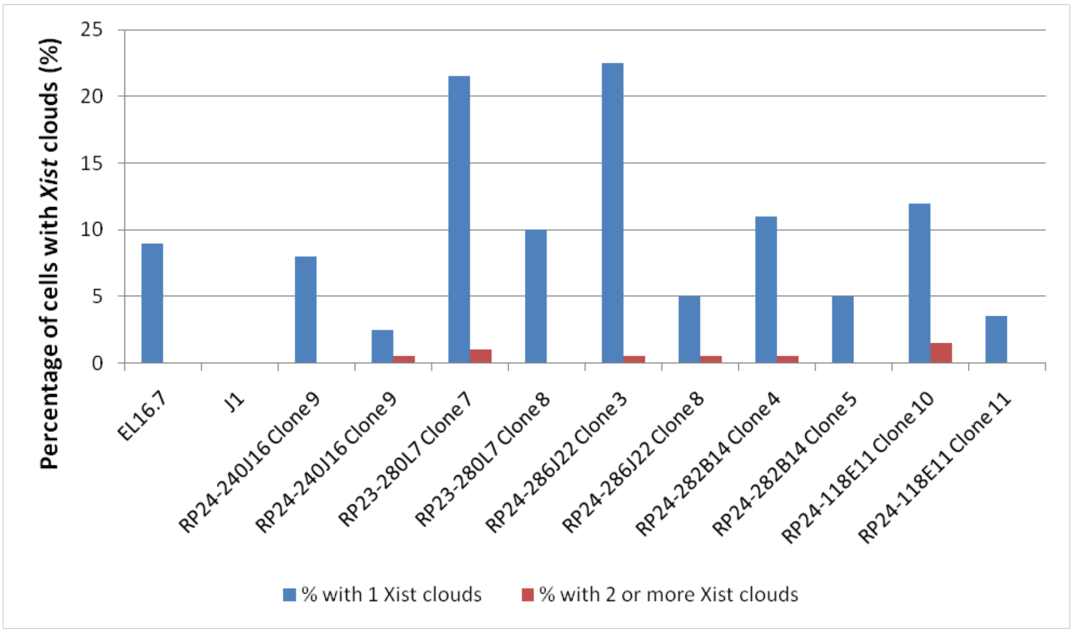


Figure 3.20 Illustration of the percentage of cells with ectopic XCI in all transgenic cell lines
Figure adapted from Ms. Khoo’s report.

3.6.2. Expansion of the screen for XCI activators

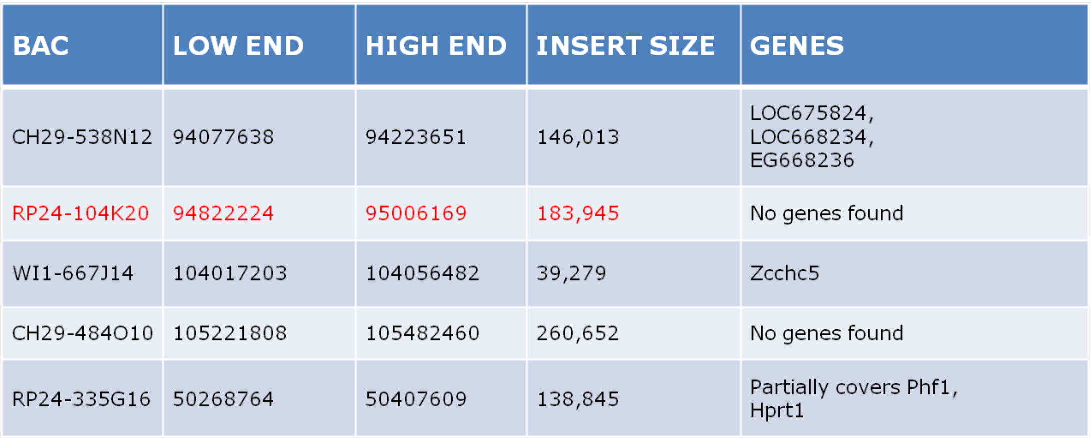
Accordingly, the screen for XCI activators was expanded into phase 2, and I continued work on this project. Four other large DNA constructs (CH29-538N12, RP24-104K20, WI1-667J14, CH29-484O10) containing inserts that cover other sequences within the HD2-HD3 breakpoint region were chosen for this purpose (see Figure 3.19 and Table 3.4). These sequences contained fewer or no known genes, and were located much further upstream or downstream of the regions already covered in the screens by Jonkers et al or Ms. Khoo. The rationale for this was to examine whether such gene-poor sequences could contribute to XCI activation, as well as to see how far afield XCI activators could be located.

At the same time, we intended to generate a transgenic cell line with autosomal integration of an insert that covered a sequence on the X chromosome. When differentiated, this cell line should be incapable of XCI. In essence, it would act as a negative control that would illustrate that our methodology does not induce ectopic XCI on its own. We chose the RP24-

335G16 BAC, which covers the *Hprt* (hypoxanthine guanine phosphoribosyl transferase) gene, and is located outside the HD2-HD3 breakpoint region, for the generation of this cell line.

Male murine J1 cells were electroporated with the RP24-335G16, CH29-484O10, CH29-538N12, and RP24-104K20 BACs, as well as the WI1-667J14 fosmid, separately, each accompanied by a neomycin-resistance cassette (1500 bp fragment of pEZ-Frt-lox-DT plasmid). After 2 weeks of selection, sufficiently mature transgenic ES cell colonies were picked, expanded in 6-well plates, frozen down and subsequently placed in liquid nitrogen.

Table 3.4 Large DNA constructs used in phase 2 of the screen for XCI activators



BAC	LOW END	HIGH END	INSERT SIZE	GENES
CH29-538N12	94077638	94223651	146,013	LOC675824, LOC668234, EG668236
RP24-104K20	94822224	95006169	183,945	No genes found
WI1-667J14	104017203	104056482	39,279	Zcchc5
CH29-484O10	105221808	105482460	260,652	No genes found
RP24-335G16	50268764	50407609	138,845	Partially covers Phf1, Hprt1

Low end and high end indicate the start and stop positions of the BAC genomic insert on the X chromosome (NC_000086.6)

3.6.3. *RP24-335G16-transfected ES cells experienced lethality or no insert integration*

As stated earlier, RP24-335G16 transgenic cells were to be used as negative controls to prove that the experimental procedure we were using did not affect XCI activation. The RP24-335G16 insert encompasses a roughly 130 kb locus on the mouse X chromosome that includes *Hprt*, a typical housekeeping gene usually used as a control marker. The RP24-335G16-transfected cells were thawed and grown in 6-well plates before being expanded into T25 flasks. Post-expansion cells suffered extensive cell death, but recovered within a few days and continued growing, permitting metaphase chromosome spreads to be made. DNA-FISH genotyping, using nick-translation-labeled fluorescent probes derived from the respective BAC, was performed on the spreads in order to detect successful insertions. Two cell lines were found to have genomic integration of the BAC insert (data not shown). Unfortunately, thawing of these clones resulted in complete lethality of both cell lines, an uncharacteristic situation.

A second electroporation of this BAC was performed using the same procedure as before. Numerous resistant ES cell colonies were formed and cell lines were generated from the picked colonies. These cell lines did not experience any lethality or growth problems pre- or post-thawing. However, during DNA-FISH genotyping, it was found that these cell lines did not have the BAC insert integrated into the genome (data not shown).

3.6.4. *WI1-667J14-transfected ES cell genotyping indicates insertion*

WI1-667J14-transfected ES cells were thawed and grown. All WI1-667J14 cell lines exhibited significant post-thawing lethality, with only a few cells from a single cell line (WI1-667J14-5) surviving. Remarkably, the few living colonies of this cell line spontaneously differentiated soon after a few days in culture, despite the presence of feeder cells plus daily feeding with growth medium containing LIF.

DNA-FISH genotyping of WI1-667J14-5 cells indicated integration of the insert, with detection of a weak yet discernible signal (Figure 3.21). RNA-FISH for ectopic *Xist* expression was not performed on these spontaneously differentiated cells, because XCI was improbable due to the lack of cell lethality in the differentiated male cells. Of course, this does not rule out the possibility that reversible XCI (Wutz and Jaenisch, 2000) occurred during the period of spontaneous differentiation.

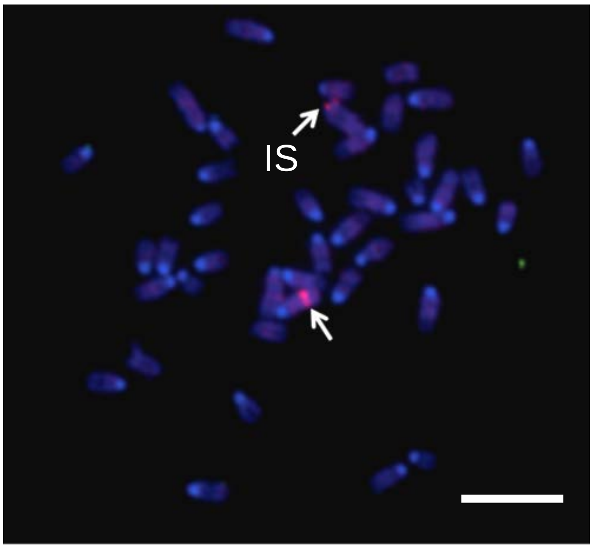


Figure 3.21 DNA-FISH genotyping of WI1-667J14-5 cell line
The red Cy3-labelled probes were made from the WI1-667J14 DNA construct**.** The transgenic integration site is indicated with “IS”. The other arrow indicates the endogenous genomic region. Chromosomes show up as blue (DAPI stain). The white scale bar represents 4 µm in the image.

3.6.5. Genotyping of ES cells transfected with CH29-484O10 shows no insertions

CH29-484O10-transfected ES cells were thawed, grown, and expanded without any unexpected cell mortality. However, DNA-FISH with CH29-484O10-derived probes on metaphase chromosome spreads of the various cell lines did not show any successful transgenic insertions (data not shown). There are no known genes that map to the region covered by the CH29-484O10 BAC. A BLAST search of this region against the mouse genome indicated a sequence with similarity (E-value: 5e-170, Max identity: 82%) to integrin alpha 4 (*Itga4*), which is located on chromosome 2.

3.6.6. DNA-FISH genotyping indicates successful CH29-538N12 and RP24-104K20 insert integration

CH29-538N12-transfected ES cells and RP24-104K20-transfected ES cells were thawed, grown and expanded successfully. DNA-FISH genotyping revealed that both groups had one cell line each with successful transgenic insertion, CH29-538N12-2 and RP24-104K20-5 (Figure 3.22). The RP24-104K20-5 cell line was found to be tetraploid (XXYY), but correspondingly had at least two transgenic insertion sites.

3.6.7. Xist RNA-FISH in differentiated CH29-538N12-2 and RP24-104K2-5 cells

When differentiated for up to 9 days, the CH29-538N12-2 transgenic ES cell line did not seem to exhibit any *Xist* RNA clouds, but instead showed clear pinpoint *Xist* signals under RNA-FISH (Figure 3.23A and B). This might indicate that the region covered by this BAC is dispensable for XCI activation. Serendipitously, these results may now be used as a negative control (replacing RP24-335G16), proving the experimental procedure used did not affect XCI activation.

In contrast, the tetraploid RP24-104K20-5 transgenic ES cell line displayed significant XCI, with about 20-25% of the cells showing *Xist* RNA cloud formation when differentiated for 6 days (Figure 3.23A and C). The percentage of cells indicating XCI at this stage far surpassed that found in typical *in vitro* differentiated female ES cells at this time point (Figure 3.20, see EL16.7 cell line). Also, remarkably, about 2-4% of the total cells were found to have inactivated two X chromosomes. Since tetraploid XXYY cells do not experience widespread XCI when differentiated (Monkhorst et al., 2008), and the two X inactivation pattern (during XCI) has been observed before in tetraploid XXXX cells (Monkhorst et al., 2008), this hinted at the possibility that the RP24-104K20 insert covers a genomic region involved in XCI activation.

The finding that the RP24-104K20 insert is important for XCI activation poses an interesting conundrum, as no known genes map to this genomic insert.

BLAST of the insert sequence against the mouse (genomic + transcript) nucleotide database indicates that part of it bears similarity to the murine putative Pol polyprotein-like (LOC100505017) predicted mRNA sequence. This indicates it resembles a retroviral Pol polyprotein open reading frame, which typically codes for a polyprotein that is subsequently cleaved to form the reverse transcriptase and integrase proteins of the retrovirus (Coffin, 1992).

In summary, the CH29-538N12 insert might be dispensable for XCI, whereas the RP24-104K20 insert may play a role in XCI activation in a yet indeterminate fashion.

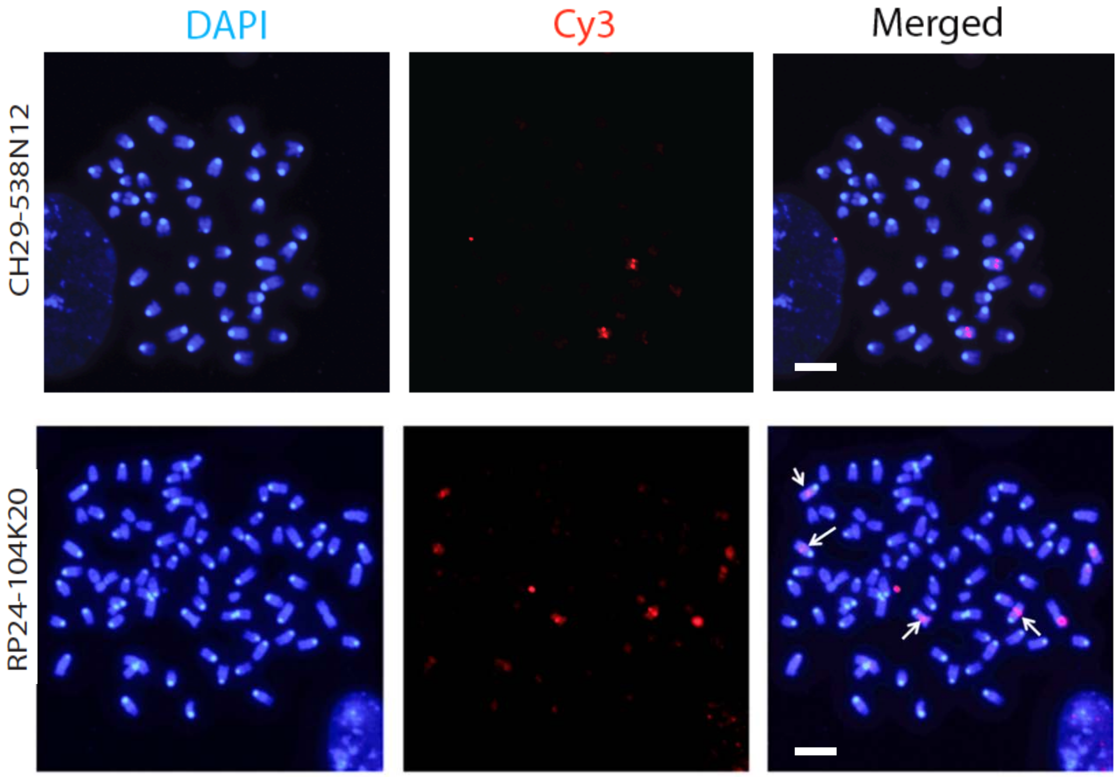


Figure 3.22 DNA-FISH genotyping of CH29-538N12-2 and RP24-104K20-5 cell lines
Both cell lines showed autosomal integration of their respective BAC inserts. CH29-538N12-2 had the typical 40 chromosomes of a murine somatic cell. RP24-104K20-5 had 80 chromosomes, making it a tetraploid cell line. It had at least two transgenic inserts as well as two endogenous regions (indicated by the arrows). White scale bars in the Merged column represent ~4 µm in each row of images.

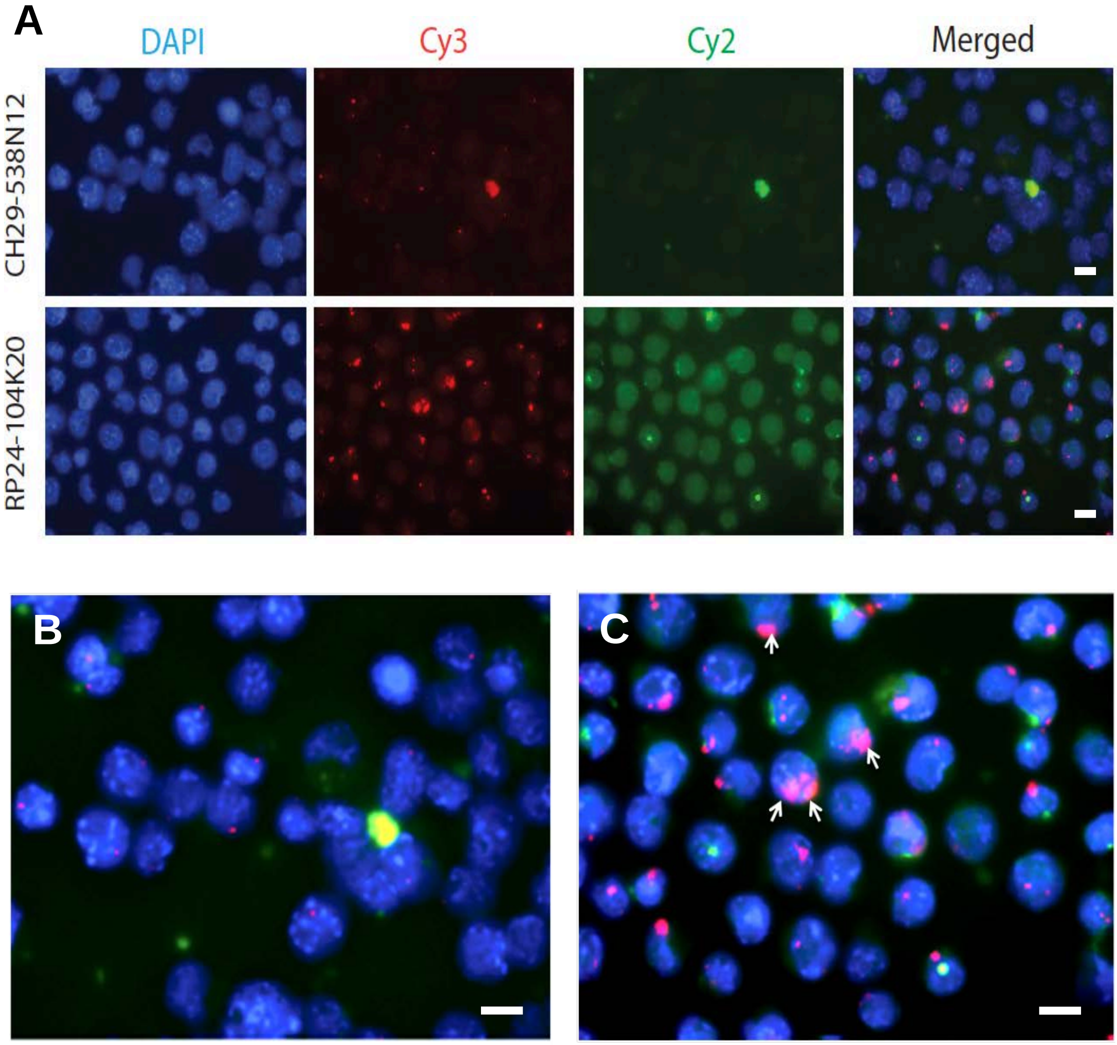


Figure 3.23 RNA-FISH for *Xist* expression in differentiated CH29-538N12-2 and RP24-104K20-5 cells

(A) The DAPI column indicates the blue DAPI-stained cell nuclei, while the Cy3 column reveals the localization of the Cy3-labeled *Xist* probe (red). The Cy2 images were used to detect autofluorescence – when the images were merged, the autofluorescent regions turned yellow while *Xist* signals remained red. **(B)** Enlarged merged image of CH29-538N12. Note punctate *Xist* RNA signals. **(C)** Enlarged merged image of RP24-104K20. White arrows indicate examples of *Xist* RNA clouds. White scale bars indicate ~8µm in each image.

4. DISCUSSION

4.1. New 5’ss/U1 recognition registers authenticated

The results from these experiments further enhance our understanding of non-canonical registers for 5’ss/U1 recognition. Combining this with prior data allows a deeper insight into the flexibility and limitations of the interaction between the 5’ss and the U1 snRNA.

4.1.1. U1 decoy analysis confirms that test 5’ss are recognized by U1 snRNA

U1-specific RNA decoys (Roca and Krainer, 2009; Roca et al., 2012) can be used to confirm that the tested 5′ss are indeed recognized by U1 and not by other U1-like snRNAs (Kyriakopoulou et al. 2006), as well as to confirm the role of U1 snRNA in test 5’ss recognition. U1 decoys are short RNAs that possess a sequence with perfect complementarity to the 5' end of U1. When the decoys bind to the snRNA, it causes a reduction in the level of free snRNPs in the cell, which in turn negatively impacts splicing (Roca et al., 2012). The expression plasmids for U1 decoy RNAs were co-transfected with the test minigenes (like *ABCC12*, *FBXL13*, or *DNAI1* for example), successfully reducing recognition of most of the test 5’ss (Luo et al, data not shown). This proved that U1 snRNA is indeed involved in the recognition of these 5’ss. Eventually, all the test minigenes will be challenged by U1 decoys in order to confirm their use of U1 recognition.

4.1.2. Asymmetric loops in 5’ss/U1 duplexes

The experimental results showed for the first time that U1 recognition of certain 5’ss may tolerate and/or require the formation of asymmetric loops, in particular asymmetric loops with one extra unpaired and unmatched nucleotide on either strand of the RNA double helix (asymmetric loop 1). Three different 5’ss recognition registers were experimentally demonstrated: asymmetric loop 1 (+3,+4), asymmetric loop 1 (+4,+5), and asymmetric loop 1 Ψ.

However, due to the nature of the predictions, asymmetric loop 1 (+3,+4) register 5’ss are strongly affected by the +5G hypothesis (discussed further below), retaining only 51 out of the predicted original 348 candidates (14.7%). Asymmetric loop 1 (+4,+5) register 5’ss fare slightly better, retaining 594 out of 653 5’ss (91.0%). Asymmetric loop 1 Ψ register 5’ss are unaffected, retaining all 115 candidate 5’ss.

The majority of the retained asymmetric loop 1 5’ss are predicted to be significantly more stable using the asymmetric loop register versus the canonical register. The retained asymmetric loop 1 (+3,+4) register 5’ss possess an average ΔΔG of -2.21 kcal/mol, with 35 5’ss (68.6%) with a predicted ΔΔG ≤ −1 kcal/mol. The retained asymmetric loop 1 (+4,+5) register 5’ss possess an average ΔΔG of -1.52 kcal/mol, with 424 5’ss (71.4%) given a ΔΔG ≤ −1 kcal/mol. Asymmetric loop 1 Ψ register 5’ss possess an average ΔΔG of -2.27 kcal/mol, with 111 5’ss (%) given a ΔΔG ≤ −1 kcal/mol. This leads us to conclude that these 5’ss should preferentially employ the asymmetric loop register over that of the canonical during 5’ss/U1 helix formation.

Other types of asymmetric loop 5’ss are predicted to exist, as shown above (Table 3.1), and should be investigated. Asymmetric loop 1 (+3/+4/+5) is a promising candidate register with 52 candidate 5’ss, with the overwhelming majority (98.1%) being much more thermodynamically stable when using the asymmetric loop register versus the canonical, and not containing any +5G-type 5’ss. This register requires the positions +3, +4, and +5 to be unpaired on the 5’ss, and both Ψ to be unpaired in the U1 5’ tail, in order to form a kinked double helix with a maximum of 9 Watson-Crick base pairs. Mutational analysis of this register would grant a deeper understanding of 5’ss/U1 interaction tolerances and show whether asymmetric loops comprising three nucleotides on one strand and two on the other can be tolerated in 5’ss recognition.

Also, longer asymmetric loop recognition registers with two or three extra unpaired and unmatched nucleotides on either side of the 5’ss/U1 helix might be possible. Of particular interest would be the asymmetric loop 2 (+3/+4/+5), asymmetric loop 2 (+4/+5/+6), asymmetric loop 3 (+3/+4/+5/+6), and asymmetric loop 3 (+4/+5/+6/+7) registers. The bulk of candidate 5’ss for these registers are predicted to have ΔΔG ≤ −1 kcal/mol, and the curated list of 5’ss account for 708, or 0.35% of all 5’ss.

4.1.3. Bulge registers with two or more bulged nucleotides confer a relatively weak energetic advantage

Experimental evidence indicates that bulges of 2 nucleotides in the 5’ss/U1 double helix can be tolerated despite the increased negative impact on thermodynamic stability (versus that of bulge 1), allowing productive splicing. *SMN1/2* minigene analysis reveals that the bulge 2 (+3,+4) register is being used over the canonical register, while *HPS4* native 5’ss testing hints at usage of the bulge (+4,+5) register. More native 5’ss predicted to use these registers need to be identified and analyzed in order to validate the usage of these registers, and the preliminary *SMN1*/2 minigene testing of bulge 2 (+4,+5) register 5’ss usage in a heterologous context needs to be completed.

Bulge 2 (+3,+4) may be more thermodynamically unstable relative to bulge 2 (+4,+5). This preliminary conjecture is based on the *SMN1/2* minigene splicing data regarding the two bulge registers, where exon 7 inclusion is significantly lower with the ideal bulge 2 (+3,+4) 5’ss, versus the preliminary data acquired with the optimal bulge (+4,+5) 5’ss. This should warrant further investigation. On the other hand, the bulge 2 (+5,+6) register might be unfeasible, as there are but a few candidate 5’ss, and their average ΔΔG is close to zero, meaning the thermodynamic stability conferred by the additional base pairs after the bulge is low. Taking these factors into consideration, the data in this thesis shows that bulge 2 registers can occur but with a weak energetic advantage.

Based on the *SMN1/2* minigene data, bulges longer than two nucleotides might not be feasible for 5’ss recognition. This is based on the fact that the test bulge 3 (+3,+4,+5) 5’ss failed to elicit exon 7 inclusion, even with use of the most optimal 5’ss sequence for this register, and after mutational inactivation of an intronic splicing silencer element. Based on their sequence, native 5’ss would presumably be even less efficient at base pairing to U1 in the bulge 3 register. However, it must be noted that several additional *cis*-acting elements other than the intronic ISS previously described have been reported to weaken the 5’ss of exon 7. These include an inhibitory terminal stem-loop structure, intronic splicing silencer N1, and a GC-rich sequence (Singh and Singh, 2011). Since the candidate bulge 3 register was not tested in its natural context, it might still be possible that productive splicing could be effected. Additionally, native *cis*-acting enhancer elements could assist in bulge 3 5’ss recognition.

Also, it is observed that bulges positioned closer to the exon-intron junction, like bulge 2 (+3,+4) for example, are significantly less stable than those located further downstream, like bulge (+4,+5). This phenomenon can also be detected in asymmetric loops, whereby asymmetric loop 1 (+3/+4) can be more easily disrupted by the same mutation versus asymmetric loop 1 (+4/+5) (Figure 3.3, lane 2 versus Figure 3.5, lane 2). Following this line of thought, perhaps bulge 3 (+4,+5,+6) might be viable for 5’ss recognition, which would cover another 352 5’ss. This necessitates further investigation. In addition, many of the longer bulges/asymmetric loops include a significant percentage of candidate 5’ss that fall under the +5G-type 5’ss category. They would therefore use the canonical register instead of the predicted non-canonical one. This is discussed further below (see section 4.1.5).

4.1.4. Rare registers and improbable registers

Some predicted registers are not likely used, due to various reasons: they either failed the experimental tests, they conferred an insignificant ΔΔG, and/or they contain too few candidate 5’ss. Examples include bulge 1 (-1), bulge 2 (-1,-2), asymmetric loop 2 (+2/+3/+4), bulge 2 (+2,+3), bulge 2 Ψ, asymmetric loop 3 (+2/+3/+4/+5), and asymmetric loop 3 (+2/+3/+4/+5/+6).

As mentioned earlier, the bulge 1 (-1) register did not pass the mutational analysis test in both *SMN1/2* (Roca et al., 2012) and UMV test minigenes (Luo and Roca, unpublished data). Additionally, all 5’ss predicted to use the bulge 2 (-1,-2) register, which is closely related to bulge 1 (-1), have very low projected thermodynamic stability advantage over the canonical register, with an average ΔΔG of 0.1 kcal/mol. This could be because the U1C protein, which is one of the protein members of the U1 snRNP (and is also capable of independently binding to the 5’ss), binds the 5’ss/U1 snRNA duplex very close to this position (Du and Rosbash, 2002; Pomeranz Krummel et al., 2009), perhaps sterically impeding the formation of bulges there.

Moreover, the asymmetric loop 2 (+2/+3/+4), bulge 2 (+2,+3), and bulge 2 Ψ registers all suffer from a dearth of 5’ss without +5G, and/or an insignificant average ΔΔG. This leads us to postulate that their usage is unlikely.

4.1.5. Implications of the +5G hypothesis

As mentioned earlier, out of the 10,248 predicted 5’ss, 4766 (46.5%) were found to have G at position +5 of the 5’ss (i.e. +5G-type 5’ss), and are thus projected to use the canonical register instead of the alternative register. In that scenario, only the other 5482 5’ss (2.72% of all 5’ss) can be considered for non-canonical register usage, a significant reduction in potential bulge/asymmetric loop candidate 5’ss.

A significant number of some of the predicted non-canonical register 5’ss may be misattributed due to this phenomenon, as the UNAfold prediction software (Markham and Zuker, 2008) did not account for ‘lone’ +5G-C4 base pairs when calculating free energy, nor did it account for U-Ψ interactions that can stabilize neighboring base pairs (see below for discussion of Ψ-U). This means that +5G-type 5’ss with +5G flanked by U/C nucleotides at position +4 of the 5’ss (+4Y), and A/C/G nucleotides at position +6 of the 5’ss (+6V), are usually classified under the non-canonical registers. According to the list of predicted bulge/asymmetric loop 5’ss, +4Y,+5G,+6V 5’ss account for 1,577 (15.4%) of the candidate 5’ss.

Out of the list of 201,541 5’ss, a total of 146,915 5’ss (72.9%) contain +5G. Of these, there are 7,807 +4U,+5G,+6V 5’ss, and 2,879 +4C,+5G,+6V 5’ss, which indicates extensive tolerance of such a motif in 5’ss recognition. Incidentally, bioinformatics analysis has shown an association between positions -1 and +5 of the 5’ss, whereby a consensus -1G allows any nucleotide at position +5, while +5G allows any nucleotide at -1. This suggests that at least one of the two strong G-C base pairs at U1 positions C4 or C9 is necessary for proper recognition (Burge and Karlin, 1997; Carmel et al., 2004; Roca et al., 2008).

4.1.6. Role of +4U in +5G-type 5’ss, and in non-canonical register 5’ss

From the experimental results, it was found that the +4U 5’ss position contributes to enhancing +5G-type 5’ss recognition. It does so possibly by forming a U-Ψ interaction which strengthens the neighboring +5G-C4 base pair in the canonical register. A prior study proposed that U-Ψ interaction may play a role in stabilizing adjacent 5’ss/U1 base pairs in yeast (Libri et al., 2002). In yeast, the 5’ss/U1 interaction is typified by the presence of a mismatch that occurs in almost all yeast introns, involving nucleotides +4U in the 5’ss and Ψ5 in U1 snRNA. The authors of this study found that the presence of this mismatch is a determinant of stability that mainly affects the off rate of the interaction. The resemblance to the +5G-type 5’ss we analyzed is certainly remarkable. Therefore, based on our results, we suggest that a similar mechanism operates in humans (and possibly other complex eukaryotes). However, in our specific case, +4U-Ψ5 is stabilizing an otherwise less stable ‘lone’ base pair at +5G, which is a novel situation.

During mutational analysis, the U at position +4 in the test 5’ss was replaced with C, which led to weakening of 5’ss recognition, as seen by the loss of exon inclusion. It is apparent that U-Ψ interaction contributes more to 5’ss/U1 recognition than any C-Ψ interaction, if that even exists. There is evidence that a U-U mismatch is more stable than a U-C mismatch in the helical portion of an internal loop (Zhong et al., unpublished data). As U and Ψ are rather similar, it is postulated that a U-Ψ mismatch will be more stable than a C-Ψ one. Furthermore, in a recent study, examination of duplexes with internal A-Ψ, G-Ψ and U-Ψ pairs revealed hydrogen bonding occurring between the bases, with A-Ψ being the strongest, then G-Ψ, followed by U-Ψ (Kierzek et al., 2014). In contrast, the same study showed that C-Ψ duplexes were the least stable, and were unable to detect any hydrogen bonding in a C-Ψ interaction. Our mutational analysis and suppressor U1 results are consistent with these observations.

Furthermore, it has not escaped our attention that many of the bulge and asymmetric loop register 5’ss can potentially utilize U-Ψ base pairing during 5’ss/U1 recognition. Regardless, this should not invalidate the fact that non-canonical registers are being used, but merely clarify the actual register being used. For example, if a 5’ss has the sequence CAG/GUUUAGUAU, instead of being an asymmetric loop 1 (+3/+4) as predicted, it may instead turn out to be a bulge register 1 (+3/+4), as the allegedly un-paired Ψ on U1 would base pair with either +3U or +4U in reality. If this is truly the case, the number of true asymmetric loop 1 (+3/+4) 5’ss would dwindle to 8, as only these have C at both positions +3 and +4, leaving no opportunity for A/G/U-Ψ base pairing. The rest have at least one U at positions +3 or +4, converting them into bulge 1 (+3) or bulge 1 (+4) register 5’ss. Similarly, there would be only 33 asymmetric loop 1 (+4/+5) 5’ss remaining, with C at positions +4 and +5. The rest have at least one U at positions +4 or +5, converting them into bulge 1 (+4) or bulge 1 (+5) register 5’ss. In any case, as the energetics of a non-canonical U-Ψ base pair are quite different from those of the canonical A-Ψ and G-Ψ base pairs, we advocate keeping these registers separate.

Additionally, considering the contribution that +4U-Ψ5 interactions provide, it stands to reason that +3U-Ψ6 interactions would also be capable of stabilizing the 5’ss/U1 helix as well as enhancing the strength of neighboring base pairs. More investigation of this premise may be necessary.

Elucidating the presence of U-Ψ base pairing in 5’ss/U1 duplexes and how it affects the thermodynamic stability and splicing outcome will further refine our predictive understanding of U1 recognition of 5’ss and 5’ss strength.

4.1.7. Importance of Ψ in U1 for recognition of 5’ss

From the predictions (Roca et al., 2012), it is readily noticeable that all the bulged/looped positions are at the 5’ss (+2 to +5) or at the 5’ end of U1 (primarily Ψ5 or Ψ6), and are limited to the middle of the helix. This is because the bulged/looped nucleotide(s) need to be flanked by adequate base pairs to fulfill energetic requirements for helix stability. Furthermore, most of the 5’ss positions that are bulged/looped out are located opposite or close to the two Ψ in the U1 snRNA 5' tail. Tm data also showed that Ψ can strengthen 5'ss/U1 helices, but no additional role in bulge/loop helices was identified (Roca et al., 2012).

It has been found that Ψ in RNA helices can coordinate a water molecule to the phosphodiester backbone of RNA, as well as stabilize base-stacking in general (Arnez and Steitz, 1994; Davis, 1995) and specifically in the context of consensus 5'ss/U1 helices (Hall and McLaughlin, 1991). Additionally, a newer study (Hudson et al., 2013) found that Ψ-A base pairs help stabilize duplexes more effectively versus U-A base pairs. Since Ψ-A base pairs retain the same Watson-Crick hydrogen bonding capacity as the original U-A pair in A-form RNA, they ascribed the stability difference to the hydrogen bonding capabilities of the repositioned imino group, as well as the novel stacking interactions provided by the altered electronic configuration of the Ψ residue. Also, the Ψ in the U2 snRNA/branch point sequence helix stabilizes base-stacking around the bulged adenosine in the pre-mRNA in addition to placing the bulge in an extrahelical conformation (Lin and Kielkopf, 2008). Thus, it is highly likely that these modified nucleotides contribute to the stability of the 5'ss/U1 bulge structure.

Our findings indicate the presence of Ψ in U1 may be an important contributor to 5'ss/U1 duplex stability. Our discovery that the U1 Ψs can also form an interaction with U in the 5'ss (in addition to A and G) means that Ψ can enhance the strength and flexibility of U1 recognition of the 5'ss. In addition, there are only two Ψ in U1 snRNA, and they are located in the 5' tail, at positions 5 and 6 (Pomeranz Krummel et al., 2009). These Ψ are conserved in both yeast and vertebrate U1 snRNA (Reddy and Busch, 1988), which may be an indicator of their significance in U1 snRNA function. The U1 Ψ seem to be adaptations in the 5' end of U1 for 5'ss recognition, and therefore maybe they provide a certain advantage in proper 5'ss selection (Roca et al., 2005). Therefore, it seems likely that preventing the isomerization of U-to-Ψ in U1 might affect 5'ss recognition and thus pre-mRNA splicing. We forecast that only weaker 5'ss might require the presence of Ψ in the U1 5' tail for proper recognition, like the 5'ss with +4U/C,+5G,+6V that was discussed earlier.

SnRNA pseudouridylation is catalyzed by the box H/ACA small nucleolar ribonucleoproteins, abbreviated as snoRNPs (Hamma and Ferre-D'Amare, 2010). These are RNA-protein complexes, each consisting of one small non-coding snoRNA – the box H/ACA RNA – which acts as a substrate-specific guide, and four common core proteins, namely the DKC1 pseudouridine synthase plus the GAR1, NHP2, and NOP10 RNP proteins (Watkins and Bohnsack, 2012). Mutations that perturb DKC1 function lead to X-linked dyskeratosis congenita (DC), a genetic condition that leads to premature aging and increased cancer susceptibility (Ruggero et al., 2003). The ACA47 and U109 snoRNAs are responsible for guiding the pseudouridylation of the fifth and sixth U1 snRNA residues respectively (Gu et al., 2005; Kiss et al., 2004). It would be interesting to see whether disrupting these snoRNAs would cause defects in U1 snRNA recognition of the 5’ss, in particular the weaker 5’ss mentioned above. Perhaps these defects contribute to DC pathogenesis.

4.1.8. Future oligonucleotide melting experiments

At this stage, we are planning oligonucleotide duplex melting experiments to further test the formation of +5G-type 5’ss duplexes with the 5’ end of U1, as well as the contribution of U-Ψ interactions. Test oligonucleotides will consist of +5G-type 5’ss sequences, with varying nucleotides at positions that correspond to position +4, +5, and +6 in the 5’ss. The complementary oligonucleotides will simply consist of the 11 nt U1 5’-end sequence, with Ψ or U at positions 5 and 6. By determining the Tm of these oligonucleotides, it will be possible to determine the contribution (or lack of) of each nucleotide to 5’ss/U1 base pairing by calculating the minimum free energy. Interpretation of the data will then allow us to see whether the +5G hypothesis phenomenon depends on *trans*-acting factors to support recognition, or if it is an intrinsic property of the 5’ss/U1 RNA-RNA interaction. As this is a protein-free RNA/RNA binding assay, it would also determine whether the effects we see are due to protein factors. Additionally, the energetic contribution of the Ψ to 5’ss/U1 binding could be obtained in these experiments.

4.1.9. U6 recognition of non-canonical 5’ss

During spliceosome assembly, U1 is displaced from the 5’ss, allowing base pairing of U6 snRNA to the intronic region of the 5’ss (Kandels-Lewis and Seraphin, 1993; Wassarman and Steitz, 1992). This event is critical for proper spliceosome assembly and splicing catalysis (Lesser and Guthrie, 1993; Staley and Guthrie, 1998).

U6 and the human consensus 5′ss can only form five Watson-Crick base pairs, suggesting modest energetic requirements for the 5′ss/U6 helix (Staley and Guthrie, 1998). Most of the tested 5’ss can increase the number of potential base pairs to the phylogenetically invariant U6 ACAGAG box (Kandels-Lewis and Seraphin, 1993; Lesser and Guthrie, 1993) by shifting the binding register (Figure 4.1) or binding via bulge registers (Figure 4.2), thereby increasing the thermodynamic stability of the interaction.

However, in previous work, U6 was found to use the canonical register while base pairing with both the consensus 5’ss and the shifted (+1) 5’ss (Roca and Krainer, 2009). Such an observation is consistent with the proposed critical role of the 5’ss/U6 RNA double helix in catalysis. Since the 5’ss/U6 helix is positioned at the spliceosomal catalytic core (Rhode et al., 2006), altering the position of this helix could impair splicing catalysis being carried out. Therefore, while U1 binding is flexible enough to recognize certain 5’ss in a non-canonical register, U6 may be required to base pair in the conventional register in order to allow the first *trans*-esterification step of splicing to occur at the correct site/position.

To test whether this altered base pairing to U6 can occur in the 5’ss that use bulge/asymmetric loop registers for U1 registers, suppressor U6 snRNAs can be used in combination with suppressor U1 snRNAs (Brackenridge et al., 2003; Hwang and Cohen, 1996; Konarska et al., 2006; Kubota et al., 2011; Lesser and Guthrie, 1993; Roca and Krainer, 2009) to rescue candidate 5’ss mutations. Such experiments will shed light on the mechanisms of splicing catalysis acting on these 5’ss.

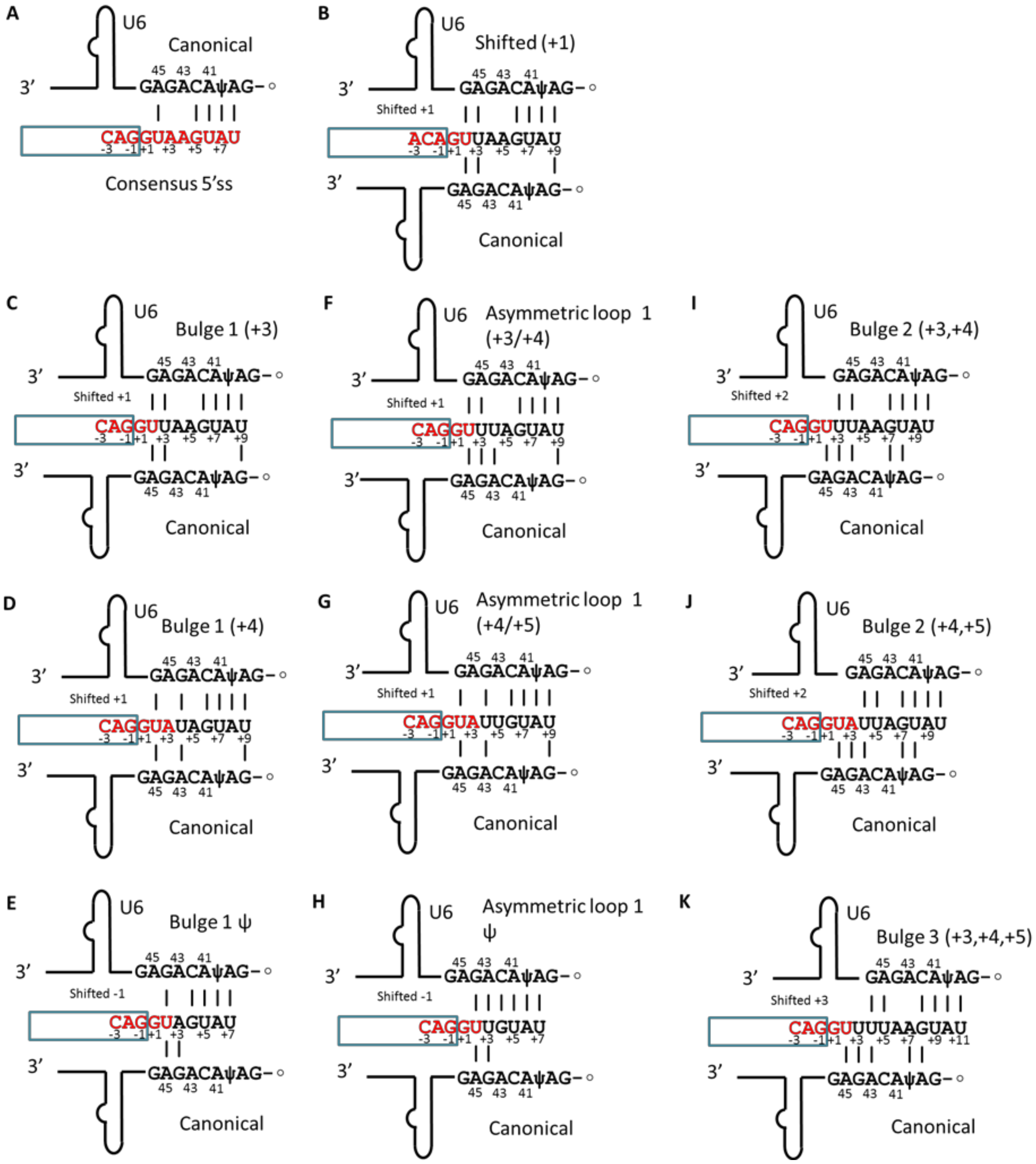


Figure 4.1 Proposed shifted U6 recognition registers for non-canonical candidate 5’ss

Each schematic represents a potential U6/5’ss shifted base pairing interaction for the various candidate non-canonical register 5’ss, apart from **(A)**, which shows the base pairing for the consensus sequence. The text label above the sequences indicates the predicted alternative U1 recognition register for each 5’ss, while the bottom label points out the canonical register of U6 binding for that 5’ss. The smaller text under the U6 stem loop indicates the proposed U6 register.

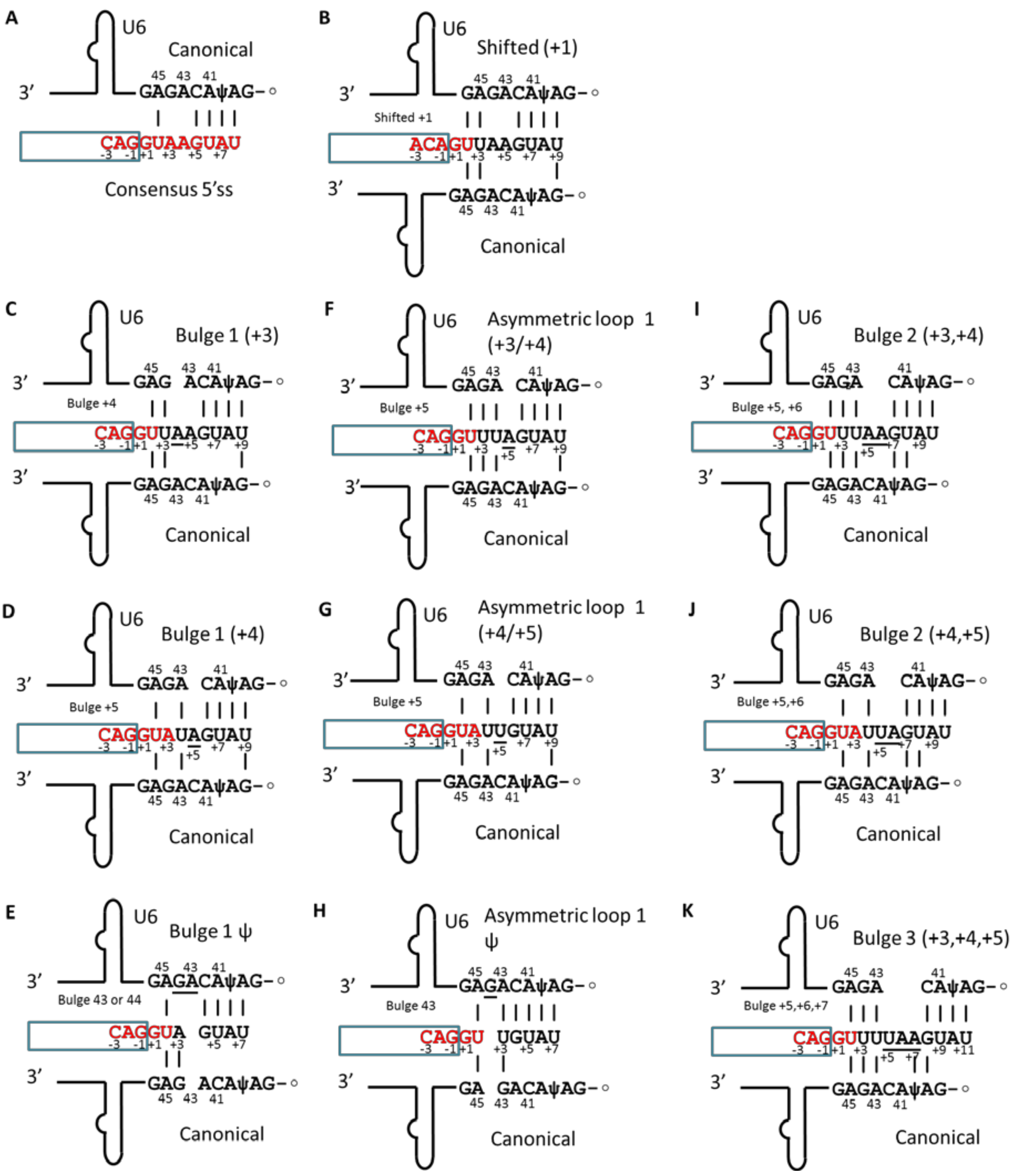


Figure 4.2 Proposed bulge U6 recognition registers for non-canonical candidate 5'ss

Each schematic represents a potential U6/5'ss bulge register base pairing interaction for the various candidate non-canonical 5'ss, apart from **(A)** and **(B)** which show the base pairing for the consensus sequence and the shifted +1 register respectively. The text label above the sequences indicates the predicted alternative U1 recognition register for each 5'ss, while the bottom label points out the canonical register of U6 binding for that 5'ss. The smaller text under the U6 stem loop indicates the proposed U6 register. Underlined nucleotides are proposed to be bulged.

4.1.10. Implications for alternative splicing and 5’ss strength predictions

Bulge/loop 5’ss are more commonly associated with alternative 5’ss events compared with canonical 5’ss, and were significantly enriched in alternative and cassette exon 5’ss (Roca et al., 2012). One example would be the alternatively spliced *CD46* gene (Purcell et al., 1991), which our lab has found to utilize two separate asymmetric loop 1 (+3/+4) register 5’ss, located on exons 7 and 8 (Tang et al., unpublished data). Both *CD46* exons 7 and 8 are cassette exons, and experiments show that if the asymmetric loop register was not used, these exons would not be included. This would lead to altered CD46 protein sequence and possibly functional shortfalls. Therefore, mutations that alter critical residues in the 5’ss or the U1 that impact recognition via these new mechanisms might affect alternative splicing outcomes, and thereby influence gene expression.

An important implication of bulge/asymmetric loop base pairing is that the length of the 5′ss motif increases with the length of the bulge/loop, causing extension of 5’ss beyond that of the canonical register such that some 5′ss are more than 11 nucleotides in length. Most of the current 5′ss scoring methods (Desmet et al., 2009; Hartmann et al., 2008; Markham and Zuker, 2008; Sahashi et al., 2007; Senapathy et al., 1990; Yeo and Burge, 2004) only consider 9 nucleotides or, in some cases, 11 nucleotides. Furthermore, the +5G hypothesis and the stabilizing influence of U-Ψ on neighboring base pairs should also be taken into account in 5’ss strength prediction algorithms. This would facilitate the development of improved algorithms to locate genes and exons in sequenced genomes, as well as to predict the effects of disease-causing mutations and SNPs that map to 5’ss.

Thus, the study of these registers should allow the development of more precise and increasingly sophisticated prediction software for 5’ss strength, 5’ss mapping, and/or splicing outcomes. This should hold important implications for the molecular classification of splicing mutations and SNPs, disease predictions, and for the study of alternative splicing.

4.1.11. Final ruminations on non-canonical registers

Building on previous studies (Roca and Krainer, 2009; Roca et al., 2012), our data expands the list of known 5'ss that are recognized by U1 via non-canonical base pairing registers. These findings further highlight the flexibility of the interaction between the 5'ss and the 5' end of U1, which permits various base pairing registers to support efficient splicing. At least 3116 human 5’ss may use a bulge/loop register that confers a substantially lower free energy than the canonical register (ΔΔG ≤ −1 kcal/mol). On the other hand, a small ΔΔG indicates that the bulge/asymmetric loop helix is roughly as stable as the non-bulge helix. Thus, it stands to reason that such 5’ss might be recognized by either canonical or bulge/asymmetric loop base pairing.

The limitations to the tractability of the 5’ss/U1 interaction are also revealed. Bulges on the exonic side of the 5’ss and bulges longer than 2 nucleotides on the intronic region of the 5’ss do not seem to be tolerated during 5’ss recognition. Additionally, the prospect of bulge/asymmetric loop usage may be negated in +5G-type 5’ss, forcing use of the canonical register.

4.2. XCI Project 1: Identification of novel genes involved in XCI via shRNA library screen

4.2.1. Transgenic cell lines generated by 'targeted' XIC integration

The targeted approach succeeded in generating many transgenic cell lines, but none of the cell lines seemed to possess the desired transgenic insert integration in the X chromosome of the Ainv15 cells. As stated earlier, the loxP site of the Ainv15 cell did not match the loxP site of the CH29-76M9-m*Tsix* BAC, preventing Cre-mediated recombination from occurring. The loxP site of Ainv15 cells was originally meant to be used with a different loxP construct, and the sequences of this loxP site were not published nor any indication given regarding the alterations made to it in the literature (Kyba et al., 2002). Future experiments should use the mutant loxP site sequence in the BAC for recombination to occur.

However, perhaps multi-copy transgenic insert integration (via homologous recombination or otherwise) occurred on the X chromosome, leading to a situation where many copies of the transgenic XIC are present in very close proximity. This would have caused an inability to resolve the signals in DNA-FISH genotyping, leading to misidentification of the mutant cells as a single signal phenotype. Quantitative PCR or Southern blotting to determine the XIC (or *Xist*) copy number in single signal transgenic CH29-76M9-m*Tsix* ES cells might shed light on the matter. If this is the case, then we might be able to utilize that cell line as part of the screen.

Also, some of the single signal transgenic cell lines may actually be the result of homologous recombination replacing the endogenous XIC with the BAC genomic insert XIC. On the genomic insert of the CH29-76M9-m*Tsix* BAC, the *Xist*/*Tsix* genes are flanked by approximately 67kb-long sequences derived from the X-chromosome. Homologous recombination of these extensive sequences with the genomic DNA seems highly plausible. In that case, in a number of the single signal phenotype transgenic CH29-76M9-m*Tsix* ES cells, the endogenous *Tsix* of the cell may have been replaced by the modified *Tsix* from the BAC. As the promoter of the *Tsix* has been disabled in CH29-76M9-

m*Tsix*, the *Tsix* inhibition of *Xist* activity might be down-regulated in these cells. Incidentally, this might explain why the 3.12.11 cell line mentioned previously was shown to exhibit a single signal when probed by DNA-FISH, but inexplicably experiences XCI during differentiation. This presents an opportunity to investigate the effects of *Tsix* knockdown using these cells. More such transgenic cell lines can be identified by performing reverse-transcription PCR using primers specific for *Tsix* on the single signal cell lines we obtained. The absence of a PCR fragment would indicate the absence of *Tsix* expression.

4.2.2. Transgene silencing in differentiated CH29-76M9-mTsix transgenic cells

The consistent transgene silencing observed, including the significant down-regulation of RFP and the sharp increase in hygromycin sensitivity detected during differentiation of the selected transgenic cell lines (Table 3.2), is most likely due to *Xist*-dependent gene silencing arising from the transgenic insertion. Since *Xist* RNA acts in *cis* to silence genes (Plath et al., 2002; Simon et al., 2013), and both the RFP and hygromycin resistance genes should tend to be located near the transgenic *Xist* gene due to the design of the CH29-76M9-m*Tsix* BAC, this seems to be a sensible assumption.

Other researchers have shown that disruption of endogenous *Tsix* expression in male murine ES cells may indeed lead to ectopic *Xist* RNA accumulation upon differentiation, but over protracted periods of culture, the *Xist* RNA accumulation is extinguished – without cell death caused by XCI silencing of critical X-linked genes (Luikenhuis et al., 2001; Sado et al., 2002). This led Sado et al. (2002) to conjecture that an X chromosome counting mechanism somehow sensed the number of the X chromosomes in the mutant cells, which then stopped and overturned the accumulation of the ectopically expressed *Xist*, reversing the XCI process before it became permanent. From this, they postulated the existence of a separate *Tsix*-independent silencing pathway for *Xist*. The same effect may also be replicated by presuming falling XCI activator levels in the differentiating cells during prolonged culture.

The observation that not all the differentiated transgenic cells were eliminated after hygromycin selection might be comparable to the situation described above by Sado et al. If it is assumed that short-lived ectopic *Xist* RNA accumulation occurs at different rates in individual cells when they are differentiated, the inactivation process might temporarily silence the hygromycin resistance gene in most cells before being reversed. As the hygromycin selection pressure was continuously applied, cells that inactivated the hygromycin resistance gene before the silencing was reversed would have been killed, leaving behind a few cells wherein silencing did not occur in time.

A later study (Ahn and Lee, 2010) argues that this phenotype is due to the use of RA to differentiate the cells. They found that RA, by negatively regulating *Oct4*, significantly affects *Xist* expression in *Tsix*-mutant male murine cells. They also showed that *Xist* RNA clouds in wild-type differentiated female cells are typically dense and well-defined, but those found in the RA-differentiated mutant males are loosely dispersed. Additionally, they discovered that the ectopic *Xist* expression seen in RA-treated male mutant cells usually does not lead to complete gene silencing on the X chromosome. It would be interesting to see if the same effect occurs in our transgenic cells.

4.2.3. Other potential uses of the CH29-76M9-mTsix transgenic ES cells

It would be interesting to determine whether autosomal insertion transgenic CH29-76M9-m*Tsix* ES cells would ‘count’ the transgenic XIC as an X chromosome, despite the transgenic XIC being located on an autosome. Would the inactivation effect be permanent or temporary? RNA-FISH for *Xist* with and without addition of RA, at different time points of differentiation of the autosomal insertion transgenic ES cells, should be able to clarify this point.

The mechanism of choice in XCI might be examined via the transgenic CH29-76M9-m*Tsix* ES cells. A previous study has shown that when expression of one of the two *Tsix* alleles in female cells is abrogated by targeted deletion, XCI is skewed towards the X chromosome carrying the mutant *Tsix* (Lee and Lu, 1999). It would be fascinating to see whether the same would occur in the autosomal insertion transgenic CH29-76M9-m*Tsix* cells. Would the mutant autosome be preferentially inactivated versus the endogenous X chromosome? Could both chromosomes be inactivated at the same time? Figure 3.18A, which seems to describe two *Xist* RNA clouds within the cell nucleus, might be an indication of the latter state. Simultaneous RNA-FISH for *Xist* cloud formation plus DNA-FISH for X-linked sequences could be performed on differentiated autosomal insertion transgenic CH29-76M9-m*Tsix* cells. This would allow us to identify the origin of the *Xist* cloud – the transgenic XIC locus on the autosome and/or the endogenous XIC locus on the X chromosome.

4.2.4. Targeting constructs for inserting Xist/XIC elements into Ainv15 ES cells

Attempts at making the ih*XIST* construct to introduce the human XIST gene (instead of the mouse *Xist*) into Ainv15 ES cells were hindered by multiple unforeseen circumstances, and the probability of acquiring the intended product appears to be extremely low due to the large size of the full *XIST* gene. Therefore, instead of trying to encompass the entire ~32kb human *XIST* gene (Flicek et al., 2014), we could try to insert the much shorter ~19kb cDNA sequence of the *XIST* RNA product (Flicek et al., 2014) instead. Step-wise homologous recombination could be done to incorporate the 11.4kb exon 1 and then the 7.4kb exon 6 into the plasmid. From there, the cDNA product covering the remaining exons could be introduced into the construct by traditional cloning methods.

Also, could it be feasible to modify the loxP sequence present in the CH29-76M9-mTsix BAC, or the imXist construct, so that it would match that of the mutant loxP site in Ainv15 ES cells? If that was done, repeating the same strategy as in electroporation 3 would allow the acquisition of transgenic cell lines with the desired genotype.

As mentioned earlier, a recent study showed that it was possible to use transgenic inducible *XIST* to silence the extra chromosome 21 in Down Syndrome cells (Jiang et al., 2013). They did so by using genome editing techniques, specifically zinc-finger nucleases (ZFNs), to insert the inducible transgene. Perhaps their technique may be adapted to the goal of inserting our transgenic sequence into the X chromosome. Other genome editing systems, like transcription activator-like effector nucleases (TALENs) could be utilized for the same purpose (Miller et al., 2011).

4.2.5. Final considerations on trangenic Xist insertion on male X chromosome

Despite multiple approaches, the goal of inserting an extra copy of the *Xist* gene into the single X chromosome of male murine ES cells (as far as we know) has not been achieved, and neither has a method of screening for XCI-linked genes been established. The attempt at targeted insertion of the transgenic XIC/*Xist* into the X chromosome of male ES cells failed despite the large number of transgenic cell lines generated, and we experienced difficulty constructing another targeting vector.

On the other hand, data obtained by leveraging the transgenic cell lines produced during the process seems to be in agreement with previous XCI research. These cells may even prove to be useful in researching the mechanisms of XCI. Also, with some tweaking of sequences and avoiding the pitfalls we experienced, it should be possible to generate an effective targeting construct to achieve transgenic *Xist* insertion into the X chromosome of the ES cells.

4.3. XCI Project 2: Expanded screen for activators involved in X chromosome inactivation

4.3.1. Prior results reveal potential sequences that are important for XCI activation

Earlier screening work by Ms. Khoo (unpublished data) indicates that the genomic regions covered by BACs RP23-280L7, RP24-285J22, RP23-282B14, and RP24-118E11 may contribute in some way to XCI activation. All these regions contain genes that are either zinc finger proteins, or involved in ubiquitination pathways. Perhaps there is a requirement for this protein functionality to be present in XCI activators.

The differentiation results of Jonkers et al. were successfully duplicated with RP24-240J16 transgenic ES cells. Interestingly, while Jonkers et al. reported that *Rnf12* BAC transgenic ES cells do not survive freeze-thawing, our RP24-240J16 transgenic ES cells do survive freeze-thawing. Perhaps subtle differences in methodology caused this discrepancy.

4.3.2. Generation of RP24-335G16 transgenic cell lines are problematic

As mentioned previously, we experienced difficulty in acquiring healthy RP24-335G16 transgenic cell lines. The reasons for this are still unclear. Perhaps extra copies of the insert sequence are prone to cause lethality in cells when integrated?

The RP24-335G16 BAC is known to cover two transcriptionally active genes, *Hprt* and *Phf6* (Plant homeodomain finger gene 6). While *Hprt* does play an important role in the cell, it is not known to exhibit lethality in cells when over-expressed (Degnen et al., 1977). On the other hand, the *Phf6* gene is conserved in mice, humans, dogs, cows, chickens, and zebrafish. Mutations in this gene lead to an X-linked mental retardation disorder (Voss et al., 2007). This may suggest that the *Phf6* gene present on the insert could be lethal when over-expressed, thereby eliminating the transgenic cells with successful insert integrations. If so, the RP24-335G16 BAC may be unsuitable for generation of a stable transgenic cell line. As of this writing, no literature exists that indicates the RP24-335G16 BAC has been successfully transfected or used to generate stable transgenic cells before.

4.3.3. WI1-667J14 cell line exhibits spontaneous differentiation

Zcchc5 (zinc-finger, CCHC domain containing 5) is the only (provisional) gene identified in the genomic region covered by WI1-667J14. *Zcchc5* is conserved in mice, rats, pigs, chimpanzees, and humans. However, no functions have previously been ascribed to *Zcchc5*, according to Genbank (Sayers et al., 2009). We can only postulate that the spontaneous differentiation observed in the cell lines might be due to the increased expression of the *Zcchc5* gene present in the transgenic WI1-667J14 insert. This hints at a possible developmental role for this gene. Of course, the anomalous phenotype might be attributed to position effects of the insertion, especially since the integration site appears to be located towards the telomeric end of a chromosome (Figure 3.21).

4.3.4. Lack of CH29-484O10 integration in cell lines

The complete absence of CH29-484O10 insert integration in the selected ES cell lines may indicate that one or more sequences in this region are lethal at increased dosage, or perhaps the transfection/integration success rate is much lower than usual. This situation might parallel the earlier RP24-66B4 electroporation experiment which also did not generate any cell lines with insert integrations, despite repeating the experiment thrice. However, as mentioned earlier, no genes have been mapped to the region covered by this BAC, and the only possible lead is that it encompasses a sequence with significant similarity to *Itga4* (Integrin alpha 4).

From Genbank (Sayers et al., 2009), it is known that the *Itga4* gene is conserved in mice, humans, chimpanzees, dogs, cows, rats, chickens, and zebrafish. *Itga4* functions to bind fibronectin, is involved in cell adhesion molecule binding, and possesses receptor activity. It is implicated in a wide variety of biological processes, in particular cell adhesion and migration, as well as participating in integrin-mediated signaling pathways. As it is an important gene, perhaps introducing similar transgenic sequences into the genome could have induced transgene silencing (by RNA interference or otherwise) that also deactivated the endogenous gene or its product, leading to cell death.

Another possibility is that the CH29-484O10 BAC encompasses an as-yet-unknown gene that is lethal when expressed at elevated levels, or that generates ncRNA that can affect the expression of other genes.

4.3.5. CH29-538N12 insert integration showed no XCI

The CH29-538N12-2 cell line was found to have autosomal insert integration (Figure 3.22), but when it was differentiated, no sign of XCI was detected (Figure 3.23). This indicates that this region might be unnecessary for XCI activation, and is promising evidence that our screening procedure does not affect XCI activation. In that case, it might be possible to use these cells as negative controls.

4.3.6. RP24-104K20 insert sequence might play a role in XCI activation

The tetraploid RP24-104K20-5 cell line contains at least two transgenic insertions (Figure 3.22). When differentiated, a significant percentage of the cells exhibit *Xist* RNA clouds (Figure 3.23). Since XCI in differentiated XXYY tetraploid cells typically resembles that of XY diploid cells, occurring in <0.3% of all cells (Monkhorst et al., 2008), this may hint at potential XCI activator activity originating from the transgenic insert. Of course, we cannot exclude the possibility that this remarkable phenotype might be due to aneuploidy, erroneous insert integration, or other technical issues.

As mentioned earlier, no known genes map to the region covered by RP24-104K20. The only sequence that shows similarity, predicted *Mus musculus* putative Pol polyprotein-like (LOC100505017), represents a retroviral or transposable element. This sequence is located on the X chromosome, albeit outside the HD2-HD3 breakpoint region. No function has yet been ascribed to this gene.

If this sequence is indeed important for XCI activation, it may continue to stoke XCI activation for longer than the other known activators during differentiation. This is because it is located much further away from *Xist* (~5Mb upstream) than *Rnf12* or even the potential activators discovered by Ms. Khoo (Figure 3.19). Therefore, the propagation of *Xist* in *cis*, and hence XCI, could take longer to affect the region covered by RP24-104K20.

4.3.7. Speculation regarding XCI activators

XCI is a complex process that hinges on a careful blend of epigenetic factors. Therefore, imbalances introduced by extra copies of a particular gene or even a specific sequence inserted in the wrong place might cause irregularities in cellular control of XCI. Consequently, the XCI activator regions discovered might act through numerous diverse pathways. They might act directly to up-regulate *Xist*, by stabilizing the *Xist* RNA, or even help promote *Xist* expression. They might also operate to inhibit or silence the inhibitors of *Xist*, like *Tsix* for example. It may even be possible that the transgenic regions/genes or the RNA or proteins derived from them are acting to soak up autosomally-derived XCI inhibitors, thus leading to a higher probability of XCI initiation.

4.3.8. Future searches for XCI activators

More data to back up the results of the expanded screen must be obtained. More cell lines with transgenic insertions need to be acquired, especially those of CH29-538N12 and RP24-104K20, to independently confirm the observed phenotype. Position effects and aneuploidy can then be dismissed as the potential cause of the ectopic XCI in the differentiated RP24-104K20 transgenic cells. It is also necessary to confirm that the ‘autosomal insertions’ seen under DNA-FISH in all the cell lines generated (including the ones made in the earlier experiments) are not due to X chromosome duplication causing incorrect genotyping.

From earlier data (Figure 3.20), independent transgenic clones with the same insert diverge in the percentage of cells that experience XCI when differentiated. This may be due to the dose-dependent nature of XCI activators: the higher the levels of the activator, the higher the percentage of cells at any one time that give rise to *Xist* clouds. Quantitative real-time PCR could be performed to confirm the difference in transgene dosage between cell lines with the same transgenic insert.

Genes that have been implicated in XCI activation could be tested by silencing these genes in female ES cells. After this, abrogation of XCI in the affected cells would positively mark that gene as important for XCI. In addition, we could try to narrow down the minimal regions required for XCI activation by fine-mapping using sequences that only cover part of the BAC inserts that have already been identified as playing a part in XCI activation.

To find more activators of XCI, BLAST could be performed to search for regions of homology to the sequences already known to grant XCI activation activity. This would allow prediction of potential XCI activators, making it easier to find the proper sequences and test them for XCI activator activity. Also, tiling arrays could be employed to perform transcription mapping of the HD2-HD3 breakpoint region. After creating a tiling array that covers the entirety of the HD2-HD3 breakpoint region, a dual-color microarray experiment could be used to detect genes in the HD2-HD3 breakpoint region that are expressed at higher levels in newly differentiated female cells versus female ES cells. Such genes would be more likely to be activators of XCI. The expression levels of these genes at different time points during differentiation of female cells could then be studied and evaluated. Since the X-encoded activators of XCI typically are themselves inactivated by XCI, genes in the HD2-HD3 region that are expressed at high levels at the beginning of differentiation but are then rapidly down-regulated as differentiation continues could be targeted for further testing. The microarray approach might also allow the general identification of more genes that are involved in XCI.

4.4. Integrated comparison of U1 snRNA and *Xist* RNA

Studying both U1 snRNA and *Xist* RNA grants a unique perspective on the distinct commonalities as well as the fascinating diversity that can be generated via the medium of ncRNA. Here we present a few comparisons between the two (Table 4.1).

Table 4.1 Comparison of U1 snRNA and *Xist* RNA

ncRNA	U1 snRNA	*Xist*
Size	Small, 164 nt.	Long, 17 kb.
Date of discovery	1968 (Hodnett and Busch, 1968).	1991 (Brown et al., 1991).
Abundance	~1 million molecules per cell.	Less than 2,000 molecules per female cell (Buzin et al., 1994).
Control of gene expression	Important for RNA splicing, which contributes to mRNA formation and hence protein expression. Inhibits polyadenylation (Gunderson et al., 1998).	Necessary for XCI to occur, which leads to X-linked gene silencing. Inhibits gene expression.
Transcription	Transcribed by RNA polymerase II (Henry et al., 1998)	Probably transcribed by RNA polymerase II.
Processing	Not spliced, not polyadenylated, capped on 5’ end with methylguanosine. Two residues pseudouridylated.	Spliced and polyadenylated (Plath et al., 2002).
Secondary structure	Extensive helix formation, highly structured, and forms internal loops that interact with other snRNP proteins (Pomeranz Krummel et al., 2009). Folds into particles.	Not much known, certain regions show conserved secondary structure (Maenner et al., 2010).
Interactions	RNA-RNA, RNA-protein.	RNA-RNA, RNA-DNA, RNA-protein.
Experimental manipulation	Alter splicing patterns by mutating residues in U1 snRNA, U1 suppressors, or on 5’ss itself.	*Xist* transgene expression causes ectopic chromosome-wide gene silencing, can be induced by *trans*-acting activators.
Potential therapeutic uses	U1 adaptor oligonucleotides can be used to tether U1 to target RNA transcripts (Goraczniak et al., 2009). Once there, U1 snRNP inhibits polyadenylation, leading to degradation of RNA	Ectopic *XIST* induction can be used for chromosome-wide silencing therapies to treat polyploidy (Jiang et al., 2013).

	transcripts, inhibiting target gene expression. Additionally, U1 suppressor therapy can also be used to correct errors in splicing arising from point mutations in the 5’ss (Hartmann et al., 2010). A possible weakness of this is that the endogenous U1 levels are already rather high, which might dilute the suppressor effect.	

From the details presented above, and indeed throughout this thesis, it is immediately obvious that these two ncRNAs are rather disparate in almost all aspects, from their size to their effect on gene expression. However, unifying themes do exist.

Importantly, both are able to influence and interact with a vast number of genes and various biological molecules in the cell. U1 snRNA, with its essential role in RNA splicing, interacts with the vast majority of pre-mRNA molecules (Roca et al., 2013), while *Xist* transcripts coordinate and recruit a myriad of gene silencing pathways in order to establish transcriptional inhibition across an entire X chromosome (Payer and Lee, 2008). The very fact that they participate in, and are indispensable for, these numerous interactions grants them the ability to control gene expression on a biologically significant scale.

This becomes apparent with their application in research and therapeutics. We are able to alter the levels of these ncRNA by adding extra exogenous copies of these ncRNAs into the cell, either by transient transfection or by permanent transgenic insertion. In doing so, we can amend the cellular environment, harnessing their ability to manipulate gene expression towards experimental and therapeutic ends – like what we did/has been done with the U1 suppressors and *Xist*. Correspondingly, silencing and knock-down of these ncRNAs can allow us to probe the pathways that they participate in.

Furthermore, transgenic modifications of these ncRNA allow us to further exploit these ncRNAs as a platform for various experimental and therapeutic purposes. Additionally, by introducing them into novel situations, new and unexpected findings can be made, just like with the +5G hypothesis. Such events indicate that much still needs to be done in order to understand the complex interplay of the myriad factors that influence the properties of not just these ncRNAs, but of all biological molecules in general. Hopefully, with this thesis, we have made more progress on that front.

5. Conclusions and Future Directions

5.1. Non-canonical 5’ss recognition registers

A total of 9 different non-canonical bulge and/or asymmetric loop 5’ss recognition registers have been validated, 3 of them in our studies. These registers affect at least 3,168 separate 5’ss, which is 1.57% of all 5’ss. We also showed evidence for usage of two other bulge 2 registers, which together account for a further 653 5’ss. Additionally, we showed preliminary evidence that bulge registers longer than 2 do not seem to be tolerated in 5’ss recognition. Moreover, we discovered that a novel mechanism requires that all 5’ss with +5G use the canonical register for 5’ss recognition, regardless of flanking mismatches or otherwise. Also, we found that +4U-Ψ5 interactions enhance 5’ss recognition, and helps stabilize the neighboring +5G-C4 base pair, a scenario that is not taken into account in the software used to predict the non-canonical registers. Applying our findings will enhance the accuracy of 5’ss strength predictions in predictive software and improve our ability to detect potentially deleterious mutations/SNPs that affect the 5’ss.

5.2. Future directions for the study of non-canonical 5’ss recognition registers

The other remaining asymmetric loop 5’ss, especially the longer asymmetric loop recognition registers with two or three extra unpaired and unmatched nucleotides on either side of the 5’ss/U1 helix could be investigated. More native 5’ss predicted to use bulge 2 (+3,+4) and bulge 2 (+4,+5) registers need to be found and authenticated, while testing of bulge 2 (+4,+5) register 5’ss usage in a heterologous context needs to be completed. Additionally, bulge/asymmetric loop 3 registers could also be tested, so as to definitely prove or disprove the usage of those registers, and thus by extension the other longer bulge/loop registers.

Oligonucleotide duplex melting experiments will further confirm the formation of +5G-type 5’ss duplexes with the 5’ end of U1, as well as the contribution of U-Ψ interactions. The role of U1 Ψ in 5'ss recognition and the +3U-Ψ6 interaction in the 5’ss/U1 helix and their contribution to 5’ss recognition could be tested. Suppressor U6 snRNA experiments should be carried out to determine whether U6 will base pair in an alternative register for these non-canonical 5'ss. Finally, the contribution of protein factors to non-canonical 5'ss recognition could also be examined.

5.3. XCI projects

5.3.1. Ectopic Xist expression: conclusions and future directions

In this project, we successfully triggered ectopic transgenic *Xist* expression in differentiating ES cells. This led to the silencing of transgenic reporter genes and antibiotic resistance cassettes in *cis*. However, antibiotic selection could not cleanly eliminate all the transgenic ES cells, and thus the desired screen could not be established. An effective and straightforward method of introducing the XIC/*Xist* transgenic sequence into the X chromosome of the male murine ES cells needs to be developed, employing any one of the plethora of molecular biology tools that have been established in recent years for this purpose. Following that, if ectopic XCI on the male X chromosome leads to complete cell lethality when differentiated, the screen for genes relevant for XCI can be established.

5.3.2. Search for XCI activators: conclusions and future directions

Based on the work performed in this project, we find that the CH29-538N12 BAC insert probably does not contain XCI activators. RP24-104K20 may well include elements important for XCI activation, but it does not encompass any known genes. These experiments will need to be repeated in order to confirm the results; much more work remains to be done if we are to find more activators of XCI and understand the mechanisms that drive the activation of XCI. As discussed above, more data to back up the results of the expanded screen must be obtained and repeat experiments done to confirm the findings. Genes that we implicated in XCI activation could be tested, and we should narrow down the minimal regions required for XCI activation. To find more activators of XCI, we could look for regions sharing homology to sequences already known to grant XCI activation activity. This would allow prediction of potential XCI activators. Also, tiling arrays could be employed to perform transcription mapping of the HD2-HD3 breakpoint region to detect genes in the HD2-HD3 breakpoint region that are expressed at higher levels in newly differentiated female cells versus female ES cells. Such genes would be more likely to be activators of XCI. The microarray approach might also allow the general identification of more genes that are involved in XCI.

6. REFERENCES

Ahn, J.Y., Lee, J.T., 2010. Retinoic acid accelerates downregulation of the Xist repressor, Oct4, and increases the likelihood of Xist activation when Tsix is deficient. BMC Developmental Biology 10, 90.

Arnez, J.G., Steitz, T.A., 1994. Crystal structure of unmodified tRNA(Gln) complexed with glutaminyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. Biochemistry 33, 7560-7567.

Avner, P., Heard, E., 2001. X-chromosome inactivation: counting, choice and initiation. Nature Reviews: Genetics 2, 59-67.

Bach, I., Rodriguez-Esteban, C., Carriere, C., Bhushan, A., Krones, A., Rose, D.W., Glass, C.K., Andersen, B., Izpisua Belmonte, J.C., Rosenfeld, M.G., 1999. RLIM inhibits functional activity of LIM homeodomain transcription factors via recruitment of the histone deacetylase complex. Nature Genetics 22, 394-399.

Barr, M.L., Bertram, E.G., 1949. A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. Nature 163, 676.

Beard, C., Li, E., Jaenisch, R., 1995. Loss of methylation activates Xist in somatic but not in embryonic cells. Genes & Development 9, 2325-2334.

Ben-Shem, A., Garreau de Loubresse, N., Melnikov, S., Jenner, L., Yusupova, G., Yusupov, M., 2011. The Structure of the Eukaryotic Ribosome at 3.0 Å Resolution. Science 334, 1524-1529.

Berget, S.M., Moore, C., Sharp, P.A., 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proceedings of the National Academy of Sciences 74, 3171-3175.

Boyle, J., 2008. Molecular biology of the cell, 5th edition by B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. Biochemistry and Molecular Biology Education 36, 317-318.

Brackenridge, S., Wilkie, A.O., Screaton, G.R., 2003. Efficient use of a 'dead-end' GA 5' splice site in the human fibroblast growth factor receptor genes. The EMBO Journal 22, 1620-1631.

Brown, C.J., Willard, H.F., 1994. The human X-inactivation centre is not required for maintenance of X-chromosome inactivation. Nature 368, 154-156.

Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R., Willard, H.F., 1991. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature 349, 38-44.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology 268, 78-94.

Busch, A., Hertel, K.J., 2012. Evolution of SR protein and hnRNP splicing regulatory factors. Wiley Interdisciplinary Reviews: RNA 3, 1-12.

Buzin, C.H., Mann, J.R., Singer-Sam, J., 1994. Quantitative RT-PCR assays show Xist RNA levels are low in mouse female adult tissue, embryos and embryoid bodies. Development 120, 3529-3536.

Carmel, I., Tal, S., Vig, I., Ast, G., 2004. Comparative analysis detects dependencies among the 5' splice-site positions. RNA 10, 828-840.

Cartegni, L., Chew, S.L., Krainer, A.R., 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nature Reviews Genetics 3, 285-298.

Cartegni, L., Hastings, M.L., Calarco, J.A., de Stanchina, E., Krainer, A.R., 2006. Determinants of Exon 7 Splicing in the Spinal Muscular Atrophy Genes, SMN1 and SMN2. The American Journal of Human Genetics 78, 63-77.

Chan, S.I., Lee, G.C., Schmidt, C.F., Kreishman, G.P., 1972. Guanine-uracil base-pairing. Biochemical and Biophysical Research Communications 46, 1536-1543.

Chow, L.T., Gelinas, R.E., Broker, T.R., Roberts, R.J., 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell 12, 1-8.

Clemson, C.M., McNeil, J.A., Willard, H.F., Lawrence, J.B., 1996. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. The Journal of Cell Biology 132, 259-275.

Coffin, J., 1992. Structure and Classification of Retroviruses. In: Levy, J. (Ed.), The Retroviridae. Springer US, pp. 19-49.

Cooper, T.A., Wan, L., Dreyfuss, G., 2009. RNA and disease. Cell 136, 777-793.

Crick, F., 1970. Central dogma of molecular biology. Nature 227, 561-563.

Csankovszki, G., Nagy, A., Jaenisch, R., 2001. Synergism of Xist Rna, DNA Methylation, and Histone Hypoacetylation in Maintaining X Chromosome Inactivation. The Journal of Cell Biology 153, 773-784.

Csankovszki, G., Panning, B., Bates, B., Pehrson, J.R., Jaenisch, R., 1999. Conditional deletion of Xist disrupts histone macroH2A localization but not maintenance of X inactivation. Nature Genetics 22, 323-324.

Davis, D.R., 1995. Stabilization of RNA stacking by pseudouridine. Nucleic Acids Research 23, 5020-5026.

Degnen, G.E., Miller, I.L., Adelberg, E.A., Eisenstadt, J.M., 1977. Overexpression of an unstably inherited gene in cultured mouse cells. Proceedings of the National Academy of Sciences 74, 3956-3959.

Desmet, F.O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M., Beroud, C., 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Research 37, e67.

Donohoe, M.E., Silva, S.S., Pinter, S.F., Xu, N., Lee, J.T., 2009. The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting. Nature 460, 128-132.

Donohue, J., Trueblood, K.N., 1960. Base pairing in DNA. Journal of Molecular Biology 2, 363-371.

Du, H., Rosbash, M., 2002. The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing. Nature 419, 86-90.

Eperon, L.P., Estibeiro, J.P., Eperon, I.C., 1986. The role of nucleotide sequences in splice site selection in eukaryotic pre-messenger RNA. Nature 324, 280-282.

Fica, S.M., Tuttle, N., Novak, T., Li, N.S., Lu, J., Koodathingal, P., Dai, Q., Staley, J.P., Piccirilli, J.A., 2013. RNA catalyses nuclear pre-mRNA splicing. Nature 503, 229-234.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Kulesha, E., Martin, F.J., Maurel, T., McLaren, W.M., Murphy, D.N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S.J., Vullo, A., Wilder, S.P., Wilson, M., Zadissa, A., Aken, B.L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T.J.P., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D.R., Searle, S.M.J., 2014. Ensembl 2014. Nucleic Acids Research 42, D749-D755.

Freund, M., Hicks, M.J., Konermann, C., Otte, M., Hertel, K.J., Schaal, H., 2005. Extended base pair complementarity between U1 snRNA and the 5' splice site does not inhibit splicing in higher eukaryotes, but rather increases 5' splice site recognition. Nucleic Acids Research 33, 5112-5119.

Gao, K., Masuda, A., Matsuura, T., Ohno, K., 2008. Human branch point consensus sequence is yUnAy. Nucleic Acids Research 36, 2257-2267.

Goraczniak, R., Behlke, M.A., Gunderson, S.I., 2009. Gene silencing by synthetic U1 Adaptors. Nature Biotechnology 27, 257-263.

Gu, A.D., Zhou, H., Yu, C.H., Qu, L.H., 2005. A novel experimental approach for systematic identification of box H/ACA snoRNAs from eukaryotes. Nucleic Acids Research 33, e194.

Gunderson, S.I., Polycarpou-Schwarz, M., Mattaj, I.W., 1998. U1 snRNP Inhibits Pre-mRNA Polyadenylation through a Direct Interaction between U1 70K and Poly(A) Polymerase. Molecular Cell 1, 255-264.

Gutell, R.R., 2012. Comparative Analysis of the Higher-Order Structure of RNA. Springer, 11-22 pp.

Hall, K.B., McLaughlin, L.W., 1991. Properties of a U1/mRNA 5' splice site duplex containing pseudouridine as measured by thermodynamic and NMR methods. Biochemistry 30, 1795-1801.

Hamma, T., Ferré-D'Amaré, A.R., 2006. Pseudouridine Synthases. Chemistry & Biology 13, 1125-1135.

Hamma, T., Ferre-D'Amare, A.R., 2010. The box H/ACA ribonucleoprotein complex: interplay of RNA and protein structures in post-transcriptional RNA modification. The Journal of Biological Chemistry 285, 805-809.

Hartmann, L., Theiss, S., Niederacher, D., Schaal, H., 2008. Diagnostics of pathogenic splicing mutations: does bioinformatics cover all bases? Frontiers in Bioscience 13, 3252-3272.

Hartmann, L., Neveling, K., Borkens, S., Schneider, H., Freund, M., Grassman, E., Theiss, S., Wawer, A., Burdach, S., Auerbach, A.D., Schindler, D., Hanenberg, H., Schaal, H., 2010. Correct mRNA processing at a mutant TT splice donor in FANCC ameliorates the clinical phenotype in patients and is enhanced by delivery of suppressor U1 snRNAs. The American Journal of Human Genetics 87, 480-493.

Hassold, T., Hunt, P., 2001. To err (meiotically) is human: the genesis of human aneuploidy. Nature Reviews Genetics 2, 280-291.

Henry, R.W., Ford, E., Mital, R., Mittal, V., Hernandez, N., 1998. Crossing the Line between RNA Polymerases: Transcription of Human snRNA Genes by RNA Polymerases II and III. Cold Spring Harbor Symposia on Quantitative Biology 63, 111-120.

Hodnett, J.L., Busch, H., 1968. Isolation and characterization of uridylic acid-rich 7 S ribonucleic acid of rat liver nuclei. The Journal of Biological Chemistry 243, 6334-6342.

Hua, Y., Vickers, T.A., Okunola, H.L., Bennett, C.F., Krainer, A.R., 2008. Antisense masking of an hnRNP A1/A2 intronic splicing silencer corrects SMN2 splicing in transgenic mice. The American Journal of Human Genetics 82, 834-848.

Huang, Y., Zhang, J., Yu, X., Xu, T., Wang, Z., Cheng, X., 2013. Molecular functions of small regulatory noncoding RNA. Biochemistry (Moscow) 78, 221-230.

Hudson, G.A., Bloomingdale, R.J., Znosko, B.M., 2013. Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. RNA 19, 1474-1482.

Hwang, D.Y., Cohen, J.B., 1996. U1 snRNA promotes the selection of nearby 5' splice sites by U6 snRNA in mammalian cells. Genes & Development 10, 338-350.

Jamison, S.F., Crow, A., Garcia-Blanco, M.A., 1992. The spliceosome assembly pathway in mammalian extracts. Molecular and Cellular Biology 12, 4279-4287.

Jiang, J., Jing, Y., Cost, G.J., Chiang, J.-C., Kolpa, H.J., Cotton, A.M., Carone, D.M., Carone, B.R., Shivak, D.A., Guschin, D.Y., Pearl, J.R., Rebar, E.J., Byron, M., Gregory, P.D., Brown, C.J., Urnov, F.D., Hall, L.L., Lawrence, J.B., 2013. Translating dosage compensation to trisomy 21. Nature 500, 296-300.

Jonkers, I., Monkhorst, K., Rentmeester, E., Grootegoed, J.A., Grosveld, F., Gribnau, J., 2008. Xist RNA is confined to the nuclear territory of the silenced X chromosome throughout the cell cycle. Molecular and Cellular Biology 28, 5583-5594.

Jonkers, I., Barakat, T.S., Achame, E.M., Monkhorst, K., Kenter, A., Rentmeester, E., Grosveld, F., Grootegoed, J.A., Gribnau, J., 2009. RNF12 is an X-Encoded dose-dependent activator of X chromosome inactivation. Cell 139, 999-1011.

Kandels-Lewis, S., Seraphin, B., 1993. Involvement of U6 snRNA in 5' splice site selection. Science 262, 2035-2039.

Kierzek, E., Malgowska, M., Lisowiec, J., Turner, D.H., Gdaniec, Z., Kierzek, R., 2014. The contribution of pseudouridine to stabilities and structure of RNAs. Nucleic Acids Research 42, 3492-3501.

Kiss, A.M., Jady, B.E., Bertrand, E., Kiss, T., 2004. Human box H/ACA pseudouridylation guide RNA machinery. Molecular and Cellular Biology 24, 5797-5807.

Konarska, M.M., Vilardell, J., Query, C.C., 2006. Repositioning of the reaction intermediate within the catalytic center of the spliceosome. Molecular Cell 21, 543-553.

Krawczak, M., Thomas, N.S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., Cooper, D.N., 2007. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Human Mutation 28, 150-158.

Kubota, T., Roca, X., Kimura, T., Kokunai, Y., Nishino, I., Sakoda, S., Krainer, A.R., Takahashi, M.P., 2011. A mutation in a rare type of intron in a sodium-channel gene results in aberrant splicing and causes myotonia. Human Mutation 32, 773-782.

Kung, J.T.Y., Colognori, D., Lee, J.T., 2013. Long Noncoding RNAs: Past, Present, and Future. Genetics 193, 651-669.

Kyba, M., Perlingeiro, R.C.R., Daley, G.Q., 2002. HoxB4 Confers Definitive Lymphoid-Myeloid Engraftment Potential on Embryonic Stem Cell and Yolk Sac Hematopoietic Progenitors. Cell 109, 29-37.

Lee, J.C., Gutell, R.R., 2004. Diversity of Base-pair Conformations and their Occurrence in rRNA Structure and RNA Structural Motifs. Journal of Molecular Biology 344, 1225-1249.

Lee, J.T., 2005. Regulation of X-Chromosome Counting by Tsix and Xite Sequences. Science 309, 768-771.

Lee, J.T., Lu, N., 1999. Targeted mutagenesis of Tsix leads to nonrandom X inactivation. Cell 99, 47-57.

Lee, J.T., Davidow, L.S., Warshawsky, D., 1999a. Tsix, a gene antisense to Xist at the X-inactivation centre. Nature Genetics 21, 400-404.

Lee, J.T., Lu, N., Han, Y., 1999b. Genetic analysis of the mouse X inactivation center defines an 80-kb multifunction domain. Proceedings of the National Academy of Sciences 96, 3836-3841.

Lee, J.T., Strauss, W.M., Dausman, J.A., Jaenisch, R., 1996. A 450 kb Transgene Displays Properties of the Mammalian X-Inactivation Center. Cell 86, 83-94.

Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L., Steitz, J.A., 1980. Are snRNPs involved in splicing? Nature 283, 220-224.

Lesser, C., Guthrie, C., 1993. Mutations in U6 snRNA that alter splice site specificity: implications for the active site. Science 262, 1982-1988.

Libri, D., Duconge, F., Levy, L., Vinauger, M., 2002. A role for the Psi-U mismatch in the recognition of the 5' splice site of yeast introns by the U1 small nuclear ribonucleoprotein particle. The Journal of Biological Chemistry 277, 18173-18181.

Lilley, D.M.J., 2011. Mechanisms of RNA catalysis. Philosophical Transactions of the Royal Society B: Biological Sciences 366, 2910-2917.

Lin, Y., Kielkopf, C.L., 2008. X-ray structures of U2 snRNA-branchpoint duplexes containing conserved pseudouridines. Biochemistry 47, 5503-5514.

Lu, Z.X., Jiang, P., Xing, Y., 2012. Genetic variation of pre-mRNA alternative splicing in human populations. Wiley Interdisciplinary Reviews: RNA 3, 581-592.

Luikenhuis, S., Wutz, A., Jaenisch, R., 2001. Antisense transcription through the Xist locus mediates Tsix function in embryonic stem cells. Molecular and Cellular Biology 21, 8512-8520.

Lyon, M.F., 1961. Gene Action in the X-chromosome of the Mouse (Mus musculus L.). Nature 190, 372-373.

Maenner, S., Blaud, M., Fouillen, L., Savoye, A., Marchand, V., Dubois, A., Sanglier-Cianférani, S., Van Dorsselaer, A., Clerc, P., Avner, P., Visvikis, A., Branlant, C., 2010. 2-D Structure of the A Region of Xist RNA and Its Implication for PRC2 Association. PLoS Biology 8, e1000276.

Markham, N.R., Zuker, M., 2008. UNAFold: software for nucleic acid folding and hybridization. Methods in Molecular Biology 453, 3-31.

Marz, M., Stadler, P., 2011. RNA Interactions. In: Collins, L. (Ed.), RNA Infrastructure and Networks. Springer New York, pp. 20-38.

Matera, A.G., Wang, Z., 2014. A day in the life of the spliceosome. Nature Reviews Molecular Cell Biology 15, 108-121.

Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H., 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of Molecular Biology 288, 911-940.

Mattick, J.S., Makunin, I.V., 2006. Non-coding RNA. Human Molecular Genetics 15 Spec No 1, R17-29.

Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J., Dulay, G.P., Hua, K.L., Ankoudinova, I., Cost, G.J., Urnov, F.D., Zhang, H.S., Holmes, M.C., Zhang, L., Gregory, P.D., Rebar, E.J., 2011. A TALE nuclease architecture for efficient genome editing. Nature Biotechnology 29, 143-148.

Monani, U.R., Lorson, C.L., Parsons, D.W., Prior, T.W., Androphy, E.J., Burghes, A.H., McPherson, J.D., 1999. A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2. Human Molecular Genetics 8, 1177-1183.

Monkhorst, K., Jonkers, I., Rentmeester, E., Grosveld, F., Gribnau, J., 2008. X Inactivation Counting and Choice Is a Stochastic Process: Evidence for Involvement of an X-Linked Activator. Cell 132, 410-421.

Navarro, P., Page, D.R., Avner, P., Rougeulle, C., 2006. Tsix-mediated epigenetic switch of a CTCF-flanked region of the Xist promoter determines the Xist transcription program. Genes & Development 20, 2787-2792.

Navarro, P., Moffat, M., Mullin, N., Chambers, I., 2011. The X-inactivation trans-activator Rnf12 is negatively regulated by pluripotency factors in embryonic stem cells. Human Genetics 130, 255-264.

Navarro, P., Chambers, I., Karwacki-Neisius, V., Chureau, C., Morey, C., Rougeulle, C., Avner, P., 2008. Molecular Coupling of Xist Regulation and Pluripotency. Science 321, 1693-1695.

Nilsen, T.W., Graveley, B.R., 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature 463, 457-463.

Nissen, P., Hansen, J., Ban, N., Moore, P.B., Steitz, T.A., 2000. The Structural Basis of Ribosome Activity in Peptide Bond Synthesis. Science 289, 920-930.

Ogawa, Y., Sun, B.K., Lee, J.T., 2008. Intersection of the RNA interference and X-inactivation pathways. Science 320, 1336-1341.

Owczarzy, R., Vallone, P.M., Gallo, F.J., Paner, T.M., Lane, M.J., Benight, A.S., 1997. Predicting sequence-dependent melting stability of short duplex DNA oligomers. Biopolymers 44, 217-239.

Panning, B., Jaenisch, R., 1996. DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes & Development 10, 1991-2002.

Patterson, D., 2009. Molecular genetic analysis of Down syndrome. Human Genetics 126, 195-214.

Payer, B., Lee, J.T., 2008. X Chromosome Dosage Compensation: How Mammals Keep the Balance. Annual Review of Genetics 42, 733-772.

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., Brockdorff, N., 1996. Requirement for Xist in X chromosome inactivation. Nature 379, 131-137.

Plath, K., Mlynarczyk-Evans, S., Nusinow, D.A., Panning, B., 2002. Xist RNA and the mechanism of X chromosome inactivation. Annual Review of Genetics 36, 233-278.

Plath, K., Fang, J., Mlynarczyk-Evans, S.K., Cao, R., Worringer, K.A., Wang, H., de la Cruz, C.C., Otte, A.P., Panning, B., Zhang, Y., 2003. Role of Histone H3 Lysine 27 Methylation in X Inactivation. Science 300, 131-135.

Pomeranz Krummel, D.A., Oubridge, C., Leung, A.K., Li, J., Nagai, K., 2009. Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. Nature 458, 475-480.

Purcell, D.J., Russell, S., Deacon, N., Brown, M., Hooker, D., McKenzie, I.C., 1991. Alternatively spliced RNAs encode several isoforms of CD46 (MCP), a regulator of complement activation. Immunogenetics 33, 335-344.

Rastan, S., Robertson, E.J., 1985. X-chromosome deletions in embryo-derived (EK) cell lines associated with lack of X-chromosome inactivation. Journal of Embryology and Experimental Morphology 90, 379-388.

Rastan, S., Brown, S.D.M., 1990. The search for the mouse X-chromosome inactivation centre. Genetics Research 56, 99-106.

Reddy, R., Busch, H., 1988. Small Nuclear RNAs: RNA Sequences, Structure, and Modifications. In: Birnstiel, M. (Ed.), Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles. Springer Berlin Heidelberg, pp. 1-37.

Rhode, B.M., Hartmuth, K., Westhof, E., Luhrmann, R., 2006. Proximity of conserved U6 and U2 snRNA elements to the 5' splice site region in activated spliceosomes. The EMBO Journal 25, 2475-2486.

Riggs, A.D., Porter, T.N., 1996. X-Chromosome Inactivation and Epigenetic Mechanisms.

Roca, X., Krainer, A.R., 2009. Recognition of atypical 5' splice sites by shifted base-pairing to U1 snRNA. Nature Structural & Molecular Biology 16, 176-182.

Roca, X., Sachidanandam, R., Krainer, A.R., 2003. Intrinsic differences between authentic and cryptic 5′ splice sites. Nucleic Acids Research 31, 6321-6333.

Roca, X., Sachidanandam, R., Krainer, A.R., 2005. Determinants of the inherent strength of human 5' splice sites. RNA 11, 683-698.

Roca, X., Krainer, A.R., Eperon, I.C., 2013. Pick one, but be quick: 5' splice sites and the problems of too many choices. Genes & Development 27, 129-144.

Roca, X., Akerman, M., Gaus, H., Berdeja, A., Bennett, C.F., Krainer, A.R., 2012. Widespread recognition of 5′ splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides. Genes & Development 26, 1098-1109.

Roca, X., Olson, A.J., Rao, A.R., Enerly, E., Kristensen, V.N., Borresen-Dale, A.L., Andresen, B.S., Krainer, A.R., Sachidanandam, R., 2008. Features of 5'-splice-site efficiency derived from disease-causing mutations and comparative genomics. Genome Research 18, 77-87.

Rogers, J., Wall, R., 1980. A mechanism for RNA splicing. Proceedings of the National Academy of Sciences 77, 1877-1879.

Ruggero, D., Grisendi, S., Piazza, F., Rego, E., Mari, F., Rao, P.H., Cordon-Cardo, C., Pandolfi, P.P., 2003. Dyskeratosis congenita and cancer in mice deficient in ribosomal RNA modification. Science 299, 259-262.

Sado, T., Li, E., Sasaki, H., 2002. Effect of TSIX disruption on XIST expression in male ES cells. Cytogenetic and Genome Research 99, 115-118.

Sado, T., Hoki, Y., Sasaki, H., 2005. Tsix Silences Xist through Modification of Chromatin Structure. Developmental Cell 9, 159-165.

Sahashi, K., Masuda, A., Matsuura, T., Shinmi, J., Zhang, Z., Takeshima, Y., Matsuo, M., Sobue, G., Ohno, K., 2007. In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5' splice sites. Nucleic Acids Research 35, 5995-6003.

Sakharkar, M.K., Chow, V.T., Kangueane, P., 2004. Distributions of exons and introns in the human genome. In Silico Biology 4, 387-393.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Yaschenko, E., Ye, J., 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 37, D5-15.

Senapathy, P., Shapiro, M.B., Harris, N.L., 1990. Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. Methods in Enzymology 183, 252-278.

Shapiro, M.B., Senapathy, P., 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Research 15, 7155-7174.

Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., Sachidanandam, R., 2006. Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Research 34, 3955-3967.

Shibata, S., Lee, J.T., 2003. Characterization and quantitation of differential Tsix transcripts: implications for Tsix function. Human Molecular Genetics 12, 125-136.

Siliciano, P.G., Guthrie, C., 1988. 5' splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. Genes & Development 2, 1258-1267.

Siliciano, P.G., Jones, M.H., Guthrie, C., 1987. Saccharomyces cerevisiae has a U1-like small nuclear RNA with unexpected properties. Science 237, 1484-1487.

Silva, J., Mak, W., Zvetkova, I., Appanah, R., Nesterova, T.B., Webster, Z., Peters, A.H.F.M., Jenuwein, T., Otte, A.P., Brockdorff, N., 2003. Establishment of Histone H3 Methylation on the Inactive X Chromosome Requires Transient Recruitment of Eed-Enx1 Polycomb Group Complexes. Developmental Cell 4, 481-495.

Silva, S.S., Rowntree, R.K., Mekhoubad, S., Lee, J.T., 2008. X-chromosome inactivation and epigenetic fluidity in human embryonic stem cells. Proceedings of the National Academy of Sciences 105, 4820-4825.

Simon, M.D., Pinter, S.F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S.K., Kesner, B.A., Maier, V.K., Kingston, R.E., Lee, J.T., 2013. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature 504, 465-469.

Singh, N.N., Singh, R.N., 2011. Alternative splicing in spinal muscular atrophy underscores the role of an intron definition model. RNA Biology 8, 600-606.

Soukup, G.A., 2001. Nucleic Acids: General Properties, eLS. John Wiley & Sons, Ltd.

Staley, J.P., Guthrie, C., 1998. Mechanical devices of the spliceosome: motors, clocks, springs, and things. Cell 92, 315-326.

Sun, H., Chasin, L.A., 2000. Multiple splicing defects in an intronic false exon. Molecular and Cellular Biology 20, 6414-6425.

Tian, D., Sun, S., Lee, J.T., 2010. The Long Noncoding RNA, Jpx, Is a Molecular Switch for X Chromosome Inactivation. Cell 143, 390-403.

Tinoco Jr, I., Bustamante, C., 1999. How RNA folds. Journal of Molecular Biology 293, 271-281.

Traub, W., Elson, D., 1966. RNA composition and base pairing. Science 153, 178-180.

Treisman, R., Orkin, S.H., Maniatis, T., 1983. Structural and functional defects in beta-thalassemia. Progress in Clinical and Biological Research 134, 99-121.

Turunen, J.J., Niemela, E.H., Verma, B., Frilander, M.J., 2013. The significant other: splicing by the minor spliceosome. Wiley Interdisciplinary Reviews: RNA 4, 61-76.

Voss, A.K., Gamble, R., Collin, C., Shoubridge, C., Corbett, M., Gecz, J., Thomas, T., 2007. Protein and gene expression analysis of Phf6, the gene mutated in the Borjeson-Forssman-Lehmann Syndrome of intellectual disability and obesity. Gene Expression Patterns 7, 858-871.

Wahl, M.C., Will, C.L., Luhrmann, R., 2009. The spliceosome: design principles of a dynamic RNP machine. Cell 136, 701-718.

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

Wassarman, D.A., Steitz, J.A., 1992. Interactions of small nuclear RNA's with precursor messenger RNA during in vitro splicing. Science 257, 1918-1925.

Watkins, N.J., Bohnsack, M.T., 2012. The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA. Wiley Interdisciplinary Reviews: RNA 3, 397-414.

Watson, J.D., Crick, F.H., 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, 737-738.

Wu, G., Yu, A.T., Kantartzis, A., Yu, Y.T., 2011. Functions and mechanisms of spliceosomal small nuclear RNA pseudouridylation. Wiley Interdisciplinary Reviews: RNA 2, 571-581.

Wutz, A., 2011. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. Nature Reviews Genetics 12, 542-553.

Wutz, A., Jaenisch, R., 2000. A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. Molecular Cell 5, 695-705.

Yeo, G., Burge, C.B., 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Journal of Computational Biology 11, 377-394.

Zhuang, Y., Weiner, A.M., 1986. A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. Cell 46, 827-835.