UNIVERSITY OF CALIFORNIA, SAN DIEGO

Exploring the Mechanisms of pre-mRNA Splicing Using Mathematical Models of Coupled Transcription and Splicing

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Jeremy Davis-Turak

Committee in charge:

Professor Alexander Hoffmann, Chair
Professor Eugene Yeo, Co-Chair
Professor Vineet Bafna
Professor Christopher Glass
Professor Tracy Johnson

2014

The dissertation of Jeremy Davis-Turak is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2014

EPIGRAPH

Nothing in biology makes sense
except in light of evolution.
—Theodosius Dobzhansky

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are so many people who helped me get through grad school and out the other side with my sanity intact. First of all the Bioinformatics grad coordinator Laura Gracia made life a lot less stressful when I started the program. All the members of the Hoffmann were a big source of support. There were countless times when I was in the lab trying my hands at wet-lab experiments that I would run around asking people for help, including Karen Schurenberg, Brbel Schrfelbauer, Bryce Alves, Masa Asagiri, Vincent Shih, Kim Ngo, Diana Rios, Kristyn Feldman, Jon Almaden, Yi Liu, Andrew Caldwell, Rachel Tsui, Riku Fagerland, Eason Lin, Bing Xia, Joyee Yao, Jesse Vargas, Rusty Lewis and Emily Chen. And of course there were plenty of computational discussions with the likes of Max Shokhirev, Marcelo Behar, Brooks Taylor, Frank Cheng, Gaju Suryavanshi, Paul Loriaux and Kate Hoff. It was a pleasure mentoring Harry Birnbaum and working with collaborators Vincent, Kim and Max. And of course Alex played a huge role in my graduate success. I am grateful of the intellectual freedom he afforded me to pursue my dissertation research (I'm also especially glad that he bought the lab an espresso machine . I have grown tremendously as a scientist, a writer and critical thinker under his mentorship. I am also thankful for Tracy's help guiding me through much of the process of publishing my manuscript, and our long conversations helped steer my focus back on track many times.

Even though none of the work appeared here, I did a rotation in Xiang-Dong Fu and continued our collaboration for several years. Fu, Jinsong Qiu, Hairi Li and Yu Zhou were great to work with, and taught me a lot. I also want to thank Kristi Fox-Walsh, Evan Merkhofer, Wei Wang and Geoff Rosenfeld.

I'd like to epecially thank all my other collaborators Karmel, Petr, Lev and Chris, for their contributions to my paper, which appears here as chapter 2: Sequence Signatures and Polymerase Dynamics Favor Co-transcriptional Splicing Genome-wide", with authors Jeremy Davis-Turak, Karmel Allison, Maxim Shokhirev, Petr Ponomarenko, Lev Tsimring, Christopher Glass, Tracy Johnson, and Alexander Hoffmann.

Outside of academics, climbing on the weekends was my favorite escape. Steve, Warren and Travis have been my partners in crime, and hopefully that will continue! (I also need to thank the rangers at Joshua Tree National Park!!!!). Spencer, Christine, Matt, and Max were always up for a game night.

Kat Drake always inspired me and supported me, and also encouraged me to have some fun once in a while! Giovanni Coppola was a great mentor prior to grad school and I'm grateful that he encouraged me to go back to school. My parents especially supported me and always believed I would succeed, and I am forever thankful for their love!

VITA

2004	B. A. in Molecular Biology with Certificate of Neuroscience, Princeton University
2011-2014	National Science Foundation Graduate Research Fellow
2014	Ph. D. in Bioinformatics and Systems Biology, University of California, San Diego

PUBLICATIONS

Alves B, Tsui R, Birnbaum H, Almaden J, Shokhirev M, Ponomarenko J, Davis-Turak J, Hoffmann A (2014). IκBϵ is a key regulator of B-cell expansion by providing negative feedback on cRel and RelA in a stimulus-specific manner. J. Immunology *192*(7) 3121-32

Shih VF, Davis-Turak J, Macal M, Huang JQ, Ponomarenko J, Kearns JD, Yu T, Fagerlund R, Asagiri M, Zuniga EI, Hoffmann A. (2012) Control of RelB during dendritic cell activation integrates canonical and noncanonical NF-κB pathways. Nat Immunol. *13*(12):1162-70.

Konopka G, Friedrich T, Davis-Turak J, Winden K, Oldham MC, Gao F, Chen L, Wang GZ, Luo R, Preuss TM, Geschwind DH (2012) Human-specific transcriptional networks in the brain. Neuron. *75*(4):601-17

Bernard A, Lubbers LS, Tanis KQ, Luo R, Podtelezhnikov AA, Finney EM, McWhorter MM, Serikawa K, Lemon T, Morgan R, Copeland C, Smith K, Cullen V, Davis-Turak J, Lee CK, Sunkin SM, Loboda AP, Levine DM, Stone DJ, Hawrylycz MJ, Roberts CJ, Jones AR, Geschwind DH, Lein ES (2012). Transcriptional architecture of the primate neocortex. Neuron. *73*(6):1083-99

Fox-Walsh K, Davis-Turak K, Zhou Y, Li H, Fu XD (2011) A Multiplex RNA-seq Strategy to Profile Poly(A+) RNA: Application to Analysis of Transcription Response and 3' End Formation. Genomics *98*(4):266-71

Rosen RF, Farberg AS, Gearing M, Dooyema J, Long PM, Anderson DC, Davis-Turak J, Coppola G, Geschwind DH, Paré JF, Duong TQ, Hopkins WD, Preuss TM, Walker LC (2008). Tauopathy with paired helical filaments in an aged chimpanzee. J Comp Neurol. *509*(3):259-70.

ABSTRACT OF THE DISSERTATION

Exploring the Mechanisms of pre-mRNA Splicing Using Mathematical Models of Coupled Transcription and Splicing

by

Jeremy Davis-Turak

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2014

Professor Alexander Hoffmann, Chair
Professor Eugene Yeo, Co-Chair

One of the main challenges in modern biology is understanding how and when genes are turned on. As our knowledge of transcription regulation has matured, bioinformatic analyses have allowed increasingly quantitative predictions of gene expression. The ultimate goal of such analyses is to predict gene expression from concentrations of proteins in the cell, based on protein-DNA networks. Yet this is a highly ambitious task in Eukaryotes, since their gene expression is determined by chromatin conformation, epigenetic factors, promoter and enhancer states, rates of transcription initiation, elongation, transcript processing and termination, and mRNA export and stability.

This thesis is focused on the co-occurrence of transcriptional elongation and pre-mRNA splicing, the process in which introns are removed from the pre-mRNA transcript. Splicing is an important regulatory step because aberrant splicing leads to either reduced or non-functional protein expression, and alternative splicing expands the repertoire of functional proteins encoded by the genome. The co-transcriptional nature of splicing implies that the kinetics of elongation in relation to splicing are important for the outcome of splicing decisions. Co-transcriptional splicing (CTS) has been extensively studied, but quantitative models of transcription networks that predict gene expression timecourses have yet to incorporate CTS considerations.

Here I constructed kinetic models of CTS. Initially, I built a model of constitutive CTS and developed methods to fit nascent RNA-seq data to the model. Fitting this model to published datasets indicated that only a subset of genes can be expected to process all of their introns co-transcriptionally. Detailed data-mining of high-throughput datasets and genomes revealed patterns of compensatory signatures in sequence, chromatin and polymerase data, suggesting an evolutionary selection towards splicing co-transcriptionally. Next I expanded the model to include alternative splicing reactions. Despite the exponential combinatorial complexity, all possible isoforms resulting from up to nine introns can be simulated. A further expansion of the model considers separate reactions at the 5 and 3 ends of introns, which allows for simulation of phenomena such as exon definition and polymerase-mediated recruitment. Together, these novel tools can be used to test quantitative predictions of genome-wide splicing outcomes, or be incorporated into larger gene expression models.

1 Introduction

1.1 Pre-mRNA splicing

Genes are the fundamental unit of of information within a cell, and DNA encodes genes along with the instructions for when to activate those genes. When genes become activated, they are first transcribed into RNA, and this messenger RNA (mRNA) strand is then translated into a polypeptide chain, which matures into a functional protein. In ancient single-celled organisms that lack a defined nucleus, transcription and translation are coupled spatio-temporally. However in eukaryotes, the nucleus separates these two processes in space and time. This separation allows for extensive editing of the mRNA message before translation.

Indeed, the processing of precursor mRNA(pre-mRNA) - including removal of introns, 5' capping and 3' polyadenylation - is a hallmark of eukaryotic gene transcription. Approximately 90% of a given human gene is composed of sequences that are not destined to be expressed as protein sequences. Prior to nuclear export, these sequences are removed by a catalytic process known as splicing, which stitches together the expressed sequences (exons), and degrades the non-expressed sequences (introns). The human genome averages 8.8 introns per gene, and the human genome contains approximately 20,000 genes, meaning there are almost a quarter million introns in the genome (Ast, 2004). Splicing is a highly dynamic and labor-intensive process that requires great specificity, but also has a non-negligible error rate, and this fact has many important consequences: one being that up to one third of human disease have a form associated with a defect in splicing (Lim et al., 2011).

'Correctly' spliced genes will contain an open reading frame (the sequences

sufficient to code for a functional protein): a start codon, a series of amino-acid specifying codons, and a stop codon. If a transcript is 'incorrectly' spliced, its resulting open reading frame, if any, will code for an aberrant protein sequence. In most cases, this protein sequence will contain a premature termination codon (PTC), and if this PTC is sufficiently far from the 3' end of the transcript, the transcript will likely get degraded by the nonsense-mediated decay (NMD) surveillance pathway (Losson and Lacroute, 1979; Lewis et al., 2003; Weischenfeldt et al., 2012). However, the dichotomy of 'correct' and 'incorrect' is an over-simplification. For example, there is widespread evidence that genomes have incorporated the NMD pathway into auto-regulation of the expression level of splicing factor proteins: often a high abundance of a splicing factor will lead to an increase in the 'incorrect splicing' of its own transcript that leads to NMD (Ni et al., 2007). Thus, these 'errors' sometimes have biological significance, and are referred to as alternative splicing (AS).

Alternative splicing does not always lead to a disfunctional protein. When a transcript differs from the most frequently observed (constitutive) splicing pattern, either a large intron is simply not removed (intron retention), or an entire exon is skipped (cassette exon splicing), or the exon boundaries of one or more exons is either lengthened or shortened (Thanaraj et al., 2004). In cases other than intron retention, the amount of exon added or subtracted can occur in one of three scenarios: a frame shift of +1, a frame shift of -1, or no frame shift (i.e., the change is a multiple of 3). If these events were completely random, all three scenarios would be equally likely: yet, cassette exons are enriched for the no-frame shift scenario, indicating that some of this 'incorrect' splicing is selected for (Resch et al., 2004). Since cassette exons whose length is a multiple of 3 would allow for a full-length protein that escapes the NMD pathway, it seems that nature has co-opted these AS events. Many of the proteins coded by alternative isoforms have different functions than those coded by constitutive isoforms, and thus AS amplifies the complexity of the genome (e.g. in humans giving rise to roughly 100,000 unique proteins from a mere 20,000 genes). Therefore, AS is widely considered an important step in the evolution of higher-order organisms (Modrek and Lee, 2002;

Nilsen and Graveley, 2010).

The question arises, why has the total intron content of genome evolved to take up 10-fold more space than the protein coding content? One possibility is that the total length of a gene is under selection so that specific timing goals can be achieved (Swinburne et al., 2008). For example, extremely long genes can take up to 16 hours to be transcribed (Tennyson et al., 1995), which could be a useful timing mechanism in development. Another possibility is that long introns allow for an increased likelihood of AS (Izquierdo and Valcárcel, 2006). Yet, this task seemingly could be accomplished with shorter introns. However, though many lower Eukaryotes have shorter introns, they exhibit little AS (Ast, 2004; Fox-Walsh et al., 2005). Therefore it has been proposed that the lengthening of introns contributed to the evolution of AS by increasing the error rate of splicing, with a concomitant increase in the regulatory logic and protein diversity of regulatory proteins (Izquierdo and Valcárcel, 2006). To explore this regulation, we need to first understand how splicing is catalyzed.

1.2 The spliceosome catalyzes the splicing reaction

The spliceosome is a large RNA-protein complex that assembles de novo on nascent strands of RNA (Matera and Wang, 2014), composed of five small U-rich RNAs, U1, U2, U4, U5 and U6, which associate with protein factors to form five small nuclear ribo-nucleic particles (snRNPs). Using the help of ATP-dependent helicase proteins, these snRNPs assemble step-wise on introns via specific base-pairing of the U-RNAs to the pre-mRNA, and help catalyze the two transesterification splicing reactions (Gornemann et al., 2005; Hoskins et al., 2011). First, the U1 snRNP binds to a consensus sequence on the upstream end of the intron (5' splice site). The U2 snRNP, chaperoned by the U2AF protein, binds to the branch-point sequence near the polypyrmidine track at downstream end of the intron (3' splice site). These two snRNPs then aggregate, forming the A-complex. Next, a complex of three snRNPs, the U4/U6.U5 tri-snRNP binds and displaces

the U1 snRNP. This re-arrangement brings the 5' and 3' splice sites close together. Once the spliceosome is catalytically active, the 2' hydroxyl group on the ribose backbone of the branch-point adenine base attacks the 5' end of the intron, cleaving the intron from the upstream exon, and creating a branched intron structure. After a further ATP-dependent rearrangement, the 3' hydroxyl group on the free exon attacks the 3' end of the intron, cleaving the branched intron from the downstream exon and ligating the two exons together. The now-free branched intron is called the lariat, and is subsequently degraded.

The specificity of a splicing reaction is partially dependent on the sequence of the pre-mRNA motifs that base-pair with the RNA components of the snRNPs. Furthremore, before two exons can be joined, U1 and U2 must recognize one another and prior to forming the A complex. Therefore, the manner in which the spliceosome components find their partners is crucial to determining the specificity of a splicing reaction (Shepard et al., 2011). In yeast, in which most intron-containing genes have only one intron, splice sites are comprised of highly circumscribed motifs: thus, most of the splicing reactions are very specific, there are very few examples of AS in yeast (Awan et al., 2013). However in the human genome, the core splicing motif is comprised of very few bases, and thus there are a great deal of sequences in every pre-mRNA strand which could theoretically act as splice sites (Ast, 2004). Somehow, the spliceosome detects a distinction between these so-called cryptic splice sites and those are constitutively spliced. Up to 95% of human genes undergo some level of AS (Pan et al., 2008; Wang et al., 2008), indicating that there is some degree of flexibility; yet overall splicing fidelity is very high in the human genome, meaning most splicing reactions do occur between constitutive splice sites (Fox-Walsh and Hertel, 2009). Thus, a central question in AS is how the spliceosome can be regulated to achieve a high degree of accuracy and yet still allow for flexible splice site choice.

1.3 Regulation of splicing

The regulation of AS is critical to many biological functions, including development and differentiation, and is also implicated in many diseases including neural disorders and cancer (Venables, 2004; Faustino and Cooper, 2003). Trans-acting splicing factors have been shown to be major players in this regulation. For example, the splicing factor gene *SF2/ASF* is a proto-oncogene (Karni et al., 2007). These splicing factors consist of RNA-binding proteins (RBPs), which generally fall into two classes: the Serine-Arginine rich protein family (SR proteins), and the heterogeneous nuclear ribonucleoproteins (hnRNPs) (Zahler et al., 1992; Huelga et al., 2012). These RBPs generally bind to specific motifs on the pre-mRNA in exons or introns, near the sites where spliceosomes bind (Xiao et al., 2007). Depending on the location of the binding, an RBP can enhance or inhibit the inclusion of a cassette exon (Barash et al., 2010). Furthermore, RBPs can act in unison (Huelga et al., 2012), setting up a combinatorial landscape for regulation.

Although splicing can occur in pre-mRNAs in the nucleoplasm that have terminated transcription, much of splicing occurs on the nascent pre-mRNA while it is still being made, in the chromatin environment. The existence of co-transcriptional splicing (CTS) has long been evident from examining nascent RNA of long genes (Singh and Padgett, 2009; Wetterberg et al., 1996). However, new insights have shown many intimate links between the splicing and transcription machineries (Alexander et al., 2010; Muñoz et al., 2010).

Of particular interest is the link between elongation rate and splicing. Elongation of the nascent strand by RNA polymerase II (Pol II) proceeds at a rate of approximately 1-10 kb/min (Veloso et al., 2014; Danko et al., 2013). This means that a typical vertebrate intron would take on the order of one minute to be synthesized. Thus, prior to the existence of the next, downstream exon, a typical cassette exon exists for about 1-2 minutes (which is the order of magnitude of the splicing half-life). If splicing really does occur immediately upon synthesis, one would expect that a delay in creating the downstream exon would increase the competitive edge for the cassette exon to be included, since the 5' splice site

of the upstream exon has no other partner but the cassette exon with which to interact (Kornblihtt, 2007). Indeed, several experiments have shown that this phenomenon occurs with experimental perturbation (de la Mata et al., 2003; Howe et al., 2003; Ip et al., 2011; Roberts et al., 1998). However, it remains unclear to what degree these kinetic considerations affect endogenous splicing decisions, in part because this process has received little detailed mathematical treatment.

Additionally, the chromatin environment plays a role in splicing outcomes (Gunderson and Johnson, 2009). Chromatin consists of DNA packed around histone protein octamers to form nucleosomes. The histone proteins in nucleosomes contain long tails whose amino acid residues can be covalently modified to include specific chemical groups, notably acetyl or methyl groups. The presence of nucleosomes, the identities of the histone proteins that comprise them, and the locations and amount of modifications to those proteins are recognized as central to the spatial organization of the genome and transcriptional regulation (Berger, 2007). More recently, evidence has emerged that directly links splicing to chromatin states (Luco and Misteli, 2011; Bieberstein et al., 2012; Luco et al., 2010; Kim et al., 2011; Hnilicová et al., 2011; Zhou et al., 2011; de Almeida et al., 2011). The existence of these interactions suggests that CTS is not merely coincidental with, but rather is functionally coupled to transcription (Batsche et al., 2006; Listerman et al., 2006; Alexander et al., 2010; Carrillo et al., 2010). Moreover the Pol II complex, which is the key processive enzyme in transcription, is also implicated in splicing control, further establishing a link between the production and the processing of the RNA message (Muñoz et al., 2010; de la Mata and Kornblihtt, 2006; McCracken et al., 1997; Misteli and Spector, 1999).

1.4 Splicing is fundamental to transcriptional control

Our working hypothesis is that splicing is an integral part of the transcription process. To understand this concept, consider that transcription factors provide information about which genes to activate at a certain time, at which

levels. However, as we saw above, incorrectly spliced transcripts are often down-regulated. Therefore splicing is fundamental in determining how much of a protein is made. Moreover, recent evidence suggests that incompletely spliced transcripts are retained in the chromatin environment (Bhatt et al., 2012; Brody et al., 2011). Therefore, splicing can also delay the release of transcripts, subsequently delaying mRNA export (Rigo and Martinson, 2009; Pandya-Jones et al., 2013; Hao and Baltimore, 2013; Carmo-Fonseca et al., 1999; Martins et al., 2011). In light of these considerations, it should not be surprising that genome-wide experiments correlating mRNA levels with the binding of transcription factors to DNA (which is a proxy for initiation at gene promoters) show lower than expected correlations, as these type of experiments largely fail to account for RNA processing including splicing and mRNA decay rates.

The overarching goal of this Thesis is to provide a mathematical basis for the integration of splicing into efforts to model gene expression. In doing so, we aimed to develop a scalable model of the splicing process that would: include CTS and AS; simulate realistic gene scenarios; create a framework for simulating the interactions of chromatin and splicing; and develop methods to interpret or fit experimental data to realistic models. The work in the next three chapters surrounds three mathematical models of co-transcriptional splicing that we developed.

1.5 Limitations of existing splicing models

Several mathematical models of splicing exist already, although none have been published that fulfill all the criteria we desire. Top-down statistical models have been used to examine inclusion rates of cassette exons in AS (Zhang et al., 2010; Barash et al., 2010). Barash et al. (2010) created a model that integrates sequence-motif data along with abundances of splicing factors to learn rules for a 'splicing code', in order to predict whether an exon is up- or down- regulated in a certain cell type. This type of analysis is useful when one doesn't know the optimal construction of the model, and where it is not tractable to include terms for the interaction of all possible parameters. However, the parameters that result from

fitting these models are not based on physical reality, and little biological insight is gained as to how the input variables can interact with one another. In contrast, kinetic models that describe physical interactions between biological molecules are useful in interpreting data and testing mechanistic hypotheses (Shih et al., 2012). For this reason, we have pursued the use of kinetic models in our study of CTS.

Several kinetic models have been used to study CTS to various extents. Only one considered CTS in relation to spliceosome assembly (but not splicing itself), and the authors concluded that the stochastic nature of Pol II elongation would contribute to CTS considerations (Murugan and Kreiman, 2012). Among other studies that modeled CTS (Aitken et al., 2011; Schmidt et al., 2011; Carrillo et al., 2010; Murugan and Kreiman, 2012), one of the most important findings was that splicing is best modeled as a multi-step process (Schmidt et al., 2011). At least one mathematical framework has previously been used to model the expected splicing patterns of multi-intron genes (Melamud and Moult, 2009). The authors explored the consequences of treating splicing noise (i.e. AS) as a function of gene length and number of introns, but did not factor CTS into their models.

To address these limitations, we developed a CTS framework that allows me to simulate variable elongation rates, multi-step splicing, and AS. The models are scalable, meaning they can be applied to genes with many introns and unique architectures. By simulating splicing across complex genomes we are therefore able to ask detailed quantitative questions.

1.6 References

Aitken, S., Alexander, R. and Beggs, J. (2011). Modelling Reveals Kinetic Advantages of Co-Transcriptional Splicing. PLoS Comput. Biol. *7*, e1002215.

Alexander, R., Innocente, S., Barrass, J. and Beggs, J. (2010). Splicing-Dependent RNA Polymerase Pausing in Yeast. Mol. Cell *40*, 582–593.

Ast, G. (2004). How did alternative splicing evolve?. Nat. reviews. Genet. *5*, 773–82.

Awan, A. R., Manfredo, A. and Pleiss, J. A. (2013). Lariat sequencing in a unicel-

lular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. Proc. Natl. Acad. Sci. United States Am. *110*, 12762–7.

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J. and Frey, B. J. (2010). Deciphering the splicing code. Nat. *465*, 53–59.

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

Berger, S. L. (2007). The complex language of chromatin regulation during transcription. Nat. *447*, 407–412.

Bhatt, D. M., Pandya-Jones, A., Tong, A.-J. J., Barozzi, I., Lissner, M. M., Natoli, G., Black, D. L. and Smale, S. T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279–90.

Bieberstein, N. I., Carrillo Oesterreich, F., Straube, K. and Neugebauer, K. M. (2012). First exon length controls active chromatin signatures and transcription. Cell Rep *2*, 62–68.

Brody, Y., Neufeld, N., Bieberstein, N., Causse, S. Z., Bohnlein, E. M., Neugebauer, K. M., Darzacq, X., Shav-tal, Y., Karla, M. and Bo, E.-m. (2011). The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. PLoS Biol. *9*, e1000573.

Carmo-Fonseca, M., Geraghty, F., Pereira, H. S., Grosveld, F., Antoniou, M. and Custodio, N. (1999). Inefficient processing impairs release of RNA from the site of transcription. The EMBO journal *18*, 2855–2866.

Carrillo, O., Preibisch, S., Neugebauer, K. M. and Oesterreich, F. C. (2010). Article Global Analysis of Nascent RNA Reveals Transcriptional Pausing in Terminal Exons. Mol. Cell *40*, 571–581.

Danko, C. G., Hah, N., Luo, X., Martins, A. L., Core, L., Lis, J. T., Siepel, A. and Kraus, W. L. (2013). Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. Mol. Cell *50*, 1–25.

de Almeida, S. F., Grosso, A. R., Koch, F., Fenouil, R., Carvalho, S., Andrade, J., Levezinho, H., Gut, M., Eick, D., Gut, I., Andrau, J.-C. C., Ferrier, P. and Carmo-Fonseca, M. (2011). Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. Nat. structural & molecular biology *18*, 977–83.

de la Mata, M., Alonso, C. R., Kadener, S., Fededa, J. P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D. and Kornblihtt, A. R. (2003). A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. Mol. Cell *12*, 525–532.

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

Faustino, N. A. and Cooper, T. A. (2003). Pre-mRNA splicing and human disease. Genes & development *17*, 419–37.

Fox-Walsh, K. L., Dou, Y., Lam, B. J., Hung, S.-P., Baldi, P. F. and Hertel, K. J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc. Natl. Acad. Sci. USA *102*, 16176–16181.

Fox-Walsh, K. L. and Hertel, K. J. (2009). Splice-site pairing is an intrinsically high fidelity process. Proc. Natl. Acad. Sci. USA *106*, 1766–1771.

Gornemann, J., Kotovic, K. M., Hujer, K. and Neugebauer, K. M. (2005). Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. Mol. Cell *19*, 53–63.

Gunderson, F. Q. and Johnson, T. L. (2009). Acetylation by the transcriptional coactivator Gcn5 plays a novel role in co-transcriptional spliceosome assembly. PLoS Genet. *5*, e1000682.

Hao, S. and Baltimore, D. (2013). RNA splicing regulates the temporal order of TNF-induced gene expression. Proc. Natl. Acad. Sci. *110*, 11934–11939.

Hnilicová, J., Hozeifi, S., Dušková, E., Icha, J., Tománková, T. and Staněk, D. (2011). Histone deacetylase activity modulates alternative splicing. PloS one *6*, e16727.

Hoskins, A. A., Friedman, L. J., Gallagher, S. S., Crawford, D. J., Anderson, E. G., Wombacher, R., Ramirez, N., Cornish, V. W., Gelles, J. and Moore, M. J. (2011). Ordered and dynamic assembly of single spliceosomes. Sci. (New York, N.Y.) *331*, 1289–95.

Howe, K. J., Kane, C. M. and Ares Jr., M. (2003). Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. RNA *9*, 993–1006.

Huelga, S. C., Vu, A. Q., Arnold, J. D., Liang, T. Y., Liu, P. P., Yan, B. Y., Donohue, J. P., Shiue, L., Hoon, S., Brenner, S., Ares Jr., M. and Yeo, G. W. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Reports *1*, 167–178.

Ip, J. Y., Schmidt, D., Pan, Q., Ramani, A. K., Fraser, A. G., Odom, D. T. and Blencowe, B. J. (2011). Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. Genome Res. *21*, 390–401.

Izquierdo, J.-M. and Valcárcel, J. (2006). A simple principle to explain the evolution of pre-mRNA splicing. Genes & development *20*, 1679–84.

Karni, R., de Stanchina, E., Lowe, S. W., Sinha, R., Mu, D. and Krainer, A. R. (2007). The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat. Struct. & Mol. Biol. *14*, 185–193.

Kim, S., Kim, H., Fong, N., Erickson, B. and Bentley, D. L. (2011). Pre-mRNA splicing is a determinant of histone H3K36 methylation. Proc. Natl. Acad. Sci. USA *108*, 13564–13569.

Kornblihtt, A. R. (2007). Coupling Transcription and Alternative Splicing. In Alternative Splicing in the Postgenomic Era, (Blencowe, B. J. and Graveley, B. R., eds), vol. 623, chapter 11, pp. 175–189. Landes Biosciences and Springer Sciences+Business Media.

Lewis, B. P., Green, R. E. and Brenner, S. E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. United States Am. *100*, 189–92.

Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J. and Fairbrother, W. G. (2011). Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc. Natl. Acad. Sci. *108*, 11093–11098.

Listerman, I., Sapra, A. K. and Neugebauer, K. M. (2006). Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. Nat. structural & molecular biology *13*, 815–822.

Losson, R. and Lacroute, F. (1979). Interference of nonsense mutations with eukaryotic messenger RNA stability. Proc. Natl. Acad. Sci. *76*, 5134–5137.

Luco, R. F. and Misteli, T. (2011). More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. Curr. Opin. Genet. & Dev. *21*, 366–372.

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

Martins, S. B., Rino, J., Carvalho, T., Carvalho, C., Yoshida, M., Klose, J. M., de Almeida, S. F. and Carmo-Fonseca, M. (2011). Spliceosome assembly is coupled to RNA polymerase II dynamics at the 3' end of human genes. Nat. Struct. & Mol. Biol. *18*, 1115–1123.

Matera, A. G. and Wang, Z. (2014). A day in the life of the spliceosome. Nat. reviews. Mol. cell biology *15*, 108–21.

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S., Wickens, M. and Bentley, D. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. Nat. *385*, 357–361.

Melamud, E. and Moult, J. (2009). Stochastic noise in splicing machinery. Nucleic Acids Res. *37*, 4873–86.

Misteli, T. and Spector, D. L. (1999). RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo. Mol. cell *3*, 697–705.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. Nat. genetics *30*, 13–9.

Muñoz, M. J., de la Mata, M. and Kornblihtt, A. R. (2010). The carboxy terminal domain of RNA polymerase II and alternative splicing. Trends Biochem. Sci. *35*, 497–504.

Murugan, R. and Kreiman, G. (2012). Theory on the coupled stochastic dynamics of transcription and splice-site recognition. PLoS Comput. Biol. *8*, e1002747.

Ni, J. Z., Grate, L., Donohue, J. P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T. a., Blume, J. E. and Ares, M. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes & Dev. *21*, 708–718.

Nilsen, T. W. and Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nat. *463*, 457–63.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J. and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. genetics *40*, 1413–5.

Pandya-Jones, A., Bhatt, D. M., Lin, C. H., Tong, A. J., Smale, S. T. and Black, D. L. (2013). Splicing kinetics and transcript release from the chromatin compartment limit the rate of Lipid A-induced gene expression. RNA *19*, 811–27.

Resch, A., Xing, Y., Alekseyenko, A., Modrek, B. and Lee, C. (2004). Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. Nucleic acids research *32*, 1261–9.

Rigo, F. and Martinson, H. G. (2009). Polyadenylation releases mRNA from RNA polymerase II in a process that is licensed by splicing. RNA *15*, 823–36.

Roberts, G. C., Gooding, C., Mak, H. Y., Proudfoot, N. J. and Smith, C. W. (1998). Co-transcriptional commitment to alternative splice site selection. Nucleic Acids Res. *26*, 1–5.

Schmidt, U., Basyuk, E., Robert, M.-C. M.-C., Yoshida, M., Villemin, J.-P. J.-P., Auboeuf, D., Aitken, S. and Bertrand, E. (2011). Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation. J. Cell Biol. *193*, 819–829.

Shepard, P. J., Choi, E.-A., Busch, A. and Hertel, K. J. (2011). Efficient internal exon recognition depends on near equal contributions from the 3' and 5' splice sites. Nucleic acids research *39*, 8928–37.

Shih, V. F.-S., Davis-Turak, J., Macal, M., Huang, J. Q., Ponomarenko, J., Kearns, J. D., Yu, T., Fagerlund, R., Asagiri, M., Zuniga, E. I. and Hoffmann, A. (2012). Control of RelB during dendritic cell activation integrates canonical and noncanonical NF-κB pathways.

Singh, J. and Padgett, R. A. (2009). Rates of in situ transcription and splicing in large human genes. Nat. Struct. & Mol. Biol. *16*, 1128–1133.

Swinburne, I. A., Miguez, D. G., Landgraf, D. and Silver, P. A. (2008). Intron length increases oscillatory periods of gene expression in animal cells. Genes & development *22*, 2342–6.

Tennyson, C. N., Klamut, H. J. and Worton, R. G. (1995). The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. Nat. genetics *9*, 1–7.

Thanaraj, T. A., Stamm, S., Clark, F., Riethoven, J.-J., Le Texier, V. and Muilu, J. (2004). ASD: the Alternative Splicing Database. Nucleic acids research *32*, D64–9.

Veloso, A., Kirkconnell, K. S., Magnuson, B., Biewen, B., Paulsen, M. T., Wilson, T. E. and Ljungman, M. (2014). Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. Genome Res. *0*, 0.

Venables, J. P. (2004). Aberrant and alternative splicing in cancer. Cancer research *64*, 7647–54.

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

Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J. S., Kristiansen, K., Krogh, A., Wang, J. and Porse, B. T. (2012). Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. Genome biology *13*, R35.

Wetterberg, I., Baurén, G., Wieslander, L. and Bauren, G. (1996). The intranuclear site of excision of each intron in Balbiani ring 3 pre-mRNA is influenced by the time remaining to transcription termination and different excision efficiencies for the various introns. RNA *2*, 641–651.

Xiao, X., Wang, Z., Jang, M. and Burge, C. B. (2007). Coevolutionary networks of splicing cis-regulatory elements. Proc. Natl. Acad. Sci. United States Am. *104*, 18583–8.

Zahler, A., Lane, W., Stolk, J. and Roth, M. (1992). SR proteins: a conserved family of pre-mRNA splicing factors. Genes & Dev. *6*.

Zhang, C., Frias, M. A., Mele, A., Ruggiu, M., Eom, T., Marney, C. B., Wang, H., Licatalosi, D. D., Fak, J. J. and Darnell, R. B. (2010). Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Sci. (New York, N.Y.) *329*, 439–43.

Zhou, H.-L., Hinman, M., Barron, V., Geng, C., Zhou, G., Luo, G., Siegel, R. and Lou, H. (2011). Hu proteins regulate alternative splicing by inducing localized histone hyperacetylation in an RNA-dependent manner. Proc. Natl. Acad. Sci. USA *108*.

2 Sequence Signatures and Polymerase Dynamics Favor Co-transcriptional Splicing Genome-wide

2.1 Abstract

Recent genome-wide studies have revealed that, to varying degrees, messenger RNA splicing occurs co-transcriptionally raising the question of whether co-transcriptional splicing (CTS) is generally functionally important. While single gene studies are often focused on trans-acting splicing factors, we hypothesized that genome-wide analyses that average-out intron-specific mechanisms might reveal kinetic determinants of CTS. To this end, we constructed a scalable mathematical model of the kinetic interplay of RNA synthesis and CTS and parameterized it with chromatin-associated and global run-on RNA-seq data. Examining vertebrate genomes, we found that protein-coding genes appear to be under evolutionary pressure to favor CTS, but via distinct determinants for different groups of genes: for example, housekeeping genes exhibited longer transcriptional read-through, higher splice site scores, but less polymerase pausing than regulated genes. Together, our findings indicate that, while mechanisms that reinforce it are diverse, regulation of CTS is intrinsic to the control of gene expression.

2.2 Introduction

Messenger RNA (mRNA) synthesis is a highly regulated process in which transcription factors and chromatin modifying factors coordinate with Pol II to produce a nascent strand of RNA. The nascent pre-mRNA is processed by 5' capping, 3' polyadenylation and pre-mRNA splicing - the removal of non-coding introns. Complete splicing is necessary for proper mRNA export, stability, and protein function. RNA processing steps can in principle be initiated and completed during the transcription process, i.e., co-transcriptionally, but may also occur post-transcriptionally (Alexander et al., 2010; Carrillo et al., 2010; Neugebauer, 2002; Pandya-Jones and Black, 2009; Shatkin and Manley, 2000; Singh and Padgett, 2009; Tennyson et al., 1995; Tilgner et al., 2012; Wetterberg et al., 1996). Nonetheless, it is now well established that a large fraction of genes undergo co-transcriptional splicing in metazoan genomes (Khodor et al., 2011; Tilgner et al., 2012).

Because of the constraints imposed on CTS by transcriptional elongation, an intron's fate may be dramatically affected by the elongation dynamics of Pol II. Indeed, a slower Pol II can result in increased use of a weak 5' splice site in reporter gene constructs (de la Mata et al., 2003; Howe et al., 2003). Splice site choice can be altered in human cell lines by removing downstream pausing sites (Shukla et al., 2011) or pharmacologically slowing down Pol II (Ip et al., 2011), and Pol II pausing generally correlates with an increase in CTS activity (Alexander et al., 2010; Batsche et al., 2006; Carrillo et al., 2010). In addition to the kinetic coupling between splicing and transcription, much of the cellular machinery for regulating transcription is also important for CTS. Spliceosome recruitment may be coordinated with transcription (Bentley, 2002; Close et al., 2012; Gornemann et al., 2005; Gunderson and Johnson, 2009; Hirose and Manley, 2000); for example, the Carboxy-Terminal Domain (CTD) of Pol II is known to recruit common factors (de la Mata and Kornblihtt, 2006), while Pol II lacking the CTD shows splicing defects (McCracken et al., 1997).

Despite clear evidence of co-transcriptional spliceosome assembly and per-

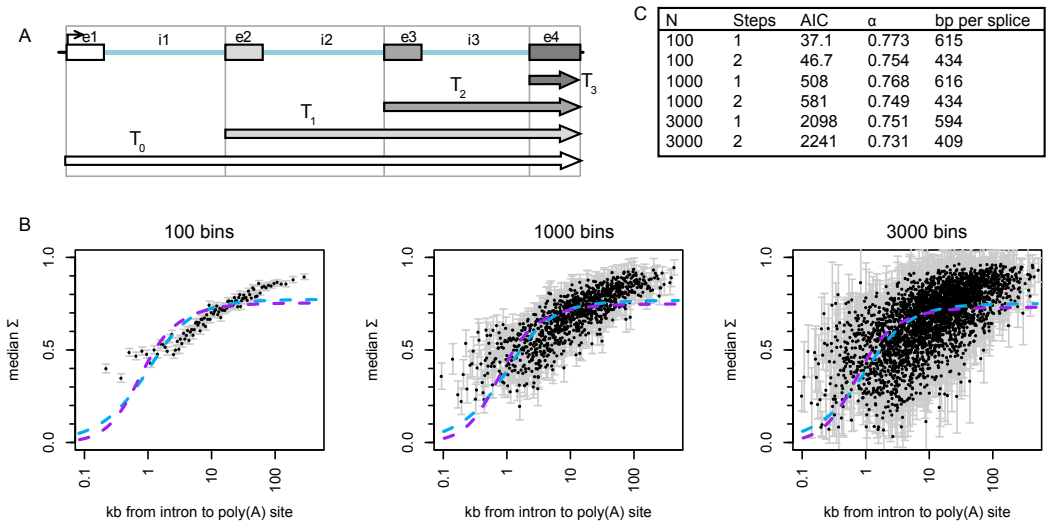
vasive CTS, some splicing occurs post-transcriptionally (Bhatt et al., 2012; Tardiff et al., 2006; Vargas et al., 2011). Which instances of CTS are functionally important, and how CTS is reinforced for specific splicing events, remain fundamentally unanswered questions. Several examples of functional CTS have emerged: CTS of the first intron may regulate chromatin modifications that reinforce transcription initiation (Bieberstein et al., 2012); during the innate immune response, CTS may facilitate rapid gene expression (Hao and Baltimore, 2013; Pandya-Jones et al., 2013). In this regard, determining the kinetics of the splicing reaction per se is critical for understanding both qualitative and quantitative aspects of gene expression (Darnell, 2013).

Here we examined how transcriptional and splicing kinetics may affect the probability of CTS. We constructed a scalable mathematical model of splicing coupled to transcriptional elongation in order to quantitatively asses how gene structure and sequence features contribute to CTS. Next generation sequencing methods have generated quantitative, chromatin-associated mRNA information (Bhatt et al., 2012; Djebali et al., 2012), and here we present methods to extract kinetic information from such datasets, thus allowing us to parameterize the kinetic CTS model based on actual transcriptome measurements.

In modeling CTS, we reasoned that while splicing of specific introns may be critically determined by trans-acting splicing factors, the common kinetic basis of CTS may be more apparent when considering cohorts of genes in which gene specific mechanisms are averaged out. This analysis revealed that gene features that contribute to CTS were over-represented. Examination of multiple genomes revealed that these gene structure and sequence features are in fact identifiable as genetic signatures that correlate with splicing dynamics to a remarkable degree. By expanding the model we were able to simulate co-transcriptional outcomes of multi-intron genes genome-wide. Our results show that while genes may differ widely in their cis-determinants of CTS, the kinetic integration of transcription and splicing is an intrinsic feature of gene expression control.

2.3 Results

2.3.1 A model to examine the contributions of gene structure and sequence features to the control of co-transcriptional constitutive splicing



N	Steps	AIC	α	bp per splice
100	1	37.1	0.773	615
100	2	46.7	0.754	434
1000	1	508	0.768	616
1000	2	581	0.749	434
3000	1	2098	0.751	594
3000	2	2241	0.731	409

Figure 2.1: Fitting CTCS models to ENCODE RNA-seq data. A: Diagram of relevant time periods of single-intron model. B: Nuclear poly(A)-depleted RNAseq. Median Σ plus S.E.M. for introns grouped into 100, 1,000 and 3,000 equally populated bins (left, middle, right, respectively) is shown as a scatterplot. Dashed lines indicate model fits to 2-parameter model (blue, one-step; purple, 2-step). C. Parameters and fit scores (AIC criteria) for all fits in B.

We first modeled co-transcriptional splicing of individual introns using a one- or two-step splicing model (cf. Schmidt et al., 2011). An intron's probability of being spliced co-transcriptionally σ is determined by its splicing rate constant and the duration of the transcriptional phase following the synthesis of the 3' splice site but prior to mRNA polyadenylation (Fig. 2.1A).

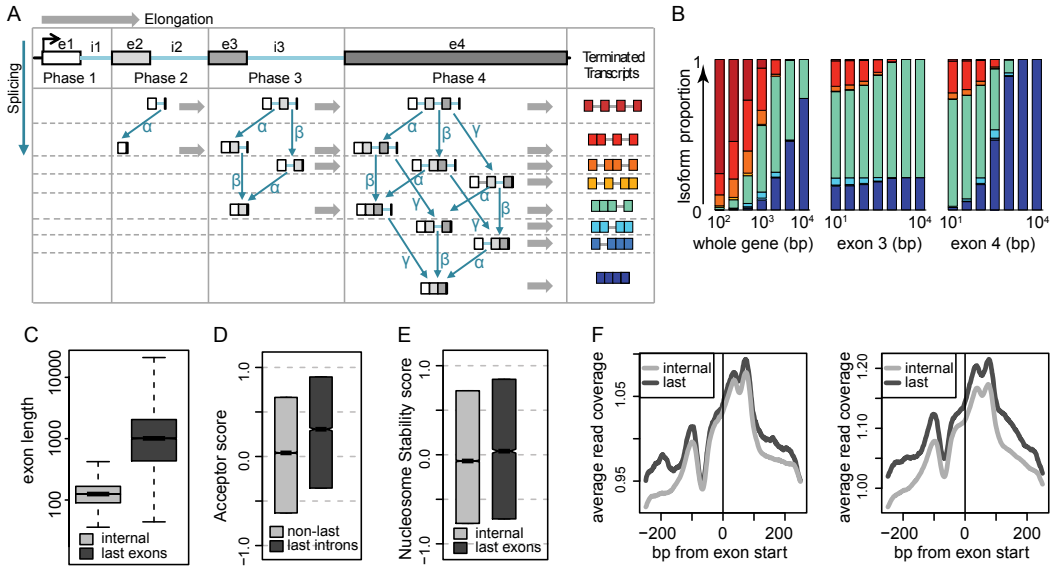


Figure 2.2: Model of co-transcriptional constitutive splicing (CTCS). A: Model schematic showing all possible reactions and species for a 3-intron gene. The 8 possible isoforms that can exist when transcription is complete are color-coded at right. B. Model simulations of the 3-intron gene. Each column represents the distribution of the 8 species after each simulation. Left: the lengths of all introns and exons were scaled up and down by a constant factor. Middle: length of exon 3 was varied; Right: length of exon 4 was varied. C: Distribution of exon lengths among last and internal exons in the human genome. D: Average splicing acceptor scores in last and non-last introns genome-wide. All boxes show the extent of the 50% inter-quartile range and the notches estimate a 95% confidence interval for the median. E: Nucleosome stability scores in the first 147 bp of last versus internal exons. F: Average genome-wide MNase-seq signal in $k_5$62 cells and GM12878 cells over internal exon starts versus last exon starts.

We next combined models of independent introns to generate a model of co-transcriptional constitutive splicing (CTCS). Our CTCS model enables simulations of multi-intron genes of any complexity and allows us to quantitatively assess the effects of genome structure and kinetic rates on genome-wide splicing outcomes (Fig. 2.2A). Using parameters fit to RNA-seq data to simulate a test gene (See Experimental Procedures, and Fig. 2.1B), the CTCS model recapitulated a central point of the kinetic theory of CTS control (Carrillo et al., 2010; Tilgner et al., 2012; Wetterberg et al., 1996): namely, that long genes (Fig. 2.2B, left), and specifically genes with long last exons (Fig. 2.2B, right) would favor CTS (because they provide

more time for splicing), whereas the length of the penultimate exon (which has no influence on the splicing time of the last intron) would be less important (Fig. 2.2B middle). These conclusions are robust to specific splicing parameter values (data not shown).

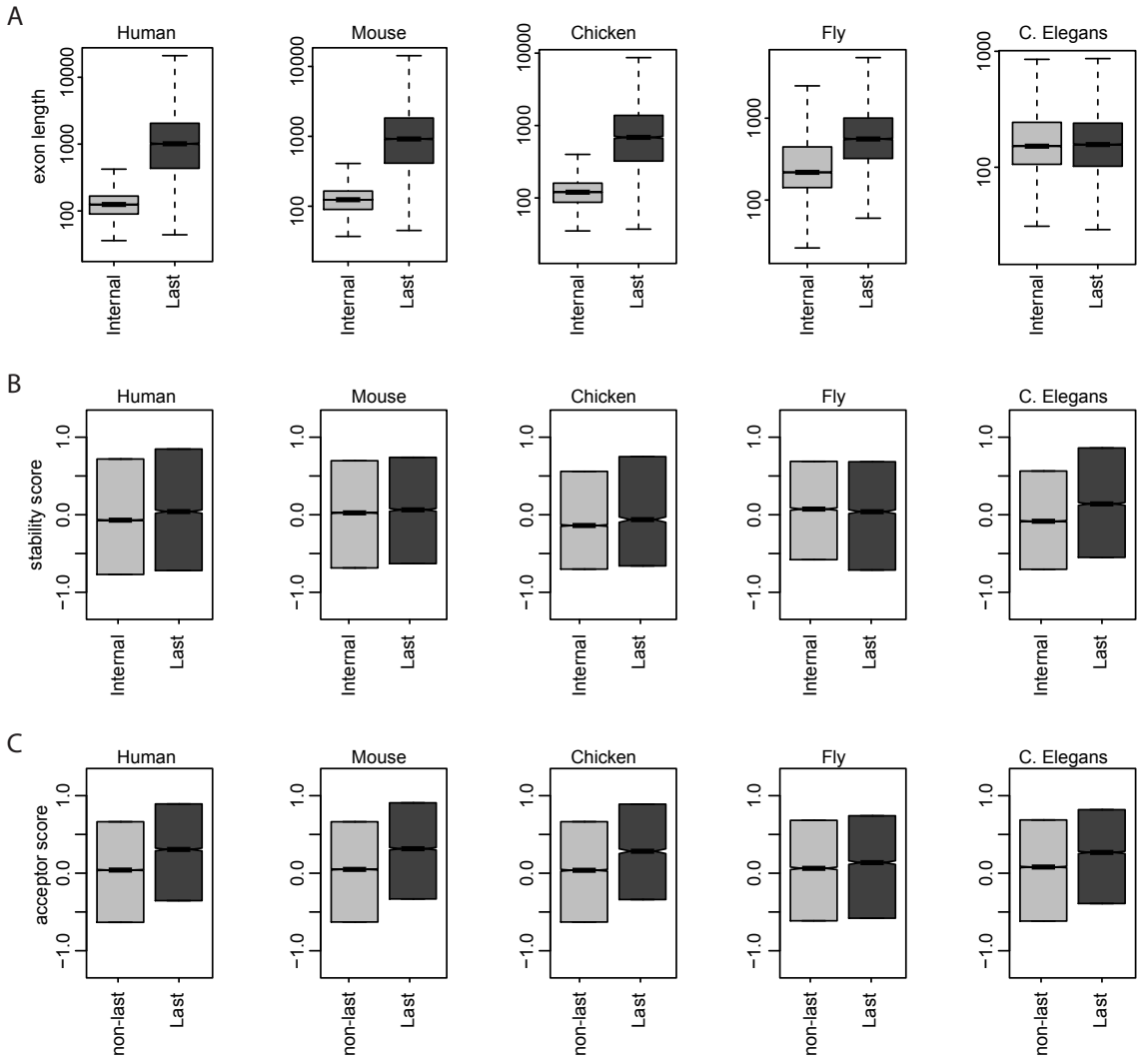


Figure 2.3: Analysis of CTS determinants in metazoan genomes. Last exon lengths (A), last intron splice site strength (B), and, last exon nucleosome stability (C) are shown for mouse, chicken, fly and worm genomes. All boxes show the extent of the 50% inter-quartile range and the notches estimate a 95% confidence interval for the median. Whiskers (show in A only) indicate the range of the data.

It has previously been suggested that long last exons may have evolved to

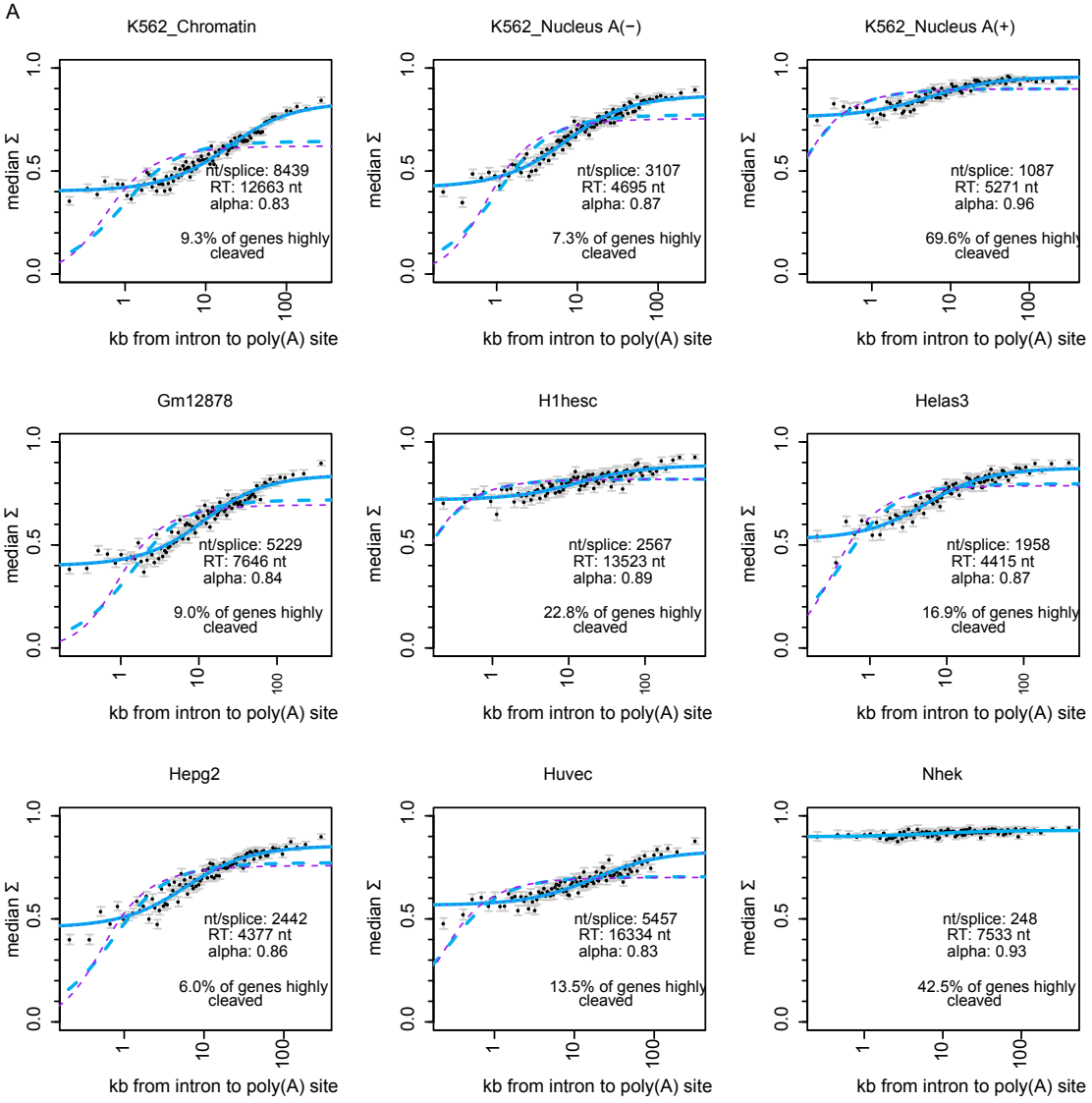
optimize CTS (Figs. 2.2C, 2.3A; Carrillo et al., 2010). We expected that if complex genomes evolved under pressure to maintain high CTS efficiency, other genomic signatures besides last exon length, that influence CTS, may be identifiable. Since our model predicted the excision of last introns to be the limiting step in determining the CTS efficiency, we compared acceptor splice site strengths across several genomes based on species-specific sequence motifs. Indeed, we found evidence for conservation of higher average acceptor scores in last introns compared to non-last introns in several vertebrate genomes (Figs. 2.2D, 2.3B). Since the presence of nucleosomes can inhibit transcription elongation (Subtil-Rodriguez and Reyes, 2010) and thus provide more time for CTS, we next tested whether nucleosome stability was enriched at 3' exons. We first evaluated nucleosome stability across several species using a simple algorithm based on biophysical considerations (Vaillant et al., 2007), and found that nucleosomes are indeed expected to be more stable at terminal exons than at internal exons (Figs. 2.2E, 2.3C). Furthermore, analysis of human cell MNase-seq data (ENCODE Project Consortium, 2011) revealed that nucleosomes are enriched in the proximity of last exons compared to internal exons (Fig. 2.2F). Interestingly, it was previously observed that nucleosomes are present in higher abundance in exons flanking weaker splice sites, both in internal and last exons (Tilgner et al., 2009), reinforcing the hypothesis that nucleosome occupancy and splice sites may balance each other to control CTS.

2.3.2 Fitting the model to genome-wide co-transcriptional splicing data reveals a role for additional time past the poly(A) site

To parameterize our model, we took advantage of existing RNA-seq measurements of purified cellular compartments in K562 cells (Djebali et al., 2012) to estimate the steady-state spliced fraction (Σ) for each intron (17,266 introns in 2,768 genes; see Experimental Procedures). For this analysis, we restricted our set of genes to those that use their most downstream annotated poly(A) site, as determined by RNA-seq of the cytoplasmic fraction (13,650 introns in 2,136 genes).

Figure 2.4: Fitting CTCS and CTCS+T_{FIT} models to ENCODE RNA-seq data. A. Median Σ for 100 bins of introns binned by distance to poly(A) site were plotted and fit to the two models. Σ was calculated from RNA-seq in different fractions of $k_5$62 cells (*top*), and in nuclear poly(A) depleted fractions of other human cell lines (*bottom two rows*). All genes were filtered for poly(A) site usage by examining RNA-seq of the cytoplasmic fraction of the respective cell type, as in the main text (see Methods section). Of the genes that passed this filter, we indicate the percent of these whose ratio of Up' reads to Up' + Down' reads (in the fraction where we measured Σ) as highly cleaved. Best-fit parameters for the three-parameter model are also shown on the figures. B. Table of parameters and goodness-of-fit tests in the $k_5$62 Nuclear poly(A)-depleted fraction, for various bin sizes. For the 100 and 1000 bin cases, the fit is also shown for the two-step splicing reactions. C Summary of fit parameters (nt/splice, *middle*; time allowance, *right*) for 100 and 1000 bins in $k_5$62 nuclear poly(A)-depleted fraction for 1-5 steps per splicing reaction. Error bars represent standard deviation of parameters. *Left:* AIC criteria of the 100 bins and 1000 bins fits (open circles and close circles, respectively) were compared to the minimum (best) fit to obtain a Fit quality' $e^{((AICmin-AIC)/2)*100]}$, which represents the percent likelihood that a given fit reduces as much of the variance as the best fit obtained.

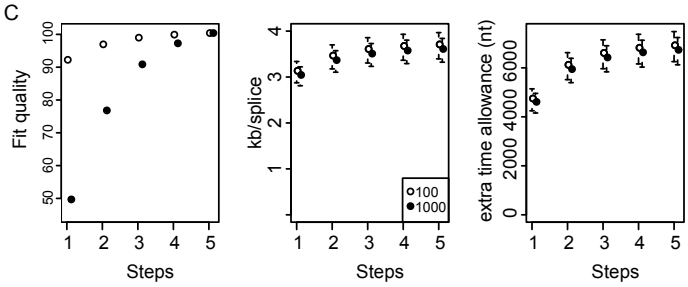
A



B

N	Steps	AIC	α	bp per splice	Extra time
100	1	10.4	0.867	3107	4695
100	2	10.3	0.871	3438	6069
1000	1	294	0.858	3016	4557
1000	2	293	0.862	3339	5892
3000	1	1626	0.849	3374	5312

C



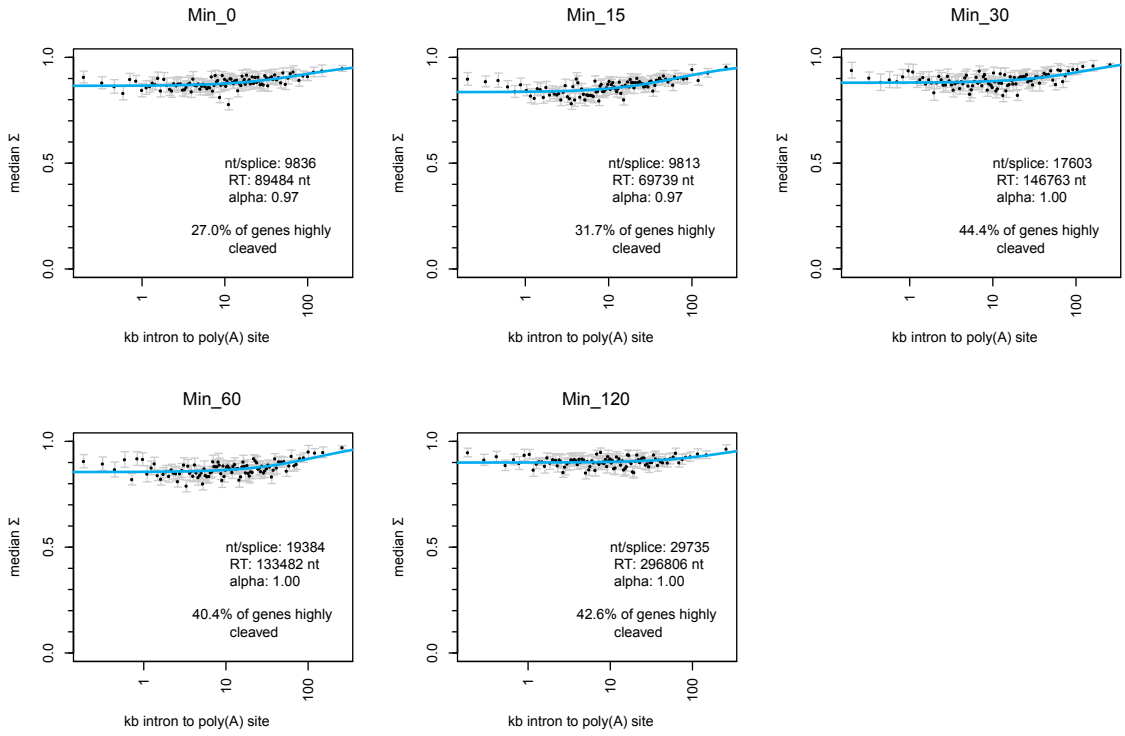


Figure 2.5: Fitting the CTCS+T_{FIT} model to mouse macrophage RNA-seq data. Median Σ for 100 bins of introns binned by distance to poly(A) site were plotted and fit to the CTCS+T_{FIT} model. Σ was calculated from RNA-seq of the chromatin fraction from Bhatt *et al.* 2012 dataset. All genes were filtered for poly(A) site usage by examining the average cleavage ratio of RNA-seq of the cytoplasmic fraction, as in the main text (see Methods section). Of the genes that passed this filter, we indicate the percent of these whose ratio of Up' reads to Up' + Down' reads (in the fraction where we measured Σ) as highly cleaved'. Best-fit parameters for the three-parameter model are also shown on the figures.

Examining non-polyadenylated nuclear transcripts, median Σ strongly correlates with distance to the poly(A) site, in K562 cells, as reported previously (Tilgner et al., 2012), and in other cell types (Fig. 2.4). By fitting our model to the median Σ of introns binned according to distance from poly(A) site, we obtained a ratio of splicing rate to elongation speed (see Experimental Procedures). We also examined RNA-seq from the chromatin fraction of mouse macrophages (Bhatt et al., 2012), but since this dataset contains genes that are polyadenylated and thus post-transcriptionally associated with chromatin (Fig. 2.5, Bhatt et al., 2012; Pandya-Jones et al., 2013), we could not derive meaningful co-transcriptional

parameters using this procedure.

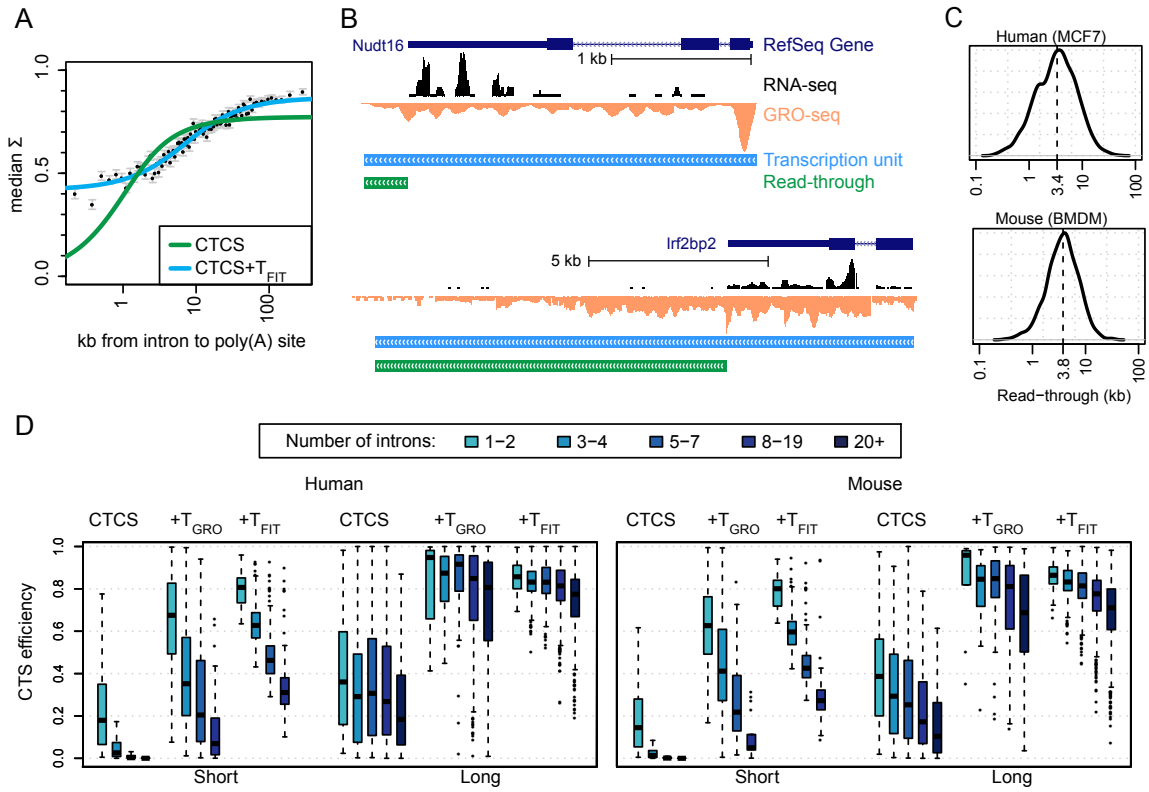


Figure 2.6: Pol II read-through may contribute to delay time following transcription of the poly(A) site. A: Deriving kinetic parameters by fitting CTCS modelf to median spliced fraction Σ of 100 equally populated bins of introns. Green and Blue lines are the fits to the CTCS or CTCS+T_{FIT} models, respectively. T_{FIT} refers to the average additional time after Pol II transcribes the poly(A) site, as determined by the model fit. B: UCSC browser tracks showing GRO-seq traces in mouse macrophages for representative genes *nudt16* and *irf2bp2*. Schematic boxes indicate the read-through lengths and transcription units determined computationally (Allison *et al.*, 2013). C. Distribution of transcriptional read-through as measured by GRO-seq in human MCF7 cells (top) and mouse macrophages (bottom). D. Simulations of CTS in all human genes (left) and mouse genes (right). Genes were split into four evenly sized groups based on total gene length [short (< 6,444 bp), medium short (6,444 - 20,252bp), medium long (20,257 - 57,229 bp), and long (> 57,229 bp)], and further subdivided by the number of introns. Boxplots of CTS efficiency for short and long groups are shown for simulations in three models: CTCS, CTCS+T_{FIT}, and CTCS+T_{GRO}. T_{GRO} refers to the additional time after Pol II transcribes the poly(A) site if transcription proceeds to the termination sites identified by GRO-seq, in individual genes for which GRO-seq measurements are available.

Fitting the CTCS model to the nuclear poly(A)-depleted K562 data resulted in a ratio of elongation rate to splicing rate of 615 bp/splicing event, equating to an intron half-life of 9 seconds if elongation is 3 kb/min (Kwak and Lis, 2013). The fit was robust to the binning procedure used (Figs. 2.1B,C), and a two-step model did not improve the fit to the model (Fig. 2.1C). However, previous studies have derived estimates of co-transcriptional splicing rates in diverse organisms ranging from a 30 second half-life to a 5-10 minute 'splicing completion time' (Aitken et al., 2011; Schmidt et al., 2011; Singh and Padgett, 2009; Tardiff et al., 2006).

A close inspection of the poly(A)-depleted nuclear RNA-seq data revealed that the CTCS model underestimated the steady-state splicing probabilities of introns proximal to the poly(A) site (Figs. 2.6A, 2.1B), similar to findings in yeast (Carrillo et al., 2010). This disconnect could be due in part to conditions that prolong association of nascent, un-polyadenylated RNA with the chromatin template beyond the time predicted by the poly(A) site (Boireau et al., 2007), such as a transcriptional pause near 3' ends (Carrillo et al., 2010), or transcriptional read-through past the poly(A) site. We therefore modified our CTCS model to include a post-poly(A) site time delay (model CTCS+T), and fit this model to the chromatin-associated RNA-seq data.

Remarkably, allowing for this additional time interval (T_{FIT}) in our model dramatically improved the fit to the data (Figs. 2.6A, 2.3). The best fit was obtained when the median time delay was equivalent to elongating 4.7 kb past the poly(A) site (see Experimental Procedures), and with a new value of 3.1 kb/splicing event for the elongation to splicing ratio. Assuming an elongation rate of 3 kb/min, these values equate to a median 3' delay of 94 seconds and a median intron half-life of 43 seconds. This second estimate of median splicing half-life is more consistent with, though still on the fast side of those previously reported (Aitken et al., 2011; Schmidt et al., 2011; Singh and Padgett, 2009; Tardiff et al., 2006). If elongation rates turned out to be slower, those half-life estimates would proportionally increase.

To investigate whether transcriptional read-through could account for the extra time observed in CTCS+T_{FIT} , we measured the extent of active transcription

Figure 2.7: Variable Pol II elongation kinetics favor CTS. A: Boxplots of the nucleosome stability score of each gene in indicated gene categories based on gene length and intron numbers. Nucleosome scores were averaged over the region of the gene that encompassed the second through final exon. B: Average MNase-seq signal over all exon starts in $k_5$62 (*left*) and GM12878 cells (*right*) in short and long genes, split up by intron number. C: Average PolS2 ChIP-seq signal in the 1 kb upstream of the poly(A) site (*left*), and gene expression, indicated by fragments per kb per million reads sequenced (FPKM) (*right*) in $k_5$62 cells for short and long genes split up by number of introns. D: Average PolS2 signal downstream of the poly(A) site. Traces are normalized to the average of the 1kb upstream of the poly(A) site for each category. E: Simulations of CTS efficiency in short and long human genes using model CTCS+T_{FIT}. Boxplots of simulations in four separate modeling conditions (See Supplemental Experimental Procedures) are shown: T_{FIT}: same as Fig. 2D; + Δ*elong*: elongation rate of each gene was modulated as an inverse function of nucleosome stability. + Δ*splice*: kinetic splicing rate was modulated so that last introns had a rate twice the speed of other introns. + Δ*both*: both elongation and splicing rates were modulated.

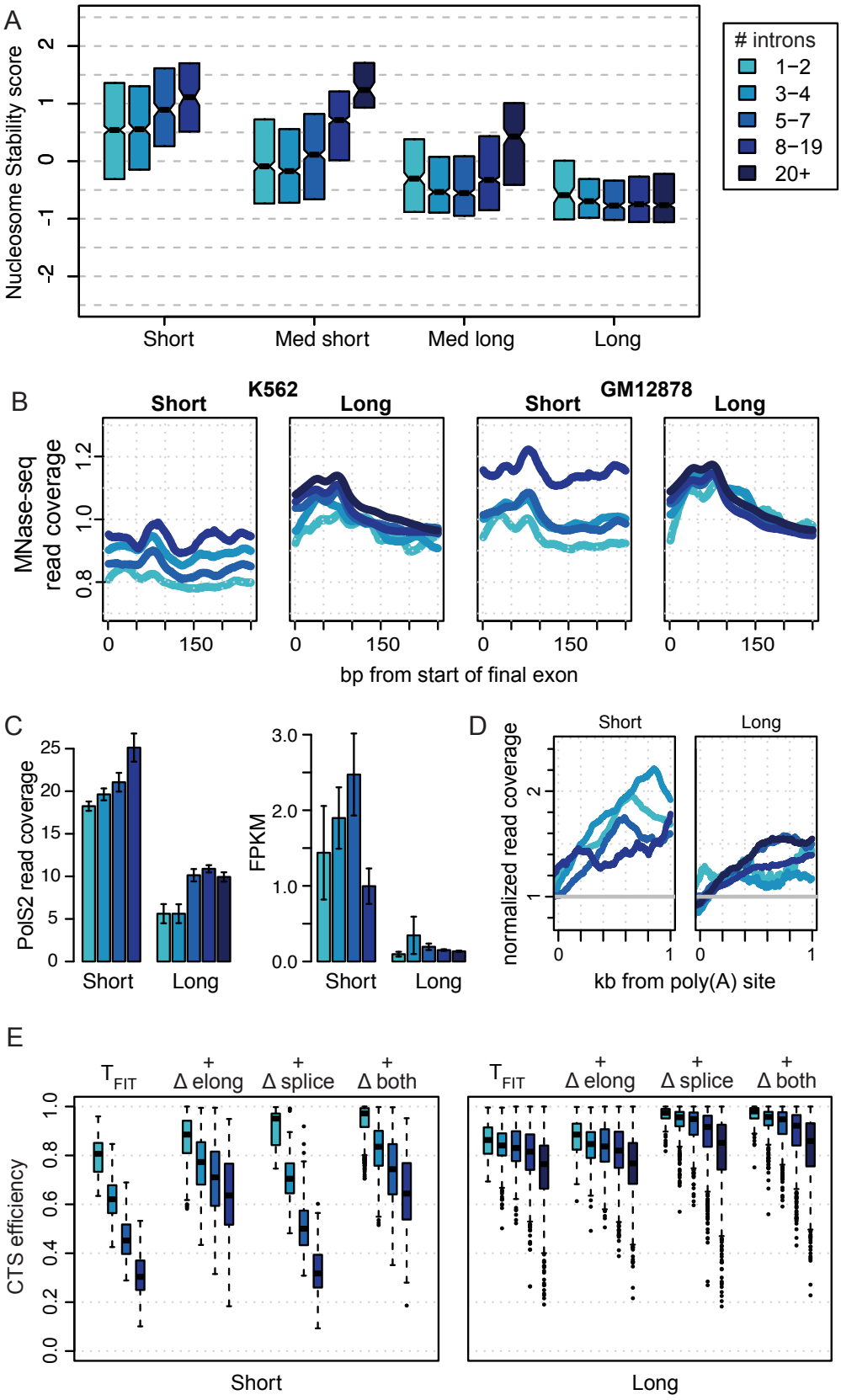
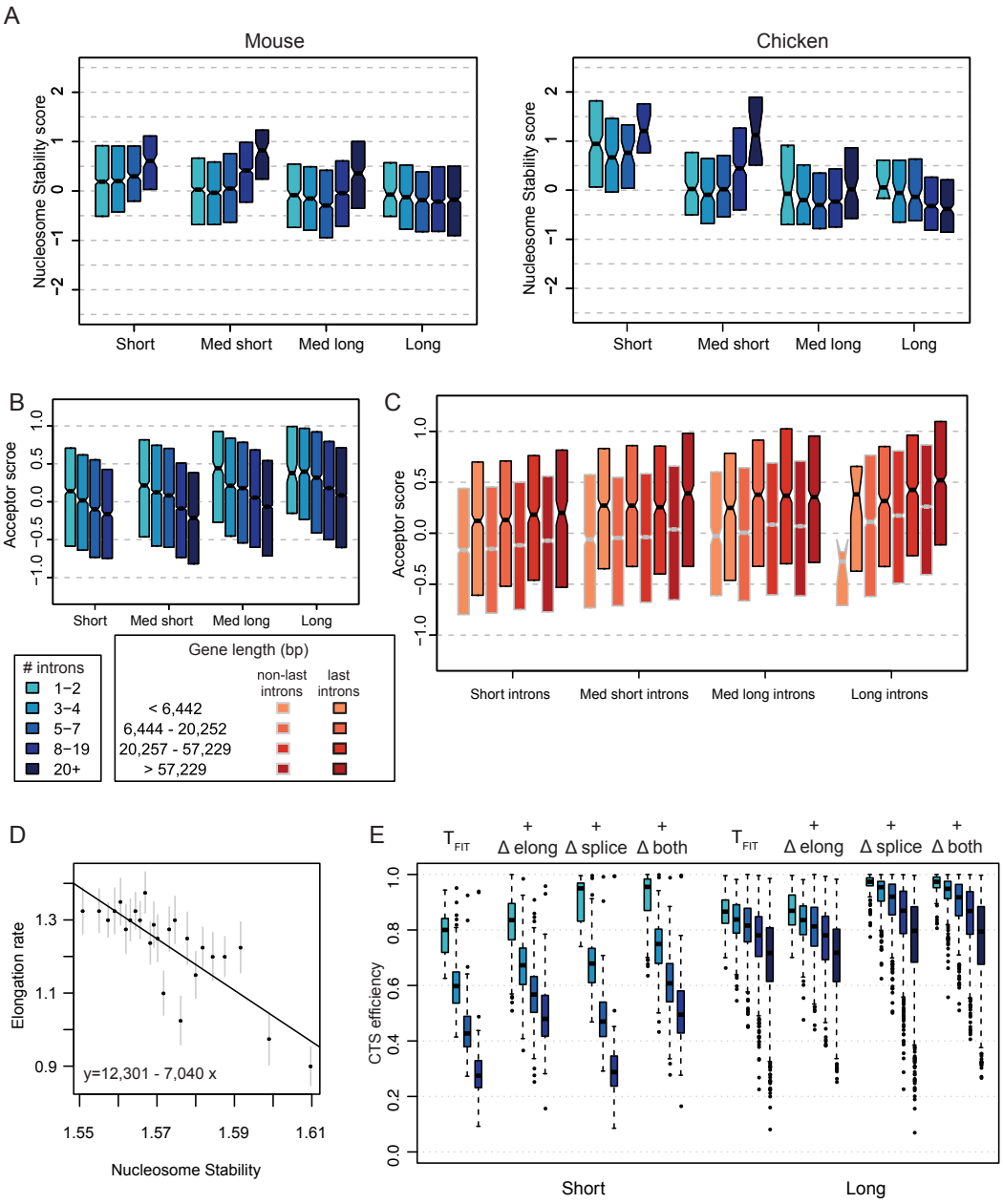

A
Nucleosome Stability score
2
1
0
-1
-2
Short
Med short
Med long
Long
introns
1−2
3−4
5−7
8−19
20+
B
K562
GM12878
Short
Long
Short
Long
MNase-seq
read coverage
1.2
1.0
0.8
0
150
bp from start of final exon
C
PolS2 read coverage
25
20
15
10
5
0
Short
Long
FPKM
3.0
2.0
1.0
0.0
Short
Long
D
Short
Long
normalized read coverage
2
1
0
1
kb from poly(A) site
E
T FIT
+ Δ elong
+ Δ splice
+ Δ both
CTS efficiency
1.0
0.8
0.6
0.4
0.2
0.0
Short
Long

Figure 2.8: Relationship between gene structure and other CTS determinants in vertebrate genomes. A: Nucleosome stability in exons was calculated for all genes in the same categories of gene length and intron count as in figure 3A, for Mouse and Chicken genomes. B: Acceptor splice site strength in humans was plotted for the same 19 categories. C: Acceptor strength of last introns (*dark outlines*) and non-last introns (*light outlines*) was plotted for different groups of introns binned by length of total gene (colors) and length of intron. Gene length groups were the same used earlier: intron length groups were composed of four equal sized groups of short introns (< 511 bp), medium short introns (511-1539 bp), medium long introns (1540-4106 bp), and long introns (> 4106 bp). All boxes show the extent of the 50% inter-quartile range and the notches estimate a 95% confidence interval for the median. D: Elongation rate (kb/min) as a function of average nucleosome stability (-Energy , AU) in $k_5$62 cells. The linear regression to the 1,166 datapoints, which is used to predict elongation rate in the rest of the genome, is depicted ($p < 2.662e\text{-}16$, $R = 0.25$). For clarity, we display the median and standard error of 25 equally-populated bins of genes. E: Boxplots of simulated CTS efficiency of long and short mouse genes. T_{FIT}: all mouse genes modeled with the basal splicing rate (3.1 kb/splice) and time allowance (4.7 kb) derived from the fit to human RNA-seq data. + $\Delta elong$: elongation rate during the 147 bp window at the beginning of each exon was modulated as an inverse function of nucleosome stability in mouse genome. + $\Delta both$: kinetic splicing rate was modulated so that last introns had a rate twice the speed of other introns. + $\Delta both$: both elongation and splicing rates were varied simultaneously.


A
Mouse
Chicken
Nucleosome Stability score
Short
Med short
Med long
Long
B
Acceptor scroe
C
Acceptor score
Short introns
Med short introns
Med long introns
Long introns
introns
1−2
3−4
5−7
8−19
20+
Gene length (bp)
non-last introns
last introns
< 6,442
6,444 - 20,252
20,257 - 57,229
> 57,229
D
Elongation rate
y=12,301 - 7,040 x
Nucleosome Stability
E
T FIT
+ Δ elong
+ Δ splice
+ Δ both
CTS efficiency
Short
Long

associated with each gene using a novel software tool (Allison et al., 2013) to analyze GRO-seq (Core et al., 2008) data. We used our previously characterized GRO-seq dataset (Kaikkonen et al., 2013) in mouse macrophages (Fig. 2.6B), and used an existing dataset of human MCF7 cells (Li et al., 2013) to measure how far pol II activity extends. As cleavage and polyadenylation may occur prior to termination of pol II activity, these measurements put an upper limit on the pre-cleavage read-through distance and transcription unit.

Most genes showed pol II activity well beyond the annotated poly(A) site (Fig. 2.6B) indicating median read-through distances in macrophages and MCF7 of 3.2 and 3.8 kb (equivalent to 68 and 76 sec), respectively (Fig. 2.6C). These data suggest that transcriptional read-through may contribute but does not fully account for the estimated delay in polyadenylation after traversing the poly(A) site.

To investigate the effect of the 3' delay T on CTS, we calculated the CTS efficiency of all human and mouse genes using the CTCS and CTCS+T models and a splicing rate of 3.1 kb/splice (Fig. 2.6D). CTS efficiency was defined as the fraction of transcripts in which all introns are removed prior to cleavage and polyadenylation (see Experimental Procedures), though some level of co-transcriptional splicing may be occurring even for transcripts that are scored as incompletely spliced. With no 3' delay T, less than 50% of transcripts were predicted to be completely spliced upon polyadenylation. Genes with many introns, especially short genes, showed even lower CTS efficiency. With the 3' delay equated to either the median fitted delay time ($+T_{FIT}$), or to the time equivalent of GRO-seq-measured read-through distances in individual genes ($+T_{GRO}$), resulted in an increase in CTS efficiency. However, CTS efficiency remained dependent on gene length and the number of introns, such that even these time delays are not sufficient to ensure that all introns are spliced in short genes, especially those with many introns.

2.3.3 Predicted CTS efficiency enhanced by selective Pol II pausing at 3' ends

Our model revealed that some genes' structures predispose their transcripts for inefficient CTS. However, if efficient CTS were selected for during the evolution of complex genomes, we would expect to find compensatory signatures of other CTS determinants. Indeed, we found that nucleosome stability of genes is markedly higher in short genes than long genes in vertebrate genomes (Figs. 2.7A, 2.8A). This trend could explain the finding that Pol II elongation rate is positively correlated with gene length (Veloso et al., 2014). Furthermore, among short genes, those with high numbers of introns had very high average nucleosome stability scores. No similar compensatory signatures were observed for splice site scores (Fig. 2.8B), which correlate with intron length and are universally stronger in last introns (Figs. 2.3C, 2.8C). We examined nucleosome occupancy in K562 and GM12878 cells using the MNase-seq data. Within short genes, nucleosome density increased with increasing numbers of introns (Fig. 2.7B).

We next tested whether we could find evidence of differential Pol II dynamics in long and short genes by examining Pol II CTD SerineS2 phosphorylation (PolS2) in K562 cells in the vicinity of poly(A) sites that were not within 1 kb of any other genes' starts or ends. In the 1kb upstream of the poly(A) site, PolS2 read densities were higher for short genes than long genes (Fig. 2.7C) (though PolS2 read densities also correlate with gene expression levels), and genes with more introns have disproportionately high PolS2 densities. Furthermore, short genes generally had prominent peaks of PolS2 signal after the poly(A) site (Fig. 2.7D), whereas long genes had lower and broader peaks. These data indicate that differential regulation of Pol II elongation could be sufficient to confer high CTS efficiency to all genes, regardless of gene structure.

To test this hypothesis we simulated all human genes (Fig. 2.7E) and mouse genes (Fig. 2.8E) with our CTCS+T_{FIT} model using variable elongation parameters. Using experimentally determined elongation rates of long genes in K562 cells (Veloso et al., 2014), we tested that elongation rates were negatively correlated

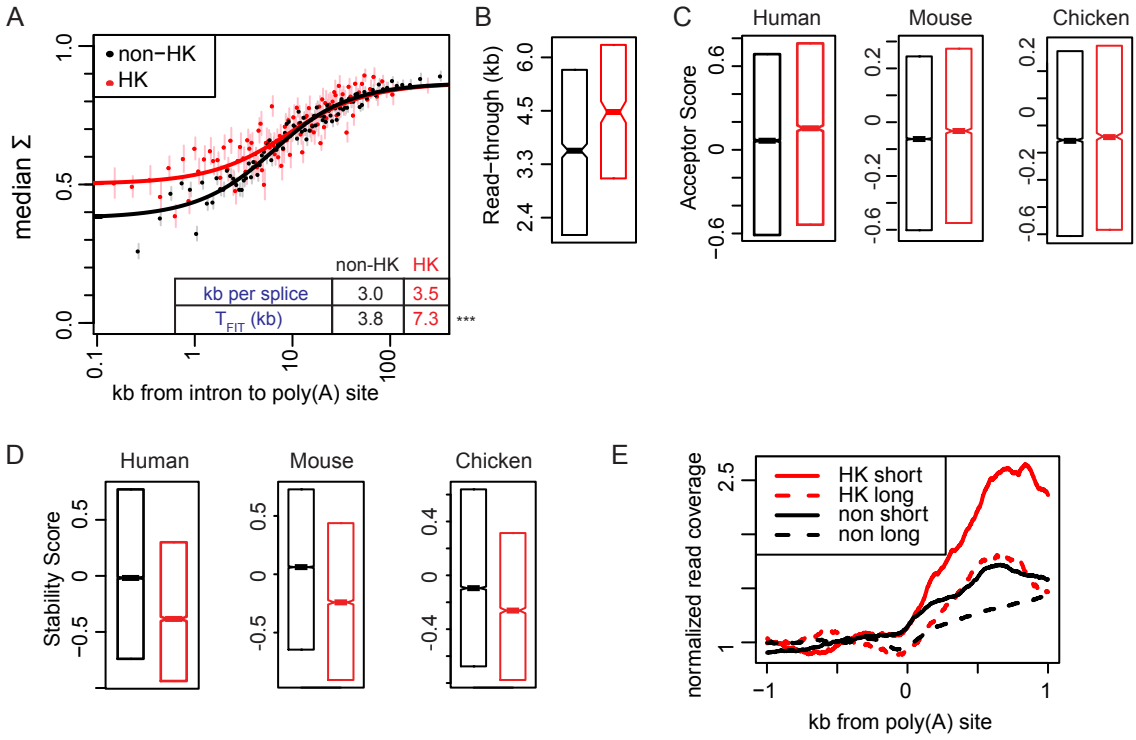


Figure 2.9: Housekeeping genes have distinct CTS determinants. A: Splicing completion in nuclear poly(A)-depleted RNA-seq is higher in housekeeping genes (*red*) than other genes (*black*). Inset shows parameter fits to CTCS+T_{FIT}. The asterisks indicate statistical significance. B: Read-through in mouse genes as measured by GRO-seq. C,D: Acceptor scores (C) and exonic nucleosome stability scores in the first 147 bps (D) in HK vs. non-HK genes based on human, mouse, and chicken genome sequence. A gene was considered a housekeeping gene if it shared the Gene Symbol of a human housekeeping gene: otherwise it was considered a non-housekeeping gene. E: Normalized PolS2 ChIP-seq signal at the poly(A) site of HK and non-HK genes, for long and short genes.

with nucleosome stability (Fig. 2.8D), and we used this correlation to extrapolate elongation rates for each gene based on the nucleosome stability scores shown in Fig. 2.7A (see Experimental Procedures). Allowing for a variable elongation parameter (+ Δelong) resulted in preferentially marked increases in CTS efficiency for short genes with many introns (Fig. 2.7E). Next, we tested the effect of having stronger splice sites in last introns (+ Δsplice). This change increased CTS efficiency in most gene categories, but exacerbated the differences between genes with many or few introns. When we took into account both variable elongation

and splicing rates (+ Δboth), there was an increase in CTS efficiency across all categories. These modeling results are consistent with a central role for elongation control in the regulation of CTS efficiency. Moreover, these data illustrate the power of modeling to elucidate the combined contributions of several factors, such as elongation, nucleosome stability, and splice site strength, for regulating co-transcriptional splicing.

2.3.4 Distinct genomic features support CTS of housekeeping genes

Since CTS efficiency depends on Pol II dynamics, we hypothesized that differentially regulated gene groups would show distinct signatures of CTS. We compared constitutively-expressed housekeeping (HK) genes with genes whose expression is more variable across cell types (non-HK; Chang et al., 2011). Nascent RNA-seq revealed that HK genes have overall higher steady-state intron spliced fraction Σ than non-HK genes, especially for introns close to the poly(A) site (Fig. 2.9A). The CTCS+T_{FIT} model, when fit to this data, predicts a significantly longer post-poly(A) site delay for HK genes (equivalent to 7.3 kb) compared to non-HK genes (3.8 kb). Interestingly, our mouse GRO-seq data, indicates that the average read-through is longer in HK genes than non-HK genes (4.4 vs. 3.6 kb, respectively: Fig. 2.9B).

Next we analyzed how the combination of other CTS determinants could contribute to the higher CTS efficiency of HK genes. The model fit resulted in a higher elongation/splicing ratio in HK genes: 3.5 kb per splice compared 3.0 kb per splice in non-HK genes (Fig. 2.9A). Interestingly, introns throughout HK genes in fact have stronger splice sites than non-HK gene introns (Fig. 2.9C): therefore the elongation/splicing ratio in HK genes is consistent with a faster elongation rate instead of a slower splicing rate. In support of this hypothesis, HK genes have on average lower nucleosome stability than non-HK genes (Fig. 2.9D). This faster elongation/splicing ratio implies that HK genes would have lower CTS efficiency were it not for a longer post-poly(A) site delay time. As transcriptional read-through measured by GRO-seq (4.4 kb) does not account for the expected delay

(equivalent to 7.3 kb), we hypothesized that transcriptional pause sites may provide additional time. Strikingly, in support of this hypothesis, HK genes have much stronger PolS2 peaks downstream of their poly(A) sites in K562 cells than non-HK genes (Fig. 2.9E).

2.4 Discussion

In this study we identified genetic structure and sequence features that selectively affect kinetic parameters underlying CTS in vertebrate genomes, and used these to construct and validate a scalable computational model of CTS for cohorts of genes. We detected conserved genomic signatures of gene structure, splicing signals and nucleosome stability that correlate with genes' inherent capacity to splice efficiently. Moreover, while chromatin plays a major role in the regulation of gene expression (Berger, 2002; Workman and Kingston, 1998), regulated genes appear to have evolved to rely more on nucleosomal control of pol II elongation to achieve high CTS efficiency than housekeeping genes, which contain more introns (Eisenberg and Levanon, 2003) but have higher splice site scores and exhibit longer Pol II read-through. As nucleosome density is a component of the regulated chromatin landscape, our observation suggests that splice patterns of non-HK genes are an integral part of their gene expression regulation.

Model simulations of all human genes, based on average parameters deduced from RNA-seq data, further suggest that many transcripts remain incompletely spliced when Pol II reaches the poly(A) site (Fig. 2.6D). However, multiple lines of evidence suggest that splicing may be completed subsequent to this event but prior to transcript release. First, we showed that nascent RNA-seq data are most consistent with an average interval of about one and a half minutes between transcription of the poly(A) site and cleavage of the pre-mRNA (Fig. 2.6A). Second, Pol II transcripts terminate well past the poly(A) site (Fig. 2.6C). Third, we find increased PolS2 occupancy indicative of pausing at the 3' end of shorter genes. And fourth, several studies have shown that even cleaved and polyadenylated but incompletely spliced mRNAs are retained on the chromatin (Bhatt et al., 2012;

Brody et al., 2011; Pandya-Jones et al., 2013). The delay in transcript release could result from the complex requirements of termination (Proudfoot, 2011), or perhaps reflects a checkpoint that prevents release of pre-mRNA transcripts (Alexander et al., 2010).

Even if the ultimate catalytic steps of splicing occur post-transcriptionally, the recruitment and assembly of splicing complexes likely occur co-transcriptionally (Brody et al., 2011; Pandya-Jones and Black, 2009; Tardiff et al., 2006; Wetterberg et al., 1996), and are therefore subject to the kinetic considerations addressed here.

A detailed quantitative delineation of post-transcriptional events will require refinement of the current model so that the chemical reaction of splicing is delineated starting with the recruitment of splicing factors (cf. Murugan and Kreiman, 2012). Similarly, the quantitative impact on CTS of other mechanisms such as splicing enhancers or suppressors and associated trans-acting factors (Barash et al., 2010; Wang and Burge, 2008), the chromatin-mediated recruitment of splicing factors, or alternative splicing may be studied by extending the current model formulation. Including these additional mechanisms will likely improve the predictive power of the model in addressing the control of individual rather than cohorts of genes, and for diverse biological scenarios that determine the cellular chromatin and splicing factor milieu. However, the present study supports the view that kinetic characteristics determined by gene structure, sequence motifs, and nucleosomes form an important basis for splicing control, even if control of individual splice patterns may be critically regulated by gene- and intron-specific mechanisms not yet considered in the current model.

2.5 Experimental Procedures

2.5.1 High-Throughput data used in this study

Encode RNA-seq in human cells lines generated from the Cold Spring Harbor Laboratories as well as and MNase-seq and ChIP-seq data generated by the 'SYDH' group were downloaded from
ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/ (ENCODE Project

Consortium, 2011).

RNA-seq in mouse from Bhatt el. al (2012) used accession GSE32916 from the Gene Expression Omnibus. GRO-seq data in human and mouse were from GSE45822 (GSM1115995 - GSM1115998) and GSE48759 (GSM1183906 - GSM1183908 and GSM118391), respectively.

2.5.2 Data Mining

We used the Biostrings (versions 1.13.19) versions 1.13.19 packages in R to download RefSeq gene sequences for hg19, and mm9 galGal4, dm3, and ce6. Nucleosome stability was predicted from genomic sequence (Vaillant et al., 2007). This algorithm calculates the relative free energy cost of bending a 73 base pair segment of double-stranded DNA in one full loop. We calculated this measure for the first 147 bp (corresponding to one nucleosome) of each exon in the genome, and for the entire gene starting with the 2nd exon. Splice site strengths were calculated by using a position weight matrix calculated from the intron sequences of each genome. We then converted nucleosome stability scores and splice site strength scores to Z scores. MNase-seq data for measuring nucleosome occupancy, as well as Pol II ChIP and Pol S2 ChIP were downloaded from the ENCODE project (ENCODE Project Consortium, 2011). We report the average number of reads mapping to each region of interest. Genes analyzed with PolS2 ChIP-seq were filtered to remove genes whose poly(A) sites were within 1kb of another gene's TSS or poly(A) site, and further filtered to include only genes expressed in the the cytoplasmic fraction of K562 cells (via RNAseq). A published list of 2,064 human housekeeping genes (Chang et al., 2011) was used to determine housekeeping genes in vertebrates. All boxplots were generated in R using the boxplot function in package 'graphics'.

2.5.3 Computational Modeling

We simulate the elongation of a single polymerase and the transformation of its associated transcript. The probability that an intron has spliced by the time

that transcription has terminated is a function of the time it takes to cleave and polyadenylate the mRNA subsequent to the intron's synthesis, and the kinetic rate of splicing. Splicing can be modeled as a series of j sequential reactions (Schmidt et al., 2011). By assuming that the time of each reaction is an independent, exponentially distributed random variable with forward rate constant $k_s * j$, we can model the probability of splicing at time t as a gamma-distributed random variable with shape j and mean $1/k_s$. Thus the probability P_i^t that an intron i has spliced by time t is the cumulative distribution:

$$P_i^t(j, k_s) = \sigma_i(t, j, k_s) = \frac{1}{\Gamma(j)} \int_0^{jk_s t} x^{j-1} e^{-x} dx \tag{2.1}$$

For a single-step reaction ($j = 1$), this equation simplifies to the exponential distribution:

$$\sigma_i(t, 1, k_s) = 1 - e^{-k_s t}$$

If we assume a constant elongation rate k_E, the total elongation time T_i downstream of intron i is proportional to the distance D_i from intron i to the poly(A) site:

$$\sigma_i(D_i/k_E, 1, k_s) = 1 - e^{-D_i k_s/k_E}$$

Splicing rate constants are reported in the manuscript as $k_E/(k_s * j)$.

2.5.4 Multi-intron Model

Each potential transcript for a gene with N introns and $N + 1$ exons can be represented as a string $S = [S_1, S_2, \ldots, S_N], S_i \in \{0, 1\}$, where $Si = 1$ if intron i has been spliced out, and 0 if it is retained. Therefore the probability of each transcript S is:

$$P(S) = \prod_{i=1}^{N} [\sigma_i(T_i, j, k_s) * I(S_i = 1) + (1 - \sigma_i(T_i, j, k_s)) * I(S_i = 0)]$$

where $I(x)$ is the indicator function. To predict the abundance of each transcript at the end of CTS, we calculate $P(S)$ for all possible transcripts: co-transcriptional splicing efficiency was defined as the abundance of the transcript S

whose introns have all been removed (all $S_i, = 1 \; \forall \; i \in \{1, 2, , N-1, N\}$). Therefore, CTS efficiency can be computed simply as:

$$CTS\ efficiency = \prod_{i=1}^{N} 1 - e^{-D_i k_s^j / k_E}$$

This is an O(n) operation and is therefore extremely fast, making this model scalable to genes of any complexity.

2.5.5 Simulations

The model gene used for Fig. 2.2D consisted of four exons of 100bp separated by introns of 1kb. Simulations were performed in R. To estimate gene-specific elongation rates (Fig. 2.7E), we fit the elongation rate of 1,166 genes measured in K562 cells in Veloso et al. (2014; their Table S1) as a linear regression of the un-normalized nucleosome stability score of the 40kb regions over which Veloso et al. measured elongation rates. We then used this function to deduce the average elongation over the entire gene body (starting with the 2nd exon, where the simulations commence) of all genes in the human and mouse genomes, assigning each gene a specific elongation rate. Perturbations of splice site strength were done by increasing the rate constant of the last introns by two-fold compare to other introns. To create a neutral effect on the overall rate of splicing, we adjusted the rates to maintain the same average $1/k_s$:

$$K^b_{fitted} * A = K^b_{adjusted} * (A-1) + 1/2 * K^b_{adjusted}$$

$$K^b_{adjusted} = K^b_{fitted} * \frac{A}{A - 1/2}$$

where $K_b = kb/splice = 1/k_s$ and $A =$ average introns per gene. Our dataset averaged 8.70 introns per gene resulting in an adjustment factor of 1.06 (3.3 kb/splice for non-last introns, 1.65 kb/splice for last introns).

2.5.6 Model fitting and RNA-seq analysis

Although we cannot directly measure σ, for each intron the splicing rate σ determines the 'spliced fraction' (Σ) of each intron, where Σ describes the proba-

bility at steady-state that an intron is spliced out. We assume that the transcripts measured are tethered to Pol II units distributed evenly along each gene. Under this assumption, the probability that a given read is spliced is:

$$\begin{aligned}\Sigma(D_i, k_s) &= \int_0^{D_i} \frac{\sigma_i(x/k_E, k_s)}{D_i} dx \\ &= \frac{1}{D_i} \int_0^{D_i} 1 - e^{-xk_s/k_E} dx \\ &= 1 - \frac{1 - e^{-D_i k_s/k_E}}{D_i k_s/k_E}\end{aligned}$$

This model is referred to as the CTCS model. We further model an additional time allowance by adding a constant distance D_0, which relates to an additional time my the relationship $D_0 = T_0 k_E$:

$$\begin{aligned}\Sigma_0(D_i, k_s) &= 1 - \frac{1 - e^{-(D_i + D_0)k_s/k_E}}{(D_i + D_0)k_s/k_E} \\ &= 1 - \frac{1 - e^{-k_s(D_i/k_E + T_0)}}{k_s(D_i/k_E + T_0)}\end{aligned}$$

This model is referred to as the CTCS+T_x model.

RNA-seq of non-polyadenylated nascent RNA captures both unspliced and spliced nascent pre-mRNA transcripts (Tilgner et al., 2012), and can be used to measure Σ. Counts of RNA-seq reads mapping to exon-exon junctions ($count^{53}$) and intron-exons junctions ($count^{50}$ and $count^{03}$) were determined using bam2ssj (Pervouchine et al., 2013). We estimated the spliced fraction Σ of each intron as follows:

$$\hat{\Sigma} = \frac{Count^{53}}{Count^{53} + (Count^{50} + Count^{03})/2}$$

Introns were omitted if any reads crossed their junctions but did not conform to either the unspliced form or the constitutive splicing form. To ensure that we were using the correct poly(A) sites in this analysis, we generated a 'cleavage ratio' by aligning RNA-seq reads from the cytoplasmic fraction of the same cell type to the poly(A) regions of genes of interest. Specifically, using bedtools (Quinlan and Hall, 2010), we computed the number of reads overlapping by at least one base

pair the regions defined by 100 bp upstream of the poly(A) site ('Up') and 100 bp downstream of the poly(A) site ('Down'). The distribution of the ratio of 'Up' reads to the sum of 'Up' and 'Down' reads is generally centered around 0.5 in un-polyadenylated fractions, but for polyadenylated fractions is it equal to one for most genes (data not shown). This indicates that most genes are cleaved at their canonical poly(A) sites: nonetheless, for the splicing assays we discarded genes that had a ratio of 'Up' reads to 'Up' + 'Down' reads in the cytoplasmic fraction of less than 0.95. To derive model parameters from the RNA-seq data we first divided up all introns into 100, 1,000 or 3,000 equally populated bins based on distance to poly(A) site. We then use the median $\hat{\Sigma}$ of introns within the bin, and the median distance to poly(A) site Di for fitting and plotting. We observed that $\hat{\Sigma}$ does not seem to reach a value of 1 (full splicing) for any groups of introns (Fig. 2.1B, 2.6A, 2.4), so we employ a correction factor α and fit $\hat{\Sigma}$ to the function

$$\hat{\Sigma} = \alpha\Sigma_0(D_i, k_s)$$

This gives us two parameters for CTCS (α and k_s/k_E) and three for CTCS+T_{FIT} (α, k_s/k_E, and D_0). When fitting this model to HK genes separately from non-HK genes (Fig 2.9A), in order to directly compare the other parameters we fixed to the value (0.87) derived for the full genome. Allowing to vary freely has a minimal effect on the parameter values slightly, and the extra time allowance is still significantly longer in HK genes (not shown). Curve fitting was performed in R with the 'port' algorithm of the 'nls' package.

2.5.7 GRO-seq

Thioglycollate-elicited macrophages were isolated from 6-8 week-old BALBc (Jackson Laboratories) female mice by peritoneal lavage 3-4 days following peritoneal injection of 2.5 ml thioglycollate. Cells were plated in RPMI medium 1640 and 10% fetal bovine serum, washed after adherence and again fed with fresh medium. The cells were then treated with 100 ng/ml of Kdo2-Lipid A or medium alone for 1 hour. Global run-on and library preparation for sequencing were done as described (Core et al., 2008).

GRO-seq reads of were mapped to the mouse (mm9) or human (hg19) genomes using bowtie (Langmead et al., 2009). Reads were gathered into overlapping segments and adjacent segments assembled into transcripts using Vespucci (Allison et al., 2013), with the relative weight of the tag density set to 10000 and weighted distance of 500. Transcripts with length ¿ 200 and RPKM greater than $1/log100 * (length\ in\ bp - 200)$ were kept and overlapped with RefSeq genes for analysis of transcription units (Fig. 2.6B). To reduce confounding signals, we removed from the analysis genes where the GRO transcript ended within 1kb of a different gene's TSS, or if their average 'cleavage ratios' (see above) were less than 0.95 in the cytoplasmic fraction of mouse macrophages (Bhatt et al., 2012) or K562 cells (Djebali et al., 2012), for the mouse and human analyses, respectively, indicating that they were cleaved at their canonical poly(A) sites.

The contents of this chapter have been submitted for publication at a peer review journal Cell Reports, as the manuscript Sequence Signatures and Polymerase Dynamics Favor Co-transcriptional Splicing Genome-wide", with authors Jeremy Davis-Turak, Karmel Allison, Maxim Shokhirev, Petr Ponomarenko, Lev Tsimring, Christopher Glass, Tracy Johnson, and Alexander Hoffmann.

2.6 References

Aitken, S., Alexander, R., and Beggs, J. (2011). Modelling Reveals Kinetic Advantages of Co-Transcriptional Splicing. PLoS Comp. Biol. *7*, e1002215.

Alexander, R., Innocente, S., Barrass, J., and Beggs, J. (2010). Splicing-Dependent RNA Polymerase Pausing in Yeast. Mol. Cell *40*, 582-593.

Allison, K.A., Kaikkonen, M.U., Gaasterland, T., and Glass, C.K. (2014). Vespucci: a system for building annotated databases of nascent transcripts. Nucleic Acids Res. *42*, 2433-47.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53-59.

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

Bentley, D. (2002). The mRNA assembly line: transcription and processing machines in the same factory. Curr. Opin. Cell Biol. *14*, 336-42.

Berger, S.L. (2002). Histone modifications in transcriptional regulation. Curr. Opin. Genet. Dev. *12*, 142-148.

Bhatt, D.M., Pandya-Jones, A., Tong, A.J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279-290.

Bieberstein, N.I., Carrillo Oesterreich, F., Straube, K., and Neugebauer, K.M. (2012). First exon length controls active chromatin signatures and transcription. Cell Rep. *2*, 62-68.

Boireau, S., Maiuri, P., Basyuk, E., de la Mata, M., Knezevich, A., Pradet-Balade, B., Backer, V., Kornblihtt, A., Marcello, A., and Bertrand, E. (2007). The transcriptional cycle of HIV-1 in real-time and live cells. J. Cell Biol. *179*, 291-304.

Brody, Y., Neufeld, N., Bieberstein, N., Causse, S.Z., Bohnlein, E.M., Neugebauer, K.M., Darzacq, X., and Shav-Tal, Y. (2011). The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. PLoS Biol. *9*, e1000573.

Carrillo, O., Preibisch, S., and Neugebauer, K. (2010). Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. Mol. Cell *40*, 571-81.

Chang, C., Cheng, W., Chen, C., Shu, W., Tsai, M., Huang, C., and Hsu, I.C. (2011). Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. PLOS ONE *6*, 1-10.

Close, P., East, P., Dirac-Svejstrup, A.B., Hartmann, H., Heron, M., Maslen, S., Chariot, A., Soding, J., Skehel, M., and Svejstrup, J.Q. (2012). DBIRD complex integrates alternative mRNA splicing with RNA polymerase II transcript elongation. Nature *484*, 386-389.

Consortium, E.P. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. *9*, e1001046.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

de la Mata, M., Alonso, C.R., Kadener, S., Fededa, J.P., Blaustein, M.a., Pelisch, F., Cramer, P., Bentley, D., and Kornblihtt, A.R. (2003). A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. Mol. Cell *12*, 525-532.

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

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F.*, et al.* (2012). Landscape of transcription in human cells. Nature *489*, 101-108.

Eisenberg, E., and Levanon, E.Y. (2003). Human housekeeping genes are compact. Trends Genet. *19*, 362-365.

Gornemann, J., Kotovic, K.M., Hujer, K., and Neugebauer, K.M. (2005). Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. Mol. Cell *19*, 53-63.

Gunderson, F.Q., and Johnson, T.L. (2009). Acetylation by the transcriptional coactivator Gcn5 plays a novel role in co-transcriptional spliceosome assembly. PLoS Genet. *5*, e1000682.

Hao, S., and Baltimore, D. (2013). RNA splicing regulates the temporal order of TNF-induced gene expression. Proc. Natl. Acad. Sci. U.S.A. *110*, 11934-11939.

Hirose, Y., and Manley, J. (2000). RNA polymerase II and the integration of nuclear events. Genes Dev. *14*, 1415-29.

Howe, K.J., Kane, C.M., and Ares, M., Jr. (2003). Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. RNA *9*, 993-1006.

Ip, J.Y., Schmidt, D., Pan, Q., Ramani, A.K., Fraser, A.G., Odom, D.T., and Blencowe, B.J. (2011). Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. Genome Res. *21*, 390-401.

Kaikkonen, M.U., Spann, N.J., Heinz, S., Romanoski, C.E., Allison, K.A., Stender, J.D., Chun, H.B., Tough, D.F., Prinjha, R.K., Benner, C.*, et al.* (2013). Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Mol. Cell *51*, 310-325.

Khodor, Y., Rodriguez, J., Abruzzi, K., Tang, C.-H., 2nd, M., and Rosbash, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes Dev. *25*, 1-12.

Kwak, H., and Lis, J.T. (2013). Control of Transcriptional Elongation. Annu. Rev. Genet.*47*, 483-508.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X.*, et al.* (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516-520.

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S., Wickens, M., and Bentley, D. (1997). The C-terminal domain of

RNA polymerase II couples mRNA processing to transcription. Nature *385*, 357-61.

Murugan, R., and Kreiman, G. (2012). Theory on the coupled stochastic dynamics of transcription and splice-site recognition. PLoS Comput. Biol. *8*, e1002747.

Neugebauer, K.M. (2002). On the importance of being co-transcriptional. J. Cell Science *115*, 1-7.

Pandya-Jones, A., Bhatt, D.M., Lin, C.H., Tong, A.J., Smale, S.T., and Black, D.L. (2013). Splicing kinetics and transcript release from the chromatin compartment limit the rate of Lipid A-induced gene expression. RNA.*19*, 811-27.

Pandya-Jones, A., and Black, D.L. (2009). Co-transcriptional splicing of constitutive and alternative exons. RNA *15*, 1896-1908.

Pervouchine, D.D., Knowles, D.G., and Guigo, R. (2013). Intron-centric estimation of alternative splicing from RNA-seq data. Bioinformatics *29*, 273-274.

Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now. Genes Dev. *25*, 1770-1782.

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

Schmidt, U., Basyuk, E., Robert, M.-C., Yoshida, M., Villemin, J.-P., Auboeuf, D., Aitken, S., and Bertrand, E. (2011). Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation. J. Cell Biol. *193*, 819-829.

Shatkin, A., and Manley, J. (2000). The ends of the affair: capping and polyadenylation. Nat. Struct. Biol. *7*, 1-5.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature *479*, 74-79.

Singh, J., and Padgett, R.A. (2009). Rates of in situ transcription and splicing in large human genes. Nat. Struct. Mol. Biol. *16*, 1128-1133.

Subtil-Rodriguez, A., and Reyes, J.C. (2010). BRG1 helps RNA polymerase II to overcome a nucleosomal barrier during elongation, in vivo. EMBO Rep. *11*, 751-757.

Tardiff, D., Lacadie, S., and Rosbash, M. (2006). A genome-wide analysis indicates that yeast pre-mRNA splicing is predominantly posttranscriptional. Mol. Cell *24*, 1-22.

Tennyson, C., Klamut, H., and Worton, R. (1995). The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. Nat. Genet. *9*, 1-7.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigo, R. (2012). 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.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcarcel, J., and Guigo, R. (2009). Nucleosome positioning as a determinant of exon recognition. Nat. Struct. Mol. Biol. *16*, 996-1001.

Vaillant, C., Audit, B., and Arneodo, A. (2007). Experiments Confirm the Influence of Genome Long-Range Correlations on Nucleosome Positioning. Phys. Rev. Lett. *99*, 218103.

Vargas, D., Shah, K., Batish, M., Levandoski, M., Sinha, S., Marras, S., Schedl, P., and Tyagi, S. (2011). Single-molecule imaging of transcriptionally coupled and uncoupled splicing. Cell *147*, 1054-65.

Veloso, A., Kirkconnell, K.S., Magnuson, B., Biewen, B., Paulsen, M.T., Wilson, T.E., and Ljungman, M. (2014). Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. Genome Res.*Epub 10.1101/gr.171405.113*

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

Wetterberg, I., Bauren, G., and Wieslander, L. (1996). The intranuclear site of excision of each intron in Balbiani ring 3 pre-mRNA is influenced by the time remaining to transcription termination and different excision efficiencies for the various introns. RNA *2*, 641-651.

Workman, J.L., and Kingston, R.E. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. Annu. Rev. Biochem. *67*, 545-579.

3 A Model of Alternative Splicing Implicates Co-transcriptionally Kinetics as Regulator Splicing Fidelity

3.1 Abstract

The human spliceosome, which catalyzes the removal of introns from pre-messenger RNA, pairs exons together with remarkable fidelity. Given that splice site motifs are notably promiscuous in the human genome, *cis* regulatory sequences, *trans* factors and kinetic considerations all play important roles in specific splice-site selection. However, the combination of these factors must also encode inherent flexibility in exon selection, since alternative splicing is a key contributor to the diversification of the human proteome. Understanding the highly complex 'splicing code' requires mathematical modeling, but current models do not combine kinetic considerations with cis or trans elements. We therefore implemented a kinetic model of alternative splicing during elongation as a tool for studying complex genomes and asking quantitative questions.

3.2 Introduction

Chapter 2 focused on the timing of CTS and the gene and sequence structures that support CTS. In the current chapter we turn our attention to alternative splicing. Alternative splicing occurs when at least two distinct pre-mRNA isoforms arise from variable splicing of identical pre-mRNA sequences. The most commonly studied form of alternative splicing, in which an exon is skipped in its entirety, is referred to as exon skipping, and the skipped exon is called a cassette exon. If a cassette exon is included, i.e. is not skipped, then the 5' splice site of the intron immediately upstream is joined to the 3' splice site bordering the cassette exon, and the 5' splice site bordering the cassette exon is joined to the 3' splice site of the immediately downstream intron. However, if the 5' splice site of the upstream intron is joined with the 3' splice site of the downstream intron, then the cassette exon is not included in the mRNA, and its pre-mRNA sequence is thus part of the lariat which is subsequently degraded .

Thousands of these cassette exons have been identified in the human genome (Thanaraj et al., 2004; Modrek, 2001). Therefore it appears that the splicing machinery is quite flexible in its ability to skip exons. In fact, the exact mechanisms of splice-site pairing is an open question. The splicing machinery faces an impressive task of having to consistently remove introns that vary in size from the dozens to millions of base pairs. Indeed 'errors' in splicing do occur in which an mRNA is spliced in such a way that codes for a non-functional protein (Weischenfeldt et al., 2012; Lewis et al., 2003). At least one molecular mechanism has evolved to 'clean up' such nonsense transcripts, nonsense-mediated decay (NMD) (Losson and Lacroute, 1979). Thus splicing is a somewhat noisy process and the kinetics of binding the various spliceosome components plays a role in this noise. Therefore alternative splicing can be modeled as a noisy process (Melamud and Moult, 2009; Schmidt et al., 2011). However, the noise is not uniform (Melamud and Moult, 2009). Furthermore, the lack of exon skipping over the greater than 100,000 exons in the human genome is quite striking, indicating that the noise level is low (Fox-Walsh and Hertel, 2009).

Despite the low level of splicing noise, alternative splicing is functionally important (Matlin et al., 2005) and highly abundant (Pan et al., 2008; Wang et al., 2008). Indeed, many alternative splicing events code for proteins that function differently from the proteins encoded by the constitutively spliced isoforms. In fact, in this manner the repertoire of approximately 20,000 human genes can be expanded to over 100,000 functional proteins. This expansion is thought to be one important factor in the evolution of complex organisms (another important factor being regulation of those proteins): for example, the 1-mm long nematode *Caenorhabditis Elegans* contains approximately 1,000 cells per individual, yet has almost the same number of genes as humans. Yet, humans have vastly more functional proteins than *C. Elegans*. Therefore it is critical to understand how the genome contains instructions for alternative splicing but still maintains a low level of splicing noise.

In many instances, specificity of alternative splicing is highly regulated. Hundreds of proteins have been classified as RNA-binding proteins (RBP), and many of these, including the SR protein family and the hnRNP family, are important in cell-type specific alternative splicing (Huelga et al., 2012; Fu, 1995). These proteins act by binding near or around splice site to either block the splice site or induce the inclusion of a particular splice site, in a complex manner (Xue et al., 2009). Many of these proteins bind to conserved sequences near the splice sites. The existence of many RBPs, cis-sequences and variation in splice site sequences all combine to produce a combinatorially complex 'splicing code'. The deciphering of the splicing code is a key goal in studying splicing, since we need to have a quantitative understanding of the code if we wish to predict splicing outcomes in a genome-wide fashion.

Computational efforts to unravel the splicing code used a machine learning approach with hundreds of input parameters (Barash et al., 2010; Zhang et al., 2010). These models are quite good at determining whether rates of individual exon inclusion will increase or decrease in the presence of different cohorts of RBPs (up to 70% accuracy). However, the models are limited in both the mechanistic predictions, since the resulting parameters are not relatable to physical variables, as

well as their application, by failing to consider of CTS kinetics. IOn the other hand, several other models have examined CTS kinetics, but do not include alternative splicing (Melamud and Moult, 2009; Schmidt et al., 2011; Aitken et al., 2011; Murugan and Kreiman, 2012). Therefore, we created a model of co-transcriptional alternative splicing.

In constructing this model, our goal was to generate quantitative predictions about splicing outcomes based on gene structure and kinetic rates of competing splicing reactions. Therefore, the scope of the model was tailored around coarser-grained considerations than the cassette exon models (Barash et al., 2010; Zhang et al., 2010), and thus offers a complementary approach. The model is intended to be disseminated such that qualitative hypotheses can be turned into quantitative predictions. In this chapter we describe the model and demonstrate its use in understanding the constraints on intron definition in *Drosophila* and human genomes.

3.3 Methods

3.3.1 Markov Chain

The model is represented as a Markov Chain (MC), and is always a directed acyclic graph (DAG). The root node represents the premature transcript with all introns present and all other nodes represent unique stages of splicing outcomes. The absorbing states in the MC are all fully spliced products . For a gene with N introns, there are $N-1$ internal exons, and thus 2^{N-1} possible fully spliced products, because each internal exon can either be included or excluded. The edges connecting the nodes represent splicing reactions. Thus each node x can be represented by an unordered series of splicing reactions that it has undergone.

3.3.2 Possible splicing reactions

Let R be an upper diagonal matrix entries of all the splicing reactions. R_{ij} represents a pairing of the 5' splice site from intron i to the 3' splice site of intron j:

$i \leq j$. A consitutive splicing reaction occurs is when $i = j$. There are $N(N+1)/2$ splicing reactions possible. We denote the identity of each node by a tuple of entries into R: e.g., a splicing intermediate in which intron 1 is constitutively spliced and the donor of intron 2 has spliced to the acceptor of intron 3 is denoted $x_{(1,1)(2,3)}$. Formally, the state space Ω of the MC $\mathbf{X}$ is the collection of states

$$\begin{aligned} x^{a}_{(i_1,j_1),(i_2,j_2),\ldots,(i_m,j_m)} : & m \leq N-1, \\ & i_k \leq j_k \leq N(N+1)/2, \forall\ k \in (1,m), \\ & a \in (1, Fibonacci(2N+1)) \end{aligned}$$

However, this is not the minimal Ω, as we further limit the state space using biophysical constraints, i.e. rules for splicing.

3.3.3 Rules for splicing reactions

The MC is constructed by enumerating all possible sequences of splicing reactions. Not all sequences are physically possible because a splice site can only be used once; furthermore we dissallow any excision of exons that have already been spliced (e.g. consitutive removal of intron 2 following by joining of Donor 1 to Acceptor 3 is not allowed). This latter rule provides for symmetry in the splicing reaction rules.

An edge edge R_{kl} will depart from node $x_{(i_1,j_1),(i_2,j_2),\ldots,(i_m,j_m)}$ and arrive at node $x_{(i_1,j_1),(i_2,j_2),\ldots,(i_m,j_m),(k,l)}$ if for each $p \in (1,m), k \leq l < i_p$ or $j_p < k \leq l$.

3.3.4 Algorithm to build MC

The exhaustive sequences of allowed splicing reactions, and thus the nodes of the MC, are built by an iterative algorithm by first considering cases with $N = 1$ and assembling all nodes before adding nodes allowed with $N = 2$, etc...

For $N = 1$, we start with the root node of the DAG. We then iterate through all legitimate moves using a recursive strategy:

function Add_moves (existing node, all new moves):

For all new moves:

If new move is allowed by current node:

If node that would be created exists already:

Add and edge between nodes

Otherwise:

Create new node connected by edge

Run add_moves(new node, all other moves):

3.3.5 Infinitesimal matrix Q

The MC is a collection of random variables in continuous time, and the transitions between each state are associated with a kinetic rate constant. Each splicing reaction may have a unique rate, and correspondingly the edge between two nodes takes on the rate of the splicing reaction that it represents. The infinitesimal matrix Q of a continuous-time MC of n nodes is an $n * n$ matrix where:

$$Q_{ab} = \begin{cases} k_s & \text{if there is an edge } a \to b \text{ and } b \neq a \\ -\sum_{c \neq a} Q_{ac} & \text{if } b = a \\ 0 & \text{otherwise} \end{cases}$$

where k_s is the kinetic rate of associated edge. We can then write the probablity of moving between nodes as follows:

$$\frac{dP}{dt} = Q * P(t)$$

which has the solution:

$$P(t) = e^{-Qt}$$

where P is an $n*n$ matrix and $P_{ab}(t)$ is the conditional probability of being in state b after time t, given that we started in state a. To solve the matrix exponentiation, we first decompose the matrix:

$$Q = VDV^{-1}$$

where D is a diagonal matrix where the diagonal entries are the eigenvalues and all others are 0, and V are the eigenvectors of Q. This transformation results in a solution that is faster to compute:

$$P(t) = VMV^{-1}$$

Where $M = e^{-Dt}$ is a diagonal matrix with:

$$M_{ij} = \begin{cases} e^{-D_{ij}t} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Note that if the ith eigenvalue of Q is 0, then $M_{ii} = 1$.

We shall later refer to the infinitesimal matrix of a MC of N introns as Q_N.

3.3.6 Simulating splicing

To simulate the splicing of a transcript, we require a vector $\boldsymbol{\beta}$ of starting probabilities, such that

$$\vec{P}_\beta(t) = \boldsymbol{\beta} e^{-Qt}$$

with β_a is the probability of starting in state a, and $\vec{P}^b_\beta(t)$ is the probability of being in state b after time t given, $\boldsymbol{\beta}$ and Q.

Simulating post-transcriptional splicing (PTS) is equivalent to letting the simulation run until $t \to \infty$. In this case, the solution can be found by setting all diagonals of M that correspond to the non-zero eigenvalues of Q, to 0.

To simulate *purely* PTS, meaning that no CTS is allowed, we use the full Q of N introns and the initial $\boldsymbol{\beta}_0$:

$$\beta_{ij} = \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{otherwise} \end{cases}$$

To simulate CTS, the simulation is broken up into phases corresponding to the time intervals during which certain subsets of splicing reactions are available. The elongation rate and the distance between splicing elements determine the

duration of each phase. For example after the first intron is transcribed, only intron 1 can be spliced while RNAPol processes through the downstream region, until intron 2 is transcribed. We shall denote this duration as T_1.

Starting first with $N = 1$, we use the initial $\boldsymbol{\beta}_0 = [10]$ and compute:

$$\boldsymbol{\theta}_1 = \vec{P}_{\beta_0}(T_1) = \boldsymbol{\beta}_0 e^{-Q_1 T_1}$$

which simplifies to:

$$\begin{bmatrix} e^{-k_1 T_1} \\ 1 - e^{-k_1 T_1} \end{bmatrix}$$

where k_1 is the rate of constitutively splicing the first intron. Note that this solution mirrors the CTCS model (Chapter 2).

Next the vector $\boldsymbol{\beta}_1$ is constructed from the results of simulating phase I:

$$\boldsymbol{\beta}_1 = [\boldsymbol{\theta}_1 \vec{0}_{y_2}]$$

where $\vec{0}_{y_2}$ is a vector of length $y_2 = rows(Q_2) - rows(Q_1)$. This vector is then used to simulate phase II:

$$\boldsymbol{\theta}_2 = \vec{P}_{\beta_1}(T_2) = \boldsymbol{\beta}_1 e^{-Q_2 T_2}$$

For a gene with N introns, this strategy is then iteratively followed until we have computed $\boldsymbol{\theta}_N$, which is the vector of probabilities of each state after CTS. Note that T_N corresponds to the time it takes to elongate the final exon. Once this CTS vector has been calculated, we simulate PTS as detailed above, starting with $\boldsymbol{\beta} = \boldsymbol{\theta}_N$, to obtain $\boldsymbol{\theta}^{complete}$, which gives us the probability of obtaining each of the 2^{N-1} fully spliced isoforms. ($\boldsymbol{\theta}^{complete}$ will be 0 for all transient states of the matrix, which represent the intermediates containing one or more introns. We obtain the following expression for $\boldsymbol{\theta}^{complete}$:

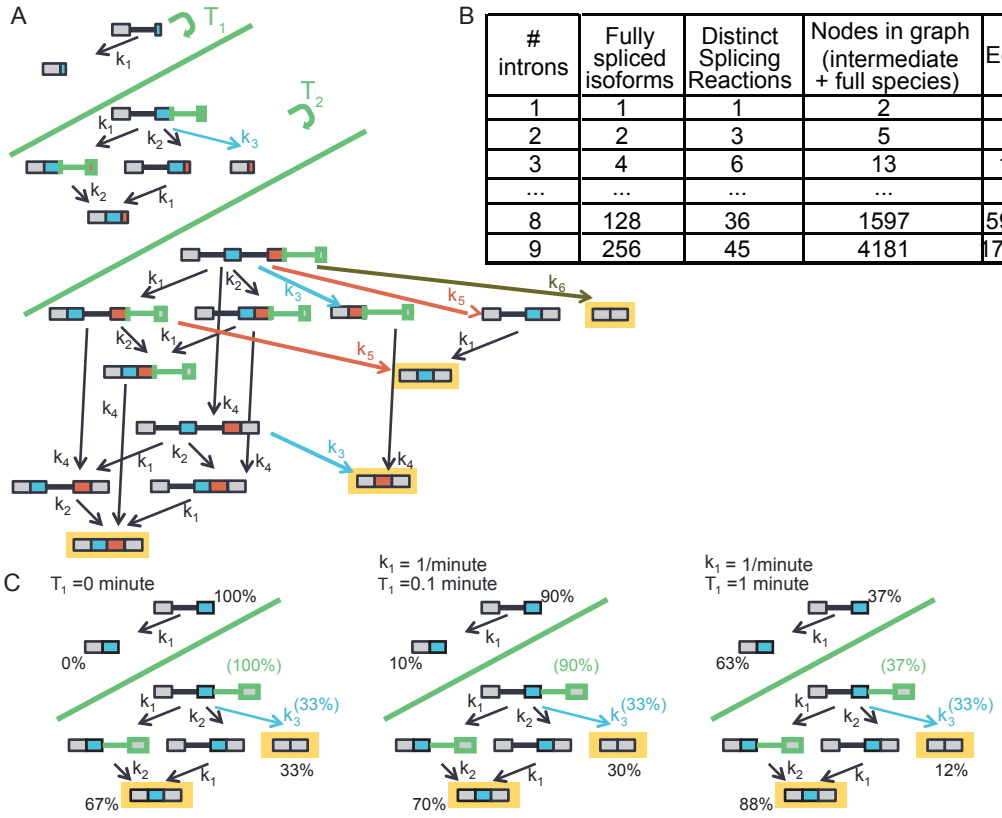
$$\boldsymbol{\theta}^{complete} =$$

$$[[[[\boldsymbol{\beta}_0 * e^{-Q_1 T_1},\ \vec{0}_{y_2}] * e^{-Q_2 T_2},\ \vec{0}_{y_3}] ... \vec{0}_{y_N}] * e^{-Q_N T_N}] * e^{-Q_N \infty}$$

3.3.7 Intron definition

DNA looping studies have demonstrated a log-linear relationship between looping distance and free energy associated with looping (Saiz and Vilar, 2006). Because calculating kinetic rates involves taking the exponential of a free energy term, kinetic rates of DNA looping therefore are linearly related to distance. Similarly, intron definition generally penalizes introns over 200 nt in length (Fox-Walsh et al., 2005). To adapt this concept to looping of pre-mRNA undergoing splicing, we also take into account the propensity of RNA to take on secondary structures via base-pairing interactions by using the square root of distance as the penalty. This formulation results in the following function for k_s of an intron as a function of its length $dist$, the default splicing rate k_s^0, and the 200-bp minimum:

$$k_s(dist) = k_s^0 * \left(\frac{200}{max(dist, 200)} \right)^{1/2}$$



# introns	Fully spliced isoforms	Distinct Splicing Reactions	Nodes in graph (intermediate + full species)	Edges
1	1	1	2	1
2	2	3	5	5
3	4	6	13	19
...	...	...	...	...
8	128	36	1597	5911
9	256	45	4181	17345

Figure 3.1: Model of co-transcriptional alternative splicing. A: Wiring diagram of CTAS model for a 3-intron gene. Light grey boxes indicate first or last exons, and colored boxes indicate internal exons. Black lines represent introns. The simulation begins after the first intron is synthesized, and time interval T_1 corresponds to the time it takes to elongate the 2nd exon and intron (phase I). During T_1, intron 1 may splice with rate k_1. Once T_1 is complete, the second intron is now available, and phase II begins (green lines separate the phases). The newly synthesized elements are outlined in thick green lines in each phase. Constitutive and alternative splicing reactions are represented by black and colored arrows, respectively. After phase II is complete, all reactions may occur, and the simulation runs until all introns are excised. The completely spliced products are highlighted with yellow boxes. B. Table of Markov chain elements for genes with 1,2,3,7,8 or 9 introns. C. Example outputs of a two-intron model. Splicing rates were all equalized and the duration of T1 was set to 0 minutes (*left*), 0.1 minutes (*right*), or 1 minute (*left*). Black percentages list the probability that the transcript is in various states at the end of phase I (*top*) or phase II (*bottom*). Green percentages represent the probability that the transcript is in the fully unspliced form at the beginning of phase II. The blue percentages indicate the fraction of flux out of the fully unspliced transcript that is directed towards the skipping isoform (in this example it is always 33% because all k_i values are equal).

3.4 Results

3.4.1 co-transcriptional constitutive splicing model

We extend the concepts in our co-transcriptional constitutive splicing (CTCS) model to implement the co-transcriptional alternative splicing (CTAS) model. We employ single-step splicing reactions and allow all possibly pairings of Donor and Acceptor sites of annotated introns (Fig. 3.1A), allowing any given internal exon can either be skipped or included. Thus for a gene with N introns, there are 2^{N-1} isoforms possible in our model (Fig. 3.1B). To simulate the co-transcriptionality of splicing, the simulation is broken up into temporal phases that correspond to the orderly synthesis of the introns by Pol II. During the first phase, only the first intron is allowed to splice (with kinetic rate constant k_1). During the second phase, both intron one and two (k_2) may splice, or the donor of the first intron may pair with the acceptor of the second intron (k_3), thus skipping exon 2 and leading to an alternative splicing outcome. During the final phase, all possible splicing reactions are allowed, and the model records both the probability of expressing each isoform as well as the expected time to finish splicing after elongating the final exon. The complexity of the model increases greatly as more introns are simulated (Fig. 3.1B), and thus we have implemented the model to handle up to 9 introns, which covers more than 50% of the human genome. An examination of the simplest CTAS model, a gene with two introns, reveals that CTS results in an increased inclusion rate of the middle exon (Fig. 3.1C). Indeed this is a fundamental mathematical property of our model, that CTS always leads to higher splicing fidelity (See Methods).

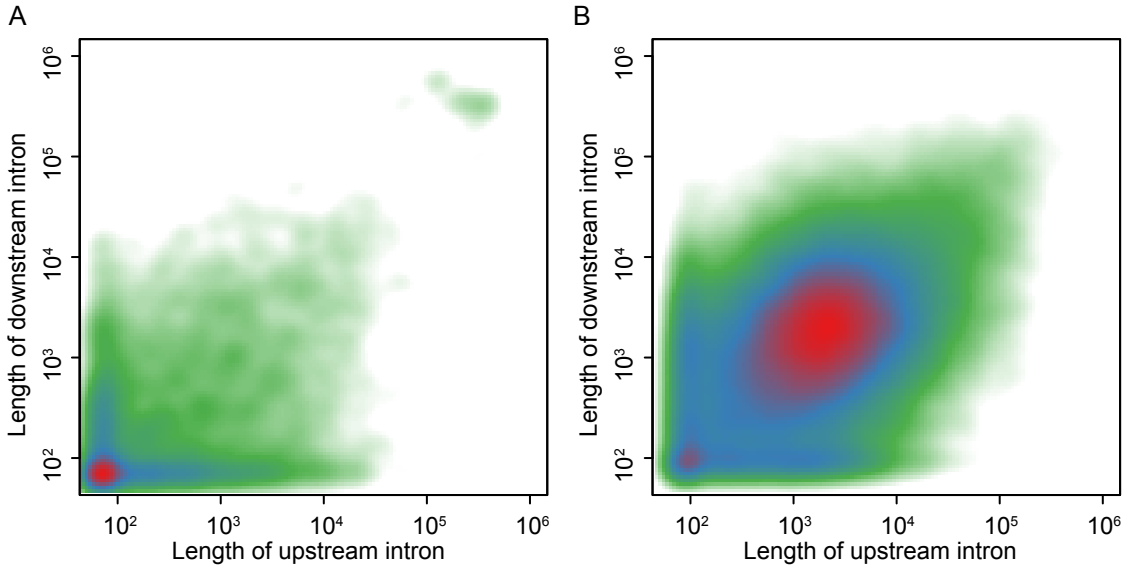


Figure 3.2: Intron sizes in human and *Drosophila* genomes. A. Density scatterplot of the lengths of the flanking introns of each internal exon in the *Drosophila* genome. Green, blue and red indicate low, medium and high density, respectively. B. Density scatterplot of the lengths of flanking introns in the human genome.

3.4.2 Mechanisms of splicing fidelity

We next sought to use our model to examine splicing fidelity in *Drosophila* and human genomes. Short introns splice significantly more effectively than longer introns (Fox-Walsh et al., 2005): this effect is referred to as 'intron definition'. Most *Drosophila* genes contain very short introns (Fig. 3.2A), and splicing specificity is largely achieved through intron definition (Fox-Walsh et al., 2005). In contrast, human genes contain much longer introns in general (Fig. 3.2B), and splicing specificity is most likely not achieved through intron definition (Fox-Walsh et al., 2005), but rather exon definition. Using the CTAS model to simulate intron definition, we expected that including intron definition would result in higher predicted splicing fidelity in *Drosophila*, but that it would have little effect in human.

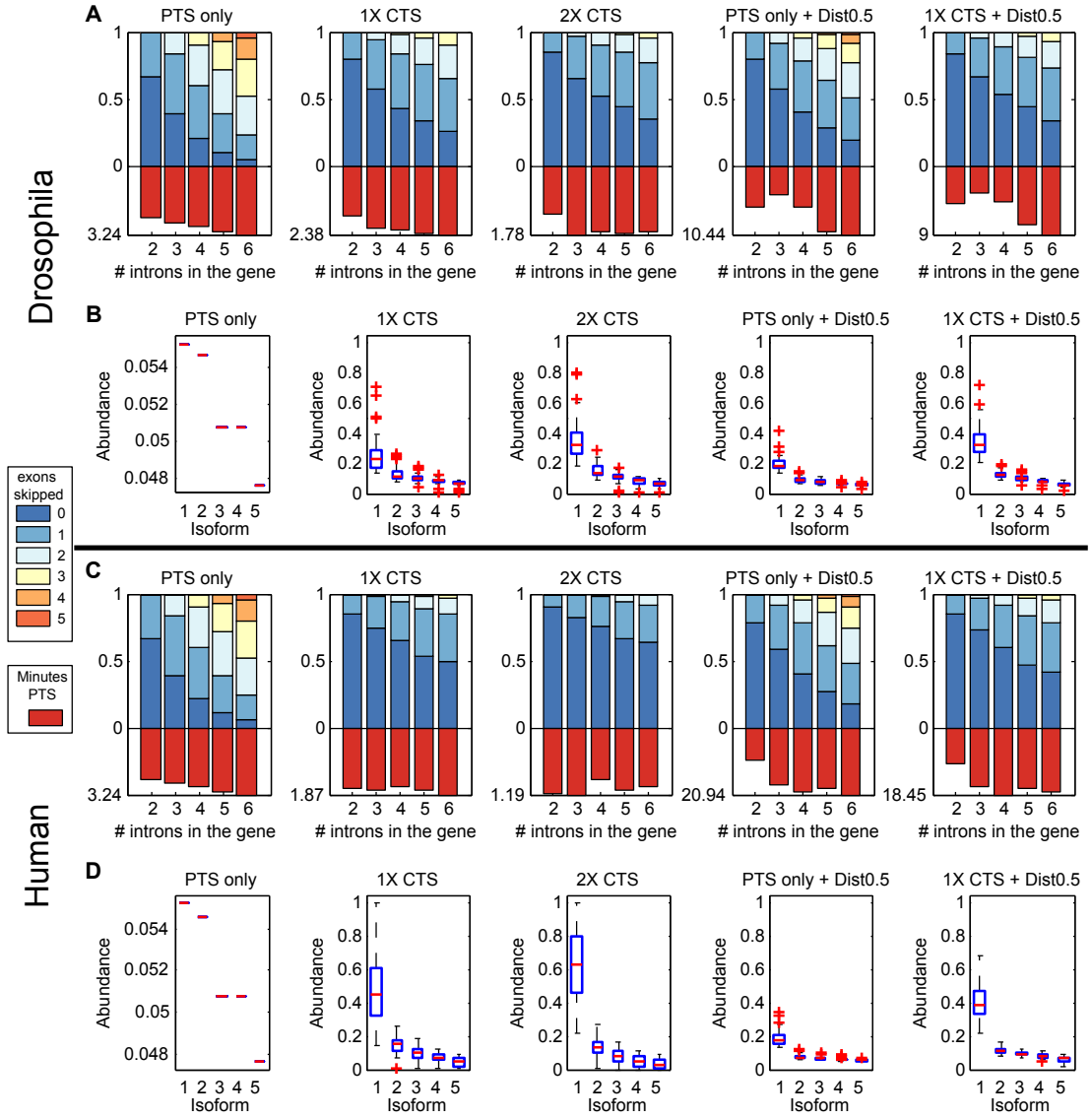


Figure 3.3: Simulations of splicing fidelity with the CTAS model. A: Simulations of *Drosophila* genes with 2-6 introns. 100 genes were selected from each category. Stacked bar plots of the average probabilities that the genes skipped 0,1,2,3,4 or 5 exons. The red downward bar is the average post-transcriptional time to finish splicing for each condition. All genes were simulated either with PTS only, or with CTS, using times derived from either an emperical splicing/elongation rate (*1X CTS*), or twice that CTS time (*2X CTS*). Splicing rates were set to all be equal, except when the intron definition distance penalty was used (*Dist0.5*). B: Distributions of the five most abundance isoforms in the 6-intron genes, from simulations in A. In each of the 100 genes, isoforms were ranked by abundance, and then boxplots were computed for the most abdunant, second most abundant, etc. C,D: Simulations of human genes with 2-6 introns (all conventions as in A,B, respectively).

We first simulated AS of a cohort of *Drosophila* and human genes, both with and without CTS, assuming that all splicing reactions had the same kinetic rate constant (Figs. 3.3A,C). Under this assumption, PTS-only resulted in the expression of a wide variety of isoforms for each gene category, including many isoforms that skip several exons (Figs. 3.3A,C: left plots). Additionally, the top-expressed isoform of each 8-intron gene accounted for only 5% of the average gene expression (Figs. 3.3B,D: left plots). However, when CTS is allowed, using experimentally derived splicing rates (Fig. 2.4B), a much smaller cohort of isoforms is expressed, and exon skipping is greatly reduced (Fig. 3.3A,C: "1X CTS"). Interestingly, exon skipping was reduced more in human genes: accordingly, the top expressed isoform in the 6-intron human genes now accounted for close to 50% of average gene expression, which corresponds quite closely with observed trends (Gonzàlez-Porta et al., 2013), but the top expressed isoform in the 6-intron *Drosophila* genes accounted for only 25% of average gene expression.

We next modeled intron definition by introducing a kinetic rate penalty for long introns wherein the splicing rate constant varied proportionally to the inverse of intron length (see Methods), and simulated our cohorts of genes again (Fig. 3.3A,C: "Dist0.5"). Upon the introduction of intron definition, with PTS we observe a moderate increase in splicing fidelity, as evidenced by less exon skipping. We also observed that the expected time to finish splicing increased to 10 minutes in *Drosophila* and 20 minutes in human. Both species show the top expressed isoform accounting for 20% of the average gene expression in the 6-intron genes (Figs. 3.3B,D). In both species, when CTS is combined with intron definition, fidelity is higher than in the PTS plus intron definition condition. However, while in *Drosophila*, adding intron definition to CTS seems to have an additive effect, adding intron definition to the CTS condition actually resulted in less fidelity than the CTS condition alone (Fig. 3.3C,D: "1X CTS" vs "1X CTS + Dist0.5").

3.5 Discussion

In this chapter, we described the CTAS model, which simulates alternative splicing during CTS. We demonstrated the model's utility by examining the landscape of splicing fidelity using naive assumptions in two genomes (Fig. 3.3). Surprisingly, this work demonstrated that CTS results in a default mode of high fidelity of splicing, and may explain aspects of the observed distribution of human isoforms. Furthermore, we found that the human genome is more prone to CTS-mediated splicing fidelity than the *Drosophila* genome, if there is no penalty for splicing long introns. Since we used human parameters in both simulations, it is possible that fitting splicing data from genome-wide data in *Drosophila* (Khodor et al., 2011) could result in a faster splicing/elongation ratio, which would indicate that *Drosophila* genes use a CTS-mediated fidelity mechanism similar to that of humans.

Additionally, we are able to vary the kinetic parameters of splicing rate as a function of intron length to simulate intron definition. Those simulations revealed that both genomes could make use of intron definition: however, intron definition was overall only advantageous for the *Drosophila* genome. Since most *Drosophila* introns are very short, the vast majority of constitutive splicing reactions fell below the 200 nt intron definition cutoff (Fig. 3.2A), but alternative splicing events would much more often be penalized by being longer than 200 nt (see Methods). Furthermore, adding this penalty for alternative reactions, on top of the moderate CTS-mediated fidelity, resulted in an additive gain of splicing fidelity (Fig. 3.3A,B: "1X CTS" vs "1X CTS + Dist0.5"). However in the human genome, even the constitutive splicing reactions were usually penalized since most introns are significantly longer than 200 nt (Fig. 3.2B), thus reducing the amount of CTS that occurs prior to the synthesis of the next, competing splice site (Fig. 3.3C,D: "1X CTS" vs "1X CTS + Dist0.5"). Although our simulations used an intron definition mechanism that has not yet been validated, and which could be incorrect, this disparity in our simulations suggests that the human splicing machinery does not have the same molecular mechanisms of intron definition as *Drosophila*, and

our findings consistent with genome-wide data indicating that intron length is a far weaker predictor of splicing fidelity in humans than it is in *Drosophila* (Fox-Walsh et al., 2005).

In summary, the CTAS model is a useful new tool for examining alternative splicing. The model is highly scalable, as it can simulate genes with up to 10 exons. Additionally, genes with more than 10 exons could be simulated by considering only the splice sites that display alternative splicing. Furthermore, the model can be easily modified to simulate variable elongation rates, similar to the CTCS model (Chapter 2). Similarly, this framework can incorporate additional features not explored in this study, for example a variable splicing rate to simulate the gene architecture-dependent noise in alternative splicing (Melamud and Moult, 2009). Future computational efforts should focus on developing fitting procedures to derive kinetic rates from biological data, as well as integrating the kinetics of CTS into models that attempt to predict inclusion rates (Barash et al., 2010).

The contents of this chapter are currently in preparation for submission to a peer-review journal.

3.6 References

Aitken, S., Alexander, R. and Beggs, J. (2011). Modelling Reveals Kinetic Advantages of Co-Transcriptional Splicing. PLoS Comput. Biol. *7*, e1002215.

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J. and Frey, B. J. (2010). Deciphering the splicing code. Nat. *465*, 53–59.

Fox-Walsh, K. L., Dou, Y., Lam, B. J., Hung, S.-P., Baldi, P. F. and Hertel, K. J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc. Natl. Acad. Sci. USA *102*, 16176–16181.

Fox-Walsh, K. L. and Hertel, K. J. (2009). Splice-site pairing is an intrinsically high fidelity process. Proc. Natl. Acad. Sci. USA *106*, 1766–1771.

Fu, X. D. (1995). The superfamily of arginine/serine-rich splicing factors. RNA *1*, 663–680.

Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J. and Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. *14*, R70.

Huelga, S. C., Vu, A. Q., Arnold, J. D., Liang, T. Y., Liu, P. P., Yan, B. Y., Donohue, J. P., Shiue, L., Hoon, S., Brenner, S., Ares Jr., M. and Yeo, G. W. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Reports *1*, 167–178.

Khodor, Y. L., Rodriguez, J., Abruzzi, K. C., Tang, C.-H. C.-H. A., Marr, M. T., Rosbash, M. and 2nd, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes & Dev. *25*, 1–12.

Lewis, B. P., Green, R. E. and Brenner, S. E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. United States Am. *100*, 189–92.

Losson, R. and Lacroute, F. (1979). Interference of nonsense mutations with eukaryotic messenger RNA stability. Proc. Natl. Acad. Sci. *76*, 5134–5137.

Matlin, A. J., Clark, F. and Smith, C. W. J. (2005). Understanding alternative splicing: towards a cellular code. Nat. reviews. Mol. cell biology *6*, 386–98.

Melamud, E. and Moult, J. (2009). Stochastic noise in splicing machinery. Nucleic Acids Res. *37*, 4873–86.

Modrek, B. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res. *29*, 2850–2859.

Murugan, R. and Kreiman, G. (2012). Theory on the coupled stochastic dynamics of transcription and splice-site recognition. PLoS Comput. Biol. *8*, e1002747.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J. and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. genetics *40*, 1413–5.

Saiz, L. and Vilar, J. M. G. (2006). DNA looping: the consequences and its control. Curr. Opin. Struct. Biol. *16*, 344–50.

Schmidt, U., Basyuk, E., Robert, M.-C. M.-C., Yoshida, M., Villemin, J.-P. J.-P., Auboeuf, D., Aitken, S. and Bertrand, E. (2011). Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation. J. Cell Biol. *193*, 819–829.

Thanaraj, T. A., Stamm, S., Clark, F., Riethoven, J.-J., Le Texier, V. and Muilu, J. (2004). ASD: the Alternative Splicing Database. Nucleic acids research *32*, D64–9.

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

Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J. S., Kristiansen, K., Krogh, A., Wang, J. and Porse, B. T. (2012). Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. Genome biology *13*, R35.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S. S. Y.-s., Zhang, C., Yeo, G., Black, D. L., Sun, H., Fu, X.-d. X.-D. D. X.-D. and Zhang, Y. (2009). Article Genome-wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping. Mol. Cell *36*, 996–1006.

Zhang, C., Frias, M. A., Mele, A., Ruggiu, M., Eom, T., Marney, C. B., Wang, H., Licatalosi, D. D., Fak, J. J. and Darnell, R. B. (2010). Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Sci. (New York, N.Y.) *329*, 439–43.

4 Model of Co-transcriptional Recruitment of Splicing Factors

4.1 Abstract

During the splicing of vertebrate genomes, the pairing of splice sites is often defined by recognition of spliceosomal units across the exons, which are usually under 200 nt, rather than across the introns, which are usually longer than 1,000 nt. 'Exon definition', as this phenomenon is referred to, can be crucial for the inclusion of an alternatively spliced cassette exon. Therefore a quantitative description of AS must take into account exon definition, and the recruitment or binding of spliceosome components to the RNA. Here we implemented a mathematical model of co-transcriptional AS that explicitly encodes dynamics at the 5' and 3' splice sites. Using this model to simulate exon definition, we show that the recruitment of the U2 snRNP, rather than the U1 snRNP, is likely to be the limiting step in inclusion of a cassette exon.

4.2 Introduction

In the previous chapter, we explored a model that focused on the combinatorial challenge of combining multiple exons to form a mature message. The CTAS model confirmed that genes with short introns can best achieve high splicing fidelity via intrinsic pairing of the 5' and 3' splice sites across the intron (intron definition), and that genes with longer introns would not need to make use of this mechanism. Because the vast majority of human genes fall into the latter category, in this chapter we turn our attention to modeling the processes thought to be involved in controlling the splicing of long introns in vertebrate genomes.

Splicing involves the binding of spliceosomal components on the 5' and 3' ends of an intron and the subsequent pairing of two splice sites, prior to the catalytic splicing reactions that remove the intron from the pre-mRNA. For an exon to be included in the mature mRNA, the splice sites bordering the exon must be paired with more distal splice sites: but if those distal splice sites are instead paired with each other, the exon will not be included in the mRNA. While exons that are separated by short a intron (< 200 nt) may be joined rapidly during intron definition, splicing across a long intron is less favored (Sterner et al., 1996; Guo et al., 1993; Talerico and Berget, 1994), presumably due to a large diffusion distance that limits the chances of the two splice sites co-localizing. However, assembly of the spliceosomes and splicing of such introns can be sped up by communication across the exon of the 3' splice site of one intron and the 5' splice site of the next intron, provided that the exon separating these introns is relatively short (Robberson et al., 1990; Talerico and Berget, 1994). This phenomenon is known as exon definition. If an exon is poorly 'defined', spliceosomes may not form at its splice sites, resulting in the distal splices sites pairing, thus skipping and leading to AS. Therefore, accurate removal of long introns separated by short exons relies heavily on exon definition.

Since most vertebrate genes contain long introns and short exons (Hawkins, 1988), understanding how and when an exon is 'defined' is intrinsic to understanding how AS is regulated. Interestingly, regulation of AS is largely accomplished

by imposing constraints on exon definition. AS events are heavily regulated by a large collection of RNA-binding proteins (RBPs), which usually bind to sequence-specific regions surrounding the splices sites and influence the inclusion of internal exons by promoting or inhibiting either the binding of spliceosome components to the pre-mRNA, or the interaction of those components to form spliceosomal complexes (Kan and Green, 1999; Zuo and Maniatis, 1996; Graveley et al., 2001). In many cases, the RBPs have dual functions as inhibitors or repressors of inclusion, depending on their binding location in relationship to the location of the splice sites. For example, when polypyrimidine tract-binding protein PTB binds near the splice site of a cassette exon, it often leads to exclusion of that exon, but when it binds near the competing splice sites of the adjacent constitutive exons, it often leads to inclusion of the alternative exon (Xue et al., 2009). These observations are consistent with PTB inhibiting the formation of spliceosomes wherever it binds: thus an exon with PTB bound nearby will have a delay in its definition, thereby slowing down its splicing and giving competitive advantage to other splicing reactions.

With much genome-wide data on the binding and functional regulatory effects of several RBPs now available (Witten and Ule, 2011; Huelga et al., 2012), mathematical models are necessary to fully understand and predict quantitative splicing phenotypes. In order to model the effect of RBPs activity on splicing output in a meaningful way, any model formulation must consider interactions taking place at different splice sites separately (Zhang et al., 2010). In addition, although machine-learning based models are useful tools (Zhang et al., 2010; Barash et al., 2010), models grounded in physical reality are necessary to extract biological insights from large datasets. Here, we extend the formulation of our existing CTAS model to include recruitment events at the 5' and 3' splice sites. The resulting model of co-transcriptional recruitment and splicing (CTRS) allows us to simulate exon definition. Additionally, this model provides a general framework for testing other important CTS events including Pol II-mediated recruitment of splicing factors (Close et al., 2012; Misteli and Spector, 1999) and cryptic splicing (Sun and Chasin, 2000; Roca et al., 2003).

4.3 Methods

4.3.1 Additional recruitment reactions

This model simulates recruitment of spliceosome components at the 5' and 3' splice sites of each intron, in addition to the actual splicing reaction. Differences between this model and the One-step splicing reaction model are:

- Temporal phases are split in two to reflect temporally separated synthesis of the upstream (5') and downstream (3') splice sites
- Splice reactions between 5' splice site s_5 and 3' splice site s_3 can proceed only if the spliceosome recruitment reactions have occurred at s_5 and s_3

4.3.2 Markov Chain

A similar DAG representation to model 1 is used. Nodes represent species of pre-mRNA, and edges represent reactions connecting species. The root node and absorbing nodes are the completely unreacted pre-mRNA, and fully spliced mRNA isoforms, respectively.

Each of the 2^{N-1} fully spliced isoforms is represented by a subset S_f of absorbing nodes. All such subsets $S_k \forall\ k \in [1, 2, ..., 2^{N-1}]$ are non-overlapping. This interpretation is necessary because for each of n exons that was skipped in a given isoform, there are 2 splice sites adjacent to that exon that were not used in any splicing reaction, yet a recuitment event may have occurred at each of those splice sites. Thus the subset S_k for isoform k that skips n exons contains 4^n nodes.

4.3.3 Possible reactions

The same $N(N+1)/2$ splicing reactions as model 1 are possible. We denote the identity of each node by a tuple of recruitment reactions states and entries into R: e.g., a splicing intermediate in which intron 1 is constitutively spliced and the

donor of intron 2 has spliced to the acceptor of intron 3, and the recruitment reactions have not occurred at the unused splice sites, is denoted

$$x_{r_1^5,r_1^3,r_2^5,r_2^{-3},r_3^{-5},r_3^3,(1,1)(2,3)}$$

where r_j^5 and r_j^5 mean recruitment at the 5' and 3' splice sites of intron j have occurred, and r_j^{-5} and r_j^{-4} mean they have not.

Formally, the state space Ω of the MC $\mathbf{X}$ is the collection of states

$$\begin{aligned} x^a_{r_1^{q^5},r_1^{q^3},\ldots,r_M^{q^5},r_M^{q^3},(i_1,j_1),\ldots,(i_m,j_m)} : & m \leq M \leq N-1, \\ & i_k \leq j_k \leq N(N+1)/2, \forall\ k \in (1,m), \\ & a \in (1, Fibonacci(2N+1)) \end{aligned}$$

where ${q^p}_j = p * i^{2+2I(\text{spliceosome has been recruited to } p\text{' splice site of intron } j)}$ and I is the identity function. The minimal Ω is obtained by employing the following constraints:

4.3.4 Rules for splicing reactions

Splice reactions between 5' splice site s_5 and 3' splice site s_3 can proceed only if the spliceosome recruitment reactions have occurred at s_5 and s_3. Identically to model 1, a splice site can only be used once, and we disallow any excision of exons that have already been spliced.

4.3.5 Simulations of recruitment and splicing

All procedures are identical to model 1, except that more fine-grained time windows are used. This modification requires us to have created two versions of the infinitesimal matrix Q, per intron, instead of one. Q_N represents the full MC of a gene with N introns, whereas Q^5_{N+1} represents the state space after having transcribed the beginning of intron $N+1$, such that the 5' splice site is available for recruitment, but not the 3' splice site.

Thus the simulations are run incrementally as before, with first Q_1^5 simulated with initial $\boldsymbol{\beta}_0 = [10]$ and time T_1^5 (the time it takes to elongation the first intron), then Q_1 simulated for time T_1 (the time it takes to elongate the 2nd exon) with $\boldsymbol{\beta}$ resulting from the simulation of Q_1^5, etc. We obtain the following expression for $\boldsymbol{\theta}^{complete}$:

$$\boldsymbol{\theta}^{complete} =$$

$$[[[[\boldsymbol{\beta}_0 * e^{-Q_1^5 T_1^5},\ \vec{0}_{y_1^3}] * e^{-Q_1 T_1},\ \vec{0}_{y_2^5}]...\vec{0}_{y_N^3}] * e^{-Q_N T_N}] * e^{-Q_N \infty}$$

$$\text{where } y_1^3 = rows(Q_1) - rows(Q_1^5)$$
$$\text{and } y_2^5 = rows(Q_2^5) - rows(Q_1)$$

$\boldsymbol{\theta}^{complete}$ describes the probability of expression of each absorbing state of the MC. Since in this version of the model, each fully spliced isoform can be represented as a subset of states, the probability of expressing isoform k is

$$P(isoform\ k) = \sum_{s=S_k} \boldsymbol{\theta}_x^{complete}$$

4.4 Results

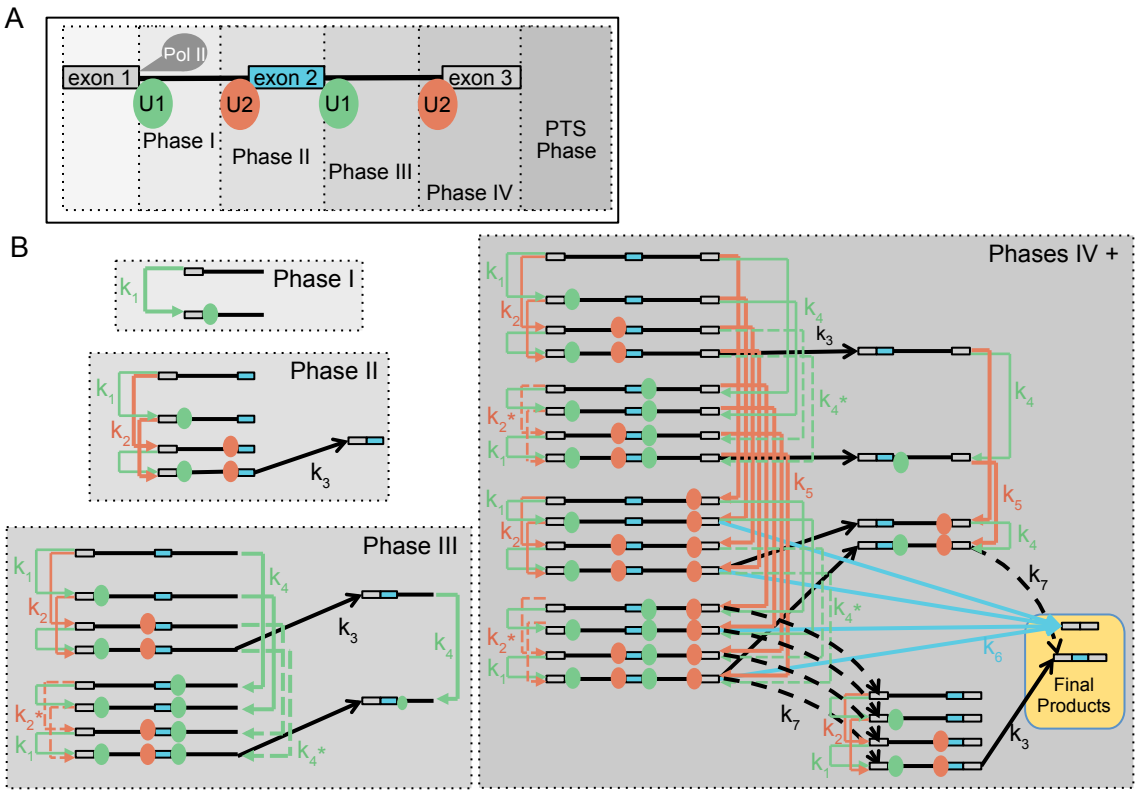


Figure 4.1: Model of co-transcriptional recruitment and splicing]. A: Temporal phases and location of spliceosome recruitment events in the 2-intron gene. B: Wiring diagram of CTRS model for a 2-intron gene. During phase I, only recruitment of the U1 snRNP to the 5' splice site of intron 1 can occur (k_1, *green arrow*). During phase II, recruitment of the U2 snRNP to the 3' splice site of intron 1 (k_2, orange arrows), as well as the excision of intron 1 from transcripts where both U1 and U2 have been recruited (k_3, *black arrow*), can additionally occur. During phase III, recruitment of U1 to intron 2 is allowed (k_4, green arrows on the right hand side). Dotted reaction lines (k_2*, k_4*)indicate reactions that could be affected by exon definition (see Main Text). During phase IV and the post-transcriptional phase, the remainder of the reactions are available: recruitment of U2 to intron 2 (k_5, orange arrows on the right hand side); excision of intron 2 (k_7, *dotted black curved arrows*); the alternative splicing reaction that removes exon 2 (k_6, blue arrows). Recruitment reactions in bold indicate those that occur proximal to Pol II, and may be subject to Pol II-mediated regulation.

Table 4.1: Markov chain elements in CTRS model for genes with 1,2 or 3 introns.

Introns	fully spliced isoforms	distinct splicing reactions	distinct recruit reactions	nodes in graph	edges in graph
1	1	1	2	5	5
2	2	3	4	29	54
3	4	6	6	181	174

In order to faithfully recapitulate AS of vertebrate genomes, we extended the scope of our previous model, which can simulate intron definition during CTAS (Chapter 3). We implemented the CTRS model, which explicitly encodes the recruitment of the U1 snRNP to the 5'ss and the U2 snRNP to the 3'ss (Fig. 4.1, for a two-intron gene). In this model, the pairing and splicing of a donor and acceptor can only occur when the recruitment events have already occurred. To simulate this model co-transcriptionally, we now require twice as many temporal phases as there are introns (Figs. 4.1A,B). During phase I, U1 can bind to the 5'ss of intron 1 (k_1). During phase II, U2 can bind to 3'ss of intron 1 (k_2), and intron 1 can be excised (k_3). During phase III, U1 can bind to the 5'ss of intron 2 (k_4). During phase IV, U2 can bind to 3'ss of intron 2 (k_5), and intron 2 can be excised (k_7), or the AS reaction can occur (k_6). The additional complexity of this model results in even greater combinatorial computation necessary to simulate increasing numbers of introns (Fig. 4.1).

4.4.1 U2 snRNP recruitment is limiting step in CTS

By adding the recruitment reactions, we now are able to simulate more complicated phenomena. Pol II-mediated regulation of recruitment is simulated by modulating the recruitment reaction rates during specific phases. For example, during Phase II, the 3'ss of intron 1 is newly synthesized and is in close proximity to Pol II. Therefore, k_2 is a candidate for Pol II-mediated regulation (Fig. 4.1B). All such candidates are indicated in Fig. 4.1B with bold arrows.

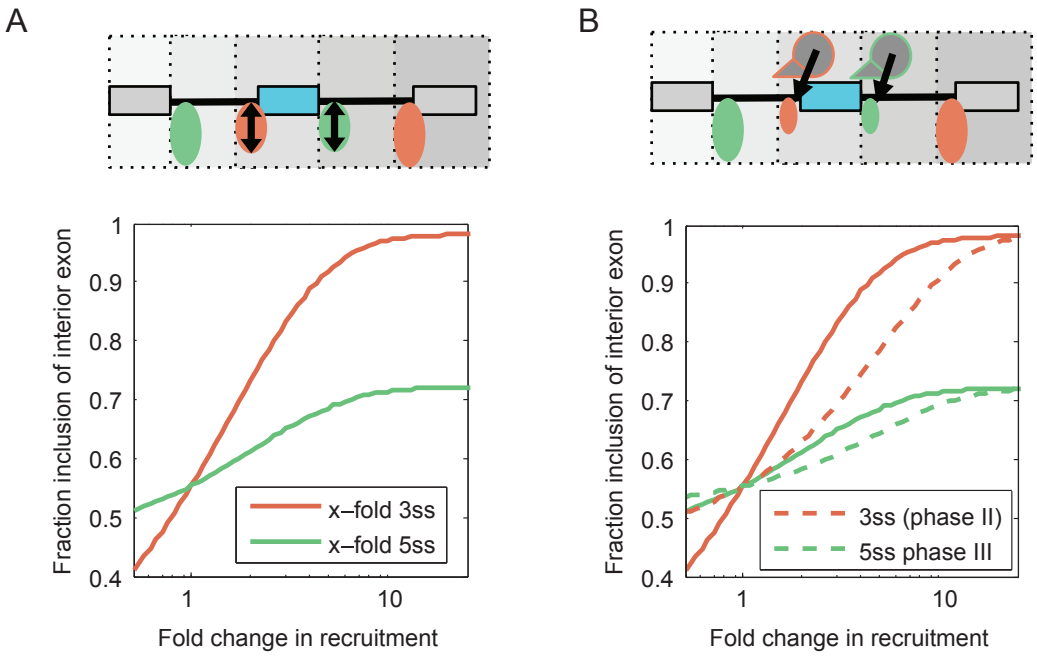


Figure 4.2: Pol II-mediated recruitment. Example simulations of exon inclusion with a CTRS model of a 2-intron gene with weak recruitment at the internal exon (k_2=k_4=1/min) compared to the other splice sites (k_1=k_5=4/min). All splices rates were fast (k_2=k_6= k_7=10/min). A: Probability of inclusion of interior exon as a function of changes to the 3'ss recruitment (k_2, orange) or 5'ss recruitment (k_4, green). B: Probability of inclusion of interior exon as a function of Pol II-mediated recruitment (dotted lines). Solid lines reproduce the results in the left panel.

We first simulated the inclusion of a cassette exon with 50% inclusion using the CTRS model, with variable recruitment rates but with all splicing reactions (k_3, k_6, k_7) set equal to each other (Fig. 4.2). The recruitment rate of U2 to the internal 3'ss had a much stronger effect on inclusion than did the rate of recruiting U2 to the internal 5'ss (Fig. 4.2A). Allowing for Pol II-mediated recruitment showed a similar trend (Fig. 4.2B), although stronger enhancements were required to achieve similar effects on the inclusion rate.

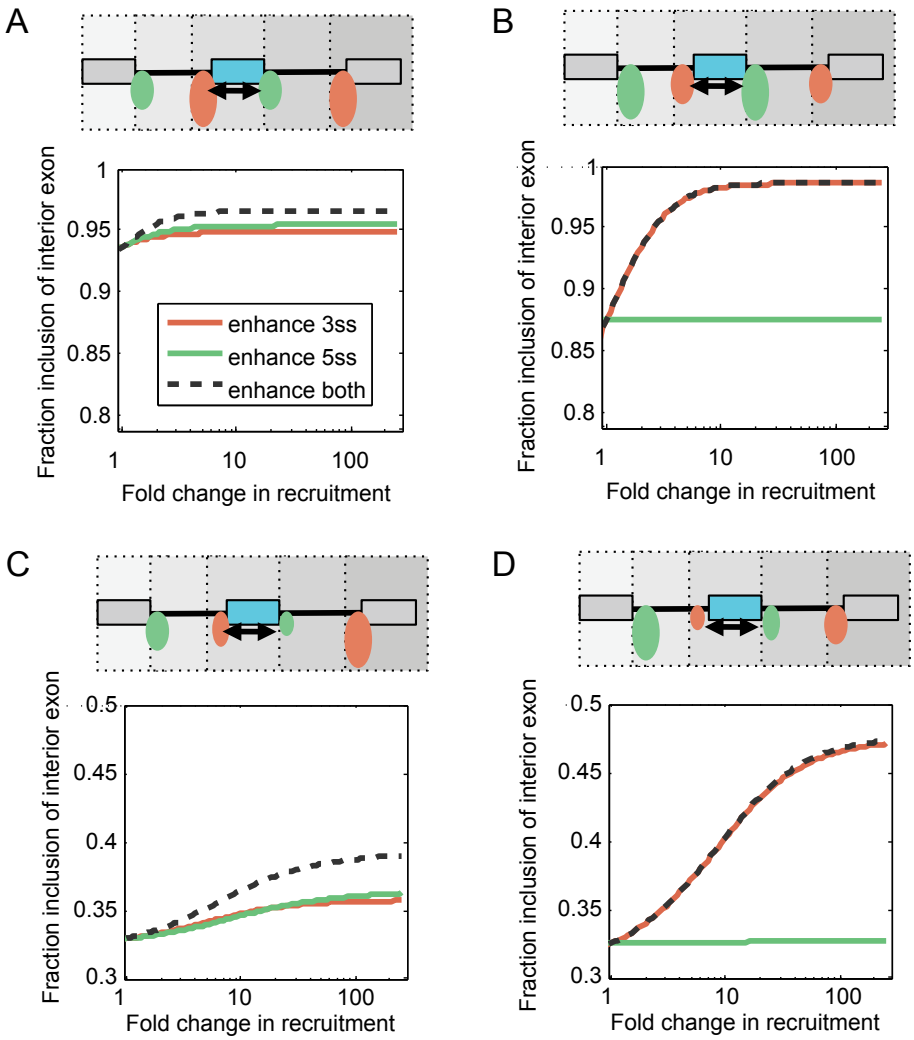


Figure 4.3: Cotranscriptional exon definition. Upper diagram in each panel shows the relative strength of each recruitment event: larger size oval means faster kinetic rate. Graphs demonstrate the probability of inclusion of interior exon as a function of fold enhancement of recruitment during exon definition (dashed lines). The orange line indicates the enhancement of k_2 when U1 is already present at the 5' of intron 2. The green line indicates the enhancement of k_4 when U2 is already present at the 3' of intron 1. Dashed black line indicates the enhancement of either snRNP when the other snRNP is present. A: U2 recruitment is faster than U1 recruitment by default and independent of intron position. B: U2 recruitment is faster than U1 recruitment by default and independent of intron position. C: U2 recruitment is faster than U1 recruitment by default and middle exon has 10-fold slower recruitment rates. D: U1 recruitment is faster than U2 recruitment by default and middle exon has 10-fold slower recruitment rates.

4.4.2 Exon definition

Exon definition, the phenomenon by which the 3'ss and 5'ss across short exons are included together in the mature RNA, is modeled by cooperativity between the two binding events. In transcripts where one snRNP has been recruited already, the binding rate of the other snRNP is increased (Fig. 4.1B, dashed green and orange arrows). We initially tested exon definition in scenarios where either the U1 recruitment was three-fold slower than the U2 recruitment rate (4.3A), or vise-versa (4.3B). In each of these scenarios, we then allowed exon definition to either enhance the U2 recruitment to the 3'ss when the U1 was present across the exon, or enhance the U1 recruitment to the 5'ss when U2 was present across the exon, or allowed both enhancements. When the U2 recruitment was fast, exon inclusion started out quite high (93%), and enhancing either or both reactions increased inclusion marginally, to about 95-96% (4.3A). When U1 was fast, inclusion started out at 87%, and enhancing U2 recruitment could guarantee almost 100% inclusion, but further increases to U1 recruitment had no effect in inclusion (4.3B).

We next tested scenarios where the internal exon's recruitment rates were 10-fold lower than in Figs. 4.3 A and B (4.3C,D respectively). In both cases, inclusion started at 33%. When U2 is faster than U1, enhancing either reaction can induce a slight change in inclusion, and enhancing both raises inclusion to almost 40% (Fig. 4.3C). However, when U1 is faster, further increases in U1 recruitment have no effect, but increasing recruitment of U2 leads to an inclusion of up to 47% (Fig. 4.3D). These result indicates that U2 recruitment, to a greater degree than U1 recruitment, is a limiting step in co-transcriptional exon definition.

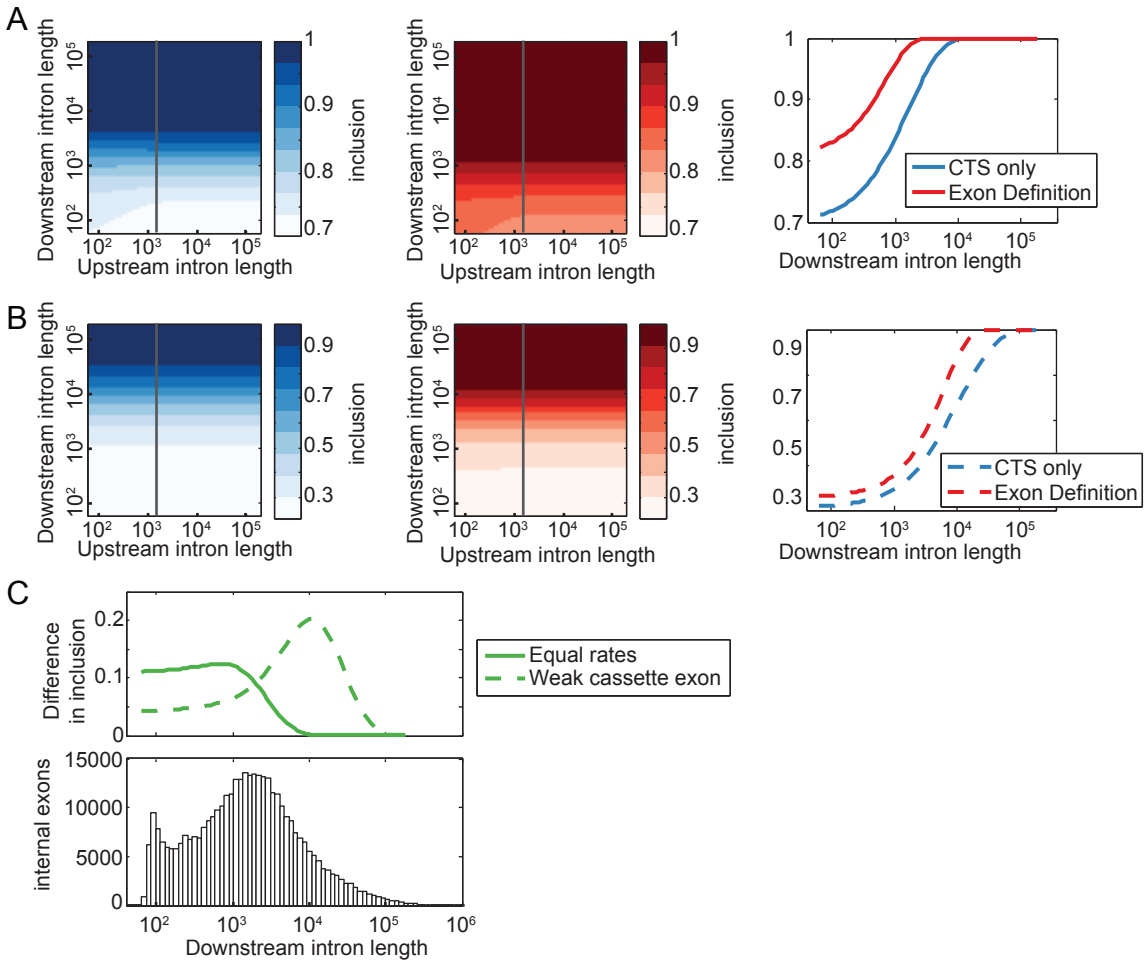


Figure 4.4: The human genome is optimally tuned for exon definition. Inclusion of a 200-bp middle exon was simulated during CTS with varying lengths of its flanking introns. U1 recruitment was set to 3-fold faster than U2 recruitment. A: The splice sites flanking the middle exon were equally as strong as the other splice sites, analogous to Fig. 4.3B). B': The splice sites flanking the middle exon were 10-fold weaker than the other splice sites (analogous to Fig. 4.3D). (*A,B left*: CTS, but not exon definition, was enabled. *A,B middle*: CTS and exon definition were enabled, such that recruitment rates were 10-fold higher across the exon when the other snRNP was present. *A,B right*: Comparison of inclusion rates with or without exon definition, when the upstream intron is 1.5 kb. C: *top*: Difference in inclusion of internal exon between exon definition scenario and normal CTS, for the simulations with equal rates (*solid line, from A*) or with the internal splice sites 10-fold weaker (*dashed line, from B*). *bottom*: histogram of intron lengths in the human genome.

To test the propensity for human genome architecture to support exon

definition, we created test cassette exons with varying lengths of flanking introns to match the observed distribution of human introns (Fig. 3.2B), which span several orders of magnitude, and simulated inclusion with or without exon definition (Fig. 4.4). Without exon definition, length of the upstream intron had little effect on inclusion, but the length of the downstream exon was sufficient to tune the inclusion within the range of 0.7 to 1 when the cassette exon had strong splice sites (Fig. 4.4A, *left*), or between 0.3 and 0.9 when the cassette exon had weak splice sites (Fig. 4.4B, *left*). Additionally, the ranges of lengths over which these changes occurred was slightly different for the two cassette exon types: for equal-strength cassette exons, the largest increase occurred when the downstream intron was between 100 and 10,000 nt, and inclusion saturated above that; for weak cassette exons, the largest increase was between 1,000 and 30,000 nt, with saturation above that.

With exon definition enabled, the inclusion rate was universally higher (Fig. 4.4A,B *middle*), with saturation occurring at shorter downstream intron lengths. To better assess the relationship between exon definition and downstream introns length, we compared the difference in inclusion between exon definition and CTS for both strong and weak cassette exons to the distribution of intron lengths in the human genome (Fig. 4.4C). The peak intron length is centered close to 1.5 kb (Fig. 4.4C, *bottom*). The peak effect of exon definition for strong cassette exons is approximately 1kb, and the peak for weak cassette exons is around 10kb (Fig. 4.4C, *top*). Our model therefore predicts that exon definition by cooperative binding can be effective in increasing inclusion over much of the genomic range of human genes. Furthermore, it suggests that weak cassette exons would benefit most from exon inclusion if their downstream intron were larger than the average intron.

4.5 Discussion

In this chapter we constructed the CTRS model of co-transcriptional spliceosome recruitment and intron splicing. This model encodes spliceosome recruitment

reactions, which allows us to model Pol II-mediated recruitment and cooperativity in exon definition. Modeling exon definition enabled us to make mechanistic predictions about AS in human genes, since most splicing events in human genes are thought to be mediated by exon definition.

The CTRS model predicts that most human exons will be included at 12% greater frequency than they would otherwise be included without exon definition, and that weak cassette exons surrounded by long introns would be included at up to 20% greater frequency. These changes are in line with changes in inclusion associated with RBPs (Xue et al., 2009). Remarkably, alternatively spliced human cassette exons tend to be flanked by long introns and have weak splice sites (Clark and Thanaraj, 2002), and are predicted to be included by means of exon definition. Thus, the model accurately predicted the biologically relevant conditions for which exon definition has the most phenotypic importance.

Additionally, the model predicted that U2 recruitment would be rate limiting, and U1 recruitment rate would be less important (re-word), regardless of whether or not a particular splicing event relies on exon definition. Interestingly, U1 is the most abundant snRNP () and does not require the existence of an intron to be recruited to the transcription site (Brody et al., 2011). Therefore it is likely that U1 is rapidly recruited to the nascent 5' splice sites, and that U2 recruitment is the limiting step in intron or exon definition, as predicted by the model. Indeed one study found that kinetics of U1 were more rapid than other snRNPs (Huranová et al., 2010). These predictions highlight the utility of the CTRS model as a tool for studying AS.

The contents of this chapter are currently in preparation for submission to a peer-review journal.

4.6 References

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J. and Frey, B. J. (2010). Deciphering the splicing code. Nat. *465*, 53–59.

Brody, Y., Neufeld, N., Bieberstein, N., Causse, S. Z., Bohnlein, E. M., Neuge-

bauer, K. M., Darzacq, X., Shav-tal, Y., Karla, M. and Bo, E.-m. (2011). The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. PLoS Biol. *9*, e1000573.

Clark, F. and Thanaraj, T. A. (2002). Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Hum. Mol. Genet. *11*, 451–464.

Close, P., East, P., Dirac-Svejstrup, a. B., Hartmann, H., Heron, M., Maslen, S., Chariot, A., Söding, J., Skehel, M., Svejstrup, J. Q. and Soding, J. (2012). DBIRD complex integrates alternative mRNA splicing with RNA polymerase II transcript elongation. Nat. *484*, 386–9.

Graveley, B. R., Hertel, K. J. and Maniatis, T. (2001). The role of U2AF35 and U2AF65 in enhancer-dependent splicing. RNA *7*, 806–818.

Guo, M., Lo, P. C. and Mount, S. M. (1993). Species-specific signals for the splicing of a short Drosophila intron in vitro. Mol. Cell. Biol. *13*, 1104–1118.

Hawkins, J. D. (1988). A survey on intron and exon lengths. Nucleic acids research *16*, 9893–908.

Huelga, S. C., Vu, A. Q., Arnold, J. D., Liang, T. Y., Liu, P. P., Yan, B. Y., Donohue, J. P., Shiue, L., Hoon, S., Brenner, S., Ares Jr., M. and Yeo, G. W. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Reports *1*, 167–178.

Huranová, M., Ivani, I., Benda, A., Poser, I., Brody, Y., Hof, M., Shav-Tal, Y., Neugebauer, K. M., Stanek, D. and Huranova, M. (2010). The differential interaction of snRNPs with pre-mRNA reveals splicing kinetics in living cells. J. Cell Biol. *191*, 75–86.

Kan, J. L. and Green, M. R. (1999). Pre-mRNA splicing of IgM exons M1 and M2 is directed by a juxtaposed splicing enhancer and inhibitor. Genes & Dev. *13*, 462–471.

Misteli, T. and Spector, D. L. (1999). RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo. Mol. cell *3*, 697–705.

Robberson, B. L., Cote, G. J. and Berget, S. M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. *10*, 84–94.

Roca, X., Sachidanandam, R. and Krainer, A. R. (2003). Intrinsic differences between authentic and cryptic 5' splice sites. Nucleic acids research *31*, 6321–33.

Sterner, D. A., Carlo, T. and Berget, S. M. (1996). Architectural limits on split genes. Proc. Natl. Acad. Sci. *93*, 15081–15085.

Sun, H. and Chasin, L. A. (2000). Multiple splicing defects in an intronic false exon. Mol. cellular biology *20*, 6414–25.

Talerico, M. and Berget, S. M. (1994). Intron definition in splicing of small Drosophila introns. Mol. Cell. Biol. *14*, 3434–3445.

Witten, J. T. and Ule, J. (2011). Understanding splicing regulation through RNA splicing maps. Trends genetics : TIG *27*, 89–97.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S. S. Y.-s., Zhang, C., Yeo, G., Black, D. L., Sun, H., Fu, X.-d. X.-D. D. X.-D. and Zhang, Y. (2009). Article Genome-wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping. Mol. Cell *36*, 996–1006.

Zhang, C., Frias, M. A., Mele, A., Ruggiu, M., Eom, T., Marney, C. B., Wang, H., Licatalosi, D. D., Fak, J. J. and Darnell, R. B. (2010). Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Sci. (New York, N.Y.) *329*, 439–43.

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

5 Conclusions

5.1 Summary

In this Thesis I have developed three computational models to study various aspects of pre-mRNA splicing. To supplement these theoretical experiments, I took advantage of several published high-throughput datasets of genomic sequences and gene architectures, RNA expression in several species, cell types and cellular compartments, positioning of nucleosome positioning and Pol II signals.

In Chapter 2, the CTCS model served as a springboard to examine interdependent patterns of gene structures, Pol II dynamics, and sequence signatures. Far from being arranged randomly, gene architectures show several correlations: first, last exons are long and last 3' splice sites are strong; short genes, and those with many introns, have a higher degree of nucleosome stability throughout their genes. Therefore, genes appear to be under selective pressure to remove their introns co-transcriptionally. These efforts were aided by published RNA-seq data. Although others have used this dataset to statistically model the relationship between splicing levels and distance to the end of the gene (Tilgner et al., 2012), our use of the kinetic model to inform the statistical model construction was the first effort to yield biological parameters from this line of research. Consequently, we were able to adapt the model by incorporating a delay time in polyadenylation. Further, by comparing model fits to RNA-seq data of HK and non-HK genes, we discovered a distinction between the mechanisms these classes of genes use to enforce CTS, and confirmed that genomic sequence signatures and Pol II dynamics upheld this distinction.

In Chapter 3, we extended the kinetic model to study AS by encoding all

possible splicing pairs among known exons. This modeling framework is the first to model AS co-transcriptionally. This construction allows us to model AS as a competition between splice sites. In this study we took a bottom-up approach to explore mechanisms of splicing fidelity (including intron definition) and ask quantitative questions about the impact of different splicing 'rules' or 'codes', which may be encoded in gene structures. This approach confirmed that as previously observed, intron definition cannot fully explain the high rates of splicing fidelity in vertebrate genomes, but that CTS actually explains most of the observed fidelity. Moreover, we also pinpointed the need for an even finer-grained model of exon definition to fully characterize splicing in complex vertebrate genomes such as human and mouse.

In Chapter 4, we explored the consequences of incorporating dynamics of spliceosome assembly at the splice sites. This model allows us to test mechanistic predictions of exon definition. In doing so, we found that the splice site dynamics at the 3' end of the intron was the limiting factor both in exon definition as well as general CTS, which agrees with previous findings. Moreover, our model predicts that human exons are well-suited for exon definition, and this is especially true for those exons that have weak 3' splice sites and long flanking introns - two properties enriched in human cassette exons. Therefore our modeling framework is well-placed to be used as a basis for studying AS in human genes. In addition this approach overlaps nicely with existing machine-learning algorithms that take advantage of cis- and trans- factors to model exon inclusion.

Several biological questions emerge from these investigations. First, it is intruiging to speculate that there a checkpoint for complete splicing prior to release of mRNAs from chromatin. Our data indicate that there is a time delay after transcribing the poly(A) site, and an increased PolS2 pause in genes predicted to be the least spliced. Interestingly, the spliceosome component U1 is already implicated in regulation of Pol II termination (Berg et al., 2012). Moreover, U1 and U2 are released from splicing complexes after successful transesterification reactions (Matera and Wang, 2014), but accumulate at transcription sites when spliceosomes are prevented from assembling past the A complex, in which U1 and

U2 remain bound to the 5' and 3' splice sites, respectively (Huranová et al., 2010). Consequently, the presence of these two snRNPs could in theory signal the presence of unspliced introns.

Second, the results from the CTAS model indicate that when the rate of CTS is high, the default mode of splicing is high-fidelity (exon inclusion); in contrast, if CTS rates are low - and therefore PTS is the dominant mode of splicing - the default mode of splicing switches to low-fidelity (exon skipping;Fig. 3.3). Therefore one would expect that in genomes operating mainly in a PTS context, regulation by splicing factors will focus on activation i.e. exon inclusion, whereas in a CTS context splicing factors may involve both activators and repressors. Exploring this hypothesis could shed light on the evolution of splicing.

5.2 Future directions

Some interesting experiments could be perfomed without changing the model construction. In chapter 4 we briefly touched on the simulations of Pol II-mediated recruitment of splicing factors. This concept could be expanded further to encode the interaction of chromatin signals and splicing.

Future explorations should combine the insights from both AS models. For example, one can use the CTRS model to obtain detailed kinetic information for individual splicing events or cassette exons. Then, these data can be plugged into the CTAS model to simulate the splicing of the entire gene. The CTRS model could be especially useful in simulation cryptic splicing, the phenomenon in which a weak splice site occasionally competes with a stronger splice site nearby, resulting in an exon being longer or shorter than normal. This simulation could be performed by specifying in the model with an intron of length 0, only one of whose splice sites (either its 5' or 3' end) has a non-zero kinetic rate constant for recruitment; in that case, the lone splice site could compete with the splice sites of other introns. This situation is not all that different from cassette exon splicing, and our model is unique in having the flexibility to simulate such a phenomenon.

5.2.1 Extending the model formulation

At least two FRAP studies have reported diffusion rates and some dynamics for various snRNPS (Huranová et al., 2010; Rino et al., 2007). If further experiment can obtain rate constants for the binding and release of snRNPs to RNA components, these rates could be incorporated into the model. Currently we use irreversible reactions, but reversible reactions can be modeled tractably in this framework, since incorporating them would not increase the size of the Q matrix (Chapters 3 and 4). Adding the backwards reaction would result in slightly different overall dynamics: the current rate constants would be roughly equivalent to the net forward rates in the reversible model, but the effects of competition between reactions may be altered because the recruitment reactions would not be permanent: thus e.g. a long delay might not result as much of an advantage for an upstream splice site.

Another useful addition would be to employ cooperativity to model intron definition (Robberson et al., 1990). This mechanism might be more biologically relevant than the looping hypothesis explored here (see Methods, Chapter 3): furthermore it may predict different results vis-a-vis the timing of CTS for long introns, and testing this prediction could distinguish which hypothesis is more accurate. This extension would be straightfoward since it simply involves altering the rates of certain recruitment events in the same manner that exon definition was employed, and would therefore require one extra rate constant per intron.

The CTRS and CTAS models could both be used to test the hypothesis that splicing fidelity increases as a function of number of introns in a gene (Melamud and Moult, 2009). This could be accomplished by artificially increasing splicing rates. However, a more appealing approach would be to explore whether we obtain similar results without artificial perturbations. Alternatively, it may be necessary to model the effect of splicing on Pol II elongation, for example by encoding a splicing-induced pause.

5.2.2 Statistical inference

In Chapter 2, our ability to fit a large RNA-seq dataset of nascent RNA splicing levels to the steady-state predictions of the CTCS model enabled us to confidently rule out one version of the model, as well as fit two parameters for a different version. Ideally we would like to make similar inferences by fitting AS data with either the CTRS or CTAS model. One challenge is that if we examine known cassette exons, we expect the contribution from trans-acting factors to be especially great. For example, our model predicts that if the cassette exon is preceded by a short intron, its inclusion rate would be high: yet it's possible that an RBP could be bound specifically to the 3' splice site of that intron, preventing intron definition from taking place. Unfortunately, measuring only one variable (exon inclusion) per AS event would prevent us from making strong inferences about individual splicing events.

A potentially more viable approach would be to examine splicing levels of induced genes with timecourse data, instead of steady-state splicing levels. By first comparing unspliced to spliced intron amounts, one could estimate the net splicing rate of an intron: this information may help differentiate between AS regulatory regimes, e.g. fast (potentially intron definition or exon definition), or slow (pausing between the last two exons).

5.2.3 Integrating with other models

The greatest advances in inferring splicing regulation will likely be from models that combine the co-transcriptional kinetics explored here, with machine-learning approaches that incorporate cis- and trans- elements specific to each event (Barash et al., 2010; Zhang et al., 2010). In these approaches, briefly, all input variables are fed into an algorithm that determines a weight for each variable, the weighted variables are summed together, and the sum is transformed into a probability between 0 and 1, via a sigmoid function (either logistic or gaussian CDF). Such machine learning algorithms can be quite accurate at classifying events. Yet, it is challenging both to interpret the value of the output parameters

in terms of physical constants, and to test for interactions between parameters in constructing the models in the first place. However, the CTS models presented here provide a good foundation for building a model that parameters like intron length and cis motifs.

One possibility is to use the variables from the other models (Barash et al., 2010; Zhang et al., 2010) to predict kinetic rates. The CTRS model for a cassette exon combines four recruitment parameters, three splicing rates and three time intervals (i.e. average elongation rates): since the other models use information about particular sequences, they are ideally suited to model the recruitment rates. To vary the other parameters, elongation rates could be estimated from nucleosome information, and splicing rates could be inferred from the intron definition paradigm. Moreover, nucleosome positioning and chromatin modifications could be considered as input variables, and the recruitment rate variations in the CTRS models would be an ideal integration point for this type of data.

5.3 References

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J. and Frey, B. J. (2010). Deciphering the splicing code. Nat. *465*, 53–59.

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

Huranová, M., Ivani, I., Benda, A., Poser, I., Brody, Y., Hof, M., Shav-Tal, Y., Neugebauer, K. M., Stanek, D. and Huranova, M. (2010). The differential interaction of snRNPs with pre-mRNA reveals splicing kinetics in living cells. J. Cell Biol. *191*, 75–86.

Matera, A. G. and Wang, Z. (2014). A day in the life of the spliceosome. Nat. reviews. Mol. cell biology *15*, 108–21.

Melamud, E. and Moult, J. (2009). Stochastic noise in splicing machinery. Nucleic Acids Res. *37*, 4873–86.

Rino, J., Carvalho, T., Braga, J., Desterro, J. M. P., Lührmann, R. and Carmo-Fonseca, M. (2007). A stochastic view of spliceosome assembly and recycling in the nucleus. PLoS computational biology *3*, 2019–31.

Robberson, B. L., Cote, G. J. and Berget, S. M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. *10*, 84–94.

Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R. and Guigo, R. (2012). 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.

Zhang, C., Frias, M. A., Mele, A., Ruggiu, M., Eom, T., Marney, C. B., Wang, H., Licatalosi, D. D., Fak, J. J. and Darnell, R. B. (2010). Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Sci. (New York, N.Y.) *329*, 439–43.