THE UNIVERSITY
of EDINBURGH

THE UNIVERSITY *of* EDINBURGH

Tuning the RNAPII elongation rate is required for optimal pre-mRNA splicing efficiency and fidelity

Vahid Aslanzadeh

APRIL, 2017

A thesis presented for the degree of Doctor of Philosophy

Declaration

I hereby declare that I am the sole author of this thesis and that the work presented here is my own. Data obtained through the collaboration with other individuals are appropriately indicated. It has not been submitted for any other degree or professional qualification.

_____

Vahid Aslanzadeh

Edinburgh, April 2017

Acknowledgements

I would like to express my sincere appreciation to Jean, for her unflagging guidance and support throughout of my PhD. Thank you Jean for providing me the opportunity to work in your lab and for your patience and giving me time to listen whenever I needed and for the suggestions. I have thoroughly enjoyed and benefited from being a student of yours. I would like to extend my gratitude to Sander Granneman for helping me to start RNAseq data analysis and guiding me through all the steps. I am very grateful to Yuanhua Huang for helping me with informatics analysis. Thanks to all Begg's lab members, past and present (David, Ema, Jane, Susana, Gonzalo, Bella, Eve and Edward) for their help and support. Thank you Gonzalo for always having time to talk and being a good friend.

I would like to especially thank Darwin Trust of Edinburgh for funding my PhD. My educational pursuits would have not been possible without your generous support. Thank you for enabling this opportunity!

I have no words to acknowledge my mom and dad. Thank you very much for all that you have given to me. Without your support I might not be the person that I am today. Finally, I express my deepest thanks to my dear wife Fariba, for her constant encouragement and support throughout my PhD. I owe much to Fariba, without whose love and understanding it would have not been possible to complete my PhD.

I dedicate this thesis to my parents and my dear wife.

Abstract

Splicing mainly occurs co-transcriptionally, suggesting that transcription and pre-mRNA splicing could be synchronized. The nature of this phenomenon suggests that transcription elongation rate may influence splicing outcomes and, indeed, there is evidence for effects on alternative splicing in mammals. To elucidate potential effects of transcription rate on splicing efficiency and fidelity, splicing of nascent transcripts was investigated in fast and slow elongating RNA polymerase II (RNAPII) mutants in *Saccharomyces cerevisiae*. High kinetic resolution 4-thio Uracil labelling of nascent RNA reveals that fast RNAPII accumulates unspliced pre-mRNA that represents reduced co-transcriptional splicing. Conversely, low levels of unspliced pre-mRNA were detected in the slow mutant due to increased co-transcriptional splicing. The highly stable association of nascent transcripts with elongating RNAPII permits co-transcriptional splicing to be measured by analysis of transcripts that co-purify with RNAPII. Measuring co-precipitation of the spliced mRNA and excised intron that are associated with RNAPII demonstrates that splicing is mostly co-transcriptional with the slow mutant, and the fast mutant reduces co-transcriptional splicing. How elongation rate affects splicing fidelity in budding yeast and whether faster and slower transcription have the opposite effect on splicing fidelity as might be predicted by the kinetic coupling model is an open question. Using deep RNA sequencing, splicing fidelity was determined in yeast transcription elongation mutants. Results show that both fast and slow transcription reduce splicing fidelity mainly in ribosomal protein coding transcripts. Analysis reveals that splicing fidelity depends largely on intron length, secondary structure and splice site score. These analyses also provide new insights regarding the effect of altering transcription rate on selection of transcription start sites. Together, these results indicate that optimal splicing efficiency and fidelity require finely-tuned transcription speed.

Lay summary

In most human genes the information (or code) in the DNA sequence is interrupted by noncoding regions called introns. A process called transcription produces RNA, which is a copy of the DNA. The RNA has to be cut and then spliced back together to remove the introns and produce an uninterrupted code for a protein. Frequently, the RNA pieces are joined in different ways, giving rise to proteins with different properties (like deleting or replacing scenes in a film to change a story). Mistakes in splicing the RNA, which can be caused by genetic defects or disease, result in defective proteins. Transcription and splicing are therefore of critical importance for controlling protein production and these processes are thought to be co-regulated. There is evidence that speed of transcription is tuned over the genes whose splicing may be regulated. The aim of this study was to investigate the effect that changing transcription speed has on RNA splicing. Yeast was used as a model organism for this study because many powerful experimental techniques are available. As these processes are highly conserved between yeast and humans, my results can also provide important insights into RNA splicing in humans. My results show that if transcription is faster or slower than normal, the amount of efficiently spliced RNA decreases or increases, respectively. In addition, most of the RNAs produced with faster and slower elongation speed have splicing mistakes that can potentially affect protein production. Overall, these findings help us to better understand the interaction between transcription and splicing and also provide us new understanding about how some splicing mistakes can occur.

Abbreviations

%	Percentage
^{0}C	Degree Celsius
5-FOA	5′Fluoroorotic acid
ATP	Adenosine tri phosphate
BP	Branch point
bp	Base pair
cDNA	Complementary Deoxy ribonucleic acid
CHCl3	Chloroform
DEPC	Diethyl pyrocarbonate
DNA	Deoxy ribonucleic acid
DNase	Deoxy ribonuclease
EDTA	Ethylenediamine tetraacetic acid
g	Gram(s)
HEPES	(4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
IgG	Immunoglobin G
IP	Immunoprecipitation
LiCl	Lithium chloride
mg	Milligram
min	Minute(s)
ml	Millilitre
mM	Millimolar
mRNA	Mature Ribonucleic acid
nM	Nanomolar
non-RP	Non-ribosomal protein
NP	40 nonidet P40 detergent
nt	Nucleotide(s)
OD600	Optical Density at 600 nm
ORF	Open Reading Frame
PAS	Polyadenylation site
PCR	Polymerase Chain Reaction
PEG	Poly ethylene glycol
PTC	Premature termination codon
qPCR	Quantitative polymerase chain reaction
RNA	Ribonucleic acid
RNAPII	RNA Polymerase II
RNase	Ribonuclease
RP	Ribosomal protein
rpm	Revolutions per minute
RT	Reverse transcription
RT-qPCR	Reverse transcriptase polymerase chain reaction

SC-Leu	Synthetic complete medium without leucine
SC-Lys	Synthetic complete medium without lysine
SDS	Sodium dodecyl sulphate
Ser2p	Phosphorylated serine 2
Ser5p	Phosphorylated serine 5
snRNA	Small Nuclear Ribonucleic acid
snRNP	Small Nuclear Ribonucleic protein particle
SS	splice site
SSDNA	Salmon sperm DNA
TAE	Tris acetate buffer
TE	Tris EDTA buffer
TSS	Transcription start site
UV	Ultra violet
YMM	Yeast Minimal Medium
YPDA	Yeast/Peptone/Dextrose/Adenine (medium)

Contents

Chapter 1 Introduction

1.1 Transcription

Transcription is the first stage of gene expression and is performed by three different RNA polymerases (I, II or III). During this process, RNA polymerase reads the genetic code from DNA and synthesizes an RNA molecule. Protein coding RNAs are polymerised by RNA polymerase II (RNAPII) and exported to the cytoplasm for protein synthesis by the translation machinery. This flow of genetic information from DNA to protein is known as the central dogma of molecular biology[1]. RNAPII is also responsible for transcription of many noncoding RNAs including small nuclear RNAs and microRNAs[2]. RNAPI synthesises ribosomal RNAs (rRNA) and RNAPIII transcribes the 5s rRNA, tRNA, U6 small nuclear RNA and some other small RNAs[3].

1.1.1 Transcription by RNA polymerase II

Transcription driven by RNAPII is a multistep controlled process that is divided into four major steps; initiation, promoter escape, elongation and termination. After assembling the RNAPII complex on the promoter region, initiation of transcription occurs with the help of general transcription factors. After synthesizing a small fragment of RNA, transcription complexes with successful initiation escape from the promoter region and enter the elongation phase. In highly expressed genes, several transcription complexes perform elongation simultaneously[4]. The transcription process is terminated after transcribing the target gene and the complex is disassembled from the DNA and recycled for further transcription cycles. Eukaryotic RNAPII is a large complex of proteins comprising 12 subunits (Rpb1-12). Three of these subunits Rpb4, Rpb7 and Rpb9 are unique to RNAPII and are not essential for the transcription process in yeast[5]. Rpb9 is essential for transcription in *Drosophila*[6]. The other 9 subunits are found in all three eukaryotic RNA polymerases. Rpb1, Rpb2, Rpb3 and Rpb11 are the core and conserved subunits that have a common structure

and function with the prokaryotic core RNA polymerase[7]. Rpb1 is the largest subunit that contains the carboxyl-terminal domain (CTD) and the trigger loop (TL).

1.1.1.1 Carboxyl-terminal domain of RNAPII

The carboxy-terminal domain (CTD) of the large subunit, Rpb1, is an essential and unique domain in RNAPII and is conserved from yeast to human. This domain is absent in the other two RNA polymerases; RNAPI and RNAPIII. In mammals, the CTD contains 52 tandem heptapeptide repeats (YSPTSPS) Tyr-Ser-Pro-Thr-Ser-Pro-Ser[8]. In yeast, the CTD is shorter, with only 26 heptapeptide repeats[8]. During the transcription cycle, the serine, threonine and tyrosine residues of the CTD undergo a series of controlled phosphorylation and dephosphorylation cycles that modulate transcription initiation, promoter escape, elongation, termination and multiple co-transcriptional processes[9–11]. Additionally, peptidyl-prolyl bond isomerisation can occur at Pro3 and Pro6 by a prolyl isomerase and the threonine and the serine residues can be glycosylated[10].

The most studied modifications of the CTD are phosphorylation and dephosphorylation of the serine residues. During the transcription process, multiple kinases and phosphatases add and remove the phosphate groups from the CTD to modulate transcription and co-transcriptional events[10]. Ser2 (Ser2p) and Ser5 (Ser5p) phosphorylations are required for expression of all genes transcribed with RNAPII, whereas Ser7 (Ser7p) phosphorylation is required for expression of specific genes[10]. Ser5p and Ser2p are predominantly placed in promoter and 3'end of the genes respectively, although some studies detected Ser5p at the 3'end of some genes as well[10]. Ser7p and Ser5p are enriched early during the transcription process but only Ser7p persists at high levels until transcription terminates[12]. The CTD modifications play crucial role in RNA biogenesis. For example, Ser7p is higher than Ser2p over non-coding genes and is required for 3'end processing of the small nuclear RNAs[13]. Ser5p is important for recruitment of the capping enzyme (guanylyl transferase) responsible for capping the 5' end of RNA[10]. Ser2p is associated with the recruitment

of histone modification, elongation, splicing, transport, 3′ end RNA-processing factors. Ser5p and Ser2p are not mutually exclusive and the Ser5p-Ser2p double mark has been shown to facilitate the recruitment of Prp40 and U2AF65[9,14].

The cycle of phosphorylation and dephosphorylation of the CTD is regulated by several enzymes (Figure 1.1). In yeast, Kin28 (human homologues Cdk7) a component of TFIIH factor, phosphorylates Ser5[15] while Ssu72 and Rtr1 have been characterized as Ser5p phosphatases [16]. Phosphorylation of Ser2p and Ser7p is carried out by Ctk1 and Bur1 (homologues of Cdk9 and Cdk12 respectively in mammals,)[10]. However, Fcp1 dephosphorylates Ser2p[17] while removal of the phosphate group from Ser7p[18] is carried out by Ssu72. The order, timing and place of phosphorylation and dephosphorisation of the CTD is important for faithful gene expression, and deciphering of the CTD code is integral to better understanding of co-transcriptional RNA processes[10,11].

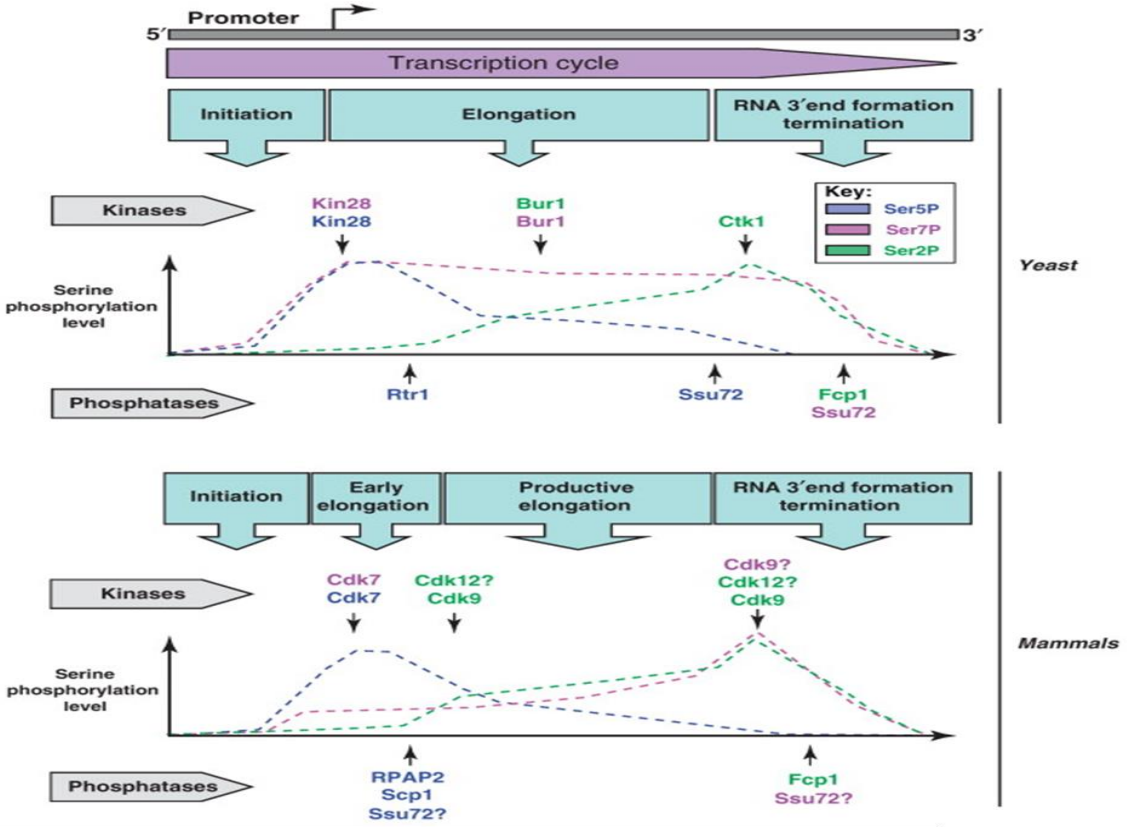


Figure 1.1: The carboxyl-terminal domain phosphorylation pattern across yeast and human protein-coding genes relative to the transcription cycle. The kinases and phosphatases responsible for establishing these patterns are noted above and below, respectively. The symbol "?" indicates an enzyme suspected to have this function. Figure is from[10].

1.1.1.2 Tigger loop of RNAPII

Trigger loop (TL) is a mobile domain of Rpb1 located in the catalytic centre of the RNA polymerase and is conserved in RNAPI, II, III and also in prokaryotic RNA polymerase[19]. TL directly interacts with the base-paired substrates through its evolutionarily conserved histidine residue (His1085 in *S. cerevisiae,* His936 in *E. coli* β′ and His1108 in human). The exact function of TL is yet to be understood but it is implicated in regulating substrate selection, translocation and catalysis[20]. It is proposed that direct interaction between the TL and a NTP substrate matched with DNA template stabilize base-pairings at the 3'-end of the nascent RNA and regulate transcription fidelity[19]. The TL was also found to play a role in intrinsic cleavage of RNA and transcriptional pausing through rearrangements in the catalytic centre[21]. Mutations in TL can affect the dynamics of TL mobility and result in changes in transcription elongation rates both positively and negatively (Figure 1.2)[20].

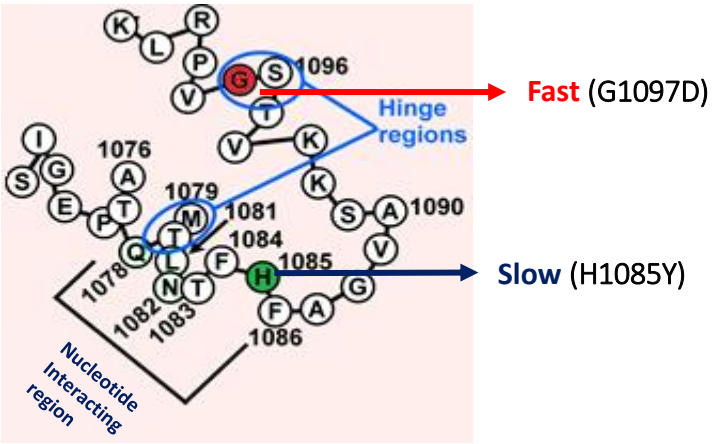


Figure 1.2: *S. cerevisiae* Rpb1 trigger loop[20]. Cartoon showing the nucleotide interacting region and the hinge regions. Mutations in the highlighted residues alter trigger loop properties and cause changes in transcription elongation rate. The yeast mutant *rpb1-G1097D* causes faster transcription while *rpb1-*H1085Y causes slower transcription.

1.1.2 Transcription initiation

The first step in transcription is transcription initiation, which starts with recognition of the promoter by the core protein recognition factors. The TATA box sequence in the promoter is then recognized by TATA binding protein (TBP), a component of the TFIID[22]. Only ~ 20% of yeast genes contain a TATA box sequence in their promoter and they are generally associated with stress response[23]. There are some differences in the factors involved in the recognition of promoters which contain or not TATA sequence, however, TFIID complex is functionally redundant at many promoters. Binding TBP to DNA induces a bend in the region[23]. After promoter recognition, the assembly of factors required for transcription initiation is facilitated by binding of the transcription activators upstream of the promoter and recruitment of chromatin remodelers. TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH are general transcription factors required for successful initiation complex formation. TFIIA and TFIIB stabilize TBP and TATA interaction and subsequently TFIIB recruits RNAPII and TFIIF to the initiation site[22]. TFIIF stabilizes TFIIB and stimulates transcription elongation. Assembly of TFIIE and TFIIH facilitate unwinding of the DNA in the promoter site and complete formation of the pre-initiation complex (PIC). Next, DNA is opened by ATP hydrolysis forming the 'transcription bubble' and RNA synthesis initiates. Phosphorylation of Ser5 and Ser7 by TFIIH (human Cdk7) and the departure of some transcription factors after transcribing a few nucleotides leads to promoter escape and entering the elongation phase[24].

1.1.3 Transcription elongation

After successful initiation, RNAPII enters the early elongation phase and synthesizes the mRNA by elongating from 5’ to 3’. Before RNAPII enters the active elongation state, in higher eukaryotes RNAPII experiences a promoter proximal pausing that is mediated by negative elongation factor (NELF) and DRB sensitivity inducing factor (DSIF)[25]. This pausing is potentially involved in effective 5’capping of the RNA, synchronous activation, checkpoint for coupling elongation, some RNA processing

functions and recruitment of multiple regulatory signals[26]. Positive transcription elongation factor (P-TEFb), facilitates the release from pausing, displacing NELF by phosphorylation[23]. P-TEFb is one of the main regulators of transcription elongation and interacts with different proteins and RNA subunits. DSIF and P-TEFb have homologs in eukaryotes ranging from yeast to human, but yeast and *Caenorhabditis elegans* appear to lack NELF, suggesting that NELF regulatory function could be restricted to a subset of eukaryotes[27]. In yeast, Bur1 and Ctk1, which are components of P-TEFb, are responsible for Ser2 phosphorylation leading to productive elongation via changing the CTD phosphorylation pattern from Ser5/Ser7 to Ser2/Ser7[23]. The processivity of transcription, stable association of RNAPII with the DNA, is regulated by multiple elongation factors such as Spt5 and TFIIS and histone chaperons like Spt6 and Spt16[23,25]. RNAPII faces multiple barriers during the elongation phase and a productive elongation requires effective histone modifications, repositioning of the nucleosomes and chromatin remodelling. RNAPII can move a few nucleotides backwards, called back-tracking, which is implicated in several critical processes, including controlling transcription elongation, pausing, termination, fidelity, and genome instability[28]. Co-transcriptional RNA processing mechanisms can also affect elongation by RNAPII[23,29].

1.1.4 Transcription termination

The final step of transcription is transcription termination that finishes transcription of the gene and prevents transcription read-through into the downstream gene. The release of protein coding RNA from transcription machinery is achieved with co-transcriptional recruitment of the cleavage and polyadenylation factors to the termination processing signals at the 3′ untranslated region (UTR) of the nascent RNA[30]. Pcf11 binds to hyper-phosphorylated Ser2 CTD and after endonucleolytic cleavage of the RNA at the cleavage and polyadenylation site, poly(A) polymerase Pap1 adds a tail of poly A repeats to the 3’ end of the cleaved RNA[31]. Currently, two models explain the dissociation of the RNA from the polymerase. According to the torpedo model, Rat1, a nuclear 5' to 3' single-stranded RNA exonuclease, is then

recruited by the CTD-interacting protein Rtt103 and degrades the downstream portion of cleaved RNA and triggers dissociation of the elongation complex[32]. Alternatively, the allosteric model suggests that polyadenylation leads to loss of elongation factors and/or conformational changes in transcription, which stimulate transcription complex dissociation. After adding a tail, poly binding protein (PABP) binds to the Poly A and protects the RNA from degradation[33]. Alternative polyadenylation that mainly occurs in higher eukaryotes also play an important role in gene regulation[34]. Transcription termination of non-coding RNAs is facilitated by the Nrd1–Nab3–Sen1 (NNS) pathway in yeast and the cap-binding complex (CBC)–ARS2 pathway in humans[32].

1.2 Pre-mRNA splicing

The primary RNA synthesized by transcription is called precursor messenger RNA (pre-mRNA). Forty years ago, Berget et al. discovered that pre-mRNA contained intervening sequences, called introns, which are flanked on either side by exon segments[35]. Splicing is the process of removing introns from pre-mRNAs and ligating adjacent exons to produce messenger RNA (mRNA). The process of splicing is catalysed by the spliceosome, a multi-megadalton ribonucleoprotein (RNP) complex. The splicing process is mainly conserved from yeast to human and this made yeast a model organism for studying splicing mechanisms. Less than 5% of *S. cerevisiae* genes (~300) contain an intron and splicing of most of these introns is required for viability[36]. Despite the low number of intron-containing genes in yeast, some of these genes are highly expressed, leading to production of approximately one third of the cellular protein coding transcripts. Although, across species the splicing mechanism is conserved, contrary to yeast, almost all of mammalian transcripts undergo splicing[37].

1.2.1 The mechanism of pre-mRNA splicing

The spliceosome removes introns from pre-mRNAs by two consecutive trans-esterification reactions[36]. The intron boundaries are defined by 5' splice site (exon-intron junction) and 3' splice site (intron-exon junction). Within the intron, there are branch-point (BP) and polypyrimidine tract sequences that are required by the splicing mechanism. The first reaction is a nucleophilic attack of the 2'-OH group of the conserved adenosine in the branch point (BP) sequence to the phosphate of the guanine nucleotide at the 5' splice site (Figure 1.3). This makes an unusual 2'-5' phosphodiester bond between BP and 5' splice site and exon-intron junction is cleaved. The products of this step are a lariat intron-exon2 intermediate and a free 3'-OH at the end of the cleaved 5'exon. Conformational rearrangement of the spliceosome components allows catalysis to proceed to the second trans-esterification reaction. In the second step, nucleophilic attack of the 3'-OH end of the released exon to the 3' splice site cleaves the intron-exon junction and the two exons are joined. The intron is released as RNA a lariat structure and then de-branched by Drb1 and finally degraded. The spliceosome is then disassembled and recycled for further splicing reactions.

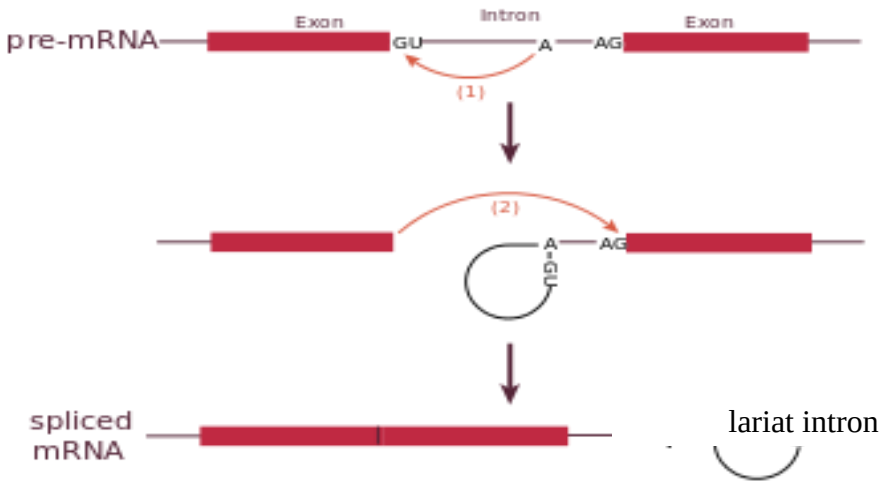


Figure 1.3: Schematic representation of the two transesterification reactions of splicing.

1.2.2 Consensus sequences define the intron boundaries

Consensus sequences determine pre-mRNA intron/exon boundaries. R/GUAUGU (where the symbol "/" denotes the cleavage site) is the consensus sequence for the 5'SS in yeast that is complementary to the U1 snRNA 5'SS base pairing region. The first two nucleotides (/GU) are conserved and their mutations prevent splicing of the intron[38]. AG/ dinucleotide at the end of introns defines the 3'SS. The vast majority of introns in eukaryotic genes fit the canonical 5'SS/GU-AG/3'SS borders. UACUAAC is the consensus BP sequence that base pairs with U2 snRNA. The last adenosine nucleotide is conserved across species and is required for the first trans-esterification reaction. The polypyrimidine tract is a U-rich sequence upstream of the 3'SS of the intron. The major class of introns with these consensus sequences are spliced with U2-dependent spliceosomes. The U12-dependent spliceosome, which is absent in *S. cerevisiae*, mediates splicing of a minor class of introns that have non-canonical consensus sequences in higher eukaryotes[39]. In addition to these consensus sequences, there are numerous *cis-acting* regulatory elements that recruit *trans-acting* factors to enhance or suppress the splicing reaction.

1.2.3 Spliceosome mediated splicing

The core spliceosome is composed of 5 small nuclear RNAs designated as snRNAs (U1, U2, U4, U5, U6), and a range of distinct proteins that associate with the snRNAs to generate small nuclear ribonucleoprotein complexes (snRNP)[36]. Several snRNA-pre-mRNA, protein-RNA and protein-protein interactions are required to catalyse the splicing reaction. Spliceosome assembly involves an ordered, stepwise assembly of snRNPs and other distinct proteins to the pre-mRNA[40]. Spliceosome assembly starts with recruitment of U1 snRNP to the 5'SS and initiating the commitment complex or early complex formation (Figure 1.4). Commitment complex formation also requires Bbp1 and Mud2 (human U2AF65) that recognize BP and nearby sequences, respectively[41]. Direct interaction of the U1 snRNP with CTD of RNAPII facilitates its recruitment to the 5'SS[14]. U1-5'SS interaction is ATP-independent and unstable, and

other auxiliary factors help its stability[42]. U2-snRNP is recruited onto the 3'SS in a co-transcriptional manner and this interaction is facilitated by U2 auxiliary factors (U2AFs) and some other proteins. ATP hydrolysis by Prp5 and Sub2 (DExD-box helicases) facilitate recognition of the BP and interaction of the U2-snRNP with U1-snRNP to form pre-spliceosome complex (A complex). Interaction between U1-snRNP and U2-snRNP across the intron is called intron definition. In higher eukaryotes where introns are often much longer than exons, splice sites are recognized via U1-U2 snRNP interaction across the exons, forming an exon definition complex[43]. Subsequently, rearrangements of this complex brings the 5'SS, 3'SS and BP of the intron in close proximity. After formation of complex A, U4/U6.U5 tri-snRNP, a 1.5-megadalton pre-assembled tri-snRNP complex is recruited, to form B complex, in a reaction mediated by Prp28 (DExD/H-box RNA helicase). Subsequently, complex B is catalytically activated (complex B*) through a series of conformational rearrangements and with the help of multiple RNA helicases including Brr2, Snu114 (a U5-snRNP component) and Prp2. During the activation of B complex, Brr2 unwinds the U4-U6 snRNA duplexes, result in forming U2–U6 snRNA structure[44]. At this step, U1 and U4 snRNPs are released from the complex. Complex B* performs the first catalytic step and nucleates the formation of complex C. Products of the first catalytic step are the excised and free 5' exon and the intron–exon lariat intermediate, which are associated with complex C. Following several ATP-dependent rearrangements, complex C carries out the second catalytic step that generates a lariat intron and joins the exons together. Function of complex C is mediated by Prp8, Prp16, Prp18 and Slu7. After the second step has occurred, the ATPase activity of Prp22 releases the spliced product from the spliceosome. Upon completion of the splicing reaction, several RNA helicases including Brr2, Snu114, Prp22 and Prp43 are responsible for promoting spliceosome disassembly to recycle its components for additional rounds of splicing[42]. Recent *in vitro* splicing studies have revealed that all splicing steps are revisable, even both catalytic steps, however, it is still unclear whether splicing can be reversed *in vivo*[42,45].

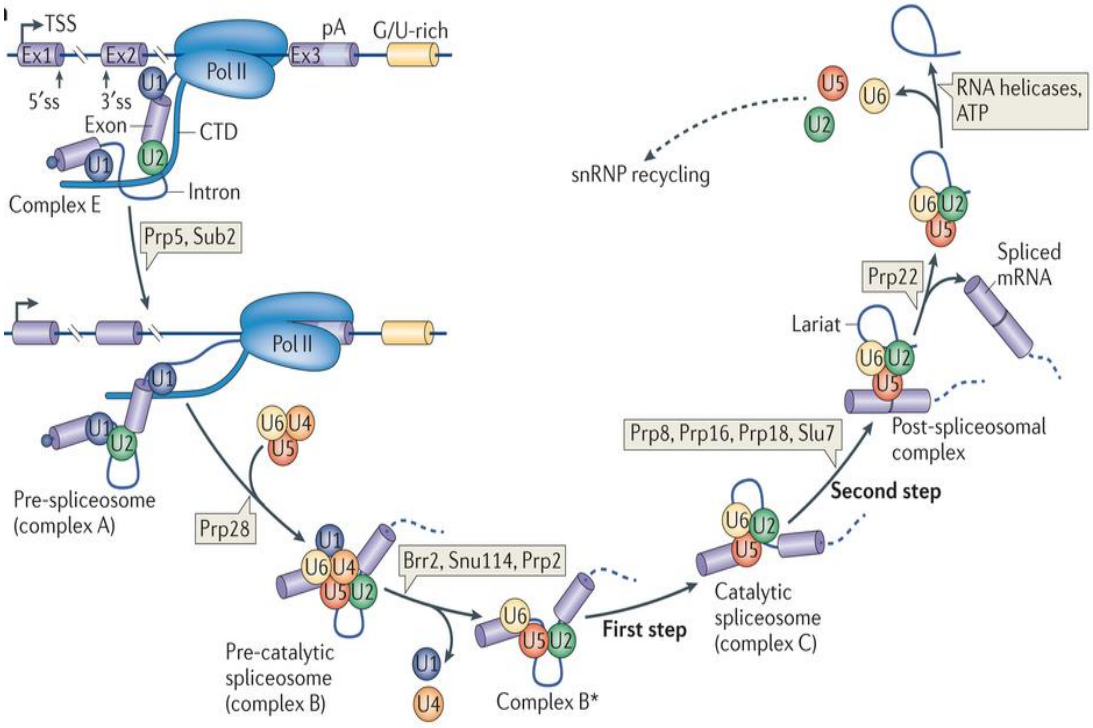


Figure 1.4: Step-wise assembly of the spliceosome and catalytic steps of splicing[42]**.** RNAPII helps co-transcriptional assembly of splicing factors.

1.2.4 Alternative splicing

Before sequencing of the human genome, it was believed that there are 40000 up to 100000 protein-coding genes in a single human cell. However, later extensive genome sequencing studies reduced this estimation to as few as 19000 protein coding genes[46]. Conversely, thanks to next generation transcriptome sequencing, the number of transcripts that undergo alternative splicing increased from 50% at 2002 to almost 100% to date[37]. Recently, by using *in silico* methods, researchers estimated that 19000 human genes can produce more than 200,000 protein coding transcripts[46,47]. This means that one gene can produce multiple isoforms of transcripts, a function that is achieved with alternative splicing of pre-mRNA. In mammals, nearly all pre-mRNAs contain more than two exons, which facilities differential alternative splicing[37]. There are five major modes of alternative splicing; exon skipping, mutually exclusive exons, alternative 5'SS, alternative 3'SS and intron retention (Figure 1.5). A combination of splice sites strength, presences of *cis* and *trans* acting elements, chromatin

environment, RNA structure, RNA transcription and degradation can determine the outcome of pre-mRNA alternative splicing. In yeast, only few genes have more than two exons and alternatives splicing is not a common event.

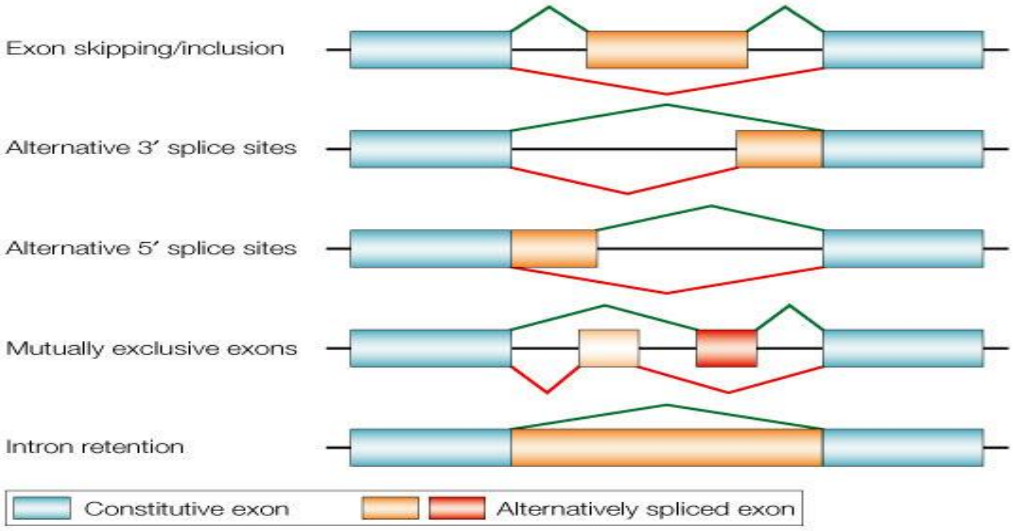


Figure 1.5: Major alternative splicing events in the metazoan transcriptome. Blue boxes indicate constitutive exons, while brown boxes indicate alternative spliced exons.

1.2.5 Co-transcriptional splicing

Based on *in vitro* studies, it was believed that splicing occurs after transcription and when 5' capping and 3' end processing of the RNA is completed. However, mounting evidence during the last two decades have changed this picture to a view that the majority of splicing events are catalysed before transcription termination; meaning that splicing is happening co-transcriptionally. The first striking evidence of splicing pre-mRNA during RNAPII transcription was shown 29 years ago by 'Miller spread' electron micrograph on *Drosophila melanogaster* embryonic gene (Figure 1.6a)[48]. Later, multiple labs demonstrated co-transcriptional recruitment of splicing factors to the nascent transcript[49]. In 2011, after more than two decades, next generation cellular total RNA sequencing (NGS) has revealed the same picture by showing the widespread co-transcriptional splicing in human liver and brain cells[50]. This study has revealed a 5'-3' slope in the read coverage, especially on the long introns, which supports the

model that splicing is carried out immediately after transcription of an intron. This 'saw-tooth' pattern is repeated over each intron due to co-transcriptional splicing of intron immediately after its transcription (figure 1.6b). Additionally, studying chromatin associated transcripts in yeast, drosophila and human cell lines further supported the prevalence of co-transcriptional splicing[51–53]. Capturing the native elongating transcripts with RNAPII immunoprecipitation followed by sequencing (NET-seq) further confirmed the extensive co-transcriptional splicing in yeast with high resolution. More recently, the innovative single-molecule intron tracking (SMIT) approach showed that spliceosome assembly and splicing can take place immediately following intron synthesis by RNAPII[54]. While splicing can also occur post-transcriptionally, splicing during transcription can add more layers of RNA processing regulation tools by allowing mutual interactions between splicing and chromatin, transcription and other RNA processing mechanisms[55]. For example, it has been shown that co-transcriptional recruitment of splicing factors to nascent RNA prevents R-loop formation and thereby enhances genomic stability, and also it can prevent RNA from degradation to allow proper maturation[56,57]. In fact, computational modelling of splicing and transcription has suggested that co-transcriptional splicing is more efficient than post-transcriptional splicing[58]. As a result of co-transcriptional splicing, recruitment of splicing factors by RNAPII and transcription elongation rate can influence splicing outcome by two mechanism, referred to as the recruitment coupling and the kinetic coupling models[59]. These two coupling models are not mutually exclusive and together explain how coupling works.

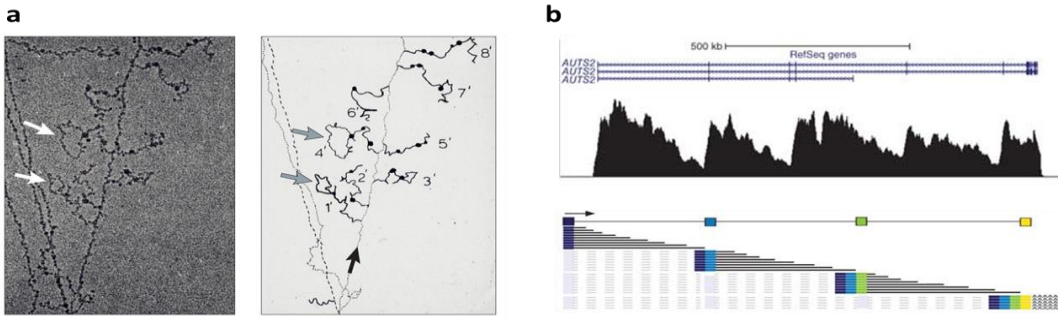


Figure 1.6: (**a**) Electron micrograph of a Drosophila melanogaster embryonic gene showing co-transcriptional splicing (left) and its interpretation (right)[48,60]**.** The dark blobs show transcription complexes. Grey and white arrows show introns that are spliced out co-transcriptionally. The black arrow indicates the direction of transcription along the DNA template. (**b**) Pattern for *AUTS2* (top) viewed in the UCSC Genome Browser. Model for co-transcriptional splicing (bottom).

1.2.6 "The recruitment model" of co-transcriptional splicing

The recruitment coupling model involves spliceosome components and splicing factor recruitment to splicing sites by the transcription machinery (Figure 1.7)[59]. The CTD is the unique feature of RNAPII that acts as a 'landing pad' for co-transcriptional recruitment of the capping, splicing and 3' end processing factors on nascent RNA[60]. In yeast, Prp40, a U1 snRNP component, was shown to bind to the phosphorylated CTD and in human PSF (PTB-associated splicing factor) and p54/NRB are splicing factors that physically bind to CTD[14]. Additionally, it has been shown that physical interaction of U2AF65–Prp19 complex with CTD promotes splicing activation[14]. A mutant RNAPII (*Ser2A*) where Ser2 phosphorylation was inhibited, reduced recruitment of U2AF65 and U2 snRNP, and as a consequence negatively affected co-transcriptional splicing[61]. This suggests that CTD Ser2p is critical for recruitment of U2AF65 to ensure efficient co-transcriptional splicing. Strong support for the recruitment model was obtained by a study in human cell lines showing CTD dependent inhibitory action of serine/arginine-rich (SR) protein SRSF3 (SRp20) in inclusion of fibronectin cassette exon 33 (E33)[59]. Some reports suggest that co-transcriptional splicing factor recruitment is not exclusively through CTD of RNAPII and splicing factors can also be recruited on nascent RNA via interaction with other transcription factors. An interesting example is PGC-1, a transcriptional coactivator that has certain motifs characteristic of splicing factors, which binds to the promoter and enhances exon 25 (E25) inclusion in fibronectin mRNA in a CTD-independent manner[62]. Transcription mediator complex may also help to recruit splicing factors at the promoter, or interact with nascent RNA splicing factors via its MED23 subunit. This subunit partially colocalizes with hnRNP L which is an alternative splicing regulator and some U1/U2 snRNPs associated factors[55]. It also has been shown that

MED23 can regulate most of the alternative splicing events that are modulated with hnRNP L[55].

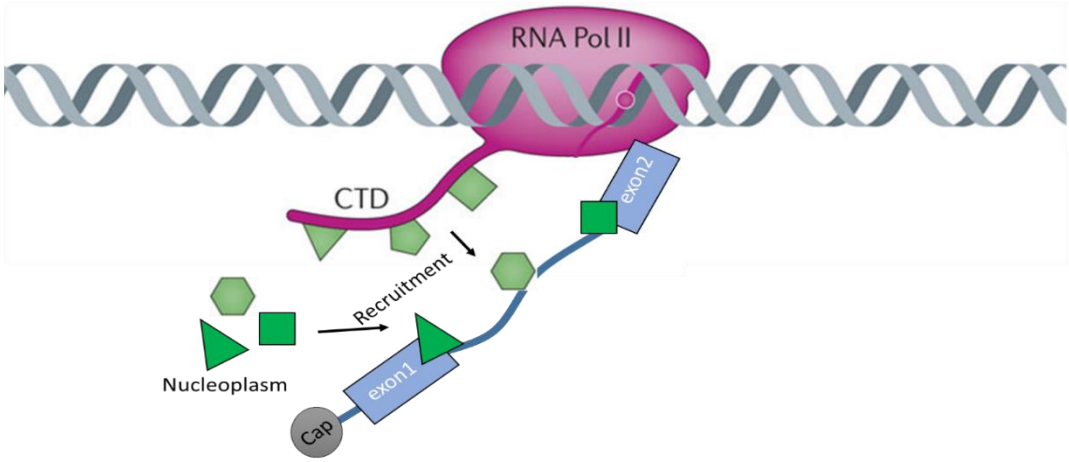


Figure 1.7: Schematic representation of the recruitment coupling model, adapted from[63]. CTD of RNAPII helps recruitment of splicing factors on pre-mRNA. Green shapes show different splicing factors.

1.2.7 "The kinetic model" of co-transcriptional splicing

This model proposes that RNAPII elongation rate can affect splicing outcome (Figure 1.8). The first evidence for the kinetic model was proposed 29 years ago by Eperon et al, where they suggested that transcription elongation rate can influence splicing through affecting nascent RNA structure[64]. Today there is a plethora of evidence showing that transcription elongation rate can affect splicing outcome both in yeast and metazoans[29,55,59,65]. For example, binding of DNA-binding protein CTCF to the downstream intron of exon 5 (E5) in *CD45* enhances E5 inclusion by acting as a road block for RNAPII[66]. DNA methylation of the intron inhibits CTCF binding and reverses the effect on E5. More direct evidences for the kinetic model were observed by using RNAPII elongation mutants or by using drugs that affect transcription elongation rate. The first evidence of kinetic coupling in budding yeast is illustrated by Howe *et al* with enhanced second exon inclusion of *DYN2* transcripts in a slow mutant RNAPII strain or when cells were treated with transcription inhibitors[67]. Similarly, inclusion of the fibronectin EDI exon was promoted in human cultured

cells that were expressing a slow RNAPII[68]. These observations have provided direct evidence for the effect of tuning transcription elongation rate on alternative splicing. More recent studies have shown that a slower elongation could also increase exon exclusion rate. For example, Dujardin *et al* reported that slower elongation causes skipping of the exon 9 in a CFTR minigene due to effective recruitment of ETR-3, a negative splicing factor to the 3'ss of the exon[69]. Likewise, RNA sequencing of human cell lines with fast and slow RNAPII also showed changes in inclusion and exclusion rates of thousands of exons[70].

Kinetic coupling also was observed globally, showing RNAPII pausing within the terminal exons of intron-containing transcripts[51]. It is suggested that RNAPII pausing over terminal exons can provide sufficient time for splicing to occur before transcription termination[51]. Similarly, high resolution kinetic assays with a reporter gene has revealed splicing-dependent RNAPII pausing in yeast[71]. This suggests that a splicing-dependent transcriptional checkpoint might exist to promote co-transcriptional splicing of the upstream intron. Consistent with these observations, it was shown that sequences required for the effect of TCERG1, a factor that associates with the Bcl-x pre-mRNA alternative splicing in drosophila, coincide with a putative polymerase pause site[72]. Recently, NET-seq has also revealed RNAPII pausing at intron-exon junctions and reported increased RNAPII density over exons that coincide with U1 snRNP occupancy[73]. These observations were explained as a model in which U1 snRNP is rapidly recruited to RNAPII pause sites through physical interaction with Ser5P CTD to promote co-transcriptional splicing. Single-molecule intron tracking (SMIT), which measures progression of splicing relative to RNAPII position along the genes genome-wide showed that the fast RNAPII transcribed significantly further than normal RNAPII when splicing occurred[54,74]. SMIT results further supported that splicing is in kinetic competition with transcription.

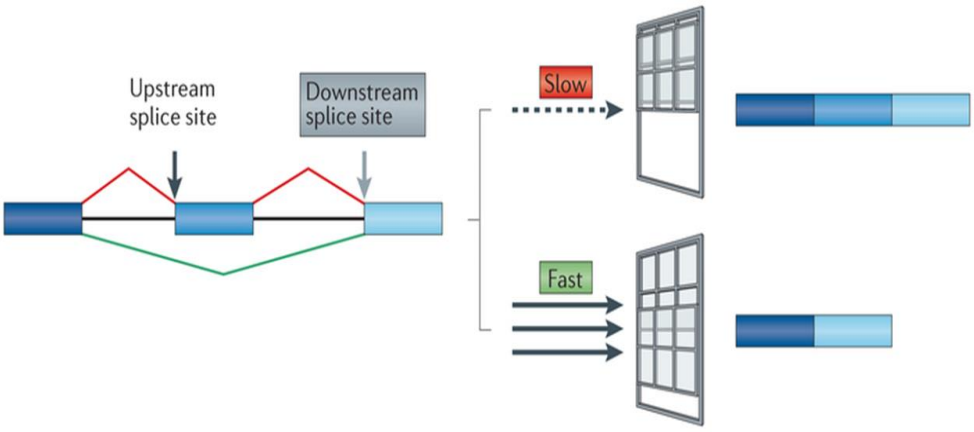


Figure 1.8: Schematic representation of the kinetic coupling model[60]. Slow transcription opens the window of opportunity for upstream events and enhances exon inclusion. Fast elongation enhances exon exclusion by allowing downstream splice sites to compete with upstream splice sites.

Chromatin and DNA sequence features are also involved in modulating transcription elongation rate over intron-containing transcripts. The analysis of genome-wide datasets has shown that RNAPII elongates faster in low GC and low-complexity DNA sequences[75–77]. Exons are GC reach and have higher nucleosome occupancy; features that are associated with slow elongation rate[77]. Additionally, some specific post-translational histone methylation signatures that are associated with reduced RNAPII elongation rate are enriched over exons. These features might be involved in reducing transcription elongation rate to facilitate splice site recognition and co-transcriptional splicing.

Overall, kinetic coupling suggests that transcription determines the temporal window of opportunity for splicing of the upstream intron or exon and selecting or rejecting the upstream splice site before competing with a downstream event. Based on the kinetic model of coupling, slow transcription expands the window of opportunity for co-transcriptional splicing and faster transcription shortens this window and reduces co-transcriptional splicing of an upstream event (Figure 1.8). Potentially, competing co-transcriptional splicing events that could be affected due to shortening or stretching of the window of opportunity by elongation rate could be

recognition of splice sites, RNA binding sites, RNA folding, the time required for spliceosome assembly and effective recruitment of splicing regulatory factors. Coordination of these events with transcription elongation rate can determine the outcome of splicing. Genome-wide studies showed that not all splicing events are kinetically coupled to transcription elongation rate and there might be specific features or conditions that make a particular splicing event sensitive to transcription elongation[78].

1.2.8 Splicing affects transcription

Coupling of transcription and splicing is not unidirectional from transcription to splicing, in fact this interaction is reciprocal and splicing activity can also regulate transcription. Studies with transgenic mice has shown that presence of an intron in a gene body promotes gene expression and transcription efficiency[79–81]. Comparing intron-containing and intronless genes also has revealed that presence of a promoter proximal splice site enhances transcription[81]. These observations suggest splicing-dependent transcription regulation. It has also shown that some splicing factors directly interact with the transcription complex and regulate its function. For example, it was shown that the snRNPs interact with TAT-SF1, and stimulate RNAPII elongation rate on an intronless gene though association with P-TEFb, a transcription elongation factor. TAT-SF1–U snRNPs interaction also stimulates *in vitro* splicing and therefore this interaction is suggested as a potential coupling factor that facilitates reciprocal activation of transcription and splicing[82]. U1 snRNA associates with TFIIH, a general transcription factor, and regulates transcription in addition to its role in splicing[83]. SC35, a splicing factor, attenuates transcriptional elongation in a gene-specific manner, modulating P-TEFb recruitment and CTD Ser2 phosphorylation[84]. In yeast, it was shown that mutating splice sites on a reporter gene abolishes transcriptional pausing at the 3' splice site[71]. Further investigations with *prp5-1* mutant, which blocks pre-spliceosome formation, has shown that RNAPII with high Ser5p CTD accumulates over the intron[85]. This extended pause was shown to be dependent on Cus2, a yeast homolog of human TAT-SF1. Mammalian NET-seq has also revealed that treatment

of cells with a splicing inhibitor pladienolide B (Pla-B), which interferes with U2 snRNP function, reduces RNAPII pausing over downstream exons[86].

Although these studies explain the mutual cross talks between splicing and transcription, still more investigation needs to be performed to determine how these links are established and regulated, which are the key coupling factors and what is the advantage of coupling for the cells.

1.2.9 Splicing proofreading and fidelity

Splicing fidelity is defined as the mechanism by which the spliceosome distinguishes optimal vs suboptimal splice sites. DExD/H-box ATPases play pivotal roles in proofreading steps and fidelity mechanisms. Two non-mutually exclusive models are proposed for splicing fidelity mechanisms[87]. In both models, DExD/H-box ATPases function as timers and/or sensors of authenticity. In the timer model, the speed of a substrate as it proceeds through the proofreading check points relative to rate of ATPase activity of proofreading factors will either promote or antagonize the substrate usage. Generally, if the substrate proceeds slower than ATPase activity, it will be rejected. For example, Prp5, Prp16 and Prp22 antagonize substrates that are slow probably due to having suboptimal introns that potentially slow down splicing[88–90]. In the sensor model, DExD/H-box ATPases reject a suboptimal substrate faster than an optimal substrate. In this model, ATPase activity either is regulated to specifically antagonize suboptimal substrate or ATPase activity will destabilize the weak substrate-spliceosome interactions. It has been shown that some of the ATPases actually destabilize less stable substrates and they also negatively or positively are regulated to proofread the splicing substrate[87].

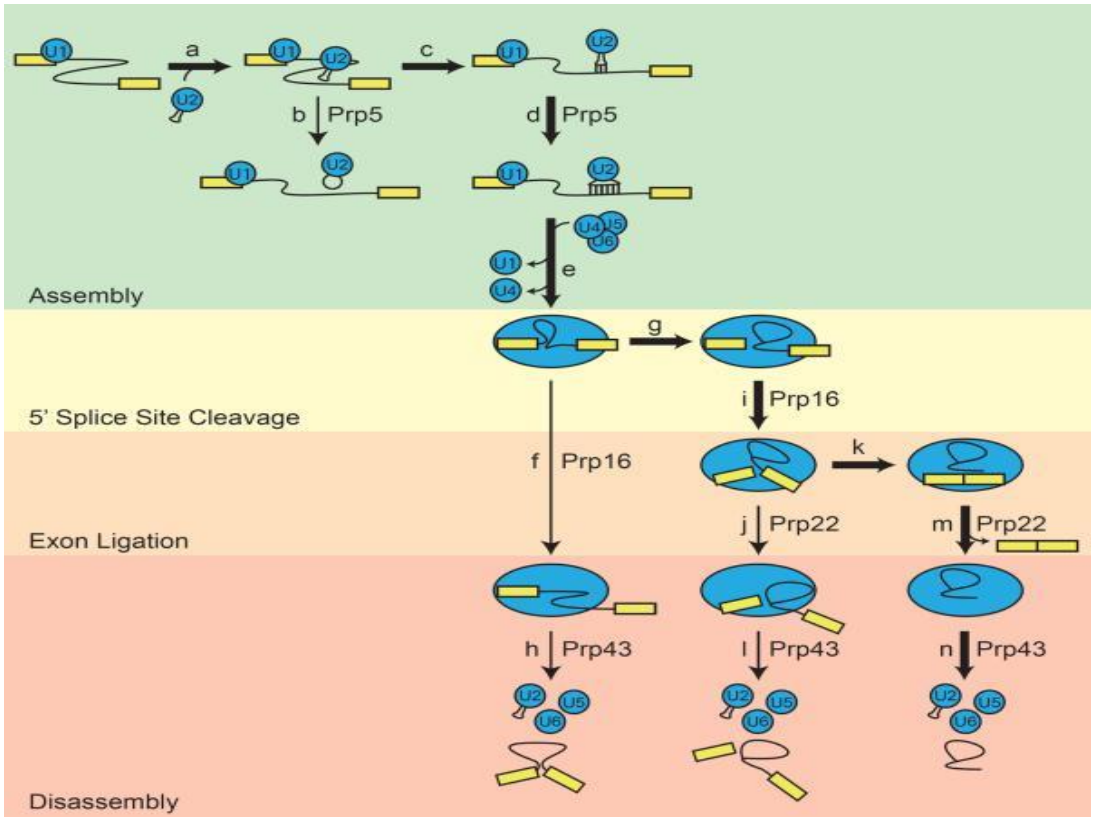


Figure 1.9: Function of spliceosomal DExD/H-box ATPases in proofreading substrates during different stages of splicing cycle[87]. Small blue circles represent snRNPs, large blue oval represents catalytically active spliceosome.

As splicing assembly occurs in stepwise fashion, there are proofreading factors that quality control splicing sequentially during the splicing process (Figure 1.9). Following U2 snRNP recruitment to pre-mRNA, ATP hydrolyses by Prp5 rejects splicing of introns with slow U2 snRNP – suboptimal branch point interactions and stabilizes optimal U2 snRNP –branch point interactions[88,91,92]. After formation of a catalytically active spliceosome (complex B*), Prp16 activity rejects substrates with suboptimal branch sites by antagonizing 5′ splice site cleavage leading to accumulation of intermediate complexes that are detected by Prp43 and eventually discarded[93,93,94]. Optimal branch point sequences lead to efficient cleavage of 5′ splice site followed by Prp16-mediated conformational changes that facilitate spliceosome to proceed to exon ligation step[95]. Prior to exon ligation, Prp22 activity rejects suboptimal slow substrates that are detected by Prp43 and eventually discarded[90,96]. Prp22 promotes the release of

the mRNA from the spliceosome after successful exon ligation of optimal substrates[97,98]. Eventually, the spliceosome is disassembled by Prp43 into its components and excised intron is released[99,100].

1.2.10 RNA degradation

In all biological systems RNA turnover plays three important roles[101]. First, differential degradation of RNA determines the half-life of a given RNA in response to cellular demands and physiological stresses. Secondly, errors can occur during all RNA maturation steps including transcription, splicing or 3′-end processing but they are largely detected by RNA surveillance systems and degraded. Thirdly, RNA degradation removes the RNA processing intermediates and by-products including excised introns or the RNA rejected from processing complexes. There are three different classes of RNA degradation enzymes: endonucleases, 5'-3' exonucleases and 3'-5' exonucleases. Xrn1 and Rat1 (Xrn2) are cytoplasmic and nuclear 5'-3' exonucleases, respectively[101]. Exosome, which is a conserved cytoplasmic and nuclear multi-protein complex, has both 3′–5′ exoribonuclease and endonuclease activity carried out by its Dis3/Rrp44 subunit. Nuclear exosome has an additional 3'-5' exonuclease called Rrp6, which is not absolutely essential for yeast cell viability[102]. The exosome requires additional co-factors such as the TRAMP complex that marks faulty mRNAs[101]. It was shown that Dis3 plays a prominent role in degradation of intron-containing transcripts[103]. Expressing a mutant Dis3 enzyme in yeast cells (*dis3 exo*$^{-}$) increased accumulation of the intron-containing transcripts in splicing proficient cells[103]. This suggests that RNA degradation competes with splicing in processing the pre-mRNAs. What is the functional significance of this competition is not clear. Recently, it has been shown that Gbp2 and Hrb1 are important pre-mRNA splicing surveillance factors that are recruited co-transcriptionally to pre-mRNA and mark faulty pre-mRNAs for degradation[104,105]. These proteins recruit nuclear export receptor Mex67 to the correct mRNAs upon completion of splicing to allow a quality controlled nuclear export. How these factors recognise whether a pre-mRNA is faulty or authentic is unknown.

Nonsense-mediated mRNA decay (NMD) is an additional surveillance mechanism coupled with translation that eliminates mRNAs that contain premature translation-termination codons (PTCs)[106]. Unspliced pre-mRNAs that are leaked to the cytoplasm can also be targeted by NMD because these pre-mRNAs more likely contain PTCs[107]. Expression of some of normal transcripts that do not have PTCs were also shown to be regulated by NMD factors[108,109]. Aberrant alternative splicing can potentially introduce PTCs in mRNA that are detected and removed by NMD; a mechanism that is often referred to as alternative splicing coupled to NMD[110,111]. Genome-wide studies in budding yeast as well as in other multicellular organisms with inactivated NMD pathway showed accumulation of the non-productive alternatively spliced transcripts that contain PTCs[111–113]. Collectively, these studies indicated that PTCs containing unspliced pre-mRNA or aberrantly spliced transcripts that are exported to the cytoplasm are generally targeted and degraded by NMD in eukaryotes.

Three core components of the NMD pathway are the proteins encoded by the *UPF1, UPF2* and *UPF3* genes. These proteins are evolutionarily conserved across eukaryotes and knocking down these factors inactivates the NMD pathway leading to accumulate mRNAs containing PTCs and truncated proteins[114]. Upf1 is an RNA helicase that has ATPase activity and is the central regulator of the NMD pathway[115]. *UPF1* is not essential for the viability of yeast or *C. elegans*; however, it is essential in *Drosophila,* zebrafish, mouse and human cells[116]. Some studies in human cells have reported that Upf1 and other NMD components are present in the nucleus and play additional roles possibly unrelated to NMD, in telomere maintenance, cell cycle progression and DNA replication[116]. To date, there is no evidence of Upf1 function in the nucleus in budding yeast.

1.2.11 Yeast, as a model organism for co-transcriptional splicing

The budding yeast *S. cerevisiae* has a small genome, short genes and also a simple constitutive splicing compared to other higher eukaryotes. Although only ~5% of

genes encode transcripts that undergo splicing, the splicing mechanism and most of the core splicing factors are largely conserved with other eukaryotes. Similar to splicing and other essential cellular processes, the transcription process is also conserved and therefore makes yeast a promising model organism for studying co-transcriptional splicing. Additionally, yeast cells are inexpensive to work with, easy to grow in the laboratory and amenable to genetic manipulations.

1.3 Aims of this study

The aims of this project were to study the consequences of altering transcription elongation rate on splicing efficiency and fidelity:

- Testing the effect of fast and slow transcription elongation rates on splicing efficiency of cellular steady state and nascent RNA
- Testing co-transcriptional splicing by immunoprecipitation of RNAPII and assessing association of the spliced mRNA and excised lariat with elongating transcription complex
- Investigating the effect of fast and slow transcription elongation rates on exon skipping and transcription start site selection in yeast
- Investigating the effect of elongation rate on splicing fidelity with RNA sequencing

References

1. Crick, F. Central Dogma of Molecular Biology. *Nature* **227,** 561–563 (1970).
2. Arimbasseri, A. G., Rijal, K. & Maraia, R. J. Comparative overview of RNA polymerase II and III transcription cycles, with focus on RNA polymerase III termination and reinitiation. *Transcription* **5,** (2013).
3. Vannini, A. & Cramer, P. Conservation between the RNA Polymerase I, II, and III Transcription Initiation Machineries. *Mol. Cell* **45,** 439–446 (2012).
4. Transcript Elongation by RNA Polymerase II. *Annu. Rev. Biochem.* **79,** 271–293 (2010).
5. Woychik, N. A. & Young, R. A. RNA polymerase II subunit RPB4 is essential for high- and low-temperature yeast cell growth. *Mol. Cell. Biol.* **9,** 2854–2859 (1989).
6. Hemming, S. A. *et al.* RNA Polymerase II Subunit Rpb9 Regulates Transcription Elongation in Vivo. *J. Biol. Chem.* **275,** 35506–35511 (2000).
7. Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat. Rev. Microbiol.* **9,** 85–98 (2011).
8. Corden, J. L., Cadena, D. L., Ahearn, J. M. & Dahmus, M. E. A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. *Proc. Natl. Acad. Sci. U. S. A.* **82,** 7934–7938 (1985).
9. Egloff, S. & Murphy, S. Cracking the RNA polymerase II CTD code. *Trends Genet.* **24,** 280–288 (2008).
10. Egloff, S., Dienstbier, M. & Murphy, S. Updating the RNA polymerase CTD code: adding gene-specific layers. *Trends Genet.* **28,** 333–341 (2012).
11. Harlen, K. M. & Churchman, L. S. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat. Rev. Mol. Cell Biol.* **advance online publication,** (2017).
12. Tietjen, J. R. *et al.* Chemical-genomic dissection of the CTD code. *Nat. Struct. Mol. Biol.* **17,** 1154–1161 (2010).
13. Kim, H. *et al.* Gene-specific RNA pol II phosphorylation and the 'CTD code'. *Nat. Struct. Mol. Biol.* **17,** 1279–1286 (2010).
14. David, C. J., Boyne, A. R., Millhouse, S. R. & Manley, J. L. The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65–Prp19 complex. *Genes Dev.* **25,** 972–983 (2011).
15. Hsin, J.-P. & Manley, J. L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.* **26,** 2119–2137 (2012).

16. Egloff, S., Zaborowska, J., Laitem, C., Kiss, T. & Murphy, S. Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes. *Mol. Cell* **45,** 111–122 (2012).
17. Clemente-Blanco, A. *et al.* Cdc14 phosphatase promotes segregation of telomeres through repression of RNA polymerase II transcription. *Nat. Cell Biol.* **13,** 1450–1456 (2011).
18. Egloff, S. Role of Ser7 phosphorylation of the CTD during transcription of snRNA genes. *RNA Biol.* **9,** 1033–1038 (2012).
19. Kaplan, C. D., Larsson, K.-M. & Kornberg, R. D. The RNA Polymerase II Trigger Loop Functions in Substrate Selection and Is Directly Targeted by α-Amanitin. *Mol. Cell* **30,** 547–556 (2008).
20. Kaplan, C. D., Jin, H., Zhang, I. L. & Belyanin, A.

Dissection of Pol II Trigger Loop Function and Pol II Activity–Dependent Control of Start Site Selection In Vivo. *PLoS Genet* **8,** e1002627 (2012).
21. Toulokhonov, I., Zhang, J., Palangat, M. & Landick, R. A Central Role of the RNA Polymerase Trigger Loop in Active-Site Rearrangement during Transcriptional Pausing. *Mol. Cell* **27,** 406–419 (2007).
22. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16,** 129–143 (2015).
23. Shandilya, J. & Roberts, S. G. E. The transcription cycle in eukaryotes: From productive initiation to RNA polymerase II recycling. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1819,** 391–400 (2012).
24. Glover-Cutter, K. *et al.* TFIIH-Associated Cdk7 Kinase Functions in Phosphorylation of C-Terminal Domain Ser7 Residues, Promoter-Proximal Pausing, and Termination by RNA Polymerase II. *Mol. Cell. Biol.* **29,** 5455–5464 (2009).
25. Yamada, T. *et al.* P-TEFb-Mediated Phosphorylation of hSpt5 C-Terminal Repeats Is Critical for Processive Transcription Elongation. *Mol. Cell* **21,** 227–237 (2006).
26. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13,** 720–731 (2012).
27. Wu, C.-H. *et al.* Molecular characterization of Drosophila NELF. *Nucleic Acids Res.* **33,** 1269–1279 (2005).
28. Nudler, E. RNA Polymerase Backtracking in Gene Regulation and Genome Instability. *Cell* **149,** (2012).

29. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16,** 167–177 (2015).
30. Pearson, E. & Moore, C. The evolutionarily conserved Pol II flap loop contributes to proper transcription termination on short yeast genes. *Cell Rep.* **9,** 821–828 (2014).
31. Richard, P. & Manley, J. L. Transcription termination by nuclear RNA polymerases. *Genes Dev.* **23,** 1247–1269 (2009).
32. Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat. Rev. Mol. Cell Biol.* **16,** 190–202 (2015).
33. Minvielle-Sebastia, L., Preker, P. J., Wiederkehr, T., Strahm, Y. & Keller, W. The major yeast poly(A)-binding protein is associated with cleavage factor IA and functions in premessenger RNA 3′-end formation. *Proc. Natl. Acad. Sci.* **94,** 7897–7902 (1997).
34. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18,** 18–30 (2017).
35. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 3171–3175 (1977).
36. Fabrizio, P. *et al.* The Evolutionarily Conserved Core Design of the Catalytic Activation Step of the Yeast Spliceosome. *Mol. Cell* **36,** 593–608 (2009).
37. Lee, Y. & Rio, D. C. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.* **84,** 291–323 (2015).
38. Horowitz, D. S. The mechanism of the second step of pre-mRNA splicing. *Wiley Interdiscip. Rev. RNA* **3,** 331–350 (2012).
39. Patel, A. A. & Steitz, J. A. Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.* **4,** 960–970 (2003).
40. Frendewey, D. & Keller, W. Stepwise assembly of a pre-mRNA splicing complex requires U-snRNPs and specific intron sequences. *Cell* **42,** 355–367 (1985).
41. Moore, M. J., Schwartzfarb, E. M., Silver, P. A. & Yu, M. C. Differential recruitment of the splicing machinery during transcription predicts genome-wide patterns of mRNA splicing. *Mol. Cell* **24,** 903–915 (2006).
42. Matera, A. G. & Wang, Z. A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **15,** 108–121 (2014).
43. Sterner, D. A., Carlo, T. & Berget, S. M. Architectural limits on split genes. *Proc. Natl. Acad. Sci.* **93,** 15081–15085 (1996).
44. Sun, J. S. & Manley, J. L. A novel U2-U6 snRNA structure is necessary for mammalian mRNA splicing. *Genes Dev.* **9,** 843–854 (1995).

45. Hoskins, A. A. *et al.* Ordered and Dynamic Assembly of Single Spliceosomes. *Science* **331,** 1289–1295 (2011).
46. Ezkurdia, I. *et al.* Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.* **23,** 5866–5878 (2014).
47. Hu, Z. *et al.* Revealing Missing Human Protein Isoforms Based on Ab Initio Prediction, RNA-seq and Proteomics. *Sci. Rep.* **5,** 10940 (2015).
48. Beyer, A. L. & Osheim, Y. N. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev.* **2,** 754–765 (1988).
49. Listerman, I., Sapra, A. K. & Neugebauer, K. M. Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. *Nat. Struct. Mol. Biol.* **13,** 815–822 (2006).
50. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18,** 1435–1440 (2011).
51. Oesterreich, F. C., Preibisch, S. & Neugebauer, K. M. Global Analysis of Nascent RNA Reveals Transcriptional Pausing in Terminal Exons. *Mol. Cell* **40,** 571–581 (2010).
52. Khodor, Y. L. *et al.* Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes Dev.* **25,** 2502–2512 (2011).
53. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22,** 1616–1625 (2012).
54. Carrillo Oesterreich, F. *et al.* Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* **165,** 372–381 (2016).
55. Naftelberg, S., Schor, I. E., Ast, G. & Kornblihtt, A. R. Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. *Annu. Rev. Biochem.* **84,** 165–198 (2015).
56. Li, X. & Manley, J. L. Cotranscriptional processes and their influence on genome stability. *Genes Dev.* **20,** 1838–1847 (2006).
57. Skourti-Stathaki, K. & Proudfoot, N. J. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes Dev.* **28,** 1384–1396 (2014).
58. Aitken, S., Alexander, R. D. & Beggs, J. D. Modelling Reveals Kinetic Advantages of Co-Transcriptional Splicing. *PLOS Comput Biol* **7,** e1002215 (2011).
59. Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* **14,** 153–165 (2013).

60. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* **15,** 163–175 (2014).
61. Gu, B., Eick, D. & Bensaude, O. CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo. *Nucleic Acids Res.* gks1327 (2012). doi:10.1093/nar/gks1327
62. Monsalve, M. *et al.* Direct Coupling of Transcription and mRNA Processing through the Thermogenic Coactivator PGC-1. *Mol. Cell* **6,** 307–316 (2000).
63. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17,** 19–32 (2016).
64. Eperon, L. P., Graham, I. R., Griffiths, A. D. & Eperon, I. C. Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? *Cell* **54,** 393–401 (1988).
65. Braberg, H. *et al.* From Structure to Systems: High-Resolution, Quantitative Genetic Analysis of RNA Polymerase II. *Cell* **154,** 775–788 (2013).
66. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479,** 74–79 (2011).
67. HOWE, K. J., KANE, C. M. & ARES, M. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. *RNA* **9,** 993–1006 (2003).
68. de la Mata, M. *et al.* A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. *Mol. Cell* **12,** 525–532 (2003).
69. Dujardin, G. *et al.* How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping. *Mol. Cell* **54,** 683–690 (2014).
70. Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28,** 2663–2676 (2014).
71. Alexander, R. D., Innocente, S. A., Barrass, J. D. & Beggs, J. D. Splicing-Dependent RNA Polymerase Pausing in Yeast. *Mol. Cell* **40,** 582–593 (2010).
72. Montes, M. *et al.* TCERG1 Regulates Alternative Splicing of the Bcl-x Gene by Modulating the Rate of RNA Polymerase II Transcription. *Mol. Cell. Biol.* **32,** 751–762 (2012).
73. Harlen, K. M. *et al.* Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. *Cell Rep.* **15,** 2147–2158 (2016).
74. Zlotorynski, E. RNA metabolism: Co-transcriptional splicing at nucleotide resolution. *Nat. Rev. Mol. Cell Biol.* **17,** 264–265 (2016).

75. Veloso, A. *et al.* Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* **24,** 896–905 (2014).
76. Danko, C. G. *et al.* Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. *Mol. Cell* **50,** 212–222 (2013).
77. Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife* **3,** e02407 (2014).
78. Ip, J. Y. *et al.* Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Res.* **21,** 390–401 (2011).
79. Brinster, R. L., Allen, J. M., Behringer, R. R., Gelinas, R. E. & Palmiter, R. D. Introns increase transcriptional efficiency in transgenic mice. *Proc. Natl. Acad. Sci. U. S. A.* **85,** 836–840 (1988).
80. Choi, T., Huang, M., Gorman, C. & Jaenisch, R. A generic intron increases gene expression in transgenic mice. *Mol. Cell. Biol.* **11,** 3070–3074 (1991).
81. Furger, A., Binnie, J. M. O., Alexandra, Lee, B. A. & Proudfoot, N. J. Promoter proximal splice sites enhance transcription. *Genes Dev.* **16,** 2792–2799 (2002).
82. Fong, Y. W. & Zhou, Q. Stimulatory effect of splicing factors on transcriptional elongation. *Nature* **414,** 929–933 (2001).
83. Kwek, K. Y. *et al.* U1 snRNA associates with TFIIH and regulates transcriptional initiation. *Nat. Struct. Mol. Biol.* **9,** 800–805 (2002).
84. Lin, S., Coutinho-Mansfield, G., Wang, D., Pandit, S. & Fu, X.-D. The splicing factor SC35 has an active role in transcriptional elongation. *Nat. Struct. Mol. Biol.* **15,** 819–826 (2008).
85. Chathoth, K. T., Barrass, J. D., Webb, S. & Beggs, J. D. A Splicing-Dependent Transcriptional Checkpoint Associated with Prespliceosome Formation. *Mol. Cell* **53,** 779–790 (2014).
86. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161,** 526–540 (2015).
87. Semlow, D. R. & Staley, J. P. Staying on message: ensuring fidelity in pre-mRNA splicing. *Trends Biochem. Sci.* **37,** 263–273 (2012).
88. Xu, Y.-Z. & Query, C. C. Competition between the ATPase Prp5 and Branch Region-U2 snRNA Pairing Modulates the Fidelity of Spliceosome Assembly. *Mol. Cell* **28,** 838–849 (2007).

89. Burgess, S. M. & Guthrie, C. A mechanism to enhance mRNA splicing fidelity: The RNA-dependent ATPase Prp16 governs usage of a discard pathway for aberrant lariat intermediates. *Cell* **73,** 1377–1391 (1993).
90. Mayas, R. M., Maita, H. & Staley, J. P. Exon ligation is proofread by the DExD/H-box ATPase Prp22p. *Nat. Struct. Mol. Biol.* **13,** 482–490 (2006).
91. Perriman, R. & Ares Jr., M. Invariant U2 snRNA Nucleotides Form a Stem Loop to Recognize the Intron Early in Splicing. *Mol. Cell* **38,** 416–427 (2010).
92. Ruby, S. W., Chang, T. H. & Abelson, J. Four yeast spliceosomal proteins (PRP5, PRP9, PRP11, and PRP21) interact to promote U2 snRNP binding to pre-mRNA. *Genes Dev.* **7,** 1909–1925 (1993).
93. Villa, T. & Guthrie, C. The Isy1p component of the NineTeen Complex interacts with the ATPase Prp16p to regulate the fidelity of pre-mRNA splicing. *Genes Dev.* **19,** 1894–1904 (2005).
94. Tseng, C.-K., Liu, H.-L. & Cheng, S.-C. DEAH-box ATPase Prp16 has dual roles in remodeling of the spliceosome in catalytic steps. *RNA* **17,** 145–154 (2011).
95. Schwer, B. & Guthrie, C. A conformational rearrangement in the spliceosome is dependent on PRP16 and ATP hydrolysis. *EMBO J.* **11,** 5033–5039 (1992).
96. Mayas, R. M., Maita, H., Semlow, D. R. & Staley, J. P. Spliceosome discards intermediates via the DEAH box ATPase Prp43p. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 10020–10025 (2010).
97. Schwer, B. & Gross, C. H. Prp22, a DExH-box RNA helicase, plays two distinct roles in yeast pre-mRNA splicing. *EMBO J.* **17,** 2086–2094 (1998).
98. Wagner, J. D., Jankowsky, E., Company, M., Pyle, A. M. & Abelson, J. N. The DEAH-box protein PRP22 is an ATPase that mediates ATP-dependent mRNA release from the spliceosome and unwinds RNA duplexes. *EMBO J.* **17,** 2926–2937 (1998).
99. Arenas, J. E. & Abelson, J. N. Prp43: An RNA helicase-like factor involved in spliceosome disassembly. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 11798–11802 (1997).
100. Prp43 Is an Essential RNA-dependent ATPase Required for Release of Lariat-Intron from the Spliceosome. Available at: http://www.jbc.org/content/277/20/17743.long. (Accessed: 2nd March 2017)
101. Doma, M. K. & Parker, R. RNA Quality Control in Eukaryotes. *Cell* **131,** 660–668 (2007).
102. Januszyk, K., Liu, Q. & Lima, C. D. Activities of human RRP6 and structure of the human RRP6 catalytic domain. *RNA N. Y. N* **17,** 1566–1577 (2011).

103. Gudipati, R. K. *et al.* Extensive Degradation of RNA Precursors by the Exosome in Wild-Type Cells. *Mol. Cell* **48,** 409–421 (2012).

104. Hackmann, A. *et al.* Quality control of spliced mRNAs requires the shuttling SR proteins Gbp2 and Hrb1. *Nat. Commun.* **5,** 3123 (2014).

105. Zander, G. *et al.* mRNA quality control is bypassed for immediate export of stress-responsive transcripts. *Nature* **540,** 593–596 (2016).

106. Brogna, S. & Wen, J. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol.* **16,** 107–113 (2009).

107. Hilleren, P. J. & Parker, R. Cytoplasmic Degradation of Splice-Defective Pre-mRNAs and Intermediates. *Mol. Cell* **12,** 1453–1465 (2003).

108. He, F. & Jacobson, A. Upf1p, Nmd2p, and Upf3p Regulate the Decapping and Exonucleolytic Degradation of both Nonsense-Containing mRNAs and Wild-Type mRNAs. *Mol. Cell. Biol.* **21,** 1515–1530 (2001).

109. He, F. *et al.* Genome-Wide Analysis of mRNAs Regulated by the Nonsense-Mediated and 5′ to 3′ mRNA Decay Pathways in Yeast. *Mol. Cell* **12,** 1439–1452 (2003).

110. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci.* **100,** 189–192 (2003).

111. Kawashima, T., Douglass, S., Gabunilas, J., Pellegrini, M. & Chanfreau, G. F. Widespread Use of Non-productive Alternative Splice Sites in Saccharomyces cerevisiae. *PLoS Genet* **10,** e1004249 (2014).

112. Ramani, A. K. *et al.* High resolution transcriptome maps for wild-type and nonsense-mediated decay-defective Caenorhabditis elegans. *Genome Biol.* **10,** R101 (2009).

113. Hansen, K. D. *et al.* Genome-Wide Identification of Alternative Splice Forms Down-Regulated by Nonsense-Mediated mRNA Decay in Drosophila. *PLoS Genet.* **5,** (2009).

114. Conti, E. & Izaurralde, E. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Curr. Opin. Cell Biol.* **17,** 316–325 (2005).

115. Peccarelli, M. & Kebaara, B. W. Regulation of Natural mRNAs by the Nonsense-Mediated mRNA Decay Pathway. *Eukaryot. Cell* **13,** 1126–1135 (2014).

116. Varsally, W. & Brogna, S. UPF1 involvement in nuclear functions. *Biochem. Soc. Trans.* **40,** 778–783 (2012).

Chapter 2 Materials and Methods

2.1 Sources of enzymes and reagents

Unless stated otherwise enzymes used in this study were purchased from New England BioLabs, Roche, Qiagen, PeqLab, Promega, Invitrogen and Sigma-Aldrich. Common reagents were mainly purchased from Fisher Scientific, New England BioLabs, Formedium and Invitrogen.

2.2 Growth Media and Common Buffers

Bacterial and yeast media were supplied by the university central facility. Self-prepared liquid media were autoclaved and cooled to room temperature before use. For solid media, 2% (W/V) agar was added to liquid media before autoclaving. When required, antibiotics were added to the media after autoclaving when media was relatively cool and stored at 4°C. All the buffers were autoclaved prior to use and stored at room temperature. Non-autoclavable buffers were filter sterilized using Nalgen Rapid-Flow 0.2 μm filter units. A list of all the media and reagents used is summarized in table 2.1.

Table 2.1

Yeast media		
YPDA	Yeast extract	1% (w/v)
	Bacto-peptone	2% (w/v)
	Glucose	2% (w/v)
	Adenine sulfate	0.003% (w/v)
YMM	Yeast nitrogen base without amino acids	0.67% (w/v)
	Glucose	2% (w/v)
Drop-out media	YMM	-
	Synthetic complete mixture (Kaiser) Drop-Out Formedium	According to manufacturer's instructions
Bacterial media		
LB	Bacto-tryptone	1% (w/v)
	Yeast extract	0.5% (w/v)
	NaCl	0.5% (w/v)
SOC	Bacto-tryptone	2% (w/v)
	Yeast extract	0.5% (w/v)

	NaCl	0.06% (w/v)
	KCl	0.02% (w/v)
	$MgCl_2$	0.1%(w/v)
	$MgSO_4$	0.12% (w/v)
	Glucose	0.4% (w/v)
Commonly used buffers		
50x TAE	Tris base	2 M
	Acetic acid	5.71 (v/v)
	EDTA	50 mM
10X TE	Tris-HCl	100 mM
	EDTA ph 8.0	10 mM
20X MOPS SDS-Page	MOPS	1M
	Tris-base	1M
	SDS	20% (w/v)
	EDTA	20 mM
A.E	NaAc pH5.3	50 mM
	EDTA	10 mM
Yeast transformation mix	40% PEG3350 (w/v)	240 µl
	1M LiAc	36 µl
	Heat denatured salmon sperm DNA	50 µg
Yeast genomic DNA extraction	dH2O	4.24 ml
	Triton-X-100	100 µl
	10% SDS	500 µl
	5M NaCl	100 µl
	1M Tris-HCl pH 8.0	50 µl
	0.5M EDTA pH 8.0	10 µl
	Total volume	5 ml
4tU labelled RNA extraction buffers		
10x NaTMg (DEPC treated)	Tris Cl pH7.0	10 mM
	NaCl	200 mM
	$MgCl_2$	25 mM
1x NaSTPMg (Stored up to 5 hrs)	NaTMg	1x
	NaPi pH6.8	100 mM
	SDS (This added last)	0.1%
	H_2O (This added first)	
Native elongating transcript extraction buffers		
Lysis buffer, 10× (Stored up to 1 year)	HEPES, pH 7.4	200 mM
	KOAc	1100 mM
	Triton X-100	5%
	Tween 20	1%
Lysis buffer, 1× (Stored up to 5 hrs on ice)	10× lysis buffer stock	1×
	$MnCl_2$	10 mM
	proteinase inhibitors (complete, EDTA-free; Roche)	1 tablet
	SUPERase.In	50 U/ml
Wash buffer (Stored up to 5 hrs on ice)	10× lysis buffer	1×
	SUPERase.In	50 U/ml
	EDTA	1 mM

TEV cleavage buffer	20X ProTEV Buffer	15 µl
	SUPERase.IN	3 µl
	0.1 M DTT	3 µl
	DPEC water	279 µl
Proteinase digestion mix	Proteinase K	0.2mg/µl
	SDS	0.5%
	TNE (Tris, Nacl, EDTA)	1x

2.3 Antibiotics

Common Name	**Company Name**	**Final concentration (µg/ml)**
Ampicillin (E.coli)	Life technologies	100
Hygromycin B	Life technologies	400
Yeast Kanamycin (G148)	Life technologies	400

2.4 Antibodies

Protein	**Beads**	**Antibodies**	**Company Name**
Rpb3 TAP tag	IgG Sepharose	Rabbit Anti-Tap-Tag	GE Healthcare

2.5 Yeast strains

Strain	**Genotype**	**Source**
JBY115	*MATa ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1Δ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3:: TAP::KlacTRP1 pCK859* ***RPB1*** *T69 corrected* CEN LEU2*	C. Kaplan *et al* 2012
JBY122	*MATa ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1hΔ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3:: TAP::KlacTRP1 pCK876* ***rpb1 G1097D*** *T69 corrected CEN LEU2*	C. Kaplan *et al* 2012
JBY123	*MATa ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1Δ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3:: TAP::KlacTRP1 pCK870* ***rpb1 H1085Y*** *T69 corrected CEN LEU2*	C. Kaplan *et al* 2012
JBY133	*MATa ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1Δ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3:: TAP::KlacTRP1* ***Δupf1::HPH*** *pCK859* ***RPB1*** *T69 corrected CEN LEU2*	This study
JBY146	*MATa ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1hΔ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3:: TAP::KlacTRP1* ***Δupf1::HPH pCK876 rpb1 G1097D*** *T69 corrected CEN LEU2*	This study
JBY147	*MATa ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1Δ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3:: TAP::KlacTRP1* ***Δupf1::HPH*** *pCK870* ***rpb1 H1085Y*** *T69 corrected CEN LEU2*	This study

JBY101	*MATa ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1Δ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3:: TAP::KlacTRP1 pCK859* ***RPB1*** *T69 corrected CEN LEU2,* ***pRS426–FUI1***	This study
JBY102	*MATa ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1hΔ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3:: TAP::KlacTRP1 pCK876* ***rpb1 G1097D*** *T69 corrected CEN LEU2,* ***pRS426–FUI1***	This study
JBY142	*MATa KY691, ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1Δ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3:: TAP::KlacTRP1 pCK870* ***rpb1 H1085Y*** *T69 corrected CEN LEU2,* ***pRS426–FUI1***	This study
JBY110	*MATa ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1Δ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3:: TAP::KlacTRP1 pCK960* ***rpb1 E1103G*** *T69 corrected CEN LEU2*	C. Kaplan *et al* 2012
JBY132	*MATa ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1Δ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3:: TAP::KlacTRP1* ***Δupf1::HPH*** *pCK960* ***rpb1 E1103G*** *T69 corrected CEN LEU2*	This study
* Kaplan et al[1] detected a polymorphism in *RPB1* plasmid encoding isoleucine at position 69 instead of threonine. This substitution was corrected (T69 corrected) before generating WT and mutant strains.		

2.6 Plasmids

Strain	Genotype	Source
pFA6a- *hphMX6*	Plasmid used for PCR amplification of Hygromycin	Beggs Lab
pRS416-*FUI1*	Yeast Uracil permease gene *FUI1* cloned into pRS416 (*LEU* selection)	Beggs Lab

2.7 Oligonucleotides

Oligonucleotides were purchased from Sigma-Aldrich.

Oligonucleotides used to construct and verify yeast strains	
UPF1-Delete-F	AGCAAGACCGAATATACTTTTTATATTACATCAATCATTGT CATTATCAACGGATCCCCGGGTTAATTAA
UPF1-Delete-R	TTTGTATCACAAGCCAAGTTTAACATTTTATTTTAACAGGG TTCACCGAAGAATTCGAGCTCGTTTAAAC
UPF1-check-F	GGCATCGTTTTAACGCACACT
UPF1-check-R	AGATGCACCTCCGGAAAAGT
RPB1-Sequ-P1	GTTTGGCCACACGTAGAGTTC
RPB1-Sequ-P2	ATCGAGGCACAATTCCTCCG
RPB1-Sequ-P3	CGCTTGTTCTTGATCGGCAG
RPB1-Sequ-P4	ACTGTGGAACGTGGATCAGG
Oligonucleotides used for RT-qPCR to determine splicing status	
ALG9_F	TAAGCTGGCATGTGCTGCATTC
ALG9_R	TTTGCATGATTCGGTTGATTGG

ALG9_1F	CAAGCTTGAAAAAGGGACAACTGG
ALG9_1R	TCTCGTTGAGTAGCCGTAATGGTACA
ALG9_2F	CGATGAATTGCAAGGCGGTAA
ALG9_2R	GAAATTAACGAGAATGTCGGCTGAA
ALG9_3F	CCGAGTATTCTATTAGATCATGGGCTTTC
ALG9_3R	GCATGCTCTTGTGATGAAAAAGTTCC
ALG9_4F	TGTTTAATCCGGGCTGGTTCC
ALG9_4R	CTGGTGGCACCTACATACAACAACA
ALG9_5F	TGAATTTCCCACTGCCTGTGC
ALG9_5R	TTGTTGAGTGAAAACGGCAATCC
ALG9_6F	AGCTGGCATGTGCTGCATTC
ALG9_6R	GGTTGATTGGCTCCGGTACG
FMP27_1F	TTCAATTGATCAAATTTATGGAAGATCCTCAAGAA
FMP27_1R	GGAACAAAAACTTGACAAATTTGAAACTCTGGAT
FMP27_2F	GAACATAAGAATCCTTAGAAAAGCCCTTTACCTCG
FMP27_2R	CCATAAGAAAGTCACTGCAAATATAAGCCACTTGT
FMP27_3F	AAGATTTGATTCCTTTTTGAGAAAACTTCTTTGGA
FMP27_3R	CCATCCTTCAGAGGATTCATAATTTCACCAATT
FMP27_4F	TTGAATCTAAATCGAAAACATCAAAGCCACG
FMP27_4R	AATTTTTGAGAGAACAAATTGGTTTCGCCA
FMP27_5F	TCTTGCCCTCTCAGAATCCAAAAA
FMP27_5R	GGGAGTGCGATTATTTGGCTGA
FMP27_6F	AGACCTAGTACCAATACAATGTTCATTCCAAACCA
FMP27_6R	CCTTGTCTGCTTTTTCGTTTTTACTTGATGTAGTG
FMP27_7F	TGAACAGCTTCAAACTTTGTATCAGTTATAAGGGC
FMP27_7R	GGGAAATTGAAAACAAAGTTAGTAACGTTAGCCAA
FMP27_8F	AAAGTAAAAAAAATAATGGTCTCTAGCGGGATCG
FMP27_8R	CCCAGTTGGTTAAGGCACCGTGCTAATAAC
ACT1 pre-mRNA F	TACATCAGCTTTTAGATTTTTCACGCTT
ACT1 pre-mRNA R	ATTCTGGTATGTTCTAGCGCTTGCACCATC
ACT1 mRNA F	TCGAAAATTTACTGAATTAACAATGGA
ACT1 mRNA R	GCAAAACCGGCTTTACACAT
ACT1 Lariat F	AGGGGCTTGAAATTTGGAAAAA
ACT1 Lariat R	GCAAGCGCTAGAACATACATAGTACA
ACT1 3’SS F	TTGCTTCATTCTTTTTGTTGCT
ACT1 3’SS R	GCAAAACCGGCTTTACACAT
RPL39 pre-mRNA F	AACACAGATAGATCAACATGGCTGTATGT
RPL39 pre-mRNA R	GGTGGTAAGGTCATTTAGATGGATGTG
RPL39 mRNA R	GTGGCAATGGTCTGTTTTGCTTC
RPL39 mRNA F	AGATCAACATGGCTGCTCAAAAGTC
RPL39 3’SS F	CGTATGTGCACGATATGTTTCCCTTT
RPL39 3’SS R	GTGGCAATGGTCTGTTTTGCTTC
RPL28 Lariat F	GAGCGCAATTATGAAAAAGAGTTACCA
RPL28 Lariat R	TTCCAAATGGAACTACATACATAGTAAAACAG
RPL28 pre-mRNA F	TCCAGATTCACTAAGACTAGAAAGCACAGA
RPL28 pre-mRNA R	TTGGTTCTTTCATTCCCTCTTCCA
RPL28 mRNA F	TCCAGATTCACTAAGACTAGAAAGCACAGA
RPL28 mRNA R	TGACCACCGGCCATACCTCT
COF1 pre-mRNA F	ATCTGGGTATGCTAAATTTCATTTGTACTCC

COF1 pre-mRNA R	AGCGAGATAAAACAGCATCATGTCAA
COF1 mRNA F	TCTGGTGTTGCTGTTGCTGATG
COF1 mRNA R	CAACGATTTCGGTTTTAGCATCG
UBC13 pre-mRNA F	AGAAATGGCATCATTACCCAAGAGAA
UBC13 pre-mRNA R	CGGCTCAGACGAAAACGTCA
UBC13 mRNA F	AGAAATGGCATCATTACCCAAGAGAA
UBC13 mRNA R	TTGCGTTTTCTGCCCTAGTAGTTTG
HRB1 pre-mRNA F	CAGGATATGTCTGATCAAGAACGAGGT
HRB1 pre-mRNA R	CATTTTACAAACTTTTCTTCACCTCCCTTA
HRB1 mRNA F	AAGTTAACAGGATATGTCTGATCAAGAACGA
HRB1 mRNA R	AGACCAATAAGTCATTGCGTTTCAAAG
ECM33 pre-mRNA F	AGTGCCTCCGCTCTAGCTGGT
ECM33 pre-mRNA R	CGAGATTTGTGAGGAAAGAGGCAAA
ECM33 mRNA F	GCCTCCGCTCTAGCTGCTAACTC
ECM33 mRNA R	TTGAGCAGTAGCAGTGGCAGAAGT
ECM33 Lariat F	CCTGTCATAGGATTAGGGCGAGT
ECM33 Lariat R	GTATGTACACATTCTCCTTTATAGTATTCCCG
RPS13 pre-mRNA F	TCGTATGCACAGTGCCGTATGTT
RPS13 pre-mRNA R	TGATTTAGCGAACTATTCAATGCAACTTT
RPS13 mRNA F	TCGTATGCACAGTGCCGGTAA
RPS13 mRNA R	AGGACAACTTGAACCAAGCTGGAG
RPS13 3'SS F	TCCAATTCCACTAAATATTACTTTAAACAGGGTA
RPS13 3'SS R	CTTGAACCAAGCTGGAGCATTTCT

2.8 Strain maintenance and growth

2.8.1 Yeast culture and preservation

Yeast liquid cultures were grown at 30°C in flasks containing YPDA medium with shaking at 180 rpm unless stated otherwise. Yeast strains containing auxotrophic plasmids were grown in YMM supplemented with the appropriate drop out powders in order to maintain the selection pressure. In the case of yeast strains containing antibiotic resistance genes integrated in the genome the suitable amount of antibiotic was added to the growth media. Yeast strains were preserved on solid medium for up to two weeks at 4°C. For long term storage, 15% (v/v) sterile glycerol was added to 1 ml of yeast culture and stored in cryogenic tubes at -80°C.

2.8.2 Growth analysis

For growth analysis, saturated overnight cultures were diluted to 0.1 OD_{600} and grown until the log phase to 0.6. OD_{600}. The exponential phase cultures were again diluted back to 0.1 OD_{600} and 100 µl of the culture pipetted into a sterile flat-bottomed 96 well plate and growth curve assayed using Sunrise Absorbance Microplate Reader (Tecan Trading AG). Plate reader was set to shake at 180 rpm and detect OD_{600} measurement every 15 minutes for 18 hours. Each growth curve experiment was carried out at least in three technical and biological replicates.

2.8.3 Spotting Assay

Saturated overnight yeast cultures were diluted to 0.1 OD_{600} and then allowed to grow until 0.6 OD_{600} was reached (this correspond to at least two doubling times). Four tenfold serial dilutions ranging from 0.2 to 0.0002 OD_{600} were made from the growing cultures and 250 µl were pipetted to a sterile flat-bottomed 96 well plate. These cells were then transferred and spotted into an appropriate sterile solid medium using pin multi-blot replicators. Plates were then incubated at 30°C and photographed after 2-3 days.

2.9 E.coli and Plasmids

2.9.1 Transformation of competent *E. coli* with pre-existing plasmids

E. coli competent cells for transformation were purchased from Agilent and stored at -80°C. A 50 µl aliquot of the competent cells was thaw on ice and split into two tubes to make a sample for plasmid transformation and a negative control. 50 ng of the plasmid was added and gently mixed with the cells. Sample was incubated on ice for 2 minutes then heat shocked at 42°C for 45 seconds. Cells were placed on ice for 2 minutes after heat shock. Next, 300 µl of SOC medium was added to the sample and

then incubated at 37°C for 1 hour with shaking. Subsequently, cells were spread on LB plate supplemented with the selective antibiotic ampicillin and incubated at 37°C.

2.9.2 Isolation of plasmid DNA from E. coli

Purification of plasmid DNA from *E. coli* was carried out using New England BioLabs plasmid DNA mini-prep kit according to manufacturer's instructions.

2.10 Yeast transformation

Yeast transformation was carried out using high efficiency lithium acetate method of Gietz and Schiestl[2]. A single yeast colony from a freshly streaked plate was inoculated into 10 ml liquid medium and then grown overnight to saturation. From the overnight culture, 50 ml of 0.1 OD_{600} of diluted cells were let to grow until log phase was reached (OD_{600} of 0.8). Cells were then pelleted by centrifugation for 2 minutes at 3500 rpm and washed twice with 10 ml of sterile distilled water. Pelleted cells were resuspended in 500 µl of 100 mM LiAc and transferred to a 1.5 ml Eppendorf tube. 100 µl of the cells were added to prepared transformation mix containing desired PCR amplified fragment or plasmid, single strand DNA from Salmon sperm and 36 µl of LiAc 1M. The mix was vortexed vigorously for 30 seconds and incubated at 30°C on a rotating wheel for 15 minutes and heat shocked at 42°C in a water bath for 30 minutes. Sample was then spun and cells resuspended in 100 µl sterile distilled water and spread on a YPDA plate and incubated overnight at 30°C. The following day, yeast were replica plated on a plate containing the appropriate selective medium and incubated at 30°C for 2-3 days. For each transformation performed, transformation mix with no PCR fragment or plasmid was used as a negative control.

2.11 Yeast DNA methods

2.11.1 Extraction of yeast genomic DNA

A single colony of yeast was inoculated in 10 ml of YPDA and grown overnight. 1.5 ml of the saturated culture was transferred to a centrifuge tube and the cells were pelleted for 5 minutes at 3500 rpm. The pellet was washed twice with distilled water. The cell pellet was resuspended in 200 µl of genomic DNA extraction buffer, 200 µl of phenol-chloroform (5:1 vv) and 300 µl zirconia beads (Thistle Scientific) were added to the sample. Sample was mechanically ground by a bench top vortex for 5 minutes to disrupt the cell wall. After centrifugation at 140000 rpm for 5 minutes, the supernatant was transferred to a new tube. DNA precipitation was carried out by adding 600 µl of 100% ethanol. The sample was incubated at -20°C for at least 1 hour and spun at maximum speed (13k rpm) for 5 minutes at 4°C. The supernatant was discarded and pelleted sample was washed twice with 70% ethanol and resuspended in 100 µl of water.

2.11.2 Standard PCR methods

PCR was performed to verify the correct genomic integration of the transformants. Genomic DNA was extracted from single colonies as described in 2.11.1 and 50 ng of genomic DNA was used for performing PCR in a 25 µl reaction mix as stated at the following table. PCR programme was adjusted according to the specific annealing temperature of the primers used and the specific DNA fragment size of interest.

PCR reaction mix	
Forward primer (100µM)	0.125 µl
Reverser primer (100µM)	0.125 µl
Taq Pol	0.5 µl
Taq Pol Buffer (10X)	2.5 µl
DMSO	1 µl
dNTP (10mM)	1 µl
Genomic DNA	50-500ng
Water	To final volume of 25 µl

2.11.3 Agarose gel electrophoresis

Agarose gel electrophoresis was performed to visualize the correct fragment size amplified by PCR. Agarose was weighed and dissolved in 1xTAE buffer by heating to make 1% gel DNA staining dye SYBR Safe was also added to the mixture. Tracking dye 0.5 µl (Bromophenol blue 1 %) was added to 5 µl of PCR reaction and samples were loaded in the gel placed in a tank submerged in 1xTAE buffer. Electrophoresis was performed at 120 V and then the gel was scanned under UV light using Syngene gel doc to visualize the DNA fragments

2.11.4 Isolation of plasmid DNA from yeast

In order to verify if the correct centromeric plasmid was contained in the appropriate yeast strains, plasmids were extracted from yeast and purified using Zymoprep yeast plasmid mini-prep kit according to the manufacturer's instructions. Finally, the correct sequence validated by Sanger sequencing.

2.11.5 DNA Sanger sequencing

The correct sequence of the plasmid was verified by Sanger sequencing. First, region of interest from plasmid was amplified by appropriate primers using BigDye® Terminator v3.1 cycle sequencing kit according to the manufacturer's instructions. Samples were then sent to the Edinburgh Genomics to undergo Sanger sequencing. SeqMan Pro tool from DNASTAR software package was used to process the chromatograms and visualize the sequences.

2.12 Yeast RNA methods

2.12.1 Rapid RNA sampling

A yeast colony was grown overnight in the appropriate liquid medium. Cells were then diluted to 0.1 OD_{600} in 50 ml medium and allowed to grow at least 2-3 doubling times (OD_{600} of 0.6-0.8). 10 ml of the culture were snap frozen by pipetting into a 15 ml falcon centrifuge tube containing chilled methanol placed in dry ice. The sample was then centrifuged immediately at 3000 g for 2 minutes at 4°C and total RNA was extracted from the pelleted cells.

2.12.2 Isolation of total RNA from yeast

For extracting total RNA, cells were resuspended in 400 µl AE buffer and transferred into a 2 ml screw cap tube. 40 µl SDS 10%, 800 µl phenol (pH=4.3) and 200 µl of zirconia beads (Thistle Scientific) were also added to the tube. The cells were lysed using a mini beads-beater (Biospec Products) for 2 minutes for three rounds. Samples were left on ice for two minutes between each lysing step. Next, samples were placed on dry ice for 5 minutes until they were solidified and centrifuged at 14000 rpm for 5 minutes at room temperature. The supernatant was transferred into a new tube containing 600 µl phenol:$CHCl_3$ 5:1, vortexed vigorously and spun for 3 minutes at 140000 rpm. The supernatant was transferred to a new tube containing 600 µl $CHCl_3$, vortexed vigorously and spun for 3 minutes at maximum speed. The supernatant was transferred to a new tube and RNA was precipitated by adding 300 µl of 10 M LiCl and incubated at -20°C for at least 30 minutes. Precipitated samples were spun for 10 minutes at 4°C and pelleted RNA was washed with 70% ethanol. RNA was finally resuspended in RNase free water or 1xTE and stored at -70°C.

2.12.3 Isolation of 4-thio-uracil labelled RNA

Strains used for this method were previously transformed with the expression plasmid pRS416-*FUI1* in order to enhance the uptake of labelled uracil from culture. Cultures

from single colonies were grown in synthetic medium lacking uracil and leucine amino acids to 0.8 OD_{600}. 100 µM of 4-thiouracil were added into 600 ml of culture shaking at 180 rpm. After 1, 1.5, 2.5, 5 and 10 minutes of labelling, 100 ml of the culture was snap frozen by pouring the culture in a beaker containing 50 ml chilled methanol sitting on dry ice. Each sample was centrifuged immediately at 3000 g for 3 min at 4 °C, the pellet was then washed twice with cold water and total RNA was extracted as described in previous paragraph. RNA was dissolved in 300 µl RNase free TE with a thermomixer at 65°C and then by pipetting up and down. The samples were biotinylated with 0.22 mg EZ-link HPDP Biotin in dark at 65°C for 15 minutes. Zeba desalting columns (Thermo Fisher Scientific) were used for desalting samples after biotinylation according to manufacturer's instructions. The samples were then precipitated with 1/3 volumes of 10 M LiCl and incubation at -20°C for 30 minutes. RNA was pelleted by centrifugation for 5 min and 14000 rpm and then washed twice with 80% ethanol to remove any remaining biotin and LiCl. 200 µl from the sample with lowest concentration of RNA was used for newly synthesised RNA purification. The same amount of RNA was used for all the other samples and the volume was equalised to 200µl with water.

Subsequently, 25 µl 10x NaTMg, 25 µl 1M NaPi pH6.8 and 2.5µl 10% SDS were added to each sample and mixed thoroughly. Throughout the following procedure, samples were kept at room temperature to avoid precipitation of SDS. 50 µl of magnetic streptavidin beads were added to a new low retention 1.5 ml tube. Samples were placed on a magnetic rack to allow beads to settle on the tube wall and the fluid was removed by aspiration. Tubes were removed from the magnet and beads washed with 200 µl of NaSTPMg and vortexed until the pellet of beads was well resuspended. Tubes were spun at 1000 g for 30 seconds, placed on the magnetic rack to allow beads to settle by the magnet and the fluid was removed with aspiration. Beads were then blocked using 200 µl NaSTPMg, 10 µl glycogen and 2.5 µl 5 mg/ml tRNA and then placed on a rotating wheel at room temperature for 20 minutes. Beads were then washed as described before.

Next, beads were added to the RNA samples and incubated at room temperature rotating for 30 minutes. The supernatant was then removed and the beads were washed as before. For eluting the RNA off the beads 100 µl of freshly made 0.7 M β-mercaptoethanol (βME) was added to the sample followed by vortexing and a brief spin to allow the beads to settle. Tubes were then placed on the magnetic rack and the supernatant was collected and transferred into a fresh 0.5ml low binding centrifuge tubes. The RNA samples were then precipitated with 2.5x volumes (280 µl) of ethanol at -20°C for 1 hour to overnight. The precipitated RNA was washed twice with 70% ethanol and dissolved in 10 µl of DEPC treated TE with RNase inhibitor. RNA quality and yield was measured by using an Agilent 2100 Bioanalyzer RNA Nano-Chip according to manufacturer's instructions.

2.13 Native elongating transcript purification

Isolation of native elongating transcript was carried out according to a protocol published by Churchman and Weissman[3] with little modifications. First, microfiltration apparatus containing a nitrocellulose filter membrane (0.45-µm, 90-mm diameter from Whatman) was set up and connected to a vacuum. 1 litre of yeast log phase growing culture was poured gradually into the filter apparatus. Filtered yeast cells were scraped off from the filter membrane with a pre-chilled metal spatula and transferred into a 50 ml conical tube containing liquid nitrogen sitting on dry ice. After liquid nitrogen vaporized, frozen cells were stored in -80°C.

Cell disruption was achieved by using a mixer mill (SPEX 6780), this system minimizes protein degradation by using a completely chilled apparatus. Liquid nitrogen was used to chill both mixer mill chamber and the metal ball used for grinding. Frozen yeast culture was ground 5 rounds for 3 min at 15 Hz with 2 minutes cooling between each round. Ground cells were scraped off from the chamber with a pre-chilled spatula and transferred to a 5 ml tube sitting in liquid nitrogen. Ground cells were stored at -80°C.

For immunoprecipitation, 0.5 ml Protein A Sepharose beads were used for each sample. Beads were washed twice by adding 10 ml of 1× lysis buffer to the slurry and supernatant was removed by centrifugation at 4°C for 2 minutes at 1000g. The washed beads were left on ice until ready to use. 1g of ground yeast was resuspended in 5 ml ice-cold 1× lysis buffer by pipetting up and down. 660 µl (660 units) of RNase-free DNase I (Promega) was added to the sample, mixed gently by inverting the tube for few times and incubated on ice for 20 minutes. Next, digested lysate was centrifuged at 20000 g, 4°C for 10 minutes. 20 µl of the supernatant was transferred to a new tube containing 20 µl 2×SDS buffer and set aside for immunoblotting to verify pull down efficiency. The remaining supernatant was added to the washed beads and incubated for 180 minutes at 4°C rotating. Subsequently, samples were spun for 2 minutes at 1000g, 4°C. 20µl of supernatant was kept as control of the unbound fraction and 20 µl of 2×SDS buffer were added to the liquid. The excess supernatant was then discarded. The beads slurry was then washed four times with 10 ml of wash buffer, resuspended thoroughly by inverting the tube and placed on a rotator for 2 minutes at 4°C. Samples were then spun for 2 minutes at 1000 × g, 4◦C and supernatant was discarded. Next, beads were resuspended in 1 ml wash buffer and transferred to a low retention 1.5ml tube, spun for 2 minutes at 1000g, 4◦C and supernatant was removed.

TAP-tagged RNAPII was eluted from beads by adding 300 µl TEV cleavage buffer and 5 µl of ProTEV protease (Promega). TEV digestion was perfomed at 18°C for 90 minutes with shaking at 180 rpm. The sample was then transferred to a Promega spin column and eluate was collected in a 2 ml tube by centrifugation. Beads were resuspended by adding 300 µl of water and mixed by pipetting up and down on the column. The column was centrifuged and the second eluate was collected and added to the first elute to obtain a final volume of 600 µl. The sample was then transferred to a new 1.5ml tube and digested with 100 µl proteinase K for 30 minutes at 37°C with gentle shaking on thermomixer. RNA was finally precipitated as described in 2.12.1.

2.14 RT-qPCR

2.14.1 Reverse transcription for splicing status

After quality control of RNA with Nanodrop or Bioanalyser, RNA sample was treated with RNase free DNase I enzyme prior to reverse transcription. For DNase I treatment, desired amount of RNA was taken and volume of sample adjusted to 8 µl with RNase free water in a PCR tube. Subsequently, 1 µl 10x DNase I buffer, 0.9 µl DNase I enzyme and 0.1 µl RNase inhibitor were added to the sample and incubated at 37°C for 30 minutes. Next, DNase I enzyme was inactivated by heating for 10 minutes at 75°C. For reverse transcription, 2.5 µl of the desired reverse primer (3µM) was added to the sample, and then heated for 3 minutes at 75°C and immediately chilled on slushy ice to allow primer annealing. From this sample, which had total volume of 12.5 µl, 5 µl was taken for RT and another 5 µl taken as negative non-RT control. 5 µl of the RT master mix was then added to each sample and incubated at 55°C for 1-2 hours. The RT master mix was prepared with 2 µl reverse transcription buffer (5X), 0.75 µl 10mM dNTPs, 0.25 µl RNAse inhibitor, 1.7 µl H2O, and 0.3 µl reverse transcriptase enzyme per sample. For the non-RT control, water was added instead of the reverse transcriptase enzyme. Prior to qPCR analysis, cDNA sample was diluted 20 times with water.

2.14.2 Quantitative real-time PCR for splicing status

qPCR was performed with Roche LightCycler 480 II instrument according to manufacturer's instructions and using the cycling illustrated in the following table. qPCR reaction was composed of 4 µl cDNA from 1:20 diluted RT reaction, 4 µl SYBR green from Agilent, 1 µl of 3 µM forward and reverse primers mix and 1 µl of H_2O to obtain a final volume of 10 µl. The non-RT samples also were subject to qPCR as negative controls. qPCR was always carried out in three technical replicates to control sample preparation and pipetting errors.

qPCR profile

Step 1:	94°C 2 min
Step 2:	94°C 10 sec
Step 3:	60°C 10 sec
Step 4:	72°C 15 sec
Step 5:	Go to step 2 for total of 40 times
Step 6:	95°C 10 sec
Step 7:	60°C 10 sec
Step 8:	Ramp to 94°C 30sec.

2.14.3 Analysis of qPCR output

Relative abundance of PCR products were measured relative to different controls depending on the type of normalization required (relative to control genes ALG9 or SCR1, exon 2, RNA from wild type, etc.) by following equation:

$$\text{Relative abundance} = 2^{-(\Delta Ct)}$$

Where Ct = the threshold cycle

$$\Delta Ct = Ct_{test} - Ct_{control}$$

2.15 Western blot

Protein samples were run on pre-cast gradient gels 4-12% NuPAGE Bis-Tris. Gel was placed in Invitrogen gel tank and filled with 1x NuPAGE MOPS SDS (Invitrgen) running buffer. Electrophoresis was performed at 170 voltage until the sample dye reached to the bottom of the gel.

Gel was disassembled from tank and proteins resolved in the gel were transferred to a nitrocellulose membrane (0.2 µm pore size) with iBlot western blotting system (Thermo Fisher Scientific) according to manufacturer's instructions. After transfer, membrane was blocked with 5% skimmed milk in 1xPBS for 1 hour or overnight at 4°C with gentle shaking. Membrane was washed with wash buffer (1xPBS + 0.1% Tween 20) three times for 10 minutes and immersed in 5% skimmed milk

containing recommended amount of the primary antibody for 1 hour at room temperature. Membrane was then washed as before and immersed in 5% skimmed milk containing recommended amount of the secondary antibody for 1 hour at room temperature. Membrane was then washed as before and proteins were detected with LI-COR Odyssey Scanner according to manufacturer's instructions.

2.16 Next-generation sequencing

Total RNA was extracted from log phase growing culture in YPDA as described before and quality controlled by Agilent 2100 Bioanalyzer RNA Nano-Chip. RNA was then shipped in a box containing dry ice to the Beijing Genomics Institute (BGI) for library preparation and sequencing on Illumina HiSeq 4000. Strand-specific, 150 bp paired-end deep sequencing was performed in two independent biological replicates for each strain (wild type, fast and slow). BGI treated the samples with DNase I and removed rRNA prior to strand specific library preparation using NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina according to manufactures instructions. After removing rRNA from the samples, total RNA was fragmented and first strand cDNA synthesis was performed with random primers. After performing the second strand cDNA synthesis, double-stranded cDNA was purified using Agencourt AMPure XP beads. Next, after end repair of the cDNA library, adaptor ligation was performed with optional NEBNext adaptor, followed by purifying the ligation reaction using AMPure XP Beads. PCR enrichment of the adaptor ligated DNA was performed according to manufactures instructions and the PCR reaction was purified using Agencourt AMPure XP beads. Library quality and quantity were assessed with a combination of qPCR and Bioanalyser. Three samples were loaded on each lane.

2.16.1 Read mapping and counting

BGI has performed initial quality controls and removed sequencing adapters and finally provided raw sequencing reads. Raw sequencing reads were then quality controlled using FastQC[4] tool. STAR aligner[5], was used for aligning the reads to the

yeast reference genome (sacCer3) using yeast GTF file (version R64-1-1.75) from ENSEMBL. STAR is a splice-aware RNA sequencing reads aligner and is widely recommended for its mapping accuracy and speed[6]. The following analysis were carried out on the Linux server (bifx-rta) provided by Wellcome Trust Centre Laboratory (COIL), University of Edinburgh.

2.16.1.1 Creating genome index

This code was used to create indexes from yeast genome that is required for STAR to run. Using or not using GTF file for building the genome index does not affect recognition of the novel sites.

```
STAR --runMode genomeGenerate --runThreadN 15 --genomeDir /path-to-genome-directory/ --genomeFastaFiles saCcer3.fa --sjdbGTFfile Saccharomyces_cerevisiae.R64-1-1.75.gtf
```

2.16.1.2 Read alignment

This code was used to map the reads to genome. Important output files were a junction file containing information about all detected junctions and their coverage, and a SAM file containing sequence alignment information.

```
STAR --runThreadN 15 --genomeDir /path-to-genome-directory/ --readFilesIn forward_reads.fq reverse_reads.fq --sjdbGTFfile Saccharomyces_cerevisiae.R64-1-1.75.gtf
```

2.16.2 Quantifying pre-mRNA fractions

In order to quantify pre-mRNA fractions, the counts for specific read classes, e.g. boundary and junction reads were obtained by using the dice-count function from the DICEseq package[7]. Among these classes, reads that only belong to pre-mRNA are boundary and intron reads, while reads that only belong to mature mRNA are junction reads. Pre-mRNA fraction was defined by this equation (pre-mRNA reads) / (pre-mRNA reads + mRNA reads).

2.16.3 Estimation of splicing error frequency (SEF) for novel alternative splicing events

Alternative splicing events or novel splicing events were defined as splicing events that are within intron containing transcripts and are not annotated in *Saccharomyces* annotation file from ENSEMBL (version R64-1-1.75). Events supported with less than five unique reads were filtered out as they could come from sequencing or mapping errors. To calculate the frequency of each alternative splicing event within a gene, we used the following equation:

$$SEF = \frac{number\ of\ reads\ that\ align\ to\ the\ novel\ splice\ junction}{number\ of\ reads\ that\ align\ to\ the\ annotated\ splice\ junction}$$

Average of SEF score from biological replicates was used for calculating the p-value between SEF score of WT and mutants by fisher's exact test.

2.16.4 Estimation of splicing error frequency (SEF) for cryptic introns in intron-less transcripts

Cryptic introns splicing events were defined as splicing events that are within intron-less transcripts and are not annotated in *S.cerevisiae* annotation file from ENSEMBL (version R64-1-1.75). Events supported with less than five unique reads were filtered out and the frequency of each splicing event was obtained from the following equation:

$$SEF = \frac{number\ of\ reads\ that\ align\ to\ the\ cryptic\ splice\ junction\ of\ the\ transcript}{RPKM\ of\ the\ transcript}$$

2.16.5 Sequence features and prediction of SEF

To predict splicing error from sequence, the following features were considered as predictors: intron length, Delta G for intron, 5'SS score, 3'SS score and frequency of 84 short sequences (1 to 3mers) from both annotated intron and novel intron. 5′SS and

3′SS scores were obtained as described[8]. The Delta G is a free energy score for RNA secondary structure, which is predicted by mfold v3.6 (http://unafold.rna.albany.edu/?q=mfold). All these features were fetched with pyseqlib (https://github.com/huangyh09/pyseqlib). In total, 176 features were used to predict the SEF scores for novel splicing events. Based on these features, a random forest regression model was trained to predict the SEF score at log2 scale, and 3-fold cross-validation was used to evaluate the prediction performance. 3.1.1 Analysis of sequence features and prediction of SEF was carried out in collaboration with Yuanhua Huang (School of Informatics, the University of Edinburgh).

2.16.6 Estimation of TSS and PAS

In order to measure the transcript start site (TSS) and polyadenylation site (PAS), the reads that were partially mapped to the 5'end and the 3'end of coding sequence were fetched. Namely, part of such read is outside to the gene body. Given these partially mapped reads, the average distance between 5' end of the reads and +1bp of the coding sequence was used to estimate the TSS (figure 2.1). TSS estimation by this strategy does not precisely map the location of TSS and might be estimated shorter due to bias introduced by the reads with shorter 5' overhang (shorter distance) from CDS. To correct for this, the reads that do not reach to the estimated TSS, 5% of the original reads with shortest 5'end distance from coding sequence were removed. In fact, he average distribution of TSS estimated via this method agrees with previous studies[9]. Similarly, this strategy was applied to estimate PAS based on distribution of 3'end of the reads. The program used here is available in DICEseq v0.2.5[7] by following command line: dice-count -a anno_file -s sam_file -o out_file --partial --TSSmethod mean. Analysis of sequence features and prediction of SEF was carried out in collaboration with Yuanhua Huang (School of Informatics, the University of Edinburgh).

Figure 2.1: Schematic representation of the approach used for estimating TSS and PAS.

References

1. Kaplan, C. D., Jin, H., Zhang, I. L. & Belyanin, A. Dissection of Pol II Trigger Loop Function and Pol II Activity–Dependent Control of Start Site Selection In Vivo. *PLoS Genet* **8,** e1002627 (2012).
2. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2,** 31–34 (2007).
3. Churchman, L. S. & Weissman, J. S. in *Current Protocols in Molecular Biology* (John Wiley & Sons, Inc., 2001).
4. Andrews, S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. (2010). Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 20th March 2017)
5. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
6. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10,** 1185–1191 (2013).
7. Huang, Y. & Sanguinetti, G. Statistical modeling of isoform splicing dynamics from RNA-seq time series data. *Bioinformatics* **32,** 2965–2972 (2016).
8. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A Sequence Logo Generator. *Genome Res.* **14,** 1188–1190 (2004).
9. Tuller, T., Ruppin, E. & Kupiec, M. Properties of untranslated regions of the S. cerevisiae genome. *BMC Genomics* **10,** 391 (2009).

Chapter 3 Studying co-transcriptional splicing of newly synthesized RNA

3.1 Acknowledgement

RNAPII elongation mutants were a gift from Craig D. Kaplan[1].

3.2 Introduction

There is extensive evidence that in both metazoans and budding yeast the process of splicing occurs as soon as the intron is transcribed and before transcription termination, i.e. co-transcriptionally[2–4]. How coupling is established and maintained between splicing and transcription is poorly understood. According to one model, referred to as the "kinetic coupling" model, variations in RNAPII elongation rate can alter the time available, or the "window of opportunity" for splicing factors to select alternative splice sites as they emerge from the RNAPII exit channel[5–10]. Consistent with this model, it was shown that RNAPII elongation rate is tuned by different elongation factors and other barriers like chromatin structure, which thereby impact splicing decisions[5–7]. Additionally, elongation rate over a gene can change multiple folds during different phases of the yeast cell cycle[11]. Probably, kinetic coupling mechanisms need to be able to adjust to fluctuations in elongation rate.

Several approaches have been designed to provide insights into co-transcriptional splicing kinetics. One way is crosslinking of the RNAPII to the DNA and immunoprecipitation of RNAPII elements bound to DNA followed by qPCR or sequencing of the isolated DNA fragments[3,12,13]. Due to presence of RNAPII in close proximity of the DNA, examining DNA fragments can provide an estimation of RNAPII coordinates on the RNA. However, this approach has limited spatial resolution and is affected by antibody specificity and also does not reveal whether RNAPII is actively engaged in transcription[14]. In human cells, deep sequencing of cellular total RNA has revealed saw-tooth profiles in read counts indicating ongoing

transcription and co-transcriptional splicing immediately after transcription of 3'SS[4]. This pattern is more obvious over the transcripts with long introns. Nuclear run-on strategy is based on the inhibition of RNAPII on the transcription initiation sites with chemical transcription inhibitors and capturing the nascent RNA from isolated nuclei after washing away the inhibitor and releasing of halted RNAPII[15,16]. This approach is used for global studying of co-transcriptional processes and mapping RNAPII positions in actively transcribed genes in yeast and mammalian cells[16,17]. Run-on assay requires extensive manipulation and isolation of the nuclei and effective re-initiation of transcription. Global co-transcriptional splicing in yeast was detected by isolating chromatin-associated RNA through chromatin fractionation and high-density tiling microarray analysis[18]. RNA extracted with chromatin fractionation does not always represent nascent RNA and also artefactual RNA interactions can occur during the fractionation procedure[19]. Native elongating transcript sequencing (NET-seq) is a breakthrough method that is used for pulling down RNAPII-nascent RNA complex and sequencing 3'ends of the nascent RNA[14]. NET-seq does not require crosslinking and is able to map the position of the RNAPII on the RNA at nucleotide resolution. This method was used recently to study co-transcriptional splicing and RNAPII pausing in yeast and mammalian cells[20,21]. Antibodies that recognize specific phosphorylated forms of RNAPII or antibodies against transcription initiation, elongation or termination factors, make NET-seq a powerful tool for studying transcripts associated with different stages of active transcription[21,22]. Efficiency of NET-seq depends on the specificity of the antibody and proper isolation of RNAPII-RNA complexes along with reducing artefactual RNA interactions. The most recent tool is the SMIT assay introduced by the Neugebauer lab[2]. SMIT determines RNAPII position with selecting genes of interest by PCR with a forward primer in exon 1 and reverse primer at RNAPII position (3'end). This approach is able to measure co-transcriptional splicing kinetics by determining position of the RNAPII when splicing occur. SMIT requires effective primer designing for a limited set of the genes that have long enough first exon or 5'UTR and additionally, SMIT could potentially overestimate the extent and speed of splicing[23].

Recently, kinetic labelling of RNA with 4-thiouracil (4tU) has proved to be an effective way to enrich for the newly synthesized RNA and used to measure the RNA synthesis, decay and splicing rates in human and yeast[24,25]. Neymotin *et al.* used *in vivo* 4tU labelling of RNA in time series followed by RNA sequencing to estimate synthesis and degradation rates for coding and noncoding transcripts[26]. With two minutes labelling intervals, Eser et al. used *in vivo* 4tU labelling and computational modelling to obtain global RNA metabolism and splicing rates in fission yeast *Saccharomyces pombe*[27]. Barrass *et al*. developed and applied extremely short 4tU labelling with high temporal resolution for measuring splicing kinetics[24]. With very short RNA labelling, as short as 60 seconds, they quantitatively measured pre-mRNA splicing speed for most of the budding yeast intron-containing transcripts. Transcripts with low stability and abundance are hardly detectable in steady state RNA and they are usually enriched in mutants with defective RNA degradation. Extremely short RNA labelling helped Barrass et al. to effectively isolate pre-mRNAs before they get fully spliced and to measure the metabolism of unstable RNAs such as CUTs, SUTs and XUTs.

Yeast transcription elongation mutants that transcribe faster or slower than the wild-type (WT) can provide a useful set of strains to investigate the effect of variations in elongation rate on kinetic coupling of splicing. Fast and slow elongation mutants may reduce and increase co-transcriptional splicing, respectively, by changing the temporal window of opportunity for splicing. A recent splicing specific microarray experiment using these mutants has shown that slower elongation increases splicing efficiency and accelerating elongation rate reduces splicing efficiency[28]. The opposite splicing phenotypes observed in this study for elongation mutants suggest that splicing efficiency is kinetically coupled to transcription. A caveat of previous studies is that assays of splicing efficiency in elongation mutants were performed using steady state RNA. Because of widespread co-transcriptional splicing and very low abundance and stability of splicing intermediates, it is challenging to measure the co-transcriptional splicing kinetics after RNA has been released from transcription complex. In fact, the

cellular steady state RNA is the consequence of multiple dynamic processes, therefore studying co-transcriptional processes requires identifying and capturing the newly synthesized RNA from the pool of total RNA. This chapter describes metabolic labeling coupled with RT-qPCR, carried out to study kinetic coupling at high temporal resolution in yeast transcription elongation mutants. It has been shown that spliced RNA and excised intron stay associated with transcription elongation complex and can be detected by immunoprecipitation (IP) of RNAPII[14]. Therefore, co-transcriptional splicing was also directly examined in elongation mutants through measuring association of the spliced RNA and lariat with RNAPII. This analysis shows that splicing fails to keep up with fast transcription, leading to accumulation of unspliced pre-mRNA, whereas in the slow mutant, nascent transcripts are more rapidly removed through splicing than in WT cells.

3.3 Results

3.3.1 Characterisation of RNAPII elongation mutants

3.3.1.1 RNAPII transcription elongation rate mutants show growth defects

The trigger loop of *RPB1*, the largest subunit of RNAPII, is a mobile and conserved domain in the catalytic centre of RNAPII that is responsible for substrate selection and catalysis[29,30]. Kaplan *et al* made a series of point mutations in the trigger loop of Rpb1 cloned in *RPB1 CEN LEU2* plasmid and transformed the mutant plasmid, *rpb1 CEN LEU,* into a *rpb1Δ* strain carrying *RPB1 CEN URA3* plasmid. The transformants were selected by their ability to grow on a plate containing 5-fluoroorotic acid. Elongation rates of these mutants were estimated using *in vitro* run-off transcription assay[1]. In this assay, the amount of full-length RNA that is transcribed during the reaction time provides an estimate of the RNAPII elongation rate *in vitro* relative to the WT. Throughout this thesis, RNAPII elongation mutants were used that transcribe faster (*rpb1-G1097D*) or slower (*rpb1-H1085Y*) than WT (*RPB1*). The Rpb1-H1085 residue, which is located in the nucleotide interacting region, interacts with NTP β-phosphate

and plays a critical role in nucleotide transfer in the catalytic centre[31]. Substitution of H1085 residue with Alanine (A), Phenylalanine (F), Aspartate (D), or Asparagine (N) is lethal[1]. G1097 is one of the trigger loop hydrophobic pocket residues that support integrity of trigger loop open state (inactive state)[1]. Mutating G1097 residue to Aspartate (D) was proposed to destabilize the open state of the trigger loop, providing a plausible mechanism for substrate-induced trigger loop closing (active state). Substitution of Rpb1-A1076, another pocket residue, with Tyrosine (T) leads to increase in the speed of transcription elongation similar to rpb1-*G1097D*[32]. Hereafter, rpb1-*H1085Y* and rpb1-*G1097D* mutants will be referred to as slow and fast RNAPII mutants, respectively.

In vitro assays have estimated transcription elongation rate as ~12 nucleotide per second (nt/s) for WT, ~50 nt/s for fast mutant and ~1.5 nt/s for slow mutant[1]. It is reported that changing transcription elongation rate suppresses growth and severity of growth correlates with the degree of change in elongation rate[1]. Therefore, to validate the phenotype of fast and slow strains used in this work and compare with previous reports[1,28], strains were grown to log phase and 10-fold serial dilutions of the cells were transferred to a SC-Leu growth plate. SC-Leu is a complete medium lacking leucine. Consistent with Kaplan et al report[1], fast and slow mutants showed growth defects that were comparable with their perturbed elongation rates (figure 3.1a). It remains to be determined whether *in vivo* elongation rates in these mutants are accurately represented by the *in vitro* elongation rates. However, estimation of the *in vivo* elongation rates of these mutants in human cell lines with genome-wide nuclear run-on sequencing (GRO-seq) showed a good agreement with *in vitro* estimates[16]. To further validate the strains, plasmids were rescued from the strains and regions of interest were amplified by PCR and samples sent to Edinburgh Genomics for Sanger sequencing. As shown in figure 3.1b, Sanger sequencing confirmed the presence of the mutations in the trigger loop. In the case of *rpb1-G1097D*, GGT (Gly1097) codon mutated to GAT (Asp) and for *rpb1-H1085Y*, CAT (His1085) codon mutated to TAT (Tyr) codon.

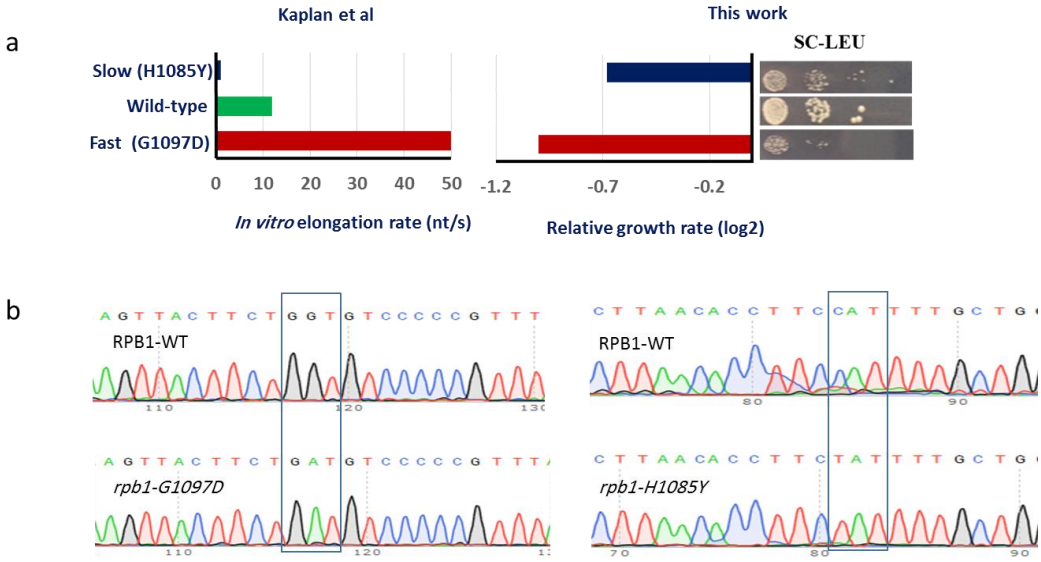


Figure 3.1: Transcription mutants and their growth phenotypes. (**a**) *In vitro* elongation rate, nucleotide per second (nt/s), quantified by Kaplan et al using transcription run-off assay. Spotting assay performed in this study shows growth on SC-Lys after 3 days (right). Relative growth rate that is shown as barplot estimated from spotting assays and normalized to the WT (**b**) Chromatograms showing Sanger sequencing results and confirmation of the mutants.

3.3.1.2 Deleting UPF1 makes fast mutant auxotroph for lysine amino acid

It has been shown that mutations that increase transcription elongation rate are sensitive to chemical inhibitors that reduce the rate of elongation[1]. To test the growth phenotype of the strains, they were grown on a plate lacking SC-Leu and containing mycophenolic acid (MPA, 20 µg/ml final concentration), a drug that reduces GTP synthesis and consequently inhibits transcription elongation. Due to the plan for doing RNA sequencing (discussed in chapter 5), *UPF1* was deleted from all three strains in order to protect mis-spliced transcripts from nonsense-mediated decay (NMD) and the growth phenotype of the *Δupf1* strains was assayed alongside wild-type *UPF1* strains. Spotting assays revealed that MPA inhibits growth of the fast (figure 3.2), whereas the slow and the WT strains were insensitive to MPA. Rpb1-*E1103G*, which also

transcribes faster than WT, was used as additional control strain. These observations confirm the expected phenotype of the elongation mutants[1,28]. Notably, the fast and slow mutants showed better growth when *UPF1* was deleted.

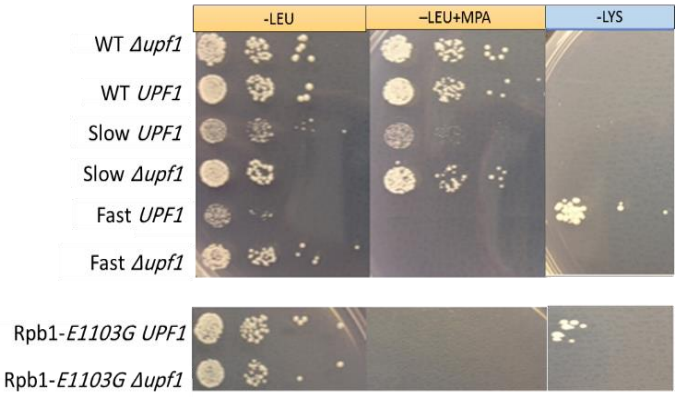


Figure 3.2: Growth in the presence or absence of MPA and also absence of Lys amino acid. After performing spotting assay as described in materials and methods, plates were incubated at 30°C and photographed after three days. –Lys plate was photographed after 4 days.

All three strains contain a transposable (Ty) element in the *LYS2* promoter (*lys2-128∂,* the Spt− phenotype) resulting in auxotrophy for lysine (figure 3.3a). Ty elements are repeated elements about 6 kilobases long and flanked by about 300bp direct repeats (δ sequences)[33]. There are around 35 Ty elements in the *S. cerevisiae* genome. Recombination of Ty elements is facilitated by δ-δ interactions[33]. The *lys2-128∂* phenotype is extremely stable, reverting to Lys^+ (ability to grow on SC-Lys) with δ-δ recombination at frequency of 2 out of 10 billion cells[33]. SC-Lys is a complete medium lacking lysine. Fast elongation suppresses the effect of this element and enables the cells to somehow transcribe the *LYS2* gene and thereby survive in the absence of lysine (figure 3.2)[1]. Conversely, when *UPF1* was deleted from these strains (*Δupf1*), the fast mutant lost the ability to grow on the SC-Lys plate. To confirm this phenotype, an additional fast-transcribing strain (*rbp1-E1103G*) was tested. As can be seen in figure 3.2, similar to the growth phenotype of the fast strain (*rpb1-G1097D*), *rbp1-E1103G* does not grow on the SC-Lys plate in the absence of *UPF1* gene (*Δupf1*). To check for expression of *LYS2* in the *Δupf1* mutants, RNA sequencing data

(generated for chapter 5) were examined for transcripts at the *LYS2* locus using the UCSC genome browser. In both fast and WT strains, there is abundant transcription initiation at the transcription start site (TSS), but this signal is lost in the Ty insertion site, indicated by 'star' (figure 3.2). Notably, antisense RNA is also elevated more than two fold in the case of the fast mutant which may interfere with *LYS2* expression on the sense strand (figure 3.3b).

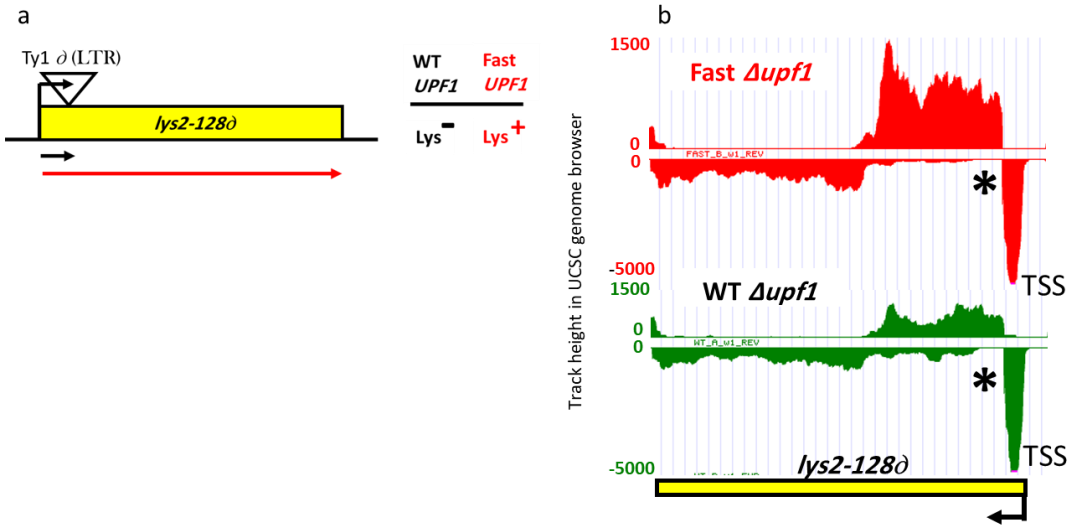


Figure 3.3: Ty1 insertion in promoter region of *LYS2* gene. (a) Insertion of the Ty1 element into the 5'end of the LYS2 coding region creates *lys2-128∂* allele. This insertion makes *RPB1* cells Lys⁻ as they only express a short non-functional transcript (black arrow), while the fast mutant allows transcription of *LYS2* from a cryptic promoter (red arrow, bottom), somewhere within the Ty1 insertion, allowing the cells to become Lys⁺. (**b**) Read coverage of the Lys2 locus in the WT and fast strains with *Δupf1* mutation. Star indicates the Ty1 insertion site.

3.3.2 Fast and slow transcription affect co-transcriptional splicing of nascent RNA

3.3.2.1 Kinetic labelling of RNA with 4tU to study co-transcriptional splicing

4-Thiouracil (4tU) is a non-toxic base analogue that can be taken up by yeast and mammalian cells and incorporated into RNA that is being actively transcribed (figure 3.4a). In order to get rapid 4tU incorporation in to the RNA during short periods of labelling, the *FUI1* permease gene was over-expressed via transforming pRS426-*FUI1* plasmid into the yeast cells. In the cell, 4tU is converted to 4tUTP and incorporated into RNA during transcription. After treating cells with 4tU for 1, 1.5, 2.5, 5 and 10 minutes, the yield of newly synthesized RNA was measured in all three strain. As can be seen in figure 3.4b, the total yield of newly synthesised RNA increases with labelling time.

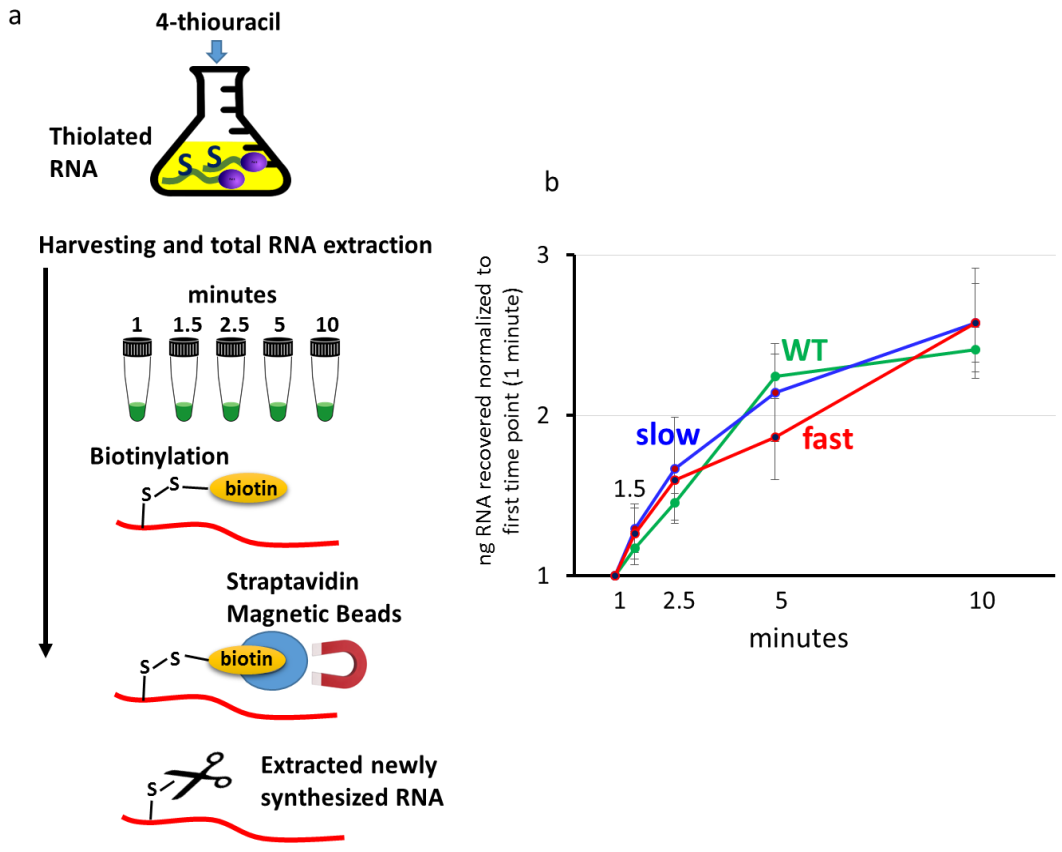


Figure 3.4: 4tU labelling diagram. (**a**) 4tU added to the growing yeast culture gets incorporated (shown with S in the flask) into the newly synthesized RNA (green). RNAPII is shown in purple in the flask. After harvesting the cells at the indicated times, total RNA is extracted and newly synthesized RNA is biotinylated and affinity-selected with streptavidin beads. Newly synthesized RNA is released from beads with beta-mercaptoethanol (scissors). (**b**) Showing total yield of the newly synthesized RNA at the indicated times (at least three biological replicates).The amount of newly synthesized RNA increases with time in all strains.

3.3.2.2 Altering transcription elongation rate influences the amount of nascent RNA that is spliced co-transcriptionally

To gain further insight into the coupling between splicing and transcription, the splicing of newly synthesised transcripts produced by RNAPII elongation mutants and WT was examined. By briefly labelling RNA with 4tU (for 1, 1.5, 2.5, 5 and 10 min), nascent, thiolated RNA was biotinylated and then affinity-selected using streptavidin beads (figure 3.4a). Analysing the nascent RNA by reverse transcription-quantitative PCR (RT-qPCR) the kinetics of splicing was followed for three well-expressed intron-containing genes, *RPL39*, *RPL28* and *ACT1*. Splicing efficiency is generally determined by measuring levels of spliced and unspliced RNAs[28]. Pre-mRNA was measured using an amplicon for exon-intron boundary (5'SS boundary) and mRNA was measured by an amplicon detecting exon-exon junction (figure 3.5a). In the fast mutant, a rapid increase in the pre-mRNA/mRNA ratio was observed for all three representative transcripts, indicating that pre-mRNA was synthesised faster than it was removed by splicing (figure 3.5b). This ratio declined gradually towards the steady-state level after 2.5 min. Conversely, with the slow mutant, transcripts were spliced before they accumulated to a significant level such that the pre-mRNA/mRNA ratio was less than for the WT RNAPII even in the case of *RPL28* transcripts that splice slowly[24]. These results clearly illustrate the "window of opportunity" effect; with the fast mutant the pre-mRNA exists for longer before it is spliced, and with the slow mutant the pre-mRNA is spliced very soon after synthesis. It should be noted that this does not mean that the rate of splicing is different in the three strains, only that it

happens later or sooner after transcription. Additionally, the pre-mRNA/mRNA ratio in total RNA was measured for 5 representative genes, finding a higher ratio than WT with the fast mutant and slightly lower for the slow mutant (figure 3.5c).

Total RNA sequencing of the human cell lines has revealed a 5′–3′ gradient in the read coverage across the introns and exons, with higher levels of reads present in the 5′ end of introns, illustrating co-transcriptional splicing[4]. Measuring 5'SS/3'SS with RT-qPCR in elongation mutants showed that fast mutant reduces and slow mutant enhances this gradient (figure 3.5d).

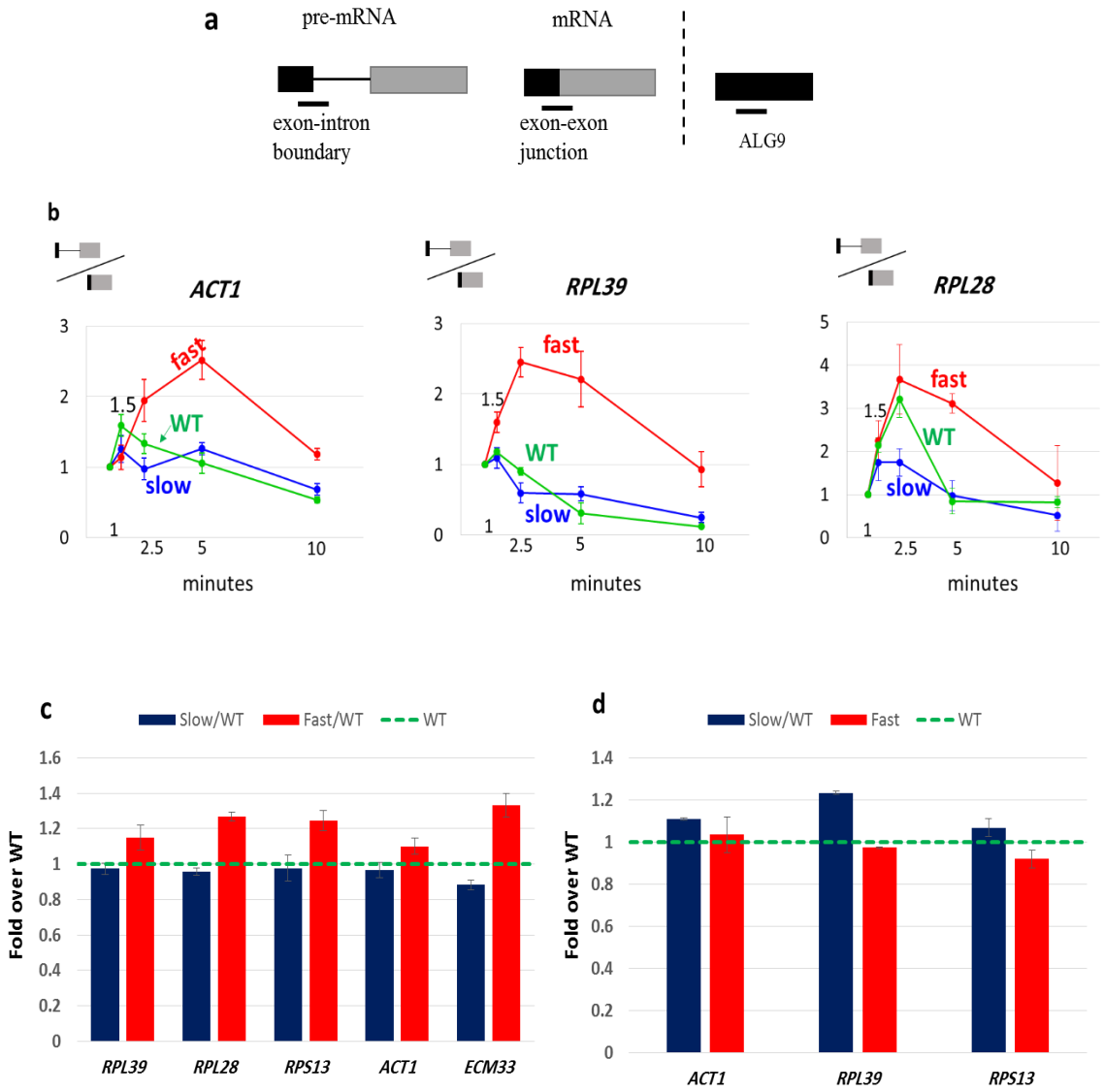


Figure 3.5: Altering RNAPII elongation rate affects RNA splicing efficiency. (**a**) Diagram showing the location of RT-qPCR amplicons (lines below) for measuring pre-mRNA (exon–intron boundary at the 5′SS) and mRNA (exon–exon junction) levels. Black boxes represent exons of intron-containing transcript or coding sequence of intronless transcript (*ALG9*). (**b**) RT-qPCR results showing pre-mRNA/mRNA ratio for *ACT1*, *RPL39* and *RPL28* in fast (red line), slow (blue line) and WT (green line) strains. To correct for amount of input RNA, values for pre-mRNA and mRNA were separately normalized to an intronless transcript (*ALG9*) in the same sample. 4tU labelling was performed for 1, 1.5, 2.5, 5, and 10 minutes (x axis), with all values plotted relative to the 1 minute value. (**c**) pre-mRNA/mRNA ratio of steady state RNA for the fast and slow mutants relative to WT (*green dashed line*) measured by RT-qPCR. Error bars represent at least three biological replicates; p-value < 0.0036. (**d**) 5'SS/3'SS ratio of steady state RNA for the fast and slow mutants relative to WT (*green dashed line*) measured by RT-qPCR. Error bars represent at least three biological replicates; p-value < 0.044.

3.4 Co-purifying nascent RNA with RNAPII

The highly stable association of nascent transcripts with elongating RNAPII permits co-transcriptional splicing to be measured by analysis of transcripts that co-purify with RNAPII[14]. DNA-RNA-RNAPII ternary complex is highly stable, allowing co-purification of the nascent transcripts directly from lysed cells by immunoprecipitation (IP) of RNAPII without crosslinking (figure 3.6a). Churchman and Weissman used *RPB3* with a 3×FLAG epitope tag for IP of the RNAPII[14]. Strains used in this study had *RPB3* with a TAP tag and the protocol was adapted accordingly to pull down transcription complex with TAP tagged Rpb3. Rpb3, the third-largest subunit of RNAPII, is essential and part of the central core of the polymerase complex. After pulling down RNAPII complex with IgG–Sepharose beads, co-precipitated RNAs were released, purified and analysed by RT-qPCR to quantify the pre-mRNA, mRNA and excised intron products of splicing (figure 3.6a). Pulldown efficiency of the elongation complex was checked by western blot (figure 3.6b).

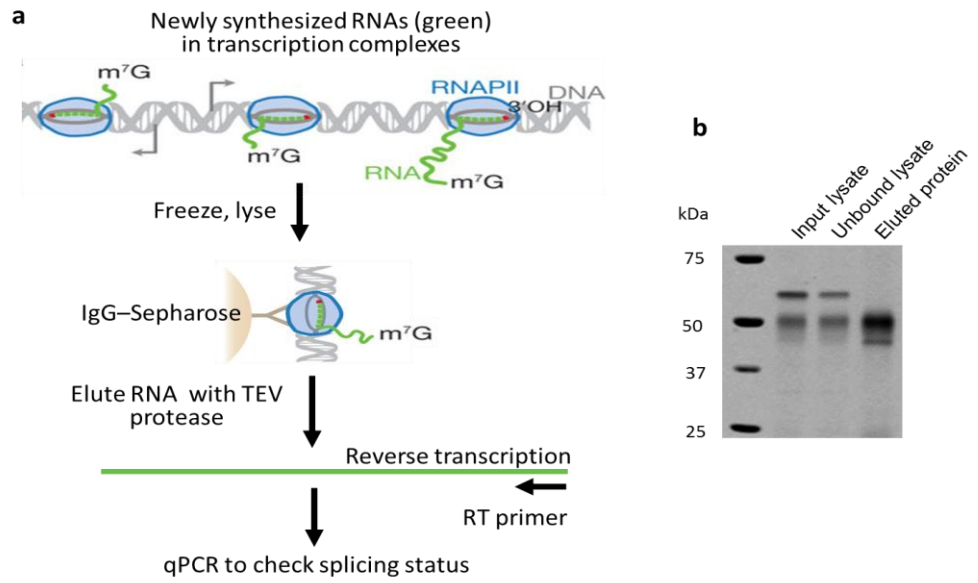


Figure 3.6: Schematic diagram of NET-RT-qPCR protocol adapted from[14]. (**a**) Yeast culture is flash frozen and lysed cryogenically. Nascent RNA (green) is co-purified with RNAPII elongation complex using IgG–Sepharose beads. RNA finally is eluted with TEV protease and reverse transcribed into cDNA and analysed with qPCR. m7G refers to the 7-methylguanosine cap structure at the 5′ end of nascent transcripts. (**b**) Western blot detecting Rpb3 immunoprecipitation samples of input lysate (56 kDa), unbound lysate and eluted protein (40 kDa) with anti Rpb3 antibody.

3.4.1 RNAPII occupancy along intronless genes

Genome-wide analysis of RNAPII density in *S. cerevisiae* by NET-seq has revealed an increase in density of RNAPII at the 5'ends of genes (figure 3.7a)[14]. In order to check the efficiency of nascent RNA pull down, it was tempting to explore the RNAPII density profile along the two well-expressed intronless genes, *ALG9* and *FMP27* and compare it with the total RNA. For this purpose, primers were designed across these genes and the amount of nascent RNA associated with RNAPII along these genes was measured by RT-qPCR. Indeed, the RNAPII density profile observed for *ALG9* and *FMP27* was similar to the published NET-seq RNAPII density profile (figure 3.7b). In all three strains, RNAPII density is higher at the 5′end near the transcription start

sites and reduces towards the 3'end of the genes. Notably, this profile was not detected when total RNA used as a control. In metazoans, RNAPII experiences a promoter-proximal pausing event at the 5'end, which causes RNAPII to pile up before entry to the active elongation state. This pause is imposed by NELF and DSIF transcription factors and released by phosphorylation of NELF, DSIF and CTD by P-TEFb[34]. In budding yeast, the 5′ pause is less prominent due to the lack of NELF[19].

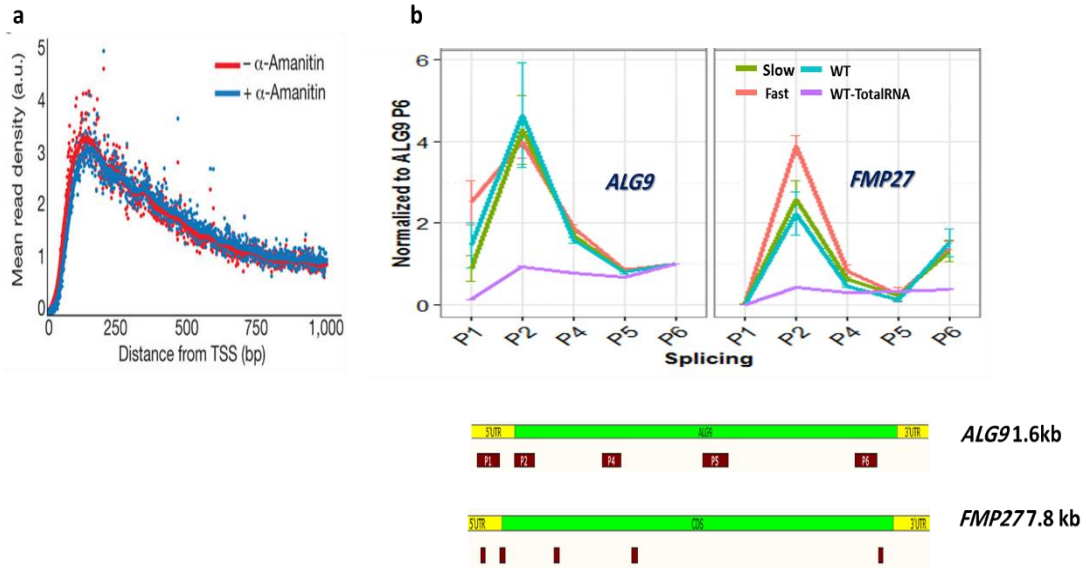


Figure 3.7: (**a**) Published metagene analysis of read counts for well-expressed genes (n = 471) in the presence and absence of α-amanitin (transcription inhibitor)[14]. (**b**) RNAPII density profile across *ALG9* and *FMP27* for the fast, slow and WT strains. WT-TotalRNA is the total RNA from WT (not purified with IP) used as a control. Positions of the primers on *ALG9* and *FMP27* are shown below the plots.

3.4.2 Excised lariat and mRNA association with RNAPII

It has been shown that the excised lariat intron and the spliced RNA remain associated with RNAPII, probably via the spliceosome[14]. Therefore, measuring the association of mRNA and lariat with the mutant polymerases could provide insights regarding the co-transcriptional splicing when transcription is faster or slower than WT. Measuring spliced exon association with RNAPII shows that, for most of the genes tested, slower

transcription increases its association with RNAPII, indicating enhanced co-transcriptional splicing, whereas faster transcription reduces co-transcriptional splicing compared to WT RNAPII (figure 3.8a). The exceptions are the two RP genes tested, *RPL39* and *RPL28*, for which the association of the spliced RNA decreases similarly with both the slow and fast mutants. *HRB1* has an unconventional BP (TACTAAT) that differs from the yeast canonical BP (TACTAAC). It appears that slower transcription opens the window of time for efficient recognition and splicing of this BP and this could explain the higher association of the *HRB1* mRNA with the slow RNAPII. As expected[14], the excised intron, in the form of a branched lariat, also stays associated with slow RNAPII and is increased significantly compared with WT (figure 3.7b). Conversely, this association reduced greatly with the fast RNAPII. As a control, lariat was also measured in total cellular RNA. Unexpectedly, association of the pre-mRNA is also reduced with the fast mutant and slightly increased with the slow mutant (figure 3.8c). Low level association of the pre-mRNAs with fast RNAPII could be due to co-transcriptional degradation of pre-mRNAs (figure 3.8d). Indeed, it has shown that *xrn1*Δ/*rat1-1* double mutant background stabilizes RNA in the fast mutant, suggesting degradation of pre-mRNAs from the 5'end, possibly due to a capping defect[35]. On the other hand, the slow mutant can potentially stabilize pre-mRNAs by allowing more efficient 5'capping, which leads to increased association of pre-mRNA with transcription complex (figure 3.8d).

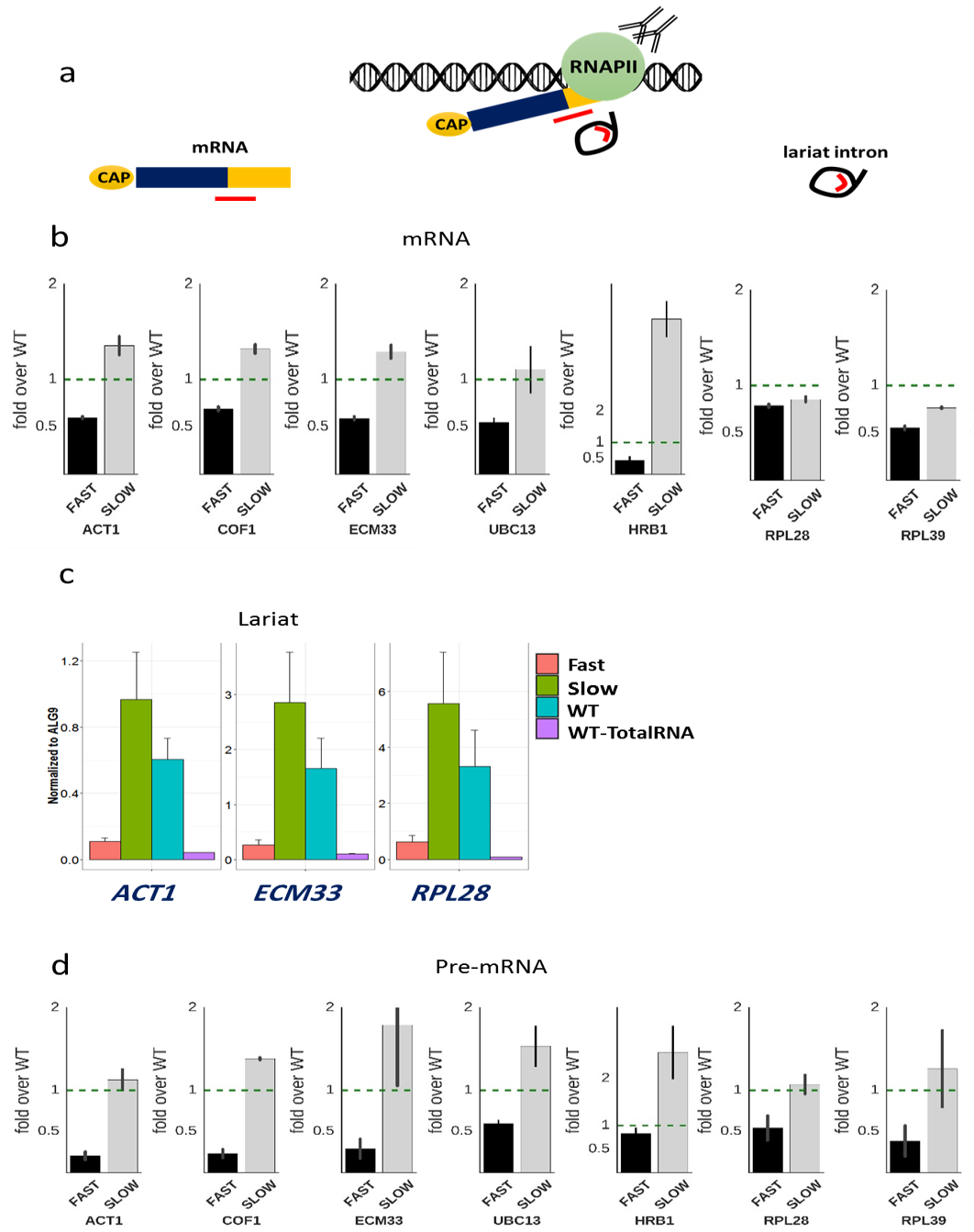


Figure 3.8: Fast elongation reduces and slow elongation enhances co-transcriptional splicing. (**a**) Diagram of capturing nascent RNA transcripts (mRNA and excised intron) directly through their association with the DNA-RNA-RNAPII ternary complex. Red lines represent location of the RT-qPCR amplicons for measuring mRNA (exon–exon junction) and excised intron levels. (b) Fold enrichment of mRNA (**b**), lariat (**c**) and pre-mRNA (**d**) in each mutant relative to WT. To correct for differences in the amount of RNA pull down, values for pre-mRNA and lariat were separately normalized to RT-qPCR values for an intronless transcript (*ALG9*). Error bars represent three biological replicates.

3.5 Discussion

3.5.1 Changes in growth phenotype of elongation mutants in the absence of Upf1

Upf1, active in the yeast cytoplasm with RNA helicase activity, is one of the key proteins in the NMD pathway, implicated in rapid turnover of RNAs containing premature translational termination codons (PTCs). Upf1 is a non-essential protein in yeast and its deletion from *RPB1* cells does not have an obvious effect on cell growth (figure 3.2), although it stabilizes RNAs containing PTCs[36]. Improved growth phenotype of RNAPII elongation mutants in the absence of Upf1 suggests that the NMD pathway is limiting for growth in these mutants. As the RNAPII TL promotes selection and catalysis of NTP substrates, it is proposed as the major contributor to transcription fidelity[37]. Thus, it seems likely that the TL elongation mutants, the fast and the slow RNAPII, increase *in vivo* nucleotide misincorporation events that could increase the likelihood of introducing PTCs in RNA transcripts. Indeed, it was shown that the *rpb1-E1103G* mutation and its analogous mutation in the *E. coli* RNA Polymerase that causes faster elongation, decreases transcription fidelity[38,39]. Upf1 activity is ATP dependent and therefore, targeting and processing all the NMD substrates accumulated due to perturbed elongation mistakes would require more energy and time. These results would seem to suggest that inactivating the NMD pathway bypasses the RNA surveillance step via NMD in the elongation mutants and promotes the cell growth at the expense of allowing accumulation of truncated proteins that could be toxic.

In spite of better cell growth on SC-Leu in the absence of Upf1, the lack of Upf1 suppresses the growth of the fast strain with *lys2-128∂* background on SC-Lys plate (figure 3.2). Why the fast mutant loses this ability in the absence of Upf1 is not clear but three points are highlighted here. First, transcription elongation mutants alter TSS selection (discussed next chapter) and it was suggested that an upstream shift in TSS somehow enables productive initiation of *LYS2* transcription from a cryptic

promoter within the Ty element[1]. This indicates that Upf1 might play a role in productive transcription initiation by modulating TSS selection. If this is true, there should be differences in the TSS selection pattern of the fast strain in the presence (fast-*UPF1*) and absence (fast-*upf1*) of Upf1. However, a comparison of TSS at the *ADH1* gene between fast-*UPF1* strain reported previously[28] and fast-*upf1* strain analysed by RNA-seq in this study (discussed next chapter) has revealed no difference. Second, as mentioned earlier, the Ty insertion in *LYS2* gene (*lys2-128∂*) creates a very stable Lys^- phenotype (suppression of growth on SC-Lys) but it has shown that *spt2* and *spt3* mutations increase the frequency of Lys^+ revertants (growth on SC-Lys) significantly[33]. The revertants of *lys2-128∂* result from δ-δ recombination, which excises most of the Ty element. Spt2 and Spt3 are involved in negative regulation of transcription. It seems possible that Lys^+ phenotype of the strains in the absence of negative elongation factors is due to hyperactivity of transcription complex that somehow promotes δ-δ recombination and excision of Ty. Accordingly, growth of the fast mutants with *lys2-128∂* background on SC-Lys likely is the result of transcription hyperactivity. Fast mutants lose this ability in the absence of Upf1, suggesting that Upf1 directly or indirectly facilitates δ-δ recombination. The first and second points appear to support the assumption that similar to human cells, Upf1 may have nuclear functions in yeast[40]. Third, it has been shown that the mRNA decapping complex, the 5′ to 3′ mRNA exonuclease, Xrn1, and the NMD pathway (Upf1, Upf2, and Upf3) stimulate post-translational steps in retrotransposition[41]. Upf1 is also required for efficient restriction of Ty1 retrotransposition by Ty1 processing complex, and deleting *UPF1* reduces retrotransposition[41,42]. These studies suggest a role for Upf1 in Ty1 retrotransposition at post-translational steps. It also has been shown that antisense RNAs interfere with posttranslational Ty1 retrotransposition processing. Looking at the RNA sequencing data reveals at least two-fold increase in antisense transcription of the *LYS2* gene with the fast mutant (Δ*upf1*) relative to the WT (Δ*upf1*) (figure 3.3). This indicates that the absence of Upf1 may lead to accumulation of antisense RNA and therefore inhibit Ty1 retrotransposition. To confirm this hypothesis, it is required

to measure antisense transcription of the *LYS2* gene in a fast-*UPF1* strain. How Upf1 is implicated in functional *LYS2* mRNA production through modulating posttranslational retrotransposition processes is not clear. More research on this topic needs to be undertaken before the association between transcription elongation mutants and Upf1 is more clearly understood.

3.5.2 Altering transcription elongation rate influences the amount of nascent RNA that is spliced co-transcriptionally

There is mounting evidence for functional links between transcription and splicing, for example, transcription elongation can affect alternative splicing, and splicing or splicing defects can affect transcription[3,5,7,13,19,21,43,44]. The functional consequences for splicing of altering transcription rate have been the subject of intensive study in mammalian cells[6,7,19,44]. Several lines of evidence presented above show that RNAPII mutations that cause faster (Rpb1-*G1097D*) or slower (Rpb1-*H1085Y*) transcription in budding yeast reduce and increase co-transcriptional splicing, and they also reduce and increase splicing efficiency, respectively.

In vivo 4tU labelling for a brief period revealed an increased pre-mRNA/mRNA ratio in the fast strain, whereas slower transcription resulted in a modest decrease in this ratio compared to WT (figure 3.5). This indicates that splicing does not keep up with a faster elongating RNAPII; i.e. there is a delay after production of the intron (actually the 5'SS in this assay) before it is removed by splicing and, conversely, when transcription is slower, splicing of newly synthesised pre-mRNA occurs sooner after production of the intron. This does not necessarily imply that the rate of splicing itself is different in the three strains. These results are in accord with single-molecule intron tracking (SMIT) that showed greater physical distance between the 3' ends of introns and the fast RNAPII compared to normal RNAPII, although they did not test a slow mutant[2]. The SMIT approach showed that, for the 87 genes tested,

90% of co-transcriptional splicing occurred, on average, by the time RNAPII was 129 bp beyond the 3'SS. Results presented here suggest that splicing occurs even closer to slow RNAPII compared to WT. In support of this, figure 3.8b shows that, most genes analysed, more spliced mRNA and excised intron lariat associates with slow RNAPII and less with fast RNAPII compared to WT. Together, these different assays show that with faster elongation less splicing occurs co-transcriptionally but more of the second exon may be available to the splicing machinery at the time when splicing takes place, whereas, with slower elongation more splicing takes place co-transcriptionally and sooner after transcription of the 3'SS. This is consistent with the window of opportunity model, which proposes that slow elongation expands and fast elongation compresses the time available for upstream splice sites to be recognised before competing downstream splice sites are produced[5–10]. In the case of single intron genes, this translates to more or less time being available for splicing to occur before 3' end processing occurs and transcription terminates, i.e. co-transcriptionally. For the slow mutant it may be that intron recognition and/or other aspects of spliceosome assembly are enhanced due to there being more time for co-transcriptional spliceosome assembly.

Measurements of splicing efficiency by RT-qPCR of total RNA revealed an inverse correlation between transcription elongation rate and splicing efficiency (figure 3.5c). This analysis further supports that the splicing efficiency is coordinated with transcription rate. Slow transcription increases the splicing efficiency possibly by allowing sufficient time for spliceosome assembly. In contrast, the spliceosome cannot process all the pre-mRNAs that are produced with faster transcription, therefore splicing efficiency decreases. A previous study with RNAPII elongation mutants also documented an anti-correlation between splicing efficiency and transcription elongation speed in budding yeast based on microarray analysis[28]. Furthermore, RT-qPCR of total RNA revealed an increase in 5'SS / 3'SS ratio with the slow mutant but in contrast, this ratio was lower for the fast mutant compared to WT (figure 3.5d). It was shown previously that 5′–3′ gradient of sequence coverage across the introns and

exons can be explained by co-transcriptional splicing and could be used for quantification of co-transcriptional splicing from RNA-seq data[4,45]. Enhancing and relaxing the 5'-3' gradient with the slow and the fast mutants, respectively, supports the notion that introns are spliced more co-transcriptionally when transcription is slower and less co-transcriptionally with the fast mutant.

References

1. Kaplan, C. D., Jin, H., Zhang, I. L. & Belyanin, A. Dissection of Pol II Trigger Loop Function and Pol II Activity–Dependent Control of Start Site Selection In Vivo. *PLoS Genet* **8,** e1002627 (2012).
2. Carrillo Oesterreich, F. *et al.* Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* **165,** 372–381 (2016).
3. Alexander, R. D., Innocente, S. A., Barrass, J. D. & Beggs, J. D. Splicing-Dependent RNA Polymerase Pausing in Yeast. *Mol. Cell* **40,** 582–593 (2010).
4. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18,** 1435–1440 (2011).
5. Saldi, T., Cortazar, M. A., Sheridan, R. M. & Bentley, D. L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J. Mol. Biol.* **428,** 2623–2635 (2016).
6. Kornblihtt, A. R. Coupling transcription and alternative splicing. *Adv. Exp. Med. Biol.* **623,** 175–189 (2007).
7. Naftelberg, S., Schor, I. E., Ast, G. & Kornblihtt, A. R. Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. *Annu. Rev. Biochem.* **84,** 165–198 (2015).
8. de la Mata, M. *et al.* A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. *Mol. Cell* **12,** 525–532 (2003).
9. Kadener, S. *et al.* Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing. *EMBO J.* **20,** 5759–5768 (2001).
10. Roberts, G. C., Gooding, C., Mak, H. Y., Proudfoot, N. J. & Smith, C. W. Co-transcriptional commitment to alternative splice site selection. *Nucleic Acids Res.* **26,** 5568–5572 (1998).
11. Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A. & Singer, R. H. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science* **332,** 475–478 (2011).
12. Lefrançois, P. *et al.* Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics* **10,** 37 (2009).
13. Chathoth, K. T., Barrass, J. D., Webb, S. & Beggs, J. D. A Splicing-Dependent Transcriptional Checkpoint Associated with Prespliceosome Formation. *Mol. Cell* **53,** 779–790 (2014).
14. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469,** 368–373 (2011).
15. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848 (2008).
16. Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28,** 2663–2676 (2014).

17. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339,** 950–953 (2013).
18. Oesterreich, F. C., Preibisch, S. & Neugebauer, K. M. Global Analysis of Nascent RNA Reveals Transcriptional Pausing in Terminal Exons. *Mol. Cell* **40,** 571–581 (2010).
19. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* **15,** 163–175 (2014).
20. Harlen, K. M. *et al.* Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. *Cell Rep.* **15,** 2147–2158 (2016).
21. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161,** 526–540 (2015).
22. Mayer, A. *et al.* Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. *Cell* **161,** 541–554 (2015).
23. Wallace, E. W. J. & Beggs, J. D. Extremely fast and incredibly close: co-transcriptional splicing in budding yeast. *RNA* rna.060830.117 (2017). doi:10.1261/rna.060830.117
24. Barrass, J. D. *et al.* Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. *Genome Biol.* **16,** 282 (2015).
25. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29,** 436–442 (2011).
26. Neymotin, B., Athanasiadou, R. & Gresham, D. Determination of in vivo RNA kinetics using RATE-seq. *RNA* (2014). doi:10.1261/rna.045104.114
27. Eser, P. *et al.* Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. *Mol. Syst. Biol.* **12,** 857 (2016).
28. Braberg, H. *et al.* From Structure to Systems: High-Resolution, Quantitative Genetic Analysis of RNA Polymerase II. *Cell* **154,** 775–788 (2013).
29. Zhang, J., Palangat, M. & Landick, R. Role of the RNA polymerase trigger loop in catalysis and pausing. *Nat. Struct. Mol. Biol.* **17,** 99–104 (2010).
30. Mejia, Y. X., Nudler, E. & Bustamante, C. Trigger loop folding determines transcription rate of Escherichia coli's RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 743–748 (2015).
31. Čabart, P., Jin, H., Li, L. & Kaplan, C. D. Activation and reactivation of the RNA polymerase II trigger loop for intrinsic RNA cleavage and catalysis. *Transcription* **5,** (2014).
32. Qiu, C. *et al.* High-Resolution Phenotypic Landscape of the RNA Polymerase II Trigger Loop. *PLOS Genet.* **12,** e1006321 (2016).

33. Simchen, G., Winston, F., Styles, C. A. & Fink, G. R. Ty-mediated gene expression of the LYS2 and HIS4 genes of Saccharomyces cerevisiae is controlled by the same SPT genes. *Proc. Natl. Acad. Sci. U. S. A.* **81,** 2431–2434 (1984).
34. Yamada, T. *et al.* P-TEFb-Mediated Phosphorylation of hSpt5 C-Terminal Repeats Is Critical for Processive Transcription Elongation. *Mol. Cell* **21,** 227–237 (2006).
35. Malik, I., Qiu, C., Snavely, T. & Kaplan, C. D. Wide-ranging and unexpected consequences of altered Pol II catalytic activity in vivo. *Nucleic Acids Res.* doi:10.1093/nar/gkx037
36. Kawashima, T., Douglass, S., Gabunilas, J., Pellegrini, M. & Chanfreau, G. F. Widespread Use of Non-productive Alternative Splice Sites in Saccharomyces cerevisiae. *PLoS Genet* **10,** e1004249 (2014).
37. Kaplan, C. D. Basic Mechanisms of RNA Polymerase II Activity and Alteration of Gene Expression in Saccharomyces cerevisiae. *Biochim. Biophys. Acta* **1829,** 39–54 (2013).
38. Kaplan, C. D., Larsson, K.-M. & Kornberg, R. D. The RNA Polymerase II Trigger Loop Functions in Substrate Selection and Is Directly Targeted by α-Amanitin. *Mol. Cell* **30,** 547–556 (2008).
39. Bar-Nahum, G. *et al.* A ratchet mechanism of transcription elongation and its control. *Cell* **120,** 183–193 (2005).
40. Varsally, W. & Brogna, S. UPF1 involvement in nuclear functions. *Biochem. Soc. Trans.* **40,** 778–783 (2012).
41. Risler, J. K., Kenny, A. E., Palumbo, R. J., Gamache, E. R. & Curcio, M. J. Host co-factors of the retrovirus-like transposon Ty1. *Mob. DNA* **3,** 12 (2012).
42. Matsuda, E. & Garfinkel, D. J. Posttranslational interference of Ty1 retrotransposition by antisense RNAs. *Proc. Natl. Acad. Sci.* **106,** 15657–15662 (2009).
43. Schor, I. E., Acuña, L. I. G. & Kornblihtt, A. R. in *RNA and Cancer* (ed. Wu, J. Y.) 1–24 (Springer Berlin Heidelberg, 2013). doi:10.1007/978-3-642-31659-3_1
44. Dujardin, G. *et al.* Transcriptional elongation and alternative splicing. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1829,** 134–140 (2013).
45. Herzel, L. & Neugebauer, K. M. Quantification of co-transcriptional splicing from RNA-Seq data. *Methods San Diego Calif* **85,** 36–43 (2015).

Chapter 4 Studying splicing efficiency and TSS selection genome-wide in elongation mutants

4.1 Acknowledgement

Statistical analyses were carried out in collaboration with Yuanhua Huang (School of Informatics, the University of Edinburgh).

4.2 Introduction

In budding yeast only about 5% of genes contain an intron, although these genes produce about 27% of total mRNA, as many intron-containing genes are highly expressed. Tuning the splicing efficiency of intron-containing transcripts plays important roles during meiosis and under diverse environmental stresses[1–3]. For example, the *MER2* gene is constitutively transcribed in mitosis as well as meiosis and is efficiently spliced only in meiosis, when meiosis-specific mRNA splicing factor, Mer1, is expressed[1]. In addition, amino acid starvation inhibits the splicing of the majority of RP transcripts, while toxic levels of ethanol inhibits splicing of a different group of transcripts[2]. These observations have demonstrated that rapid changes in splicing efficiency can allow fast changes in gene expression in response to environmental stresses.

Splicing efficiency is traditionally presented as the amount of pre-mRNA divided by the amount of mRNA for each gene. RT-qPCR is widely used to measure pre-mRNA and mRNA levels, with primers spanning the exon-intron boundary and exon-exon junction respectively. Despite the high sensitivity of the RT-qPCR approach, it is only feasible to measure splicing efficiency for a limited number of genes. To measure splicing efficiency in RNAPII elongation mutants, Braberg et al performed splicing-specific microarray analysis of steady state RNA[4]. The splicing

microarray chip contain three different oligonucleotide probes that allow for the specific quantification of differences in the levels of the pre-mRNA, mRNA and exon2 (measuring total transcript) of intron-containing transcripts between WT and transcription elongation mutants. Splicing efficiency was measured by analysing the behaviour of pre-mRNA and mRNA against the background changes in the total level of transcript. This approach revealed that fast transcription reduces and the slower transcription increases splicing efficiency for most genes[4]. There are several lines of evidence showing that changes in RNAPII elongation rate alter alternative splicing decisions in mammalian cells and on a reporter gene in budding yeast[5–9]. Although the frequency of alternative splicing is low in yeast (discussed next chapter), studies have shown that RNAPII may slow down or pause to enhance co-transcriptional splicing[10–13]. Indeed, microarray analysis demonstrated that a slow mutant favours and a fast mutant reduces co-transcriptional splicing efficiency likely via stretching or shortening the time window available for spliceosome assembly with the slow and fast mutants, respectively. Here, sequencing of RNA was used to carry out a comprehensive splicing efficiency analysis at the genome-wide scale in yeast RNAPII elongation mutants. The classic way of measuring splicing efficiency from RNA-seq data is based on comparing read counts of intronic (representing pre-mRNA), exon-exon junction (representing mRNA) and also exonic (representing total transcript) regions. Here, pre-mRNA fraction (pre-mRNA reads / (pre-mRNA reads + mRNA reads) was used to compare splicing efficiency between strains with the DICEseq package[14]. This has allowed new insights into co-transcriptional pre-mRNA splicing and showed that fast transcription mostly reduces splicing efficiency of the RP transcripts and the slow mutant reduces splicing efficiency for most of the transcripts.

In addition, RNA has enabled investigation into the effects of altering elongation rate on exon skipping and TSS and PAS selection genome wide. There are many studies with human cells showing that transcription elongation rates tune the rate of alternative exon inclusion and exclusion in transcripts with multiple exons[15,16]. In contrast, yeast only has 10 genes with three exons, including *DYN2*, and enhanced

exon skipping of *DYN2* with a slow RNAPII was the first evidence of the effect of elongation rate on modulating exon skipping in buddying yeast[6]. Exon skipping was detected by RT-qPCR in other transcripts in yeast strains with inactivated RNA decay pathways but their physiological importance mostly remains unknown[17,18]. `This chapter describes an RNA-seq analysis to investigate the effect of fast and slow transcription on exon skipping for more genes. Furthermore, recent studies showed that RNAPII elongation mutants alter TSS selection pattern likely due to changes in efficiency of nucleotide selection and incorporation rate[4,20]. TSS pattern was examined transcriptome wide for the first time here with RNA-seq data, showing that the fast mutant shifts TSS upstream at most of genes and the slow mutant shifts TSS slightly downstream, consistent with previous primer extension analysis[4,20]. Also, RNA-seq data provide new insights regarding the effect of altering transcription rate on PAS selection.

4.3 Next generation RNA sequencing

To investigate splicing efficiency and fidelity genome-wide (fidelity is discussed in next chapter) in RNAPII elongation mutants, high depth RNA sequencing was performed on total RNA. Library preparation and next generation RNA sequencing was carried out at the Beijing Genomics Institute (BGI). After receiving RNA sequencing data from BGI, raw sequencing reads were quality controlled with FastQC[21]. All samples had good quality. STAR aligner was used to align the reads to the *S. cerevisiae* genome (sacCer3). STAR is ultra-fast and finds exon-exon junctions with high precision mapping strategy[22]. RNA sequencing was performed in two biological replicates. On average, 85 million unique paired-end reads were mapped successfully to the genome per sample (Table1). As shown in figure 4.1, the majority of the reads mapped to protein coding transcripts in all three strains. However, comparing the fast mutant with WT strain shows that expression of the protein coding

genes reduced almost 5%, whereas, expression of non-coding RNAs increased proportionally.

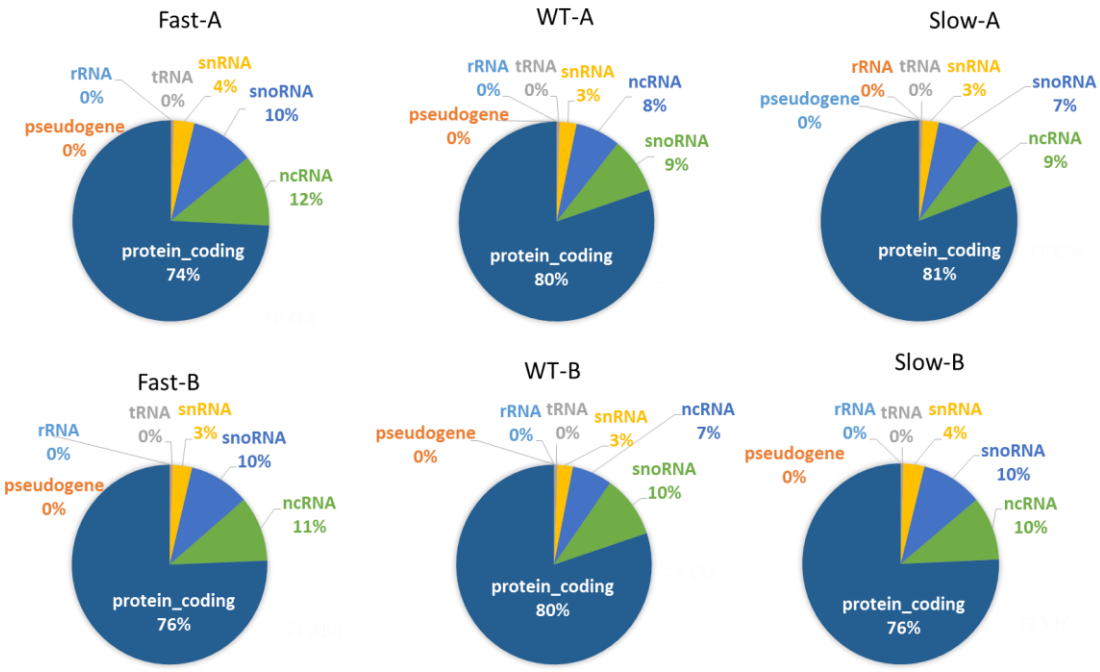


Figure 4.1: Percentage of the reads mapped uniquely to pseudogenes, rRNAs, tRNAs, snRNAs, ncRNAs, snoRNAs and protein coding genes in each sample. ENSEMBL GTF file, version R64-1-1.75 was used for generating this figure.

Table 4.1: Showing total uniquely mapped reads in each sample

Sample	Input reads	Mapped reads	Average fragment Size
WT-A	92558381	82437911	297.98
WT-B	88679441	76543116	298.23
FAST-A	102505775	92443485	298.25
FAST -B	102770598	92603013	298.23
SLOW-A	91390436	73547492	298.14
SLOW-B	115494388	102656123	298.29

4.4 Global changes in co-transcriptional splicing efficiency in RNAPII elongation mutants

As discussed in chapter 3, 4tU labelling, and IP of RNAPII showed that the fast strain reduces co-transcriptional splicing but the slow mutant increases co-transcriptional splicing. Additionally, RT-qPCR of steady state RNA revealed that the fast mutant increases and the slow mutant decreases pre-mRNA/mRNA ratio. Overall, these results indicated that the fast and slow mutants decrease and increase co-transcriptional splicing efficiency, respectively. To investigate how co-transcriptional splicing efficiency changes genome-wide in elongation mutants, splicing efficiency was quantified for all strains using RNA sequencing with the DICEseq package[14]. The transcripts that did not have enough intron or junction coverage, such as meiosis specific transcripts, were excluded from further analysis.

It has been suggested that RP transcripts are more co-transcriptional than non-RP transcripts[23,24]. Thus, splicing efficiency was measured separately for RP and non-RP transcripts. Analysis showed that the fast mutant increases pre-mRNA/mRNA ratio for many RP transcripts but the slow mutant reduces this ratio for both RP and non-RP transcripts (figure 4.2a). Reanalysis of published[4] data from a splicing-specific microarray analysis of steady state RNA from the same fast RNAPII mutant as used here, and a different slow mutant, showed the same effect (figure 4.2b). Furthermore, analysing pre-mRNA and mRNA of several genes in the mutants revealed a negative correlation, so that fast mutant leads to accumulation of pre-mRNA and consequently reduction in mRNA levels, but more mRNA and less pre-mRNA were detected with the slow mutant compared to WT (figure 4.2c).

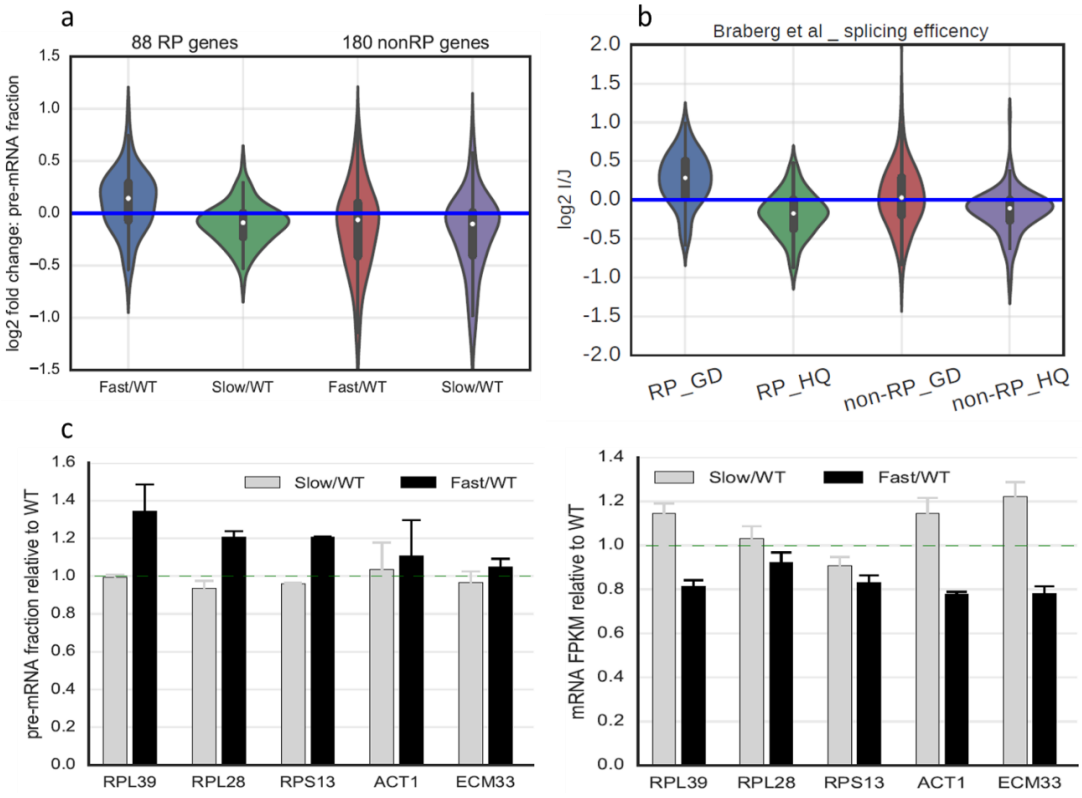


Figure 4.2: (**a**) pre-mRNA fraction for the fast and the slow mutant in RP and non-RP transcripts, measured by DICEseq measured in collaboration with Yuanhua Huang from RNA sequencing data. Y axis shows pre-mRNA fraction in log2 scale normalized to WT. Horizontal blue line in Y axis is WT level. (**b**) Splicing efficiency of RP and non-RP transcripts measured by splicing-sensitive microarrays. Data were downloaded from Braberg et al (2013) supplementary file. G1097D (GD) is fast mutant and H1097Q (HQ) is slow mutant. Y axis shows intron over junction ratio in log2 scale normalized to WT. Horizontal blue line in Y axis is WT level. (**c**) pre-mRNA fraction (*left panel*, pvalue < 0.00498) and mRNA levels (*right panel*, pvalue < 0.00244) of five example genes measured by DICEseq from RNA sequencing data.

4.5 Exon skipping in transcripts containing two introns

At least 95 percent of mammalian transcripts undergo alternative splicing but only 10 genes in *S. cerevisiae* contain two introns and evidence of alternative splicing was found for few of these by RT-PCR[17,18]. To search for evidence of exon skipping in RNA-seq data, reads that spanned the exon1-exon3 junction were extracted for each

sample and the frequency of exon skipping was measured relative to number of reads that aligned to exon1-exon2 junction (figure 4.3a). Analysis showed a 3-fold increase in skipping of the *TAD3* second exon in the fast mutant (figure 4.3b). Exon skipping is detected with lower frequency in other two-intron transcripts but with insufficient read density to draw a clear conclusion (figure 4.3c). There is no sign of exon skipping for the highly expressed *RPL7A* and *RPL7B*, the only RP transcripts with three exons (figure 4.3c).

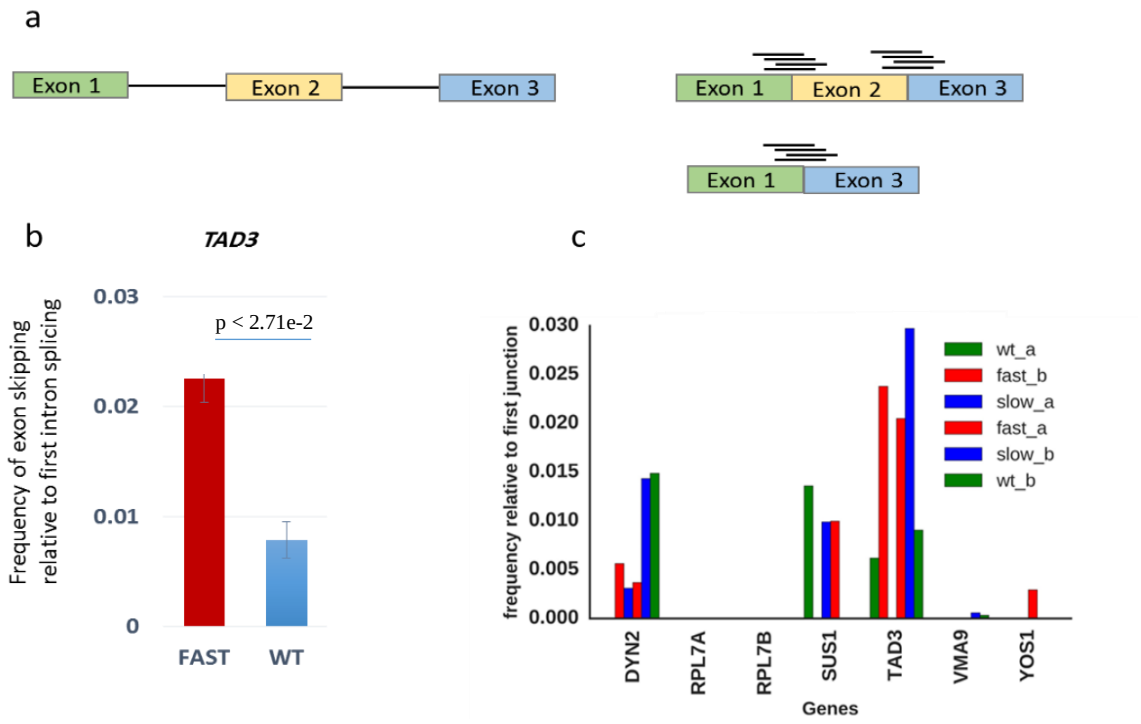


Figure 4.3: Fast mutant enhances exon skipping. (**a**) Diagram showing second exon skipping of *TAD3* transcript, joining the first and third exons. Short, black lines represent junction reads. (**b**) Fast mutant increased second exon skipping of the *TAD3* transcript more than two-fold. (**c**) Exon skipping in other transcripts with two introns. Skipping reads were not detected for *RPL7A* and *RPL7B*.

4.6 Splicing polarity

Fededa *et al* reported splicing with 5' to 3' polarity in human cells, such that inclusion rate of the upstream exon affects splicing efficiency of the downstream exon, but not vice versa[19]. Furthermore, inhibition of transcription elongation by 5,6-dichloro-1-

beta-D-ribobenzimidazole (DRB) treatment enhanced polarity, and when transcription was driven by the "more elongating" α-gb promoter, polarity was abolished[19,25]. A similar relationship was identified between the first and second introns in the yeast *SUS1* transcript[18]. To investigate splicing polarity in yeast transcripts that contain two introns, frequency of the first (E1) and second (E2) intron retentions were obtained by dividing intron read counts over junction read counts (figure 4.4a). Notably, fast transcription shows reduction in E_1/E_2 ratio ($p < 2.11e\text{-}2$, t-test) for transcripts with higher E_1/E_2 ratio in WT (*SUS1*, *RPL7B*, *YOS1*, *VMA9*), but there was not a significant change ($p > 0.05$) in this ratio for transcripts with lower E_1/E_2 ratio in WT (*RPL7A*, *DYN2*, *TAD3*) (figure 4.4b). These results indicate that the fast mutant tends to abolish polarity and relax the E_1/E_2 intron retention towards 1:1 ratio for the genes that have higher E_1/E_2 ratio in WT. In the slow mutant, retention of the first intron increased significantly ($p < 3.67e\text{-}2$) in *DYN2*, *SUS1*, *TAD3*, and *VMA9* relative to the second intron, but E_1/E_2 ratio in *RPL7A*, *RPL7B*, and *YOS1*, did not change significantly ($p > 0.05$) relative to WT.

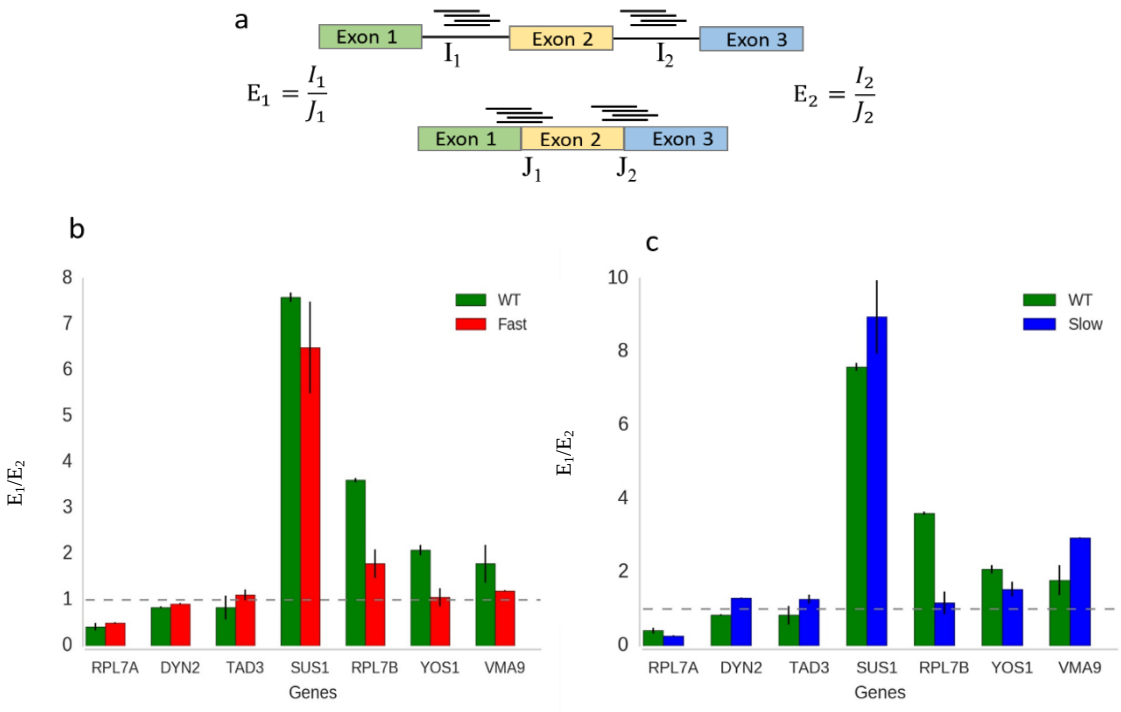


Figure 4.4: The fast mutant abolishes and the slow mutant enhances the splicing polarity. (**a**) Diagram showing the method used for measuring intron retention of the first and second introns. Short, black lines represent intronic or exonic reads. (**b**) Showing relative intron retention of the first intron to the second intron in 7 transcripts in WT (green) and the fast mutant (red). (**c**) Showing relative intron retention of the first intron to the second intron in 7 transcripts in WT (green) and the slow mutant (blue). Error bars denote two biological replicates.

4.7 Altering elongation rate affects transcription start site and polyadenylation site selection

It was shown previously by primer extension analysis that mutations that alter transcription elongation rate also alter transcription start site (TSS) selection *in vivo* for four tested genes[4,20]. For example, fast RNAPII lead to upstream shifts in TSS selection at *ADH1*, whereas a slow mutant shifted TSS downstream at this gene[4,20]. To investigate the effect of fast and slow elongation on TSS genome-wide, the distribution of the 5'ends of sequence reads that were partially aligned to 5' ends of coding sequences (CDS) was used to estimate the TSS distance from +1 bp of each CDS. The average distribution of TSS estimated via this method for the WT strain agrees with previous studies[4,26]. Consistent with primer extension analysis of RNA 5′-ends for *ADH1*[20], the fast and the slow mutant shifted TSS selection upstream and downstream, respectively, for *ADH1* and genome-wide (figure 4.5). Notably, the effect of fast transcription, shifting TSS upstream, is stronger than TSS shifting downstream in the slow mutant. Moreover, TSS at RP genes are shifted further upstream by the fast mutant compared with non-RP genes, indicating that TSS selection at highly expressed RP genes is more sensitive to alteration in transcription speed. Significantly, this shift is further enhanced at the RP intron-containing genes (figure 4.5). Figure 4.5c shows that RP genes have shorter 5'UTRs, which is consistent with previous reports[26]. Analysis of TSS usage also reveals that fast transcription robustly shifts TSS upstream at small nucleolar RNA (snoRNA) genes, but slow elongation does not show this effect (figure 4.6).

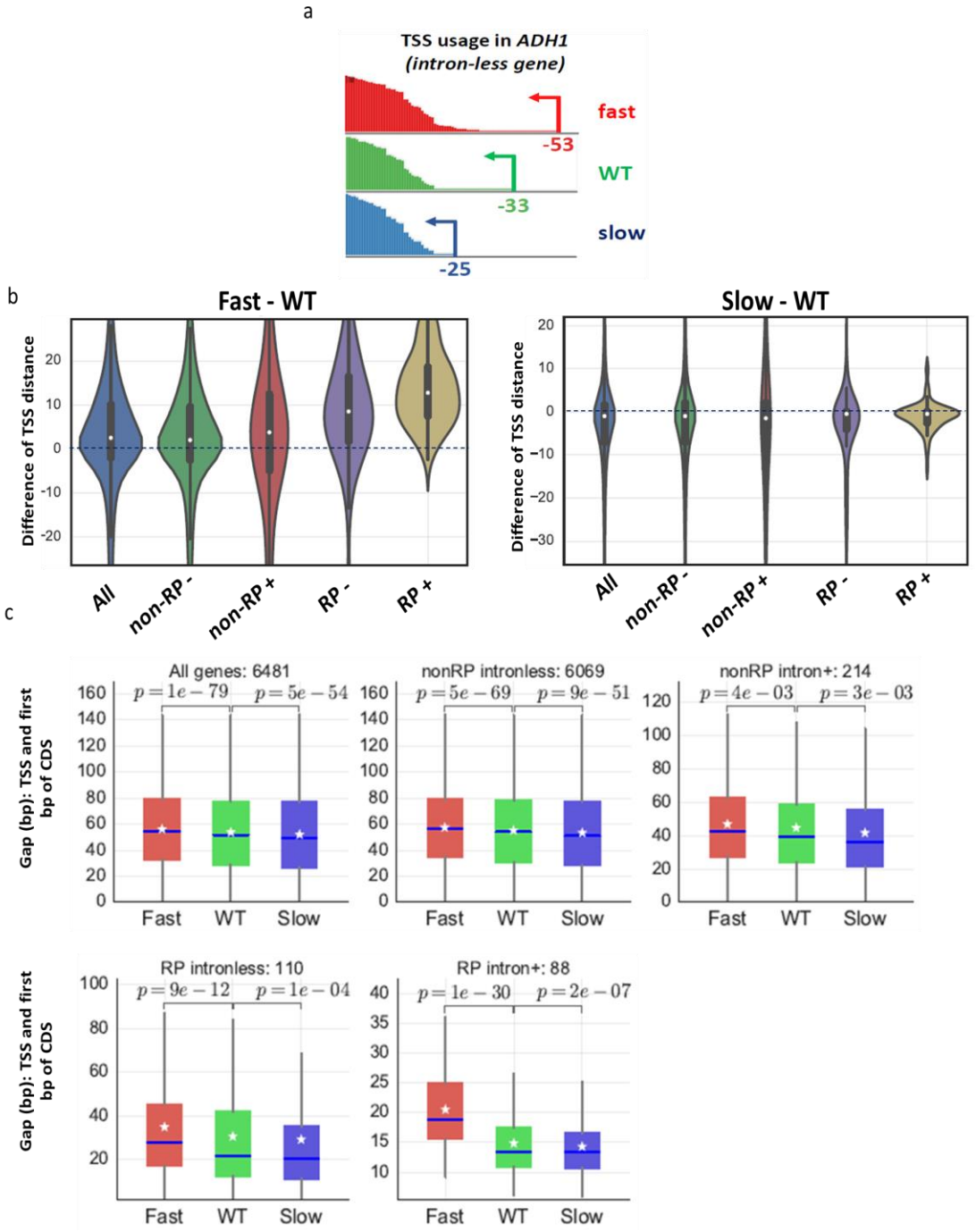


Figure 4.5: Elongation mutants change TSS selection. TSS changes estimated in collaboration with Yuanhua Huang from RNA sequencing data (**a**) Fast mutant shifts TSS upstream and slow mutant shifts it downstream at the *ADH1* locus. (**b**) Left panel shows difference in TSS distance between fast and WT strains from +1 bp of coding sequence for all yeast protein coding genes (All, 6481 genes), non-RP intronless genes (non-RP-, 6069 genes), non-RP intron-containing genes (non-RP+, 214 genes), RP

intronless genes (RP-, 110 genes) and RP intron-containing genes (RP+, 88 genes). Right panel shows difference in TSS distance between slow and WT strains. (**c**) Distance of TSS from first bp of CDS of different groups of genes for fast, slow, and WT strains. *Star* inside the boxes represents mean and blue line is the median. p-values are obtained using t-test.

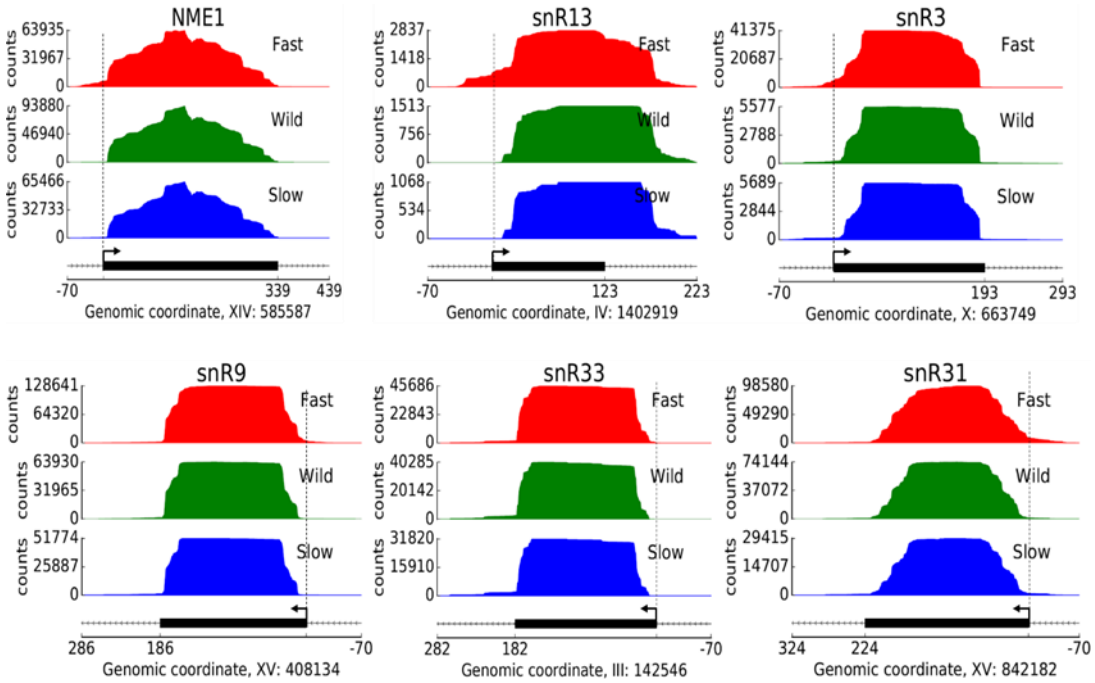


Figure 4.6: Density plots showing read counts and coverage for 6 different snoRNAs. Y axis shows read counts. Vertical dashed line is an arbitrary line to compare TSS profile of the three strains. Sashimi plots were generated by the MISO package[27].

The formation of the 3' ends of transcripts is an essential process, but the effect of altering transcription elongation rate on efficient and accurate cleavage and polyadenylation of protein coding transcripts is poorly understood. Here, the distribution of the 3'ends of sequence reads that were partially aligned to the 3' ends of CDS was used to estimate the distance from the last nt of each CDS to the cleavage/polyadenylation sites (PAS). The PAS distribution identified in this way for the WT strain agrees well with previous findings[28]. Analysis showed that, for most of the genes, slower transcription shifts PAS faintly upstream (shortening the 3'UTR) (figure 4.7a), whereas the fast mutant does not change the PAS distribution. In the case of 3'end processing sites for snoRNAs, analysis demonstrates that the elongation

mutants do not affect 3'end processing sites in snoRNAs and, in all the snoRNAs analysed here, the profile of 3'end processing sites in elongation mutants coincides with the WT profile (figure 4.6).

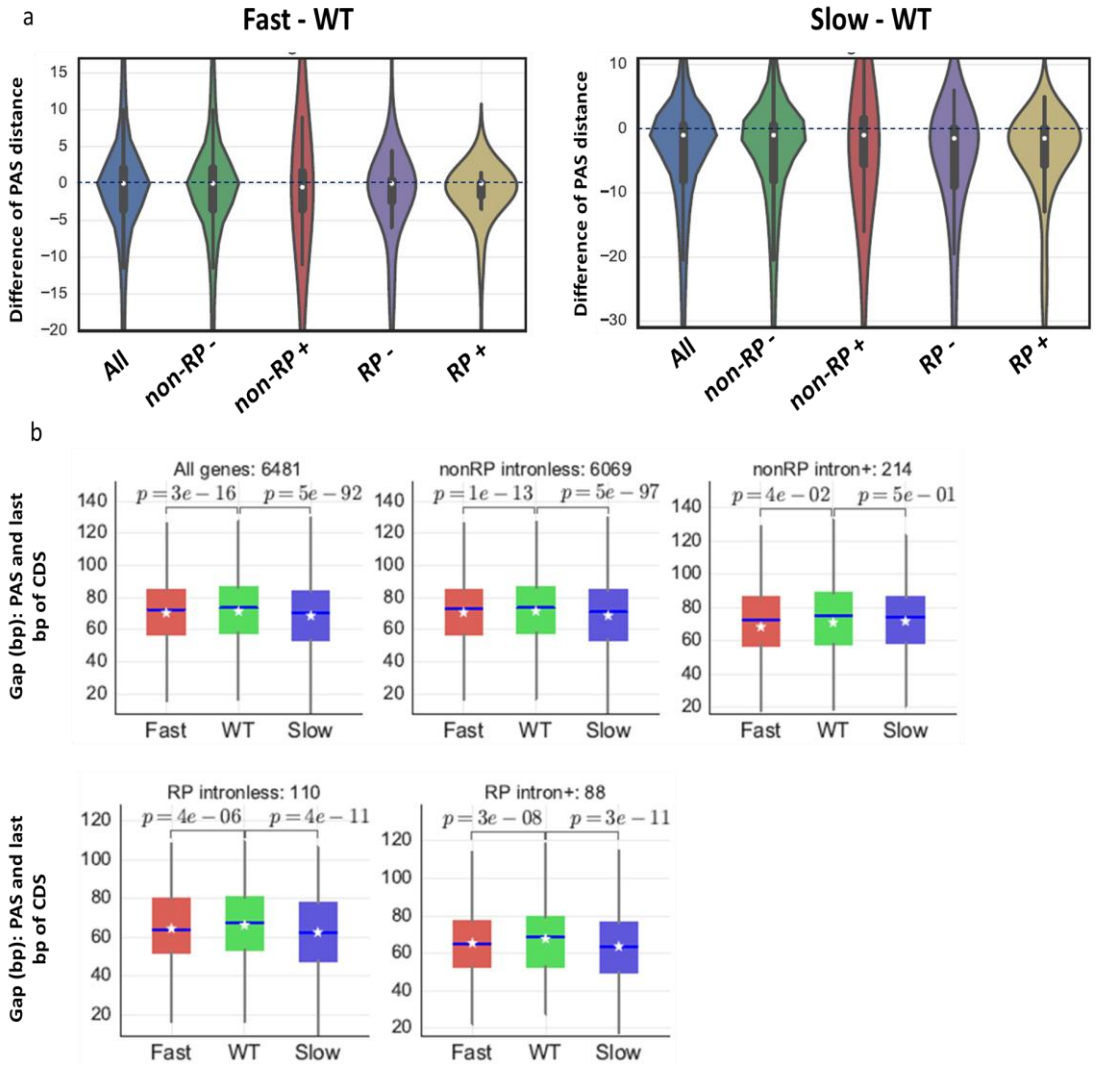


Figure 4.7: The slow mutant change PAS selection. PAS changes estimated in collaboration with Yuanhua Huang from RNA sequencing data. (**a**) Left panel shows difference in PAS distance between fast and WT strains from last bp of coding sequence for all yeast protein coding genes (All, 6481 genes), non-RP intronless genes (non-RP-, 6069 genes), non-RP intron-containing genes (non-RP+, 214 genes), RP intronless genes (RP-, 110 genes) and RP intron-containing genes (RP+, 88 genes). Right panel shows difference in PAS distance between slow and WT strains. (**b**) Distance of PAS from last bp of CDS of different group of the genes for fast, slow, and WT strains. *Star* inside the boxes is representing mean and blue line is the median. p-values are obtained using t-test.

4.8 Discussion

Recent high-throughput studies revealed that RP transcripts tend to be spliced faster, more efficiently and more co-transcriptionally than non-RP transcripts[29,30,24]. Nascent RNA analysis in chapter 3 showed that changes in elongation rate affect the amount of splicing that is co-transcriptional, with slower elongation resulting in more co-transcriptional splicing and vice versa. In this chapter, measurements of splicing efficiency by RNA-seq of total RNA showed that elongation rate affects splicing efficiency transcriptome-wide (figure 4.2). Analysis reveals that the overall reduced splicing efficiency in the fast mutant is due to an effect predominantly on RP transcripts, whereas the slow mutant improves splicing efficiency for both RP and non-RP transcripts (figure 4.2). Reanalysis of published microarray data[4] showed the same effect**.** Together, these observations suggest that splicing of RP transcripts is more functionally coupled to transcription than is splicing of non-RP transcripts, and may indicate that it is beneficial to splice co-transcriptionally. These results argue against the notion that most transcripts interact equally with the spliceosome, suggesting instead that transcript identity is an important factor in determining splicing efficiency. Moreover, analysis show that this is not simply due to there being more or less unspliced pre-mRNA because of different decay rates in the mutants, as the level of spliced mRNA is elevated in the slow mutant and reduced in the fast mutant compared to WT (figure 4.2). Different levels of pre-mRNA and mRNA in the mutants is in contradiction to the proposed RNA buffering mechanism in which RNA degradation compensates for different synthesis rates[31]. In that case, the cytoplasmic exonuclease Xrn1 showed the strongest effect on RNA level. It would be interesting to investigate whether such a buffering mechanism also operates in the nucleus, where RNA splicing takes place. One possible explanation for higher pre-mRNA and less spliced mRNA in the fast mutant at steady state (figure 4.2) could be that post-transcriptional splicing is less efficient than co-transcriptional splicing. Interestingly, stochastic modeling based on high-resolution kinetic analyses of transcription and splicing during induction of an intron-containing reporter gene in budding yeast revealed co-transcriptional

splicing to be the most efficient splicing pathway, due in part to a positive feedback mechanism for co-transcriptional second step splicing[12]. However, other interesting possibilities are that reduced splicing efficiency in the fast mutant is a consequence of inefficiently capped pre-mRNAs, changes in transcription start site, changes in pre-mRNA or mRNA stabilities, nascent RNA folding, chromatin remodelling, nucleosome occupancy, or failure of RNAPII to pause for splicing[10,32,33].

Using *rat1* and *rrp6* mutants, Egecioglu *et al* characterized second exon-skipped species from the *MATa1*, *DYN2*, *SUS1* and *YOS1* genes[17]. They argued that these species mainly are targeted by nuclear RNA surveillance and not by NMD. Because exon skipping is a rare event for these genes, its physiological importance was thought to be subtle and remained unknown. However, in 2011, two papers reported that *SUS1* alternative splicing responds to the cellular and environmental conditions and inefficient splicing of the pre-mRNA leads to defects in H2B deubiquitination, temperature sensitivity, mRNA export and cell growth[18,34]. Heat stress downregulates *SUS1* expression by preferentially retaining the first intron that is subject to *NMD* due to introducing a premature termination codon[18]. Interestingly, truncated protein encoded from the retained intron isoform partially suppressed *sus1* phenotypes. Exon skipping was detected with low abundance, approximately 5% of total, and was insensitive to heat stress. Deleting *MUD1* or the *LEA1* or *MSL1* genes that encode U2snRNP components enhanced the exon skipping, indicating that this event is regulated[18]. It has been shown that *SUS1* cDNA only suppresses the *sus1* phenotype partially, suggesting its alternatively spliced isoform is important for proper cellular function[18]. These results indicate that intron retention and exon skipping are regulated for *SUS1* and tight regulation of its alternative splicing is critical for the cells. Consistent with previous reports, RNA sequencing data analysis detected exon skipping for *VAM9*, *YOS1*, *SUS1*, *TAD3*, and *DYN2*. *TAD3* showed the highest number of the reads supporting exon skipping in WT, which was enhanced by fast transcription (figure 4.3). In the case of the slow mutant, there were insufficient reads to determine how slow transcription affects *TAD3* second exon usage. Additionally, in both

biological replicates there were not enough supporting reads for other genes to conclude how the fast and slow mutants affect their exon skipping/inclusion rate (figure 4.3). Not detecting exon-skipping species in steady state RNA could be due to their fast degradation by RNA surveillance mechanisms. Therefore, it would be better to assess the effects of altering transcription rate on exon skipping in nascent yeast RNA or using strains with inactivated RNA degradation pathway(s). Notably, no sequence reads were detected to indicate exon skipping for *RPL7A* or *RPL7B*, the only RP transcripts with two introns. It is possible that exon skipping does not occur on these transcripts, or only in certain physiological conditions. Alternatively, exon skipping might be tightly regulated in RP transcripts, which suggests a better splicing quality control for RP transcripts, a topic discussed in the next chapter.

Analysis showed that fast transcription relaxes splicing polarity in yeast transcripts containing two introns that have relatively higher first intron retention (*SUS1*, *RPL7B, YOS1, VMA9*) (figure 4.4b). This is indicating that faster transcription allows both introns to be exposed simultaneously to the spliceosome, which results in the relaxing of splicing polarity, i.e. differences between intron retention of the first intron and second intron becomes smaller (figure 4.4b). This shows the existence of splicing polarity in budding yeast and supports the idea that fast transcription decreases or eliminates polarity and splicing polarity is determined co-transcriptionally in yeast similar to mammalians. Not detecting a significant change in E_1/E_2 ratio for the transcripts (*DYN2* and *TAD3*) that this ratio is close to 1:1 ratio could be because these transcripts do not have splicing polarity or have a weak polarity. The fast mutant does not relaxes polarity for *RPL7A* that has higher second intron retention rate in WT. This is suggesting that fast transcription does not affect polarity for the transcripts that have higher second intron retention (less efficiently spliced). Further work is required to establish this and poses an interesting question for future research. If splicing polarity is determined co-transcriptionally, it is expected that the slow mutant would enhance splicing polarity and therefore reduces intron retention of the upstream intron relative to the downstream intron. However, slow transcription dose not significantly changes

E_1/E_2 ratio for *RPL7A*, *RPL7B* and *YOS1* and unexpectedly increased intron retention of the first intron relative to the second intron for *DYN2*, *SUS1*, *VMA9* and *TAD3* (figure 4.4c). It could be possible that changes in splicing polarity in the elongation mutants is a consequence of inefficiently capped pre-mRNAs, changes in transcription start site, changes in pre-mRNA or mRNA stabilities, nascent RNA folding, chromatin remodelling, nucleosome occupancy, or failure of RNAPII to pause faithfully for splicing. More research on this topic needs to be undertaken before the association between RNAPII elongation rate and splicing polarity is more clearly understood.

Recent primer extension analyses revealed that fast transcription shifts TSS upstream at *ADH1*, *HIS3*, *PMA1* and *GAL1* and slow transcription shifts TSS downstream at *ADH1* and *GAL1*, creating longer and shorter 5'UTRs, respectively[4,20]. There is accumulating evidence that 5'UTRs play important roles in post-transcriptional gene regulation and yet the role of 5'UTRs in regulating co-transcriptional processes is largely unknown[26,35,36]. According to the scanning model, start site selection occurs in a polar fashion, with abortive upstream initiation leading to increased productive downstream start site selection, and vice versa[20,37]. As described by Kaplan et al, fast RNAPII mutants with increased nucleotide incorporation activity make normally abortive upstream initiations become productive, whereas abortive downstream initiations become productive with a slow RNAPII that has reduced catalytic activity[20]. For the first time, RNA sequencing here provides new insights regarding the effect of altering transcription rate on selection of TSS genome-wide (figure 4.5). TSS shift mainly upstream with fast mutant and the strongest effect on TSS selection is observed with RP transcripts, shifting TSS further upstream. Slow transcription slightly shifts TSS downstream and there is not any difference between RP and non-RP transcripts (figure 4.5). The fact that the fast mutant has stronger effect on TSS than the slow mutant is compatible with previous reports[4,20]. From four tested genes by Kaplan et al[20], the fast mutant shifted TSS upstream in all four genes but the slow mutant shifted TSS downstream only with two genes[20]. Why TSS shifted further upstream in RP transcripts with the fast mutant is not clear and it might be correlate

with their higher expression. However, no correlation is detected between expression level of the transcripts and TSS selection in this study. It is also unclear why the fast mutant exerts a stronger effect on TSS selection of RP intron-containing genes. However, this also begs the question of whether changes in TSS exert a significant impact on splicing efficiency. It has been reported that mutations in general transcription factors TFIIB (sua7-3) and TFIIF (tfg2Δ261-273) that cause downstream and upstream TSS selection, respectively, do not alter splicing efficiency significantly[4]. Transcriptome-wide analysis here, for the fast mutant shows stronger TSS shifts for RP intron-containing transcripts that also show the strongest effect on splicing efficiency (figure 4.2 and Figure 4.5). It could be simply a coincidence that the fast transcription enhances upstream TSS selection in RP intron-containing transcripts and also reduces their splicing efficiency. However, further work is required to establish whether TSS selection and splicing efficiency are interdependent. Interestingly, the fast mutant shifts TSS selection upstream strongly with snoRNAs demonstrating that the effect of altering RNAPII elongation rate on TSS selection in mRNAs and snoRNAs is similar (figure 4.6 and Figure 4.5). Whether this would affect snoRNAs processing or their functions is an interesting question for future research. In the case of PAS selection, small changes were observed with the slow mutant moving cleavage/polyadenylation a few bases upstream compared to WT (figure 4.7). As 3' end formation is coupled with transcription, it seems likely that slow transcription enhances selection of upstream PAS by allowing more time for their recognition[38]. Use of alternative PAS could potentially affect stability, localization and translation efficiency of transcripts[39]. In contrast, for all snoRNAs analysed, the processed 3' ends are the same in the elongation mutants as in WT. Overall, these cases support the view that TSS selection in protein coding and snoRNAs is more sensitive to alterations in transcription elongation rate than selecting of PAS or processed 3' ends at these genes.

References

1. Engebrecht, J., Voelkel-Meiman, K. & Roeder, G. S. Meiosis-specific RNA splicing in yeast. *Cell* **66,** 1257–1268 (1991).
2. Pleiss, J. A., Whitworth, G. B., Bergkessel, M. & Guthrie, C. Rapid, Transcript-Specific Changes in Splicing in Response to Environmental Stress. *Mol. Cell* **27,** 928–937 (2007).
3. Bergkessel, M., Whitworth, G. B. & Guthrie, C. Diverse environmental stresses elicit distinct responses at the level of pre-mRNA processing in yeast. *RNA N. Y. N* **17,** 1461–1478 (2011).
4. Braberg, H. *et al.* From Structure to Systems: High-Resolution, Quantitative Genetic Analysis of RNA Polymerase II. *Cell* **154,** 775–788 (2013).
5. de la Mata, M. *et al.* A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. *Mol. Cell* **12,** 525–532 (2003).
6. Howe, K. J., Kane, C. M. & Ares, M. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. *RNA* **9,** 993–1006 (2003).
7. Ip, J. Y. *et al.* Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Res.* **21,** 390–401 (2011).
8. Dujardin, G. *et al.* How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping. *Mol. Cell* **54,** 683–690 (2014).
9. Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28,** 2663–2676 (2014).
10. Alexander, R. D., Innocente, S. A., Barrass, J. D. & Beggs, J. D. Splicing-Dependent RNA Polymerase Pausing in Yeast. *Mol. Cell* **40,** 582–593 (2010).
11. Oesterreich, F. C., Preibisch, S. & Neugebauer, K. M. Global Analysis of Nascent RNA Reveals Transcriptional Pausing in Terminal Exons. *Mol. Cell* **40,** 571–581 (2010).
12. Aitken, S., Alexander, R. D. & Beggs, J. D. Modelling Reveals Kinetic Advantages of Co-Transcriptional Splicing. *PLOS Comput Biol* **7,** e1002215 (2011).

13. Harlen, K. M. *et al.* Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. *Cell Rep.* **15,** 2147–2158 (2016).
14. Huang, Y. & Sanguinetti, G. Statistical modeling of isoform splicing dynamics from RNA-seq time series data. *Bioinformatics* **32,** 2965–2972 (2016).
15. Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* **14,** 153–165 (2013).
16. Naftelberg, S., Schor, I. E., Ast, G. & Kornblihtt, A. R. Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. *Annu. Rev. Biochem.* **84,** 165–198 (2015).
17. Egecioglu, D. E., Kawashima, T. R. & Chanfreau, G. F. Quality control of MATa1 splicing and exon skipping by nuclear RNA degradation. *Nucleic Acids Res.* **40,** 1787–1796 (2012).
18. Hossain, M. A., Rodriguez, C. M. & Johnson, T. L. Key features of the two-intron Saccharomyces cerevisiae gene SUS1 contribute to its alternative splicing. *Nucleic Acids Res.* **39,** 8612–8627 (2011).
19. Fededa, J. P. *et al.* A Polar Mechanism Coordinates Different Regions of Alternative Splicing within a Single Gene. *Mol. Cell* **19,** 393–404 (2005).
20. Kaplan, C. D., Jin, H., Zhang, I. L. & Belyanin, A. Dissection of Pol II Trigger Loop Function and Pol II Activity–Dependent Control of Start Site Selection In Vivo. *PLoS Genet* **8,** e1002627 (2012).
21. Andrews, S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. (2010). Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 20th March 2017)
22. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10,** 1185–1191 (2013).
23. Barrass, J. D. *et al.* Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. *Genome Biol.* **16,** 282 (2015).
24. Wallace, E. W. J. & Beggs, J. D. Extremely fast and incredibly close: co-transcriptional splicing in budding yeast. *RNA* rna.060830.117 (2017). doi:10.1261/rna.060830.117

25. Schor, I. E., Acuña, L. I. G. & Kornblihtt, A. R. in *RNA and Cancer* (ed. Wu, J. Y.) 1–24 (Springer Berlin Heidelberg, 2013). doi:10.1007/978-3-642-31659-3_1
26. Tuller, T., Ruppin, E. & Kupiec, M. Properties of untranslated regions of the S. cerevisiae genome. *BMC Genomics* **10,** 391 (2009).
27. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7,** 1009–1015 (2010).
28. Ozsolak, F. *et al.* Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation. *Cell* **143,** 1018–1029 (2010).
29. Barrass, J. D. *et al.* Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. *Genome Biol.* **16,** 282 (2015).
30. Carrillo Oesterreich, F. *et al.* Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* **165,** 372–381 (2016).
31. Sun, M. *et al.* Global Analysis of Eukaryotic mRNA Degradation Reveals Xrn1-Dependent Buffering of Transcript Levels. *Mol. Cell* **52,** 52–62 (2013).
32. Saldi, T., Cortazar, M. A., Sheridan, R. M. & Bentley, D. L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J. Mol. Biol.* **428,** 2623–2635 (2016).
33. Malik, I., Qiu, C., Snavely, T. & Kaplan, C. D. Wide-ranging and unexpected consequences of altered Pol II catalytic activity in vivo. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx037
34. Cuenca-Bono, B. *et al.* SUS1 introns are required for efficient mRNA nuclear export in yeast. *Nucleic Acids Res.* **39,** 8599–8611 (2011).
35. Mignone, F., Gissi, C., Liuni, S. & Pesole, G. Untranslated regions of mRNAs. *Genome Biol.* **3,** reviews0004.1-reviews0004.10 (2002).
36. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* **352,** 1413–1416 (2016).
37. Kuehner, J. N. & Brow, D. A. Quantitative Analysis of in Vivo Initiator Selection by Yeast RNA Polymerase II Supports a Scanning Model. *J. Biol. Chem.* **281,** 14119–14128 (2006).

38. Adamson, T. E., Shutt, D. C. & Price, D. H. Functional Coupling of Cleavage and Polyadenylation with Transcription of mRNA. *J. Biol. Chem.* **280,** 32262–32271 (2005).
39. Elkon, R., Ugalde, A. P. & Agami, R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* **14,** 496–506 (2013).

Chapter 5 Splicing fidelity is sensitive to transcription elongation rate

5.1 Introduction

More than 95% of human protein-coding transcripts are alternatively spliced and splicing patterns can be regulated distinctively during different developmental, tissue-specific, or pathology-specific stages[1,2]. Splicing quality control mechanisms are thought to monitor the configuration of splicing complexes and whether optimal versus suboptimal splice sites are appropriately distinguished within the splicing machinery[3]. Low specificity in recognition of splice sites would cause insertions, deletions, and frame shifts in mRNA. Recent findings have opened new insights, showing that alternative splicing could also be important for the expression of some budding yeast genes. For example, intron retention in *PTC7* transcripts and alternative splicing of the 3' end intron of *FES1* create mRNAs that code for different protein isoforms[4,5]. Also, alternative 5'SS use in *SRC1* transcripts results in different cellular localization of the resulting Src1 protein isoforms, and alternative 3'SS splicing upon heat shock regulates expression of the *APE2* gene[6,7]. Although, alternative splicing events in yeast that give rise to functional transcripts are rare, there are hundreds of non-productive events that introduce stop codons in mRNAs, which may couple with NMD to tune overall gene expression by down-regulating the nonsense spliced isoform[8–10].

In budding yeast, selecting the optimal splice sites to generate the correct spliced isoforms is determined by splicing factors (e.g. U1 and U2 snRNPs and associated proteins) and the selection is monitored by fidelity mechanisms. DExD/H-box ATPases are implicated in promoting fidelity at distinct steps in the splicing process to reject and discard suboptimal substrates[3,11]. Genetic mutations that disrupt ATPase activity of these factors was shown to increase the productive splicing of suboptimal substrates[11]. As most splicing occurs co-transcriptionally, quality control mechanisms that promote splicing fidelity should also occur co-transcriptionally.

While there is a robust body of evidence demonstrating that transcription elongation rate affects alternative splicing decisions in higher eukaryotes, only one study in yeast showed the effect of transcription rate on alternative splicing of a cassette exon using a *DYN2* reporter gene[2,12,13]. *DYN2* has two introns and mutating the branch point (BP) of the first intron enhances second exon skipping indicative of rejecting the suboptimal BP splicing by fidelity mechanisims[12]. Exon skipping was partially prevented with slower transcription showing that slower transcription reduces BP fidelity allowing to utilize suboptimal BP[12]. These observations support the hypothesis that the kinetic coordination of transcription elongation and splicing is important for splicing fidelity, in this case for correct recognition of the BP in yeast and rejecting suboptimal splicing substrates. Transcription-dependent BP fidelity raises the question of whether fidelity of selecting optimal vs suboptimal splice sites in yeast is adjusted by transcription elongation rate. A splicing-dependent transcriptional pause at the 3′ ends of introns has been proposed to correspond to a splicing fidelity checkpoint[14]. Furthermore, a branch point mutation caused RNAPII accumulation on the intron of a reporter gene suggesting the existence of a transcriptional elongation checkpoint that may promote splicing fidelity[15]. It is not known whether these quality control steps occur faithfully when transcription is faster or slower than normal.

How elongation rate affects splicing fidelity in budding yeast raises the question of whether faster and slower transcription have the opposite effect on splicing fidelity as might be predicted by the kinetic coupling model. Using deep RNA sequencing, for the first time, splicing fidelity was determined in yeast transcription elongation mutants in which *UPF1* was deleted in order to protect mis-spliced transcripts from nonsense-mediated decay (NMD). Results show that both fast and slow transcription reduce splicing fidelity mainly in ribosomal protein (RP) coding transcripts. Analysis reveals that splicing fidelity largely correlates with intron length, secondary structure and splice site score.

5.2 Filtering RNA sequencing data

After aligning the reads to the genome, all detected exon-exon splice junctions were divided into two groups, known and novel. The known junctions are those that are annotated in GTF file (ENSEMBL, version R64-1-1.75) and the novel junctions result from splicing events that are not annotated in the GTF file. To reduce noise and remove genes with low coverage some genes were excluded from further analysis. Mitochondrial genes containing group I or group II introns were excluded because they are not spliced by spliceosomes. Meiosis specific genes that are not expressed during vegetative growth were also filtered out from further analysis. In addition, genes containing more than one intron (10 genes in total) were excluded to simplify initial data analysis. Next, from the remaining intron-containing genes, those that did not have novel events or had less than 5 unique supporting reads for novel junctions were also filtered out. Table 5.1 lists the excluded genes and the reason for their exclusion.

Table 5.1 excluded genes	
HAC1	Not spliced by spliceosome
AI2	Mitochondrial genes
AI3	
AI4	
AI5_ALPHA	
BI2	
BI3	
BI4	
COX1	
COB	
AI5_BETA	
TAD3	have two introns
DYN2	
SUS1	
YOS1	
RPL7A	
RPL7B	
VMA9	
AML1	
HOP2	
EFM5	
AMA1	Meiosis-specific, only significantly expressed during meiosis
BUD25	

DMC1	
GMC2	
LDS2	
MEI4	
MND1	
PCH2	
REC102	
REC107	
REC114	
SAE3	
SPO1	
SPO22	
YHR218W	low coverage or no novel event
YLR464W	
YRF1-6	
YPL109C	
YDR535C	
ATG38	
YKR005C	
ASC1	
NOG2	
YLL067C	
YSF3	
YLR202C	
MRK1	
YRF1-3	
YIL177C	
YEL076C-A	
YBL111C	
YPR170W-B	
YML133C	
HFM1	
YBR219C	
QCR9	
YHL050C	
IMD4	
YJL225C	
YLL066C	
OSW2	
GIM4	
YBR220C	
MPT5	
YJR079W	
VMA10	
YRF1-7	

TEF4	
CGI121	
YPR202W	
YHR097C	

5.3 Altering RNAPII elongation rate affects splicing fidelity

After filtering, 244 novel splicing events were retained that had at least 5 unique reads in both replicates of one of the strains, and the splicing error frequency (SEF) was calculated for each novel event (figure 5.2a, see also Materials and Methods). Figure 5.1a shows 4 types of alternative splicing events studied in this chapter; alternative upstream 5' splice site (u5'SS), alternative downstream 5' splice site (d5'SS), alternative upstream 3' splice site (u3'SS), and alternative downstream 3' splice site (d3'SS). Figure 5.1b shows the number of novel junction reads detected in fast and slow mutants compared with wild type. It indicates that increasing transcription speed promotes use of novel junctions by the splicing machinery.

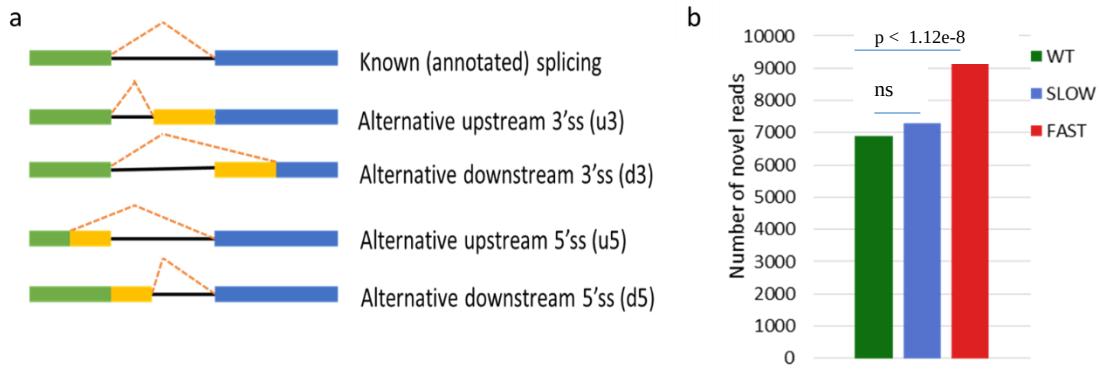


Figure 5.1 Alternative splicing events and their read counts in the mutants. (**a**) Different types of the alternative splicing events. (**b**) Total number of novel reads in each strain. To compare total amount of novel reads in each strain, reads normalized to sequencing depth.

Further analysing the novel events showed a marked difference in the average SEF of RP transcripts compared to that of non-RP transcripts in the WT strain (figure 5.2b). The SEF of RP transcripts ranges from 1 in 100 000 up to 1 in 250, while in the case of non-RP it ranges from 1 in 100 up to 7 in 100. Therefore, the observed average splicing error rate for RP transcripts is orders of magnitude lower than that of non-RP transcripts despite the higher expression level of RP genes. To investigate if novel splicing events are the consequence of random splicing errors, correlation between mRNA abundance and SEF was examined in all three strains. This shows that the SEF and the mRNA abundance are anti-correlated (figure 5.2c), indicating that the detection of alternative splicing events in high depth RNA-seq data is not an indirect result of the higher number of reads for transcripts with higher expression.

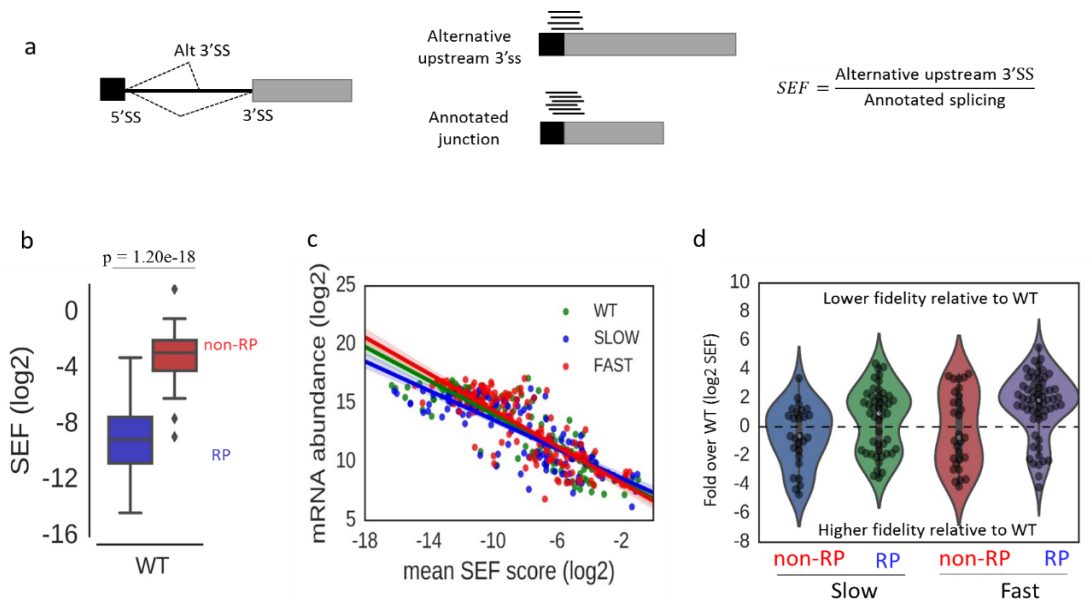


Figure 5.2: Changes in splicing error frequency in fast and slow strains. (**a**) Diagram showing how SEF of novel splicing events was measured using an alternative upstream 3'SS event as an example. (**b**) Distribution of the splicing error frequency in RP (blue box) and non-RP (red box) intron-containing transcripts in WT strain. The p-value was obtained by t-test. (**c**) Negative correlation between mRNA (both RP and non-RP) abundance and average splicing error frequency in fast (red), slow (blue) and WT (green). mRNA abundance was estimated from the number of reads aligned to the exon-exon junctions. Pearson correlation coefficient and p-value for each strain were: fast (R2=-0.82, p=1.02e-33), slow (R2=-0.81, p=1.80e-31) and WT (R2=-0.85, p=5.53e-38). (**d**) Violin plot showing distribution of

SEF of non-RP and RP intron-containing transcripts in fast and slow mutants normalized to WT. This plot includes all novel splicing events whose SEF was significantly different in mutants relative to WT (Fisher's exact test ($p < 0.01$). Points above dashed line (zero) are novel events with higher SEF than WT (reduced fidelity), points below dashed line are novel events with lower SEF than WT (improved splicing fidelity).

Using Fisher's exact test ($p < 0.01$) to determine which novel splicing events occur more or less frequently in the fast mutant relative to WT, 59 were identified in RP transcripts, of which 47 have significantly increased SEF but only 12 have a lower SEF (figure 5.2d). This reveals an ~4:1 ratio of reduction in splicing fidelity (higher SEF) over increase (lower SEF) in the fast mutant ($p < 5.16e-05$, Wilcoxon test). In comparison, 29 novel events in non-RP transcripts show significant changes in their SEF relative to WT, with a 1:1 ratio of reduction (15 events) versus increase (14 events) in fidelity observed for this group ($p > 0.05$). With the slow mutant 49 novel events were identified in RP transcripts, with 30 of these having higher SEF and 19 lower SEF, which reveals an ~1.5 ratio of reduction in splicing fidelity over increase ($p < 4.89e-02$) (figure 5.2d). In addition, 27 novel events were detected in non-RP with higher (12 events) or lower (15 events) SEF, which was similar to the 1:1 ratio in fidelity changes observed with the fast mutant ($p > 0.05$) (figure 5.2d). Notably, fast and slow transcription did not necessarily reduce or increase utilization of a specific event in opposite direction. For instance, utilization of an alternative downstream 5'SS in *RPL37A* increased only with fast transcription. However, both fast and slow mutants increased use of an alternative upstream 3'SS in *RPL26B* (figure 5.3a). Collectively, these results demonstrate that fidelity of splicing RP intron-containing transcripts is more sensitive to changes in transcription elongation rate than for non-RP transcripts. There are some novel events that occur with much higher frequency, for example in *IWR1* and *SPT14*, in which the alternative (annotated and novel) events occur with almost equal frequency, suggesting that they could potentially produce two major isoforms (figure 5.3b).

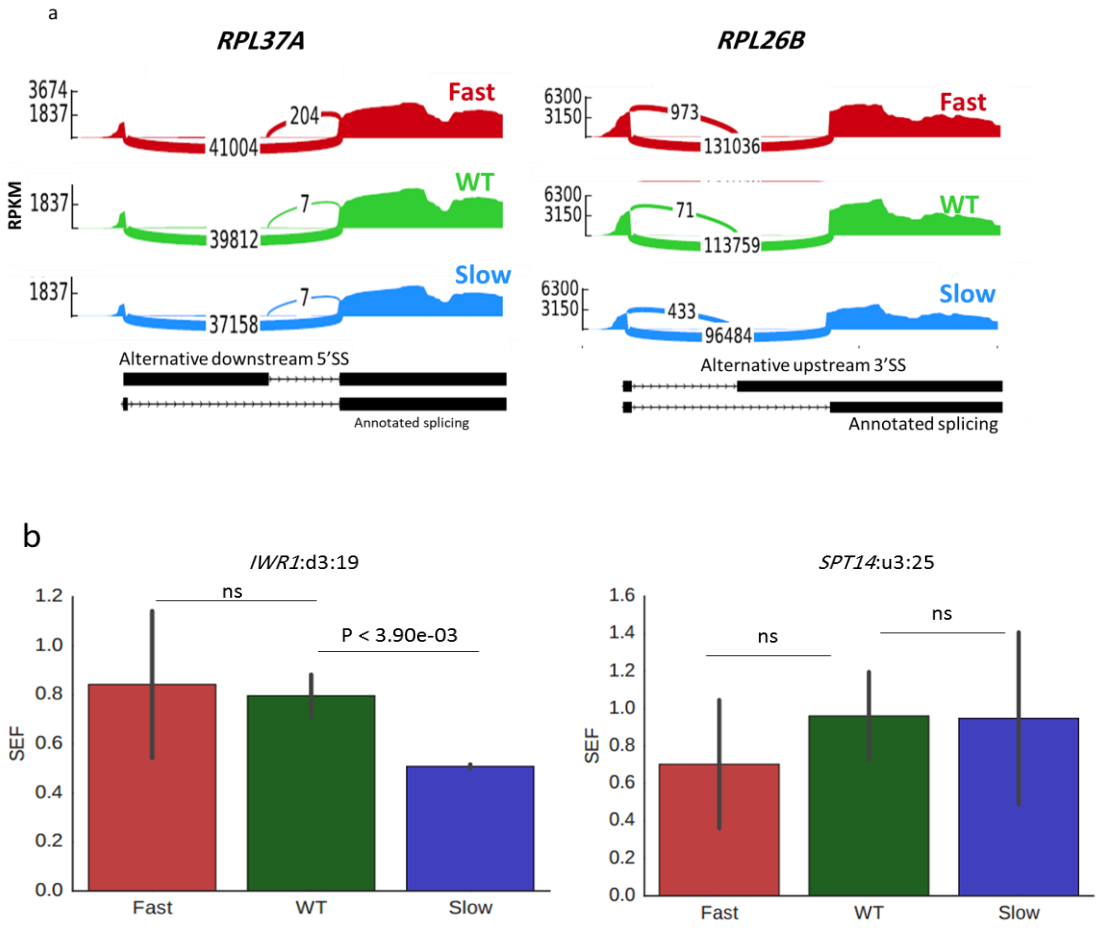
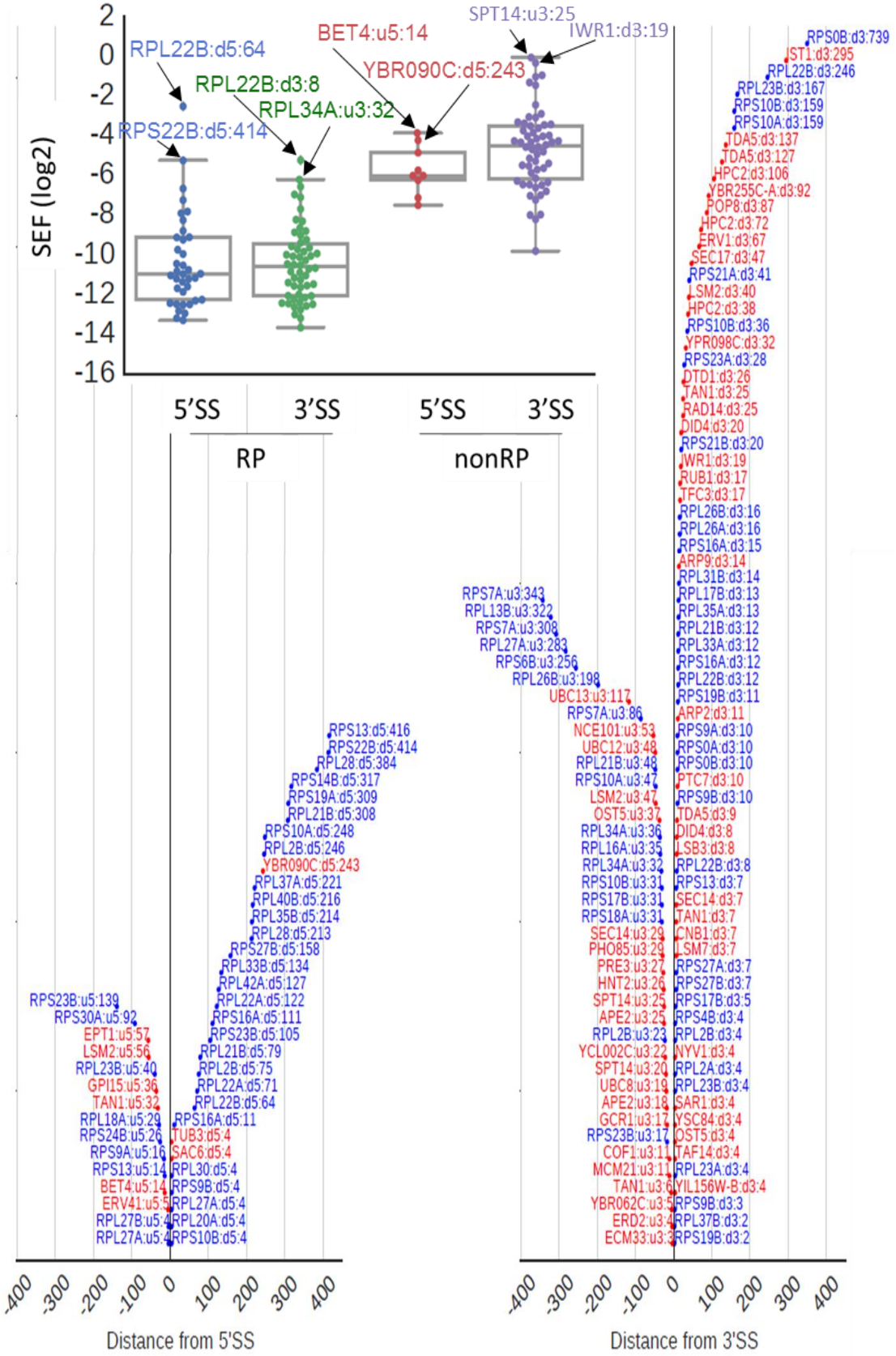


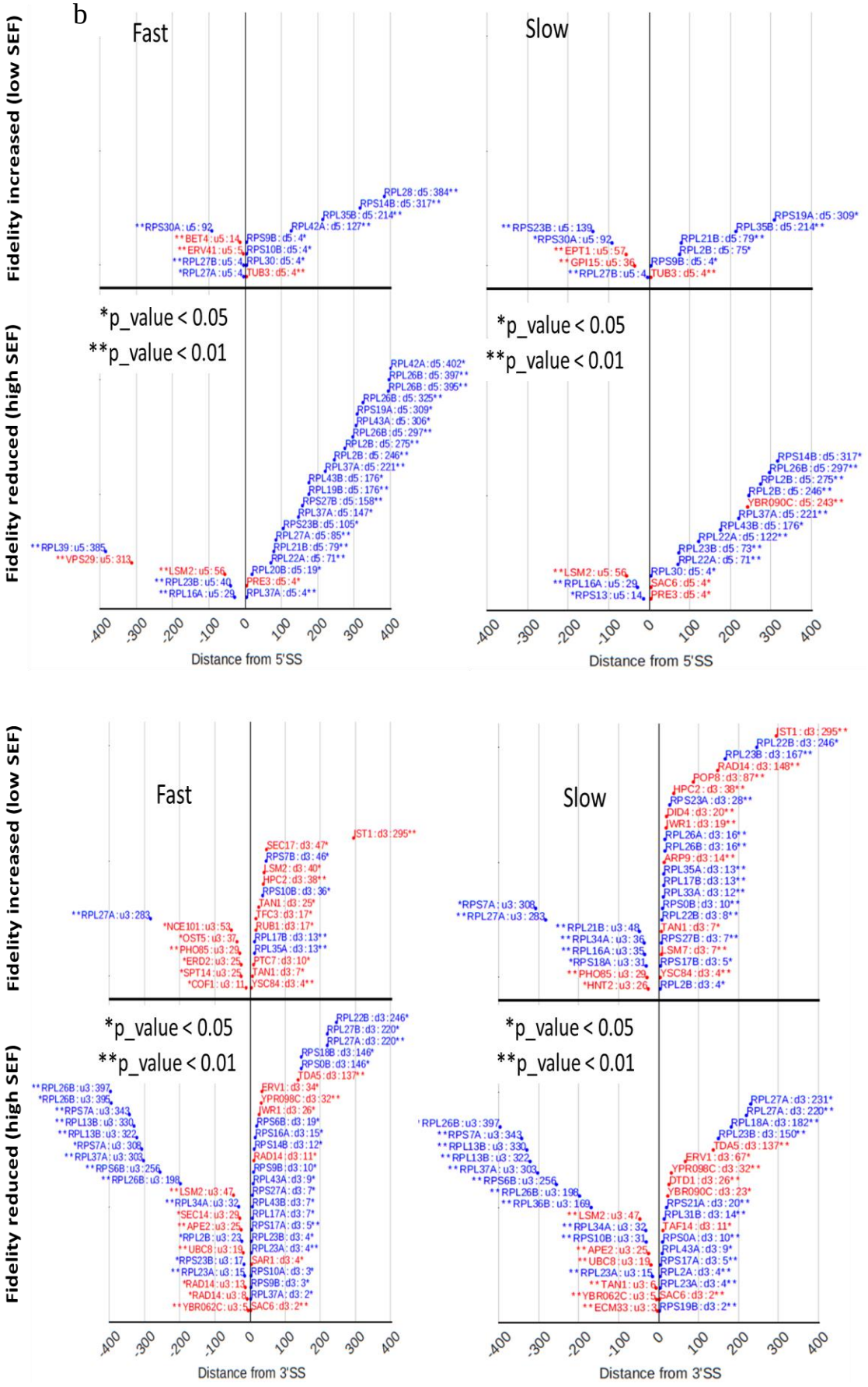
Figure 5.3: (**a**) Use of novel splice sites in *RPL37A* and *RPL26B*. Sashimi plots were generated with MISO package[16]. Arcs represent junction reads connecting first and second exon borders. Resulting isoforms drawn at the bottom of the plot. Y axis shows Reads Per Kilobase of transcript per Million mapped reads (RPKM) in each sample. (**b**) High splicing error frequency of two alternative splicing events in *IWR1* and *SPT14*. *IWR1*:d3:19 is a novel 3'SS 19 nucleotides downstream of the annotated 3'SS. *SPT14*:u3:25 is a novel 3'SS 25 nucleotides upstream of the annotated 3'SS. Error bars represent two biological replicates.

5.4 Distance of alternative novel splice sites from annotated splice site

The position distribution of the novel alternative splicing events in the WT strain shows that they occur less than ~400 nucleotides from the annotated splice sites (figure 5.4a). Strikingly, most of the alternative 3' splice sites (3'SSs) were found in a very short distance from the annotated 3'SS and there were only half as many alternative 5'SSs. Although there are twice more alternative 3'SS events, as can be seen in boxplot in figure 5.4a, alternative 3'SSs and 5'SSs have similar SEF distributions in RPs and non-RPs. Figure 5.4b shows the positions of novel splicing events whose SEF is significantly different in the fast or slow mutant compared to the WT. As can be seen in this figure, the number of distinct novel 5'SSs and 3'SSs with higher SEF than the WT (lower fidelity) is greater in the fast than the slow mutant. This suggests there is more reduction in splicing fidelity with the fast mutant as mentioned earlier (see figure 5.2). Because the first exon is generally short in yeast, the majority of the u5'SS events occur less than ~50 nucleotides distance from the annotated 5'SS (a5'SS). However, *RPL39* and *VPS29* have u5'SS occurring 358 and 319 nucleotides upstream of a5'SS, respectively, with the fast mutant (figures 5.4a,c). This indicates that transcription started upstream in these genes with the fast mutant, resulting in utilization of these cryptic u5'SSs. Unlike most of the novel 3'SS events that are very close to annotated 3'SS likely due to optimal distance constraint from branch point (BP) (BP-3'SS)[7], d5'SS events spread in a window of 400 nucleotides from a5'SS (figure 5.4a,b). This is probably due to proximity of the BP to the a3'SS that gives more space (5'SS-BP) for d5'SS events to occur. The majority of the novel alternative splicing events introduced premature translation termination codons in the coding region of the spliced RNA, which would normally be recognized by the NMD system leading to early transcript degradation.

a


SEF (log2)
2
0
-2
-4
-6
-8
-10
-12
-14
-16
RPL22B:d5:64
RPL22B:d3:8
RPS22B:d5:414
RPL34A:u3:32
BET4:u5:14
YBR090C:d5:243
SPT14:u3:25
IWR1:d3:19
5'SS
3'SS
5'SS
3'SS
RP
nonRP
RPS13:d5:416
RPS22B:d5:414
RPL28:d5:384
RPS14B:d5:317
RPS19A:d5:309
RPL21B:d5:308
RPS10A:d5:248
RPL2B:d5:246
YBR090C:d5:243
RPL37A:d5:221
RPL40B:d5:216
RPL35B:d5:214
RPL28:d5:213
RPS27B:d5:158
RPL33B:d5:134
RPL42A:d5:127
RPL22A:d5:122
RPS16A:d5:111
RPS23B:d5:105
RPL21B:d5:79
RPL2B:d5:75
RPL22A:d5:71
RPL22B:d5:64
RPS16A:d5:11
TUB3:d5:4
SAC6:d5:4
RPL30:d5:4
RPS9B:d5:4
RPL27A:d5:4
RPL20A:d5:4
RPS10B:d5:4
RPS23B:u5:139
RPS30A:u5:92
EPT1:u5:57
LSM2:u5:56
RPL23B:u5:40
GPI15:u5:36
TAN1:u5:32
RPL18A:u5:29
RPS24B:u5:26
RPS9A:u5:16
RPS13:u5:14
BET4:u5:14
ERV41:u5:5
RPL27B:u5:4
RPL27A:u5:4
-400
-300
-200
-100
0
100
200
300
400
Distance from 5'SS
RPS7A:u3:343
RPL13B:u3:322
RPS7A:u3:308
RPL27A:u3:283
RPS6B:u3:256
RPL26B:u3:198
UBC13:u3:117
RPS7A:u3:86
NCE101:u3:53
UBC12:u3:48
RPL21B:u3:48
RPS10A:u3:47
LSM2:u3:47
OST5:u3:37
RPL34A:u3:36
RPL16A:u3:35
RPL34A:u3:32
RPS10B:u3:31
RPS17B:u3:31
RPS18A:u3:31
SEC14:u3:29
PHO85:u3:29
PRE3:u3:27
HNT2:u3:26
SPT14:u3:25
APE2:u3:25
RPL2B:u3:23
YCL002C:u3:22
SPT14:u3:20
UBC8:u3:19
APE2:u3:18
GCR1:u3:17
RPS23B:u3:17
COF1:u3:11
MCM21:u3:11
TAN1:u3:6
YBR062C:u3:5
ERD2:u3:4
ECM33:u3:3
RPS0B:d3:739
ST1:d3:295
RPL22B:d3:246
RPL23B:d3:167
RPS10B:d3:159
RPS10A:d3:159
TDA5:d3:137
TDA5:d3:127
HPC2:d3:106
YBR255C-A:d3:92
POP8:d3:87
HPC2:d3:72
ERV1:d3:67
SEC17:d3:47
RPS21A:d3:41
LSM2:d3:40
HPC2:d3:38
RPS10B:d3:36
YPR098C:d3:32
RPS23A:d3:28
DTD1:d3:26
TAN1:d3:25
RAD14:d3:25
DID4:d3:20
RPS21B:d3:20
IWR1:d3:19
RUB1:d3:17
TFC3:d3:17
RPL26B:d3:16
RPL26A:d3:16
RPS16A:d3:15
ARP9:d3:14
RPL31B:d3:14
RPL17B:d3:13
RPL35A:d3:13
RPL21B:d3:12
RPL33A:d3:12
RPS16A:d3:12
RPL22B:d3:12
RPS19B:d3:11
ARP2:d3:11
RPS9A:d3:10
RPS0A:d3:10
RPS0B:d3:10
PTC7:d3:10
RPS9B:d3:10
TDA5:d3:9
DID4:d3:8
LSB3:d3:8
RPL22B:d3:8
RPS13:d3:7
SEC14:d3:7
TAN1:d3:7
CNB1:d3:7
LSM7:d3:7
RPS27A:d3:7
RPS27B:d3:7
RPS17B:d3:5
RPS4B:d3:4
RPL2B:d3:4
NYV1:d3:4
RPL2A:d3:4
RPL23B:d3:4
SAR1:d3:4
YSC84:d3:4
OST5:d3:4
TAF14:d3:4
RPL23A:d3:4
YIL156W-B:d3:4
RPS9B:d3:3
RPL37B:d3:2
RPS19B:d3:2
-400
-300
-200
-100
0
100
200
300
400
Distance from 3'SS


b
Fast
Slow
Fidelity increased (low SEF)
Fidelity reduced (high SEF)
*p_value < 0.05
**p_value < 0.01
Distance from 5'SS
Distance from 3'SS

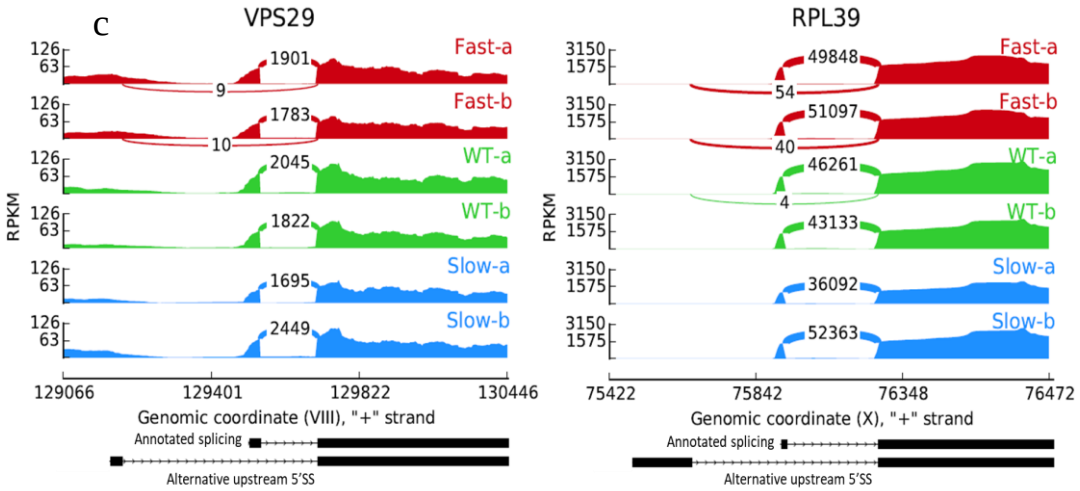


Figure 5.4: (**a**) Distribution of novel splice sites that have at least 5 unique reads in the WT strain, plotted relative to the annotated 5'SS (left plot) or 3'SS (right plot). d3 is downstream and u3 is upstream novel 3'SSs, and d5 is downstream and u5 is upstream novel 5'SSs. Last number at each point is the distance in bases from annotated splice site. Boxplot is showing SEF distribution of the alternative 5'SSs and 3'SS in RPs and non-RRs. Two events with the highest SEF are highlighted for each group. (**b**) Distribution of novel splice sites whose SEF is significantly different in the fast and slow mutants relative to WT. * is p_value < 0.05, ** is p_value < 0.001. (**c**) Use of U5'SS in *RPL39* (-385 bp from annotated 5'SS) and *VPS19* (-319 bp from annotated 5'SS) increased with fast transcription. Sashimi plots were generated as describe for figure 5.3.

Meyer et al[7] have suggested that spliceosomes can use any 3'SS located within an optimal distance (10 to 50 nucleotides (nt)) downstream of the BP, and that achieving and selecting the optimal 3'SS is strongly influenced by RNA structures that make optimal 3'SS more accessible while masking suboptimal 3'SSs. Here, the results show that novel upstream 3'SSs are mainly restricted to within the annotated BP-3'SS region and they occur in a window of ~50nt from the annotated 3'SS (figure 5.4a). Novel upstream 3'SS events that occur further upstream of the annotated BP probably make use of suboptimal BPs further upstream. For example, there are two upstream novel 3'SS events in *RPL13B* located 330nt and 322nt from annotated 3'SS. These events probably make use of a cryptic BP (AACTAAT) which is detected by lariat sequencing[7] and located within the optimal distance from these novel 3'SS, 21 and 13 nucleotides, respectively (figure 5.5a). Interestingly, the SEF of these novel events in *RPL13B* and similar events in other genes were increased by both faster and slower

transcription, most likely due to increased utilization of cryptic BPs further upstream of the annotated BP (figure 5.4b). Howe et al showed that mutations that create a suboptimal BP sequence in the first intron of *DYN2* caused second exon skipping and that the defect was partially alleviated by slower transcription[12]. This suggests that slow transcription provides more time for the assembling spliceosome to select the suboptimal BP (within the first intron) before transcription of the competing downstream BP (within the second intron), indicative of reduced BP fidelity. However, results presented here suggest that both fast and slow mutants reduce BP fidelity and thereby increase utilization of suboptimal BPs further upstream of the annotated BP leading to splicing of proximal cryptic 3'SSs. Meyer et al have also reported that there is a cryptic 3'SS between BP and annotated 3'SS of *RPS23B* and secondary structures mask this 3'SS and make annotated 3'SS more accessible to the spliceosome[7]. Interestingly, the fast mutant enhances use of this suboptimal 3'SS, suggesting that faster transcription disrupts the structure of pre-mRNA in this region, making suboptimal 3'SS more accessible for splicing (figure 5.5b). However, the elongation mutants did not increase SEF of an alternative 3'SS event in *APE2* (*APE2*:u3:18), which was shown to be spliced with higher frequency upon heat shock, probably because of unstable RNA structure in BP-3'SS region in higher temperature[7,17]. Instead, both faster and slower transcriptions increased SEF of a further upstream alternative 3'SS event in *APE2* (*APE2*:u3:25) (figure 5.5e)

Downstream novel 3'SSs occur very close to the annotated 3'SS (figure 5.4a), presumably constrained by the optimal distance between BP and 3'SS; if they occur further downstream their distance from BP will get longer than optimal and could negatively affect spicing. The frequency of 'AG' dinucleotides (potential 3'SSs) on the second exon confirms that the lower frequency of selecting further downstream novel 3'SS events is not due to lower frequency of 'AG' dinucleotides (figure 5.5c). Further downstream novel 3'SSs can use cryptic BPs. For example, lariat sequencing detected two additional BPs in *TDA5* downstream of the annotated BP. These

suboptimal BPs can modulate splicing of suboptimal downstream 3'SS events occurring 127 and 137 nucleotides from annotated 3'SS (figure 5.5d).

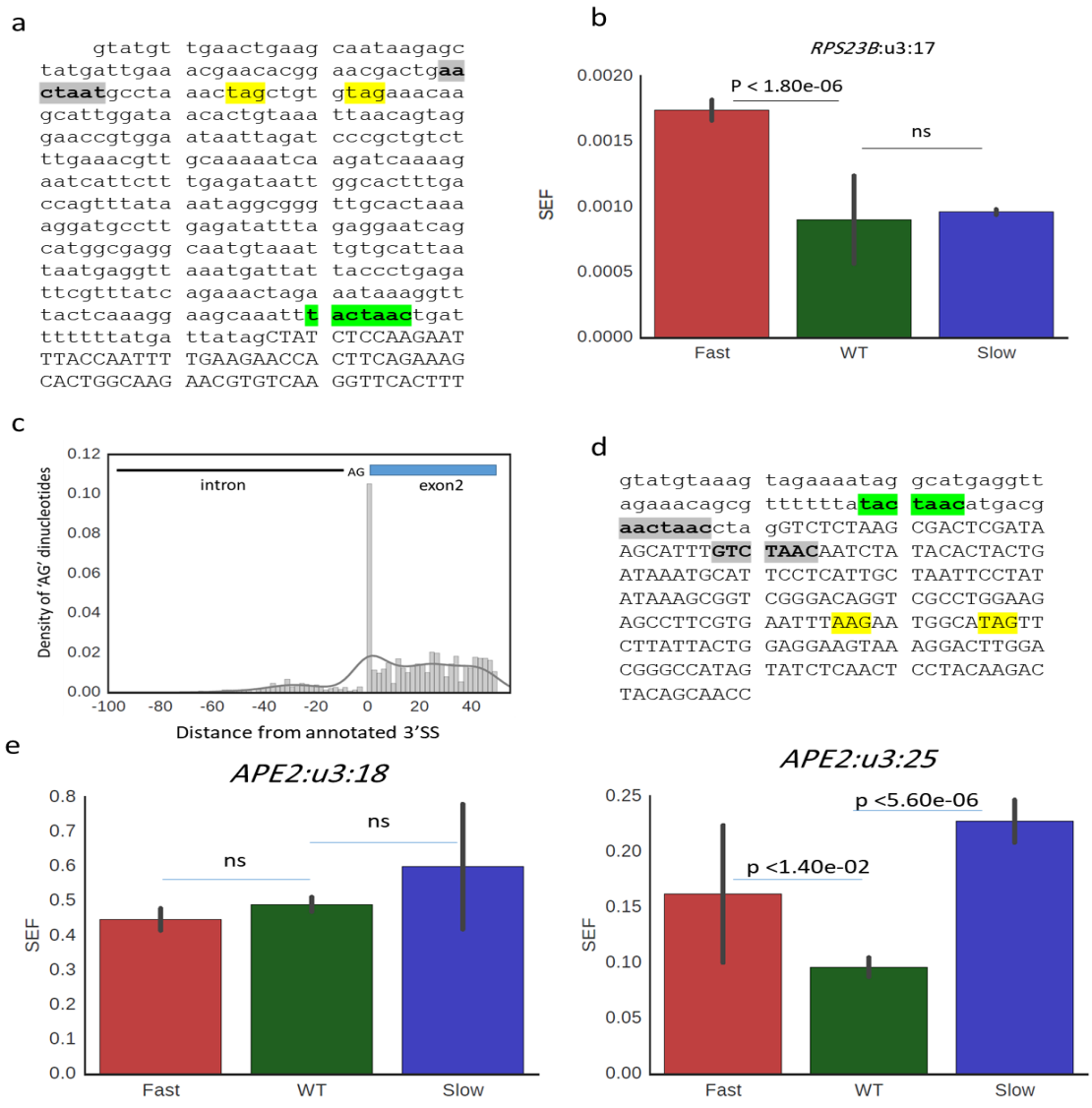


Figure 5.5: (**a**) Position of the cryptic BP and novel splice sites at 5'end of the intron in *RPL13B*. Annotated BP is highlighted with green, cryptic BPs identified with lariat sequencing[18] highlighted with grey, intron sequence is shown in lowercase letters, novel 3'SS are highlighted with yellow, partial sequence from exon2 is shown in uppercase letters. (**b**) *RPS23B*:u3:17 shows enhanced use of novel 3'SS shifted 17 nucleotides upstream of annotated 3'SS. Error bars are for two biological replicates. (**c**) Density of 'AG' dinucleotides in the region between annotated BP position and 50 nucleotides into exon2 of all intron-containing transcripts. The highest density belongs to the 'AG' of the annotated 3'SS. (**d**) Position of the cryptic BPs and novel 3'SSs at *TDA5*. (**e**) *APE2*:u3:18 shows use of novel

3’SS shifted 18nts upstream of annotated 3’SS (left) and *APE2*:u3:25 shows use of novel 3’SS shifted 25nts upstream of annotated 3’SS (right). Error bars are for two biological replicates.

5.5 Detecting introns in genes known to be intronless

Approximately 95% of *S. cerevisiae* genes are intronless. Thanks to the depth of RNA sequencing, cryptic introns supported by at least five unique reads were identified in 140 normally intronless transcripts. Analysis reveals that the fast RNAPII increases splicing of these cryptic introns ($p < 4.45e-02$, Wilcoxon test) but the slow mutant does not affect use of cryptic introns significantly ($p > 0.05$). Splicing of cryptic introns mainly introduces premature translation termination codons (PTCs), leading to degradation by NMD. Notably, some cryptic introns found in the WT strain were insensitive to alternation in transcription speed (figure 5.6b). A good example is *PRP5,* which is required for prespliceosome formation and annotated as an intronless gene in ENSEMBL and *Saccharomyces* genome database (SGD) but RNA-seq data revealed a 167 nt intron with canonical splice sites (5’SS/GTATG , AGCAG/3’SS) and branch point (TACTAAC) sequence (figure 5.6c). The 5’SS of the intron begins two nucleotides upstream of the coding sequence, which means splicing of the intron would disrupt the conventional start codon (ATG) according to the current ENSEMBL annotation. Schreiber et al[19] reported the existence of this novel intron based on much fewer reads. The spliced form is not the only expressed isoform, as figure 5.6c shows, there are many reads that detect unspliced *PRP5* transcripts. The spliced isoform could have alternative functions or could act as a tool to down-regulate gene expression in specific physiological conditions via coupling with NMD. Interestingly, in addition to the main intron (intron 1 with highest coverage), there are two u3’SS (3 and 8 nucleotides upstream) and two d3’SS (12 and 23 nucleotides upstream) in *PRP5*. u3’SS with eight nucleotides shift is the second major isoform that could potentially disrupt the open reading frame of the main spliced isoform.

Interestingly, splicing of the intron reported at the 3’ end of *FES1*[4] is affected differentially with elongation rate. Two-fold induction of the *FES1* shorter isoform (using a promoter-proximal polyadenylation site) with the fast mutant and a slight

increase in splicing of the intron (generating a longer isoform) with the slow mutant suggest that *FES1* splicing is regulated co-transcriptionally (Fig. 5.6d). This is consistent with preferential induction of the shorter isoform due to increased transcription upon heat shock with Hsf1[4]. Interestingly, slight increase in splicing of the intron (selecting 5'SS and recruitment of the spliceosome instead of selecting the downstream polyadenylation site) with the slow mutant is consistent with kinetic competition observed in selecting upstream polyadenylation sites with slower transcription (see chapter 4)

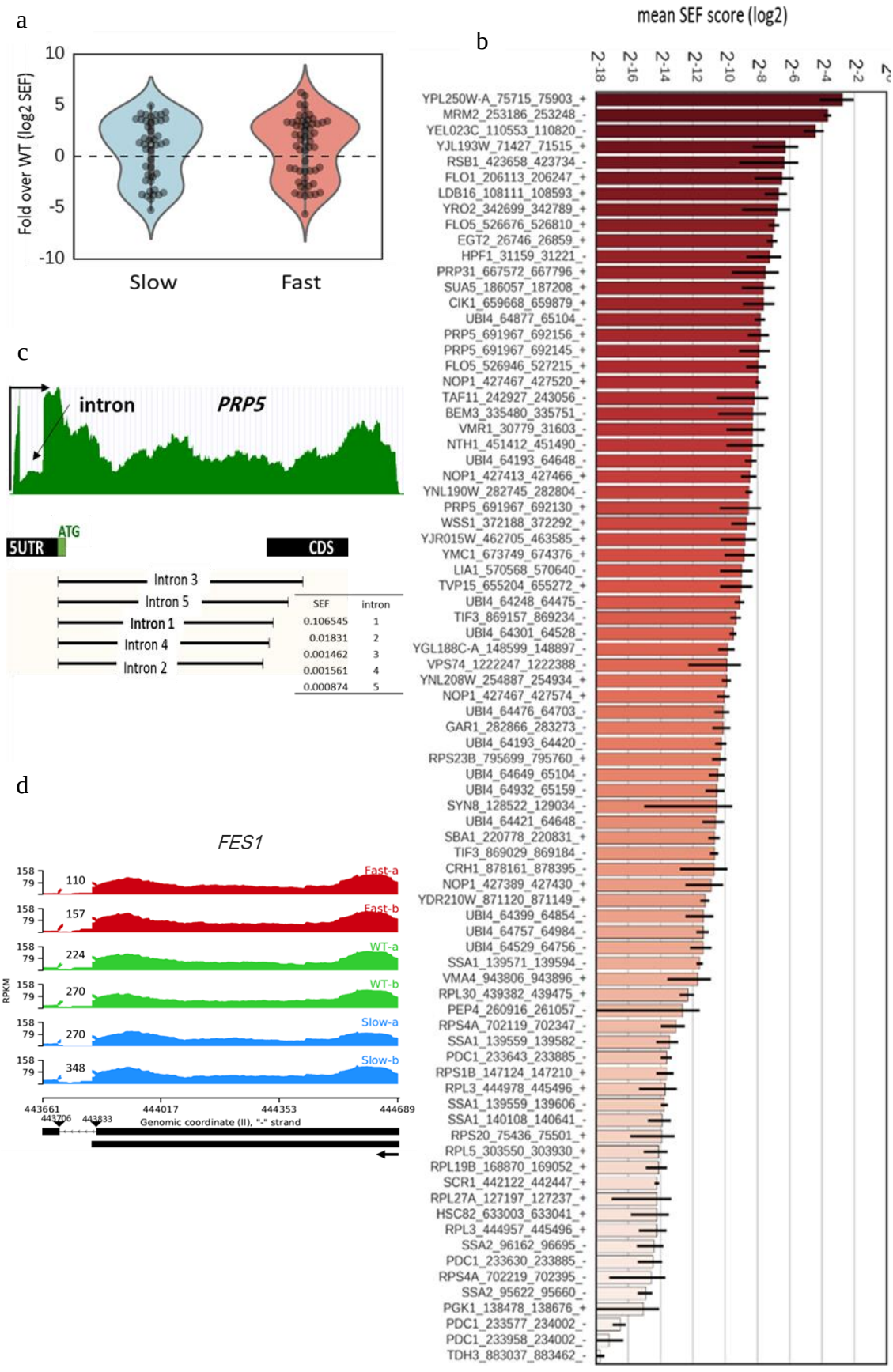

a
Fold over WT (log2 SEF)
10
5
0
-5
-10
Slow
Fast
b
mean SEF score (log2)
2-18
2-16
2-14
2-12
2-10
2-8
2-6
2-4
2-2
YPL250W-A_75715_75903_+
MRM2_253186_253248_-
YEL023C_110553_110820_-
YJL193W_71427_71515_+
RSB1_423658_423734_-
FLO1_206113_206247_+
LDB16_108111_108593_+
YRO2_342699_342789_+
FLO5_526676_526810_+
EGT2_26746_26859_+
HPF1_31159_31221_-
PRP31_667572_667796_+
SUA5_186057_187208_+
CIK1_659668_659879_+
UBI4_64877_65104_-
PRP5_691967_692156_+
PRP5_691967_692145_+
FLO5_526946_527215_+
NOP1_427467_427520_+
TAF11_242927_243056_-
BEM3_335480_335751_-
VMR1_30779_31603_-
NTH1_451412_451490_-
UBI4_64193_64648_-
NOP1_427413_427466_+
YNL190W_282745_282804_-
PRP5_691967_692130_+
WSS1_372188_372292_+
YJR015W_462705_463585_+
YMC1_673749_674376_+
LIA1_570568_570640_-
TVP15_655204_655272_+
UBI4_64248_64475_-
TIF3_869157_869234_-
UBI4_64301_64528_-
YGL188C-A_148599_148897_-
VPS74_1222247_1222388_-
YNL208W_254887_254934_+
NOP1_427467_427574_+
UBI4_64476_64703_-
GAR1_282866_283273_-
UBI4_64193_64420_-
RPS23B_795699_795760_+
UBI4_64649_65104_-
UBI4_64932_65159_-
SYN8_128522_129034_-
UBI4_64421_64648_-
SBA1_220778_220831_+
TIF3_869029_869184_-
CRH1_878161_878395_-
NOP1_427389_427430_+
YDR210W_871120_871149_+
UBI4_64399_64854_-
UBI4_64757_64984_-
UBI4_64529_64756_-
SSA1_139571_139594_-
VMA4_943806_943896_+
RPL30_439382_439475_+
PEP4_260916_261057_-
RPS4A_702119_702347_-
SSA1_139559_139582_-
PDC1_233643_233885_-
RPS1B_147124_147210_+
RPL3_444978_445496_+
SSA1_139559_139606_-
SSA1_140108_140641_-
RPS20_75436_75501_+
RPL5_303550_303930_+
RPL19B_168870_169052_+
SCR1_442122_442447_+
RPL27A_127197_127237_+
HSC82_633003_633041_+
RPL3_444957_445496_+
SSA2_96162_96695_-
PDC1_233630_233885_-
RPS4A_702219_702395_-
SSA2_95622_95660_-
PGK1_138478_138676_+
PDC1_233577_234002_-
PDC1_233958_234002_-
TDH3_883037_883462_-
c
intron
PRP5
5UTR
ATG
CDS
Intron 3
Intron 5
Intron 1
Intron 4
Intron 2
SEF
intron
0.106545
1
0.01831
2
0.001462
3
0.001561
4
0.000874
5
d
FES1
RPKM
158
79
110
Fast-a
157
Fast-b
224
WT-a
270
WT-b
270
Slow-a
348
Slow-b
443661
443706
443833
444017
444353
444689
Genomic coordinate (II), "-" strand

Figure 5.6: Cryptic splicing of normally intronless transcripts. (**a**) Violin plots show the occurrence of cryptic splicing events within normally intronless transcripts in the fast and slow mutants normalized to WT. This plot includes all novel splicing events with significantly different SEF in mutants relative to WT (Fisher's exact test ($p < 0.01$). Points above the dashed line (zero) are splicing events with SEF greater than WT (reduced fidelity). Points below the dashed line are the events with improved splicing fidelity compared to WT. (**b**) Intronless genes containing a cryptic intron whose SEF was insensitive to altered transcription elongation rate. SEF is the average of all biological replicates. Black bars indicate standard deviation. (**c**) Read density plot showing position of the intron in *PRP5*. (**d**) Splicing of *FES1* intron. Numbers show coordinates on minus strand of chromosome II. Intron (solid line) is 127 nucleotides, and dashed lines with down triangle sign shows where splicing occurs at 3'end of the transcript. Splicing of the intron generates a longer isoform.

5.6 Second intron detection within intron-containing genes

Cryptic introns were also found within the second exons of some intron-containing transcripts (Table 5.2). For example, a cryptic intron in the second exon of *RPL30* was spliced with a frequency of 0.7 percent relative to the annotated intron. The cryptic intron has GTAAG as 5'SS which spliced mainly to GAAAG/ 3'splice site or 4 base pairs shorter to TTGAG/ 3'SS (Table 5.2). Upstream of the 3'SS of the novel intron, the best match to the branch point consensus is the suboptimal CGCTAAC. Existence of this intron was reported with much fewer reads[19]. It has been shown that Rpl30 binds to *RPL30* pre-mRNA and prevents U2-snRNP association with the branch point while allowing 5' splice site recognition by U1-snRNP[20]. Blocking splicing of the annotated intron in *RPL30* could favour splicing of the downstream novel intron that would disrupt the open reading frame. Gould et al reported finding an unusual AT-AC intron nested inside the annotated intron of *RPL30*[18], but neither this study nor two other studies[9,19] found evidence for such an intron. AT-AC intron splicing takes place in metazoans by a distinct splicing machinery which yeast does not have[21]. In addition, the sequences of the 5' and 3' splice sites of the AT-AC intron reported in *RPL30*[18] deviate considerably from consensus AT-AC intron splice sites[22]. The second example of a cryptic novel intron within a second exon is a 62 nucleotides long novel intron in the second exon of *RPS23B* with a frequency of 0.07 percent. This event and also the

novel splicing event in the second exon of *RPL30* would disrupt the open reading frame. Previously, only *RPL30* was reported to have a novel intron in the second exon. Here other genes like this are reported (Table 5.2). The existence of these novel introns could provide a tool for the cells to repress expression of a particular gene by directing spliceosome machinery towards non-productive splicing and therefore promoting the nuclear turnover of spliced RNA[36].

Table 5.2

gene	start	end	5'SS - 3'SS	strand	length	Fast	Slow	WT
RPL19B	168870	169052	GTTTG - GAAAG	+	183	10	5	8
RPL27A	127197	127237	GTCAC - TGAAG	+	41	10	1	7
RPL27A	127148	127264	GTTGT - AGAAG	+	117	31	24	74
RPL27A	127176	127264	GTTCC - AGAAG	+	89	1	10	13
RPL30	439382	439475	GTAAG - TTGAG	+	94	45	21	38
RPL30	439382	439479	GTAAG - GAAAG	+	98	389	178	340
RPS16B	306806	307073	GTACG - TTTAG	-	268	15	33	5
RPS23B	795699	795760	GTCGT - CCAAG	+	62	92	88	92
RPS4A	702191	702442	GTAAG - TTTAG	-	252	16	4	18
RPS4A	702119	702140	GTCTT - ACAAG	-	22	7	9	1
RPS4A	702219	702395	GTTAC - ACAAG	-	177	11	1	5
RPS4A	702438	702500	GTCAA - GTAAG	-	63	22	1	9
RPS4A	702087	702227	GTTCA - GTAAG	-	141	24	30	6
RPS4A	702438	702587	GTTAG - GTAAG	-	150	66	17	84
RPS4A	702119	702347	GTCAA - ACAAG	-	229	14	16	17
RPS4A	702180	702347	GTCAA - TCAAG	-	168	6	1	15
RPS4A	702180	702418	GTAAG - TCAAG	-	239	6	1	8
RPS4A	702119	702313	GTAAG - ACAAG	-	195	6	2	21

Table 5.2: List of additional introns detected within second exons of the intron-containing genes. Fast, slow and WT columns show read counts (normalized to depth of sequencing) mapped to each novel junction in each strain. Length column is the length of the novel intron.

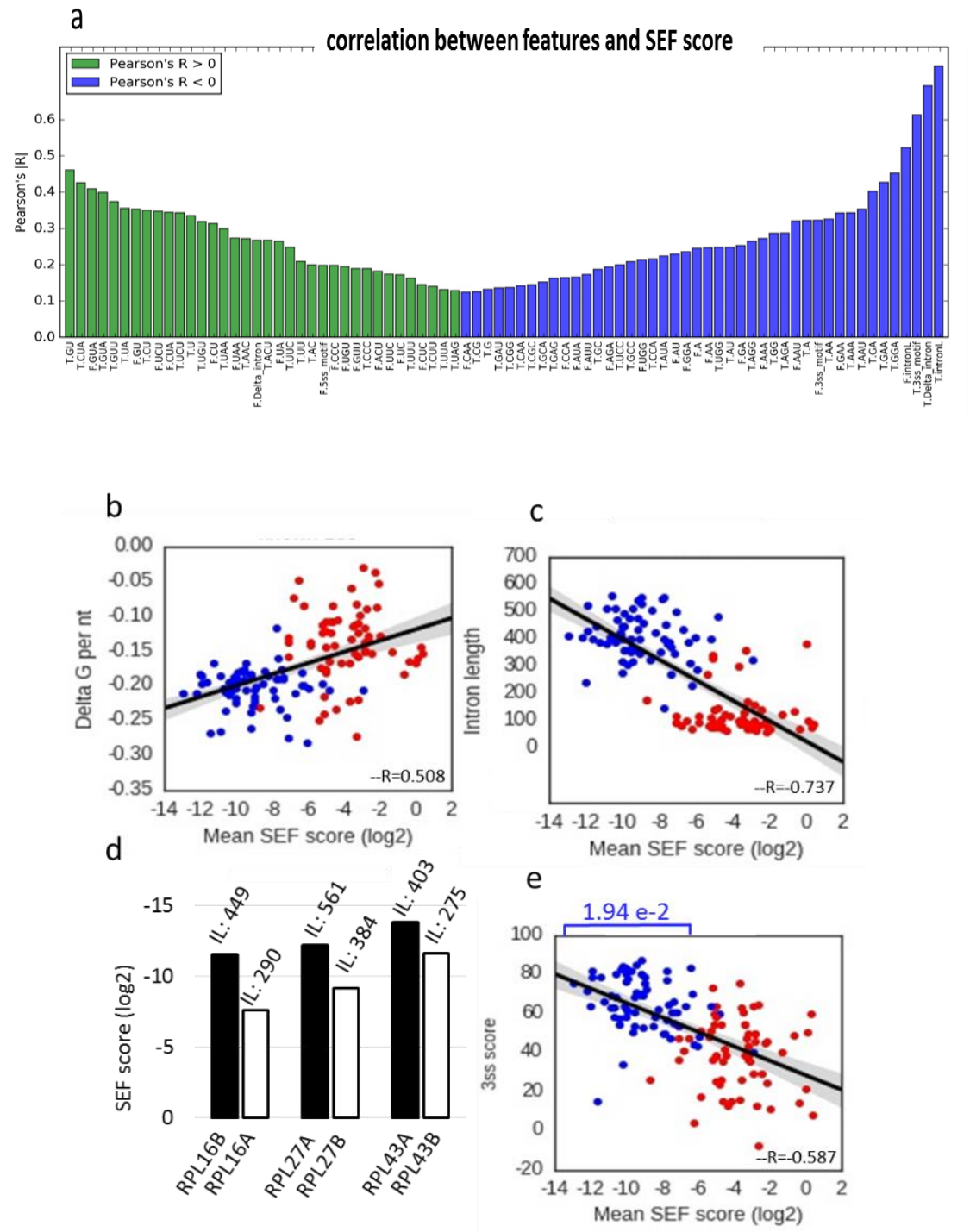
5.7 Features associated with splicing fidelity

In order to identify features that might be responsible for modulating splicing fidelity, the correlation between splicing error rate in all three strains and various transcript features, including intron length, predicted intron secondary structure (i.e., free folding energy, normalized by the intron length), 5'SS and 3'SS scores, and frequency of 1- to 3-mers in the intron were investigated. Figure 5.7a shows features that were identified to have significant correlation ($p < 0.05$) with splicing error rate. In the case of intron secondary structure, transcripts with less stable structure (higher free folding energy)

have higher SEF ($p < 6.13e-10$) (figure 5.7b). There is also a significant negative correlation between the intron length and SEF indicating that shorter introns are more susceptible to splicing errors ($p < 1.19e-23$) (figure 5.7c). This leads to the prediction that within pairs of RP paralogs that have introns of different lengths (but very similar or identical exons), the one with a shorter intron will have higher SEF than the other. Indeed, this is the case for *RPL16A/B, RPL27A/B* and *RPL43A/B* (figure 5.7d). As might be predicted, 3'SS score anti-correlates with splicing error rate ($p < 2. 94e-13$) (figure 5.7e). Notably, the effect on splicing error rate of having a poor 3'SS score is also significantly greater within RP genes ($p < 1.94$ e-2). Taken together, greater SEF (lower fidelity) correlates with shorter introns, less stable intron secondary structure and poor 3'SS score. Additionally, analysis shows that novel introns in RP transcripts whose use increases with fast or slow mutants have significantly higher predicted free energy (figure 5.7f). In contrast, events that are less frequent with the RNAPII mutants do not have different free energy compared to annotated introns. This suggests that selection of novel introns with less stable secondary structure increases with changing transcription elongation rate. It is well stablished that budding yeast introns have highly conserved splice sites. Sequence logos were generated for 5' and 3' splice sites of novel introns, finding that the novel introns predominantly use suboptimal splice sites (figure 5.7g). The polypyrimidine tract preceding the novel 3'SS also shows a relaxation from the consensus sequence, highlighting the significance of the polypyrimidine tract sequence in correct recognition of the 3'SS.

Using features discussed here and other features listed in figure 5.7a to predict the splicing error rate by a random forest regression model reveals a good correlation ($R = 0.804$) between observed and predicted splicing error rate, suggesting that these features play important roles in splicing fidelity (figure 5.7h). An interesting question was, can these features distinguish novel and known introns from a pool that have both kind of introns? To answer this question, novel and known intron sequences were pooled together and a random forest classifier used to differentiate novel and known introns. Figure 5.7h shows a good sensitivity and specificity for classifier for

differentiating novel and known introns from a mixed pool (AUC=0.782, much better than a random guess which is 0.5).



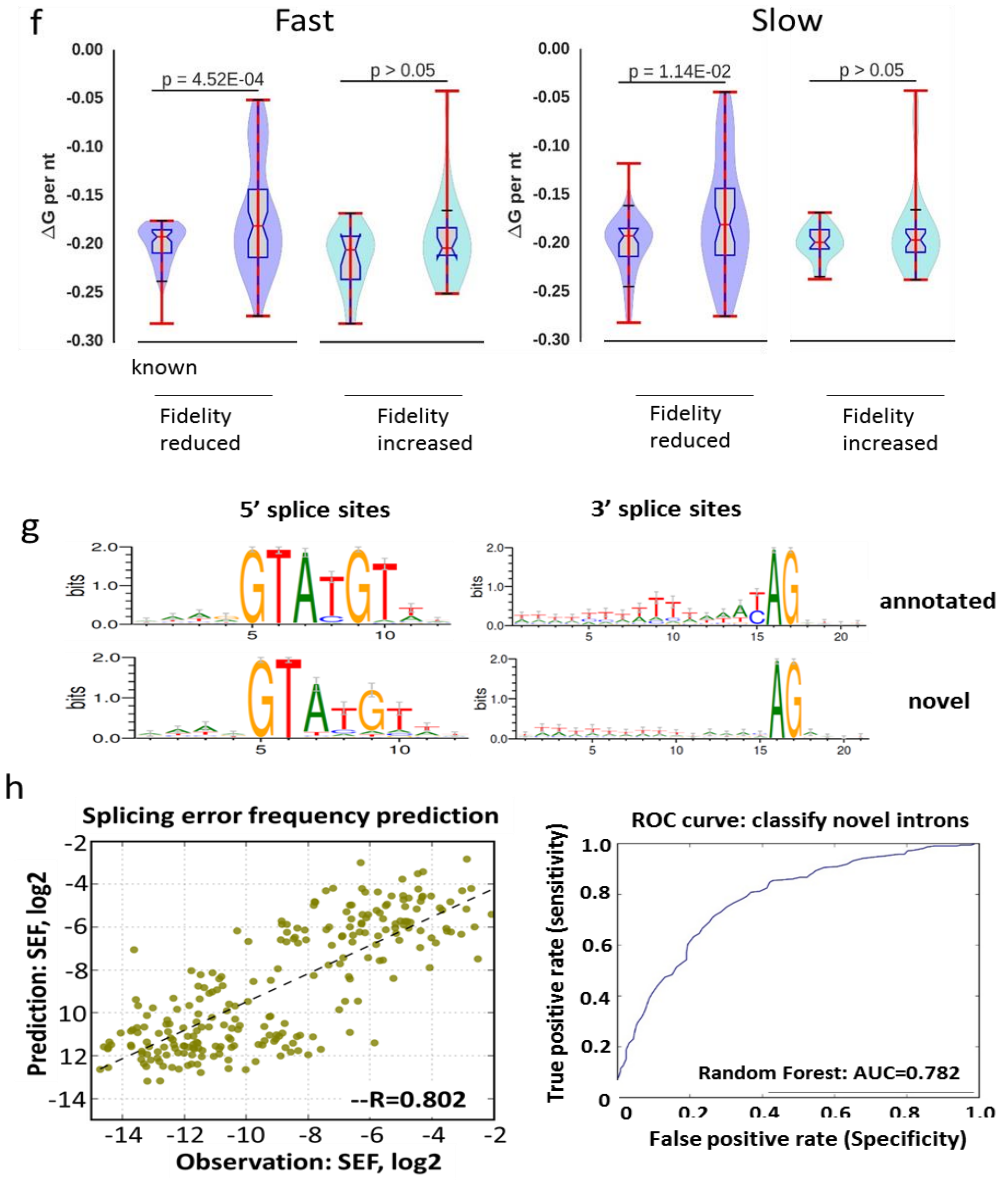


Figure 5.7: Features associated with splicing error rate. Analysis of features associated with splicing fidelity and measuring correlation between observed and predicted features were carried out in collaboration with Yuanhua Huang (**a**) Pearson's correlations between splicing fidelity and sequence patterns shows the significantly correlated features ($p < 0.05$) to SEF. Green represents positive correlation and blue represents negative correlation. "T" and "F" prefixes are used to show whether features come from annotated or novel introns, respectively. (**b**) Red points represent novel events in non-RP transcripts and blue points denote RP transcripts. Positive correlation between delta G of intron (see methods) and average SEF of the transcripts. Delta G of intron was divided by the intron length to achieve delta G per nucleotide. More negative delta G values represent more structured introns. (**c**) Negative correlation between intron length and average SEF. (**d**) Average SEF of three pairs of paralogs.

Paralogs with shorter intron length (IL) have higher splicing error frequency. (**e**) Negative correlation between 3'SS score (see methods) and average SEF. The score expresses how similar the splice sites are to the budding yeast 3'SS consensus. (**f**) Correlation between splicing fidelity in fast and slow mutants and intron delta G per nucleotide of known and novel RP introns. The p-values were obtained by t-test. (**g**) Sequence logos were generated for 5' and 3' splices sites of all novel alternative splicing events. For annotated introns, sequence logos were generated only from splice sites of transcripts that had novel alternative splicing events. (**h**) Correlation between observed and predicted SEF.

5.8 A comprehensive database of alternative and cryptic splicing events for budding yeast

Merging the list of unannotated (according to ENSEMBL) splicing events found here with those reported in most recent studies creates a database of alternative and cryptic splicing events for budding yeast. Including 21 alternative splicing events that have been reported repeatedly[9,18,19] and hundreds of new splicing events that are not reported in those dataset (figure 5.9). To build a more comprehensive database, alternative splicing events that are detected in other studies with RT-PCR, RNA-sequencing or predicted with computational methods also need to be included[17,23–25].

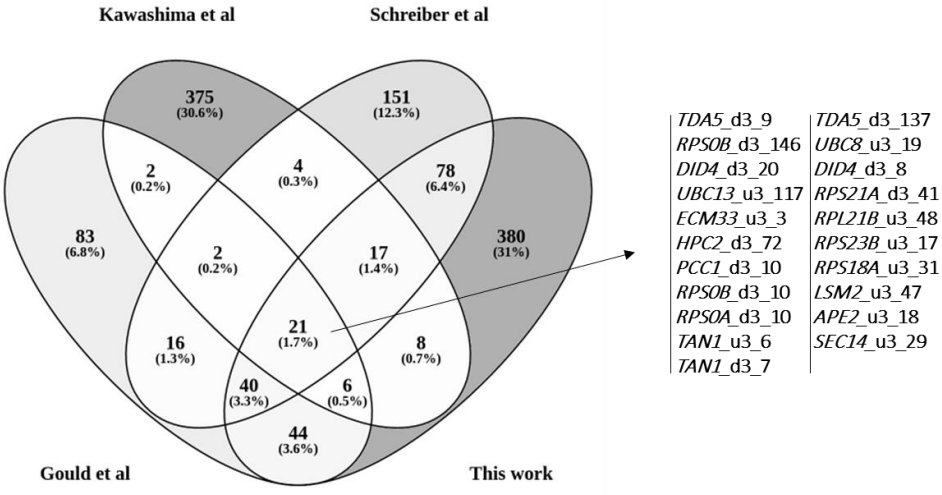


Figure 5.8: Venn diagram showing overlap between all novel splicing events with at least 5 unique reads found in this work and three published lists of alternative splicing events in budding yeast (see main text for references). The novel splicing events detected in this work that were also reported in all three previous studies are listed on the right.

5.9 Discussion

Results here show that RNAPII elongation mutants affect alternative splicing and splicing fidelity transcriptome-wide in budding yeast. Significantly, results show that the relationship between transcription elongation and splicing fidelity is not readily explained by the commonly accepted definition of the window of opportunity; both fast and slow transcription reduce fidelity of RP transcripts. In chapter 4, it is shown that elongation rate affects splicing efficiency, with the efficiency of splicing RP transcripts being significantly more sensitive to changes in elongation rate than non-RP transcripts. Accordingly, results presented in this chapter show that although RP transcripts are highly abundant and, in general, are spliced with high fidelity (figure 5.2), their splicing fidelity is reduced by changes in transcription speed, whether faster or slower. Together, these observations suggest that splicing of RP transcripts is more functionally coupled to transcription than is splicing of non-RP transcripts, and may indicate that it is beneficial to splice co-transcriptionally. Results also indicate that high fidelity correlates with high splice site score, longer and more highly structured introns; all features that are typical of RP introns (figure 5.7). It seems likely that RP introns have evolved for maximum splicing fidelity to ensure optimal expression of these abundant and vital components of the translation machinery. In contrast, non-RP transcripts are, on average, spliced with lower fidelity than RP transcripts but are less sensitive to changes in transcription elongation rate, when fidelity can be increased or reduced with the same likelihood. The observation that splicing fidelity of non-RP introns is less sensitive to changes in splicing speed may indicate that their splicing is already at the lower acceptable limit of fidelity, or of permissible flexibility to allow the selection of alternative splice sites.

The evidence that transcription rate affects splicing fidelity suggests that, at normal elongation rate, many splicing errors are either avoided or eliminated co-transcriptionally. It seems less likely that transcription rate would affect post-transcriptional degradation of mis-spliced transcripts. Indeed RP transcripts, whose

splicing fidelity is generally higher than for non-RP transcripts are more co-transcriptionally spliced, and increased elongation rate strongly reduced their splicing fidelity. Reduction in splicing fidelity of the RP transcripts in the RNAPII fast mutant might be explained by a window of opportunity model, in that there is less time for the splicing machinery to discriminate between optimal vs suboptimal splice sites, making errors more likely. On the other hand, the finding that RP splicing fidelity is also reduced in the slow mutant, albeit less strongly (figure 5.2), cannot be so easily explained, and suggests that additional factors affect splicing fidelity. Similar to the observation with the slow mutant that showed both reduction and increase in splicing fidelity, it has been shown that in human cell lines, slow transcription both decreases or increases inclusion of a particular alternative exon, opposite to the window of opportunity prediction[26]. These results support the idea that either splicing quality control mechanisms are acting differently in RPs and non-RPs or there are specific features controlling splicing fidelity in these genes. Similar to splicing efficiency (discussed in chapter 4) distinct effect of elongation rate on splicing fidelity in RP and non-RP transcripts indicates that transcript identity is an important factor in determining splicing fidelity. These results demonstrate that shorter intron length, lower stability of intron secondary structure and poor 3'SS score significantly correlate with higher SEF (figure 5.7). Furthermore, analyses show that RP transcripts with lower splicing fidelity in the fast and slow mutants have introns with less stable secondary structure (figure 5.7). Altogether, these observations suggest that offering more time for co-transcriptional spliceosome assembly by slowing down the elongation rate does not guarantee identifying and rejecting the wrong splice sites. Slower elongation can also expand the time window for recruitment of splicing suppressors or even perturb NTP-dependent kinetic proofreading mechanisms and eventually reduce overall splicing fidelity[3,27]. Additionally, changing the elongation rate was reported to change multiple properties of nascent RNA and the chromatin environment around active transcription sites, affecting alternative splicing decisions[28–30]. For example, it has been suggested that co-transcriptional pre-mRNA folding kinetics can be subject to modulation by transcription elongation rate[7,31–33],

therefore it could be envisioned that increase in utilization of suboptimal splice sites with both faster and slower transcriptions is due to incorrect and unstable pre-mRNA folding. This is consistent with alternative upstream 3'SS selection upon heat shock in *APE2*, an event that creates a mRNA 18nts longer than the annotated one[17]. It is proposed that high temperature interferes with correct and stable RNA structure formation in the BP-3'SS region of *APE2* and thereby makes alternative 3'SS more accessible for splicing[17,34]. Notably, altered elongation rates did not change use of this event in *APE2* but use of another alternative upstream 3'SS, which creates a mRNA 25nts longer than the annotated one increased with the fast mutant and more significantly with the slow mutant (figure 5.5e). Additionally, in *RPS23B* selection of an upstream 3'SS event whose use is determined by pre-mRNA folding[7] increased significantly with the fast mutant (figure 5.5b). These observations seem to suggest that transcription elongation rate modulates the interplay between pre-mRNA folding kinetics and alternative splice sites selection such that altering transcription rate compromises RNA folding and correct splice sites selection. Therefore, a finely-tuned coordination in time and space between transcription and splicing is required to achieve maximum splicing fidelity. Altogether, the findings presented here strongly argue that pre-mRNA splicing and its coupling with transcription can play an important role in regulating gene expression even in a system that lacks the infrastructure necessary for alternative splicing.

As discussed in chapter 4, elongation mutants also alter TSS selection. It is not known whether TSS usage can affect alternative splicing events, however there are studies showing the possible effect of promoter identity and 5'UTR structure on alternative splicing in human and yeast. Cramer *et al* demonstrated a functional link between promoter structure and alternative splicing of the EDI exon in the human fibronectin gene[35]. Likewise, swapping the *RPL22B* promoter with the *GAL1* promoter in yeast affected splice site selection[9]. Observations in chapter 4 and here indicate that the fast mutant not only causes stronger TSS shifts for RP intron-containing transcripts but also has the strongest effect on their splicing fidelity. In particular, the fast mutant

enhanced use of a suboptimal alternative 5'SS in *RPL39* and *VPS19* transcripts and shifted their TSS upstream compared with WT (figure 5.4). Overall, these results suggest that promoter identity and TSS selection likely influence co-transcriptional splicing fidelity, but further studies are needed to more directly investigate the effect of TSS selection on splicing fidelity.

Additionally, enhanced splicing of cryptic introns residing in intron-less transcripts were identified with fast elongation rate (Fig. 5.6). This is presumably a consequence of fast transcription reducing spliceosome fidelity. Splicing of cryptic introns was proposed as an additional layer in dynamic regulation of gene expression referred to as spliceosome-mediated decay[36]. When and how cells ensure controlled expression of each cryptic isoform in intronless genes is an interesting question for future investigations. An interesting example presented here is the effect of elongation rate on splicing of the intron in *FES1* transcript. Fast elongation reduces the spliced isoform of the *FES1* transcript, indicating that faster transcription reduces co-transcriptional recognition/splicing of the intron at the 3'end of the *FES1*. However, slight increase of the spliced isoform with the slow mutant suggests that slower transcription opens the window of opportunity for recognition and splicing of the intron relative to WT. Reduced splicing of *FES1* with the fast mutant could be due to ineffective recruitment of the splicing factors or cleavage and polyadenylation before the spliceosome completes splicing of the intron. It seems likely that incompletely spliced transcripts are degraded and are not spliced post-transcriptionally; otherwise, there should not be differences in the expression of the splice isoform between the fast and WT strain. However, more work is required to establish this. These results are compatible with observations of Gowda et al, showing that expression of the spliced isoform of *FES1* is modulated via transcription[4].

The observation that there are hundreds of unannotated alternative splicing events in budding yeast suggests that the *S. cerevisiae* intron database is not complete and there are many spliced isoforms that could play important functional roles in the

yeast life cycle. Performing deep RNA sequencing under different physiological conditions could reveal more hidden alternatively spliced isoforms in intron-containing and intronless genes.

References

1. Kalsotra, A. & Cooper, T. A. Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* **12,** 715–729 (2011).
2. Naftelberg, S., Schor, I. E., Ast, G. & Kornblihtt, A. R. Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. *Annu. Rev. Biochem.* **84,** 165–198 (2015).
3. Semlow, D. R. & Staley, J. P. Staying on message: ensuring fidelity in pre-mRNA splicing. *Trends Biochem. Sci.* **37,** 263–273 (2012).
4. Gowda, N. K. C. *et al.* Cytosolic splice isoform of Hsp70 nucleotide exchange factor Fes1 is required for the degradation of misfolded proteins in yeast. *Mol. Biol. Cell* **27,** 1210–1219 (2016).
5. Juneau, K., Nislow, C. & Davis, R. W. Alternative Splicing of PTC7 in Saccharomyces cerevisiae Determines Protein Localization. *Genetics* **183,** 185–194 (2009).
6. Mishra, S. K. *et al.* Role of the ubiquitin-like protein Hub1 in splice-site usage and alternative splicing. *Nature* **474,** 173–178 (2011).
7. Meyer, M., Plass, M., Pérez-Valle, J., Eyras, E. & Vilardell, J. Deciphering 3′ss Selection in the Yeast Genome Reveals an RNA Thermosensor that Mediates Alternative Splicing. *Mol. Cell* **43,** 1033–1039 (2011).
8. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci.* **100,** 189–192 (2003).
9. Kawashima, T., Douglass, S., Gabunilas, J., Pellegrini, M. & Chanfreau, G. F. Widespread Use of Non-productive Alternative Splice Sites in Saccharomyces cerevisiae. *PLoS Genet* **10,** e1004249 (2014).
10. Johansson, M. J. O., He, F., Spatrick, P., Li, C. & Jacobson, A. Association of yeast Upf1p with direct substrates of the NMD pathway. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 20872–20877 (2007).
11. Query, C. C. & Konarska, M. M. Splicing fidelity revisited. *Nat. Struct. Mol. Biol.* **13,** 472–474 (2006).

12. HOWE, K. J., KANE, C. M. & ARES, M. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. *RNA* **9,** 993–1006 (2003).
13. Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* **14,** 153–165 (2013).
14. Alexander, R. D., Innocente, S. A., Barrass, J. D. & Beggs, J. D. Splicing-Dependent RNA Polymerase Pausing in Yeast. *Mol. Cell* **40,** 582–593 (2010).
15. Chathoth, K. T., Barrass, J. D., Webb, S. & Beggs, J. D. A Splicing-Dependent Transcriptional Checkpoint Associated with Prespliceosome Formation. *Mol. Cell* **53,** 779–790 (2014).
16. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7,** 1009–1015 (2010).
17. Yassour, M. *et al.* Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl. Acad. Sci.* **106,** 3264–3269 (2009).
18. Gould, G. M. *et al.* Identification of new branch points and unconventional introns in Saccharomyces cerevisiae. *RNA* **22,** 1522–1534 (2016).
19. Schreiber, K., Csaba, G., Haslbeck, M. & Zimmer, R. Alternative Splicing in Next Generation Sequencing Data of Saccharomyces cerevisiae. *PLOS ONE* **10,** e0140487 (2015).
20. Macías, S., Bragulat, M., Tardiff, D. F. & Vilardell, J. L30 Binds the Nascent RPL30 Transcript to Repress U2 snRNP Recruitment. *Mol. Cell* **30,** 732–742 (2008).
21. Russell, A. G., Charette, J. M., Spencer, D. F. & Gray, M. W. An early evolutionary origin for the minor spliceosome. *Nature* **443,** 863–866 (2006).
22. Sheth, N. *et al.* Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* **34,** 3955–3967 (2006).
23. Davis, C. A., Grate, L., Spingola, M. & Ares Jr, M. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res.* **28,** 1700–1706 (2000).

24. Juneau, K., Palm, C., Miranda, M. & Davis, R. W. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 1522–1527 (2007).
25. Plass, M., Codony-Servat, C., Ferreira, P. G., Vilardell, J. & Eyras, E. RNA secondary structure mediates alternative 3′ss selection in Saccharomyces cerevisiae. *RNA* **18,** 1103–1115 (2012).
26. Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28,** 2663–2676 (2014).
27. Hopfield, J. J. Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *Proc. Natl. Acad. Sci.* **71,** 4135–4139 (1974).
28. Buratti, E. & Baralle, F. E. Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. *Mol. Cell. Biol.* **24,** 10505–10514 (2004).
29. Saldi, T., Cortazar, M. A., Sheridan, R. M. & Bentley, D. L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J. Mol. Biol.* **428,** 2623–2635 (2016).
30. Dujardin, G. *et al.* Transcriptional elongation and alternative splicing. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1829,** 134–140 (2013).
31. Eperon, L. P., Graham, I. R., Griffiths, A. D. & Eperon, I. C. Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? *Cell* **54,** 393–401 (1988).
32. Perales, R. & Bentley, D. 'Co-transcriptionality' - the transcription elongation complex as a nexus for nuclear transactions. *Mol. Cell* **36,** 178–191 (2009).
33. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* **15,** 163–175 (2014).
34. Pérez-Valle, J. & Vilardell, J. Intronic features that determine the selection of the 3′ splice site. *Wiley Interdiscip. Rev. RNA* **3,** 707–717 (2012).
35. Cramer, P., Pesce, C. G., Baralle, F. E. & Kornblihtt, A. R. Functional association between promoter structure and transcript alternative splicing. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 11456–11460 (1997).
36. Volanakis, A. *et al.* Spliceosome-mediated decay (SMD) regulates expression of nonintronic genes in budding yeast. *Genes Dev.* **27,** 2025–2038 (2013).

Chapter 6 Closing remarks

Splicing is coupled to transcription machinery and occurs simultaneously with transcription like many other pre-mRNA processing functions. The main focus of this project was to investigate the importance of kinetic coupling between transcription and splicing for efficiency and fidelity of splicing in budding yeast (figure 6.1).

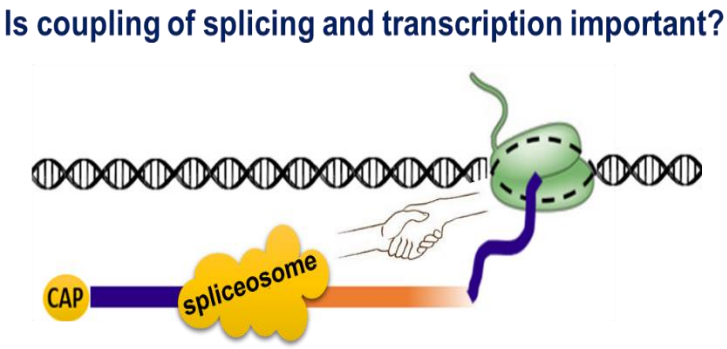


Figure 6.1: Schematic representation of coupling between transcription and splicing.

Various physiological stimuli such as UV light, airborne particulate matter and light/dark effects have been reported to affect alternative splicing by modulating transcription elongation rate[1–3]. In addition, epigenetic changes and promoter identity were shown to play a role in regulating alternative splicing by changing transcription elongation rate[4–6]. This study extends the evidence for functionally coupled transcription and splicing by demonstrating that changes in elongation rate also impact co-transcriptional splicing efficiency and fidelity in a gene-specific manner, providing a tuneable system to enhance or suppress expression of certain genes. Results show that fast transcription leaves splicing behind and therefore pre-mRNA accumulates specially for RP genes, however, slow transcription enhances co-transcriptional splicing transcriptome-wide. Interestingly, results showed that both fast and slow transcription reduce splicing fidelity mainly in RP genes (figure 6.2). The finding that although slower transcription increases splicing efficiency it can also reduce splicing fidelity, suggests that RNAPII speed may have evolved to optimise splicing efficiency

as well as to allow flexibility for making alternative splice site choices in yeast, as in mammals. Recent high-throughput studies revealed that RP transcripts tend to be spliced faster, more efficiently and more co-transcriptionally than non-RP transcripts[7,8,9], and analyses presented in this thesis show that altered transcription rates, especially faster elongation, affect the efficiency and fidelity of splicing RP transcripts significantly more than non-RP transcripts. Together, these observations suggest that splicing of RP transcripts may be more functionally coupled to transcription than is splicing of non-RP transcripts, and may indicate that it is beneficial to splice co-transcriptionally. Ribosomal proteins are abundant and vital components of the translation machinery, and RP gene expression is coordinated to match cell growth rate according to the requirement for more or fewer ribosomes in rapidly or slowly dividing cells respectively[10]. Overall, the transcription and splicing of RP transcripts seem to be tuned to meet the demands of high expression without compromising quantity and quality of the spliced transcripts. It was previously proposed that there is an optimal transcription rate for splicing different introns in mammalian cells[11]. In view of the greater sensitivity of RP splicing to transcription speed, it can be speculated that, in budding yeast, the RNAPII elongation rate has evolved in tune with RP transcript splicing, to optimise the expression of these highly important genes, while allowing greater flexibility in splice site choice for non-RP genes. As the *S. cerevisiae* spliceosome represents an evolutionarily conserved core of the splicing machinery in higher eukaryotes, the effects of transcription elongation rate on splicing fidelity observed in yeast may be closely related to effects on alternative splicing decisions in higher eukaryotes. Important goals now are to determine how the transcription and splicing machineries communicate, and how these interactions are regulated to ensure the appropriate outcome.

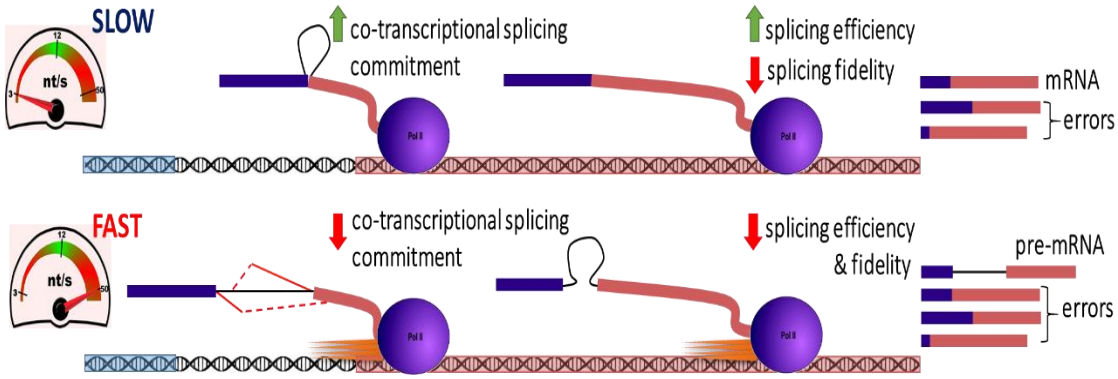


Figure 6.2: Schematic representation of the co-transcriptional splicing and consequences of fast and slow transcription on splicing efficiency and fidelity, adapted from[12]. Slow transcription enhances co-transcriptional splicing commitment and therefore increases splicing efficiency. However, fast transcription reduces co-transcriptional splicing commitment causing pre-mRNA accumulation due to reduced splicing efficiency. In the case of splicing fidelity, both fast and slow transcription reduce splicing fidelity (increased use of suboptimal splice sites). The effect of fast transcription in reducing splicing fidelity is stronger than with the slow mutant.

References

1. Muñoz, M. J. *et al.* DNA Damage Regulates Alternative Splicing through Inhibition of RNA Polymerase II Elongation. *Cell* **137,** 708–720 (2009).
2. Petrillo, E. *et al.* A chloroplast retrograde signal regulates nuclear alternative splicing. *Science* **344,** 427–430 (2014).
3. Buggiano, V. *et al.* Effects of airborne particulate matter on alternative pre-mRNA splicing in colon cancer cells. *Environ. Res.* **140,** 185–190 (2015).
4. Saldi, T., Cortazar, M. A., Sheridan, R. M. & Bentley, D. L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J. Mol. Biol.* **428,** 2623–2635 (2016).
5. Cramer, P., Pesce, C. G., Baralle, F. E. & Kornblihtt, A. R. Functional association between promoter structure and transcript alternative splicing. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 11456–11460 (1997).
6. Kawashima, T., Douglass, S., Gabunilas, J., Pellegrini, M. & Chanfreau, G. F. Widespread Use of Non-productive Alternative Splice Sites in Saccharomyces cerevisiae. *PLoS Genet* **10,** e1004249 (2014).
7. Barrass, J. D. *et al.* Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. *Genome Biol.* **16,** 282 (2015).
8. Carrillo Oesterreich, F. *et al.* Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* **165,** 372–381 (2016).
9. Wallace, E. W. J. & Beggs, J. D. Extremely fast and incredibly close: co-transcriptional splicing in budding yeast. *RNA* rna.060830.117 (2017). doi:10.1261/rna.060830.117
10. Warner, J. R. The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* **24,** 437–440 (1999).
11. Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28,** 2663–2676 (2014).
12. Moehle, E. A., Braberg, H., Krogan, N. J. & Guthrie, C. Adventures in time and space. *RNA Biol.* **11,** 313–319 (2014).