Modelling Splicing

Hagen Tilgner

TESI DOCTORAL UPF / 2011

Barcelona, April 2011

Modelling Splicing

Hagen Tilgner

Memòria presentada per optar al grau de Doctor la Universitat Pompeu Fabra.

Aquesta Tesi Doctoral ha estat realitzada sota la direcció del

Dr. **Roderic Guigó i Serra** al Centre de Regulació Genòmica (CRG), dins del

Programa de Bioinformàtica i Genòmica

Hagen Tilgner **Roderic Guigó**

Barcelona, April 2011

To my family, Susanne, Hans, Dieter & Paul,
for making me the person wanting to do this.
And to Luisa,
without whom I would have given up.

cover design by Luisa Lente (luisa.lente@yoyo.es)

Acknowledgments

Five years in Barcelona and now it is after midnight, which will not make it easy to find the right words for all the people that deserve to be thanked. I owe a lot, to a lot, that is for sure. First of all of course my family, and in this context above all my father, for I think it is due to him that I could never imagine to be anything but a footballer (and despite of what I wrote in “primary school”, it was Van Basten whom I wanted to be, not Matthäus) or a scientist. The former being beyond my talents I became a scientist.

To my mother and Dieter, who made sure that the above aspirations did not totally take over and that I can, maybe not often, pronounce a phrase that neither contains the word “significant”, “p-value” nor “hatrick”.

To Paul, who made sure that a sufficient amount of alcohol was introduced into my life and early enough to shield me from total nerdness.

To Olivia, who, despite a distance of thousands of kilometers, has been the best friend in thousands of occasions. Most importantly, when things were not alright! And for “monsieur, ca va pas ca”s and “malaka”s when needed!

To Friedrich, a friend over so many countries and continents and the one and only Bulaia. Space-less and timeless.

To Kim, and also Laia for first inviting me to Barcelona - how one invitation can change a life ... incredible.

To Ari and Philipp, for welcoming me on my next trips to Barcelona.

To Bet and Ramon, for being the first friends I had in Barcelona.

To Cata, for offering me my first flat in Barcelona, for exploding eggs at 4am and for sleeping on the toilet at six but most importantly for bringing Luisa into my life. And to Lunes for teaching me to question every causality I thought unquestionable, exemplified by the "castration implies no masturbation"-causality.

To Kriszti & Pere, Martin & Renia, Helena, Jessica & Henrik, Tamara & Rasa and Miquelangel, the group of friends who have accompanied me throughout this PhD and who have been the closest in Barcelona, from very early on during our PhD classes to now, when it is almost over and all of us are moving around the world to find "new Barcelonas". It is you guys, with whom I have always felt most at ease! Trips to your home towns and marriages in Hungary, Sweden, Greece and Caius remain unforgettable memories. Kriszti, the closest of the closest, my (almost) Catalan sister, Martin my role model, although very different from me and yet similar ... probably typical for an Austrian/German connection.

To the PhD-students with whom I shared the courses with ... it always touches me when hearing about how it went for you guys, where you'll go and what you'll do. The above gang, Anna, Mireya, Gerrard and David V. - the ones with whom I interacted most.

To the "Flassaders"-Bar, Luisa, "El Pais", and the Barcelona sun for the most comfortable weekend mornings.

To Frank, Holger, Tobi, Felix & Uli, Martin, Timmi and later on Sascha. For unforgettable football games and Champions league nights, for simply having a beer and enjoying to speak some German.

To Micha and Thomas, for making long evenings at the lab fun with music ... and for making the cab cheaper afterwards. And to Andrea for being the

best of Micha and Thomas.

To Medya, whose friendship meant 10 times more, than the time we actually spent together ... and for the concept of “south Sweden”.

To Jana, Robert, Andre, Eduardo, Giovanni, Blanka, Romina & Gianni, for an unforgettable volleyball-season.

To Rory and Pedro, for always having food when it was needed. Rory for making the phrase “the blond guy who works on weekends” ambiguous.

To Andreas, Sajani and Julia, to Olivia and Boris, to Swantje, to Fritz, to Anja, Nadji and Michael, to Mutlu, to Florian & Quinn, to Emmi & Sibylle, to Friedrich & Loyse for not forgetting me and making sure that one giri was surrounded by others - although those giris were much less giri than this giri.

To the french connection Antoine, Thomas, Sylvain, Sylvain, Julien, Sarah and David M. for keeping the french in me alive.

To all the football crowd, with whom I spent many of my best moments in Barcelona, including the German fraction and the French, and the girls, Jana and Anna, who taught me that with brains, tactic and technique, one can overcome brute force. And Mafalda “iron foot” who proved that the same can be achieved with brute force, also.

To Julien, who almost made it to being my best flat mate ever, almost, for in some, but few respects he could not compete ... with Luisa.

To the Valcárcel lab people, for welcoming someone from the dark side and for putting up with my presentations filled with words like “significant” ... and “p-value”. I should have used the word “hatrick” more often.

To Oscar and Judith, who always found a solution, also on weekends. Oscar, for being the fastest tongue in Barcelona - it is due to you that I know that my Spanish is alright and my Catalan still requires work!

To Christoforos and Sonja, who offered me a way out of splicing simulation.

To Nuria, for finally not throwing me into the pool.

To Ramon, for getting me drunk the night before a CSH talk and for making sure I get up the morning after. To MP for the support on that meeting and to Sara for the aspirin.

To the Gracia crowd, Kriszti and Pere, Antoine & Sylvain and Thomas, Camilla and Francesca, Sonja, Anne and Eric - I have never and probably will never again live in a place with so many friends so close. A great feeling!

To Camilla, and Kriszti for taking advantage of my naivety and organizing a birthday surprise. Camilla, for tirelessly trusting a bioinformatician.

To Elias & Victor for a wonderful invitation to Navarra.

To Rosemarie, for being physically twice as old as me, but mentally twice as young - and for a family-like environment in Barcelona.

To Romina, who has been a guardian angel, justifying every word of the term.

To Mariano, for with no one else, I spent as much time in front of an open door - looking exactly for this door. Hopefully we will do that in NY, not in science. And to Pao & Mariano, for wonderful times in Barcelona and hopefully in BA, and a biology book that has touched my heart!

To David, for the coolest RAP in town.

To Francisco, Joao and Pedro, whose Portuguese conversation almost made work almost like home. To Joao for hundreds of kilograms of unpaid chewing gum.

To Juan, whose advice has been crucial in so many moments in this thesis, and with whom things started to go well. Finally, if I can consider myself a biologist now, than a very large portion is due to Juan.

To Roderic, for it is good to work for someone you consider a nice guy. When things did not go anywhere during year two, I came to hate his lack

of time. In the end, he has been a very good boss and an even better person, as proved in many occasions.

Almost finally, to Pep, Enrique, Robert, Sergi Castellano, Genís, whose template made writing this thesis much quicker. An extra thanks to Pep for sitting 4 hours until midnight with me until all layout/respect-UPF-rules issues were solved, even though wife and daughter were waiting at home. Things like that make you feel that there is a special bond between past and present members of the lab.

And finally to Luisa, for whom I could and actually should write pages. For love, for support, for a drawing that is and will be known more widely than this thesis. Above all, without you I would have given up or ended up in a psychiatric hospital. And I love you very much.

Finally to many unnamed people, that my mind cannot find right now, but that I will remember and regret, the very moment this goes into print.

Abstract

Splicing of RNA molecules is the process, by which intervening sequences ("introns") in the primary transcript are excised, and the remaining sequences ("exons") are concatenated to form the mature RNA. Recent evidence shows that almost all spliced genes are affected by alternative splicing. Here, we define the minimal length of RNA oligomers that can sensibly be called splicing factor binding sites. Then, we explore the capacity of these oligomers to predict complete exon-intron structures. We highlight those oligomers that are most informative for this and show, that equal accuracy as in previous approaches can be achieved with less RNA oligomers. The observation, that this approach falls short of accurately predicting the entire exon-intron structure, led us to investigate determinants linked to co-transcriptional splicing. We show that nucleosomes are preferentially positioned on exons and hypothesize that they play a role in splicing decisions. We then introduce the "completed splicing index" and conclude that co-transcriptional splicing is very wide-spread in humans. Furthermore co-transcriptional splicing exhibits links to chromatin organization. In the light of these results, we go on to monitor chromatin changes on differentially included exons in pair-wise tissue comparisons. We find a variety of histone marks, but not all, showing significantly different behavior on up- and downregulated exons. The most prominently appearing marks are H3K9ac and two lysine 4 methylation states.

Resum

L'*Splicing* de les molècules d'ARN és el procés pel qual les seqüències interposades ("introns") s'eliminen, i les seqüències restants es concatenen per a formar l'ARN madur. La investigació recent mostra que gairebé tots els gens amb *splicing* es veuen afectats per *splicing* alternatiu. Aquí, en primer lloc definim la longitud mínima d'un oligomer d'ARN per a funcionar com a lloc d'unió d'un factor d'*splicing*. A continuació, explorem la capacitat d'aquests oligomers per a predir estructures completes exó-intró. Destaquem els oligomers que són més informatius per a això, i demostrem que la mateixa precisió com en enfocaments anteriors es pot aconseguir amb menys oligomers. L'observació de que aquest enfocament és lluny de predir amb exactitud tota l'estructura exó-intró ens va portar a investigar els factors que juguen un paper en l'*splicing* co-transcripcional. Demostrem que els nucleosomes es col.loquen preferentment en els exons i plantegem la hipòtesi que juguen un paper en les decisions de l'splicing. A continuació, introduïm el "completed splicing index" i concluem que l'*splicing* co-transcripcional és molt generalitzat. A més, l'splicing co-transcripcional mostra vincles amb l'organització de la cromatina. A la llum d'aquests resultats, es van supervisar els canvis de la cromatina en exons diferencialment inclosos en dos teixits. Hem descobert una varietat de marques de les histones, però no totes, mostrant un comportament significativament diferent en els exons més inclosos i més exclosos. Las marques més destacades que apareixen són H3K9ac i dos estats de metilació de lisina 4.

Preface

Since the discovery of the helix structure, DNA has been widely recognized as a major determinant of life, even beyond scientific circles. In fact phrases like “It’s in your DNA” have become part of everyday language and conversation. Two other types of molecules, central to the biology of the cell, have not been appreciated that much by the wider public: RNA and proteins. Many people would probably associate proteins with nutrition, eggs (at least in German) or muscles, and not associate any meaning to the word “RNA”.

By scientists RNA has for a long time been appreciated as a vehicle of information between DNA and protein. On the one hand, this underscores the importance of RNA, as ALL information flow from DNA to protein must necessarily pass through an RNA molecule. On the other hand, this is tremendously underestimating the importance of RNA: First, we are learning more about RNA molecules whose purpose is not to code for proteins. Second, before coding RNAs get translated into protein, they undergo a variety of processes. Among them splicing can be seen as especially important, because it discards large parts of the primary RNA molecule. Furthermore, splicing can be “regulated”, that is, carried out with different results in different tissues, for example. Therefore, it is of importance to understand in

which ways RNA molecules are spliced and what determines how they are spliced.

Initially, we have explored the determinants of splicing that lie within the RNA sequence (see chapter 2), identifying the most important of these determinants. The observation, that these determinants were insufficient to predict splicing accurately in all cases, led us "to think out of the box" - the box being the RNA sequence.

A large body of research has shown that splicing can occur while the DNA molecule is copied into RNA. Furthermore, chromatin organization strongly influences this copying mechanism. Hence, we, along with other groups, contributed to this field by showing the links between intragenic chromatin organization and splicing (see chapters 3.1 and 3.2). Initially, we focused on the most basic aspect of chromatin organization, nucleosomes, and some few of their histone-tail-modifications within one cell type. Then, we have investigated the links between chromatin regulation and splicing regulation on a genomic level, based on the finding that "co-transcriptional" splicing appears to be very wide-spread in humans. Our results show that alternative splicing changes are frequently accompanied by chromatin changes.

Hagen Tilgner

Barcelona, April 2011

Contents

List of Tables

List of Figures

PART I

General Introduction

Chapter 1

General Introduction

Summary

94% of human genes contain introns and almost all of these multi-exon genes undergo alternative splicing (Wang et al., 2008). Thus, the splicing mechanism is central for shaping the RNA (and protein) population of the cell. Splicing mis-regulation has been connected to a variety of diseases (see Cooper et al. (2009) for review) and represents therefore an important subject in the development of therapeutic approaches. We are, however, currently not able to use all determinants of splicing in order to predict complete exon-intron structures, meaning that we still fall short of understanding why RNA molecules are spliced the way they are.

1.1 Sequence elements and their partners

The discovery of splicing

SPLICING OF RNA SEQUENCES WAS INITIALLY DISCOVERED, when mRNA molecules of adenovirus 2 mRNAs were hybridized to single strand DNA from its gene and visualized using electron micrographs. Besides double strand DNA-RNA hybrids, loops were also observed that corresponded to single strand DNA. Hence, consecutive RNA sequence hybridized to non-consecutive DNA sequence, showing that not all DNA sequence was reflected in RNA transcripts (Berget et al., 1977; Chow et al., 1977). Figure 1.1 shows one of the most important figures leading to this discovery (taken from Berget and co-workers (Berget et al., 1977)). Shortly after, “split genes” were also reported in vertebrates (Breathnach et al., 1977).

Discovery of Primary Sequence Elements and Their Role

Fairly soon it became clear that the sequences surrounding the removed RNA-sequences ("introns”) were not random but rather showed GT and AG dinucleotides at the 5’- and 3’-end of introns (Breathnach et al., 1978; Catterall et al., 1978). This finding, although exceptions exist, is nowadays referred to as the "GT-AG rule”. Figure 1.2 (which is taken from Breathnach et al. (1978)) illustrates the initial data, on which the “GT-AG”-rule is

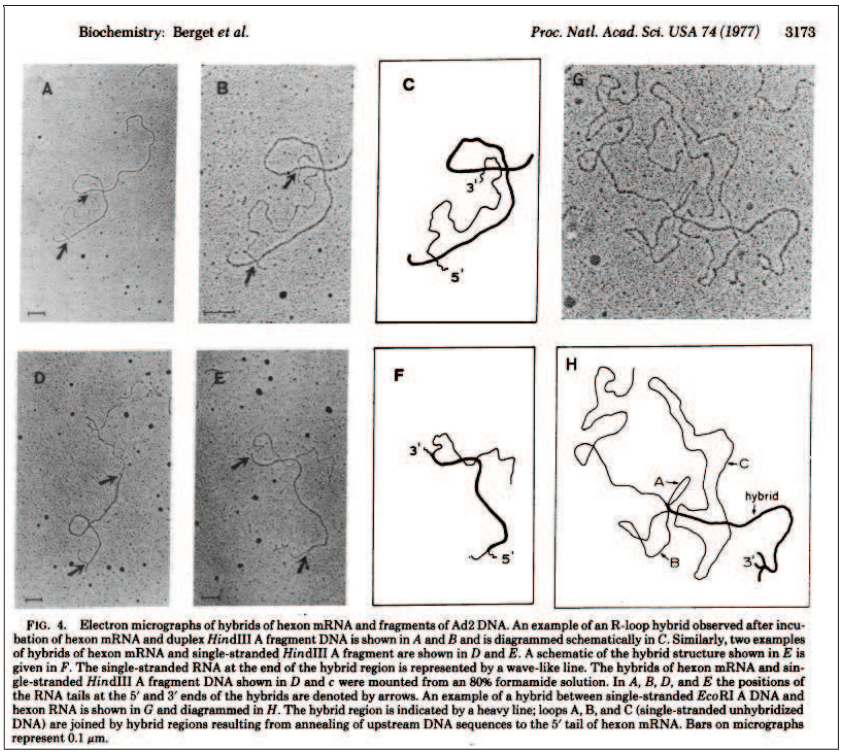


FIG. 4. Electron micrographs of hybrids of hexon mRNA and fragments of Ad2 DNA. An example of an R-loop hybrid observed after incubation of hexon mRNA and duplex *Hin*dIII A fragment DNA is shown in *A* and *B* and is diagrammed schematically in *C*. Similarly, two examples of hybrids of hexon mRNA and single-stranded *Hin*dIII A fragment are shown in *D* and *E*. A schematic of the hybrid structure shown in *E* is given in *F*. The single-stranded RNA at the end of the hybrid region is represented by a wave-like line. The hybrids of hexon mRNA and single-stranded *Hin*dIII A fragment DNA shown in *D* and *c* were mounted from an 80% formamide solution. In *A*, *B*, *D*, and *E* the positions of the RNA tails at the 5′ and 3′ ends of the hybrids are denoted by arrows. An example of a hybrid between single-stranded *Eco*RI A DNA and hexon RNA is shown in *G* and diagrammed in *H*. The hybrid region is indicated by a heavy line; loops A, B, and C (single-stranded unhybridized DNA) are joined by hybrid regions resulting from annealing of upstream DNA sequences to the 5′ tail of hexon mRNA. Bars on micrographs represent 0.1 μm.

Figure 1.1 Discovery of splicing: Figure by Berget and co-workers (Berget et al., 1977): Of special interest are subfigures G and H, where one can see both DNA-RNA hybrid sequences as well as single strand DNA loops indicating the existence of introns.

based.

Investigation of introns known at the time subsequently led to the discovery, that although the very strict consensus at acceptors and donors is limited to the GT and AG dinucleotides, a less strict but more widely spread consensus could be defined. While exonic nucleotides were involved, the larger part of this consensus was located on the intronic side, for both the acceptor and the donor. In this respect, a strong signal, now referred to as the polypyrimidine tract (ppy-tract), was a stretch of pyrimidine nucleotides upstream of the acceptor (Seif et al., 1979; Rogers and Wall,

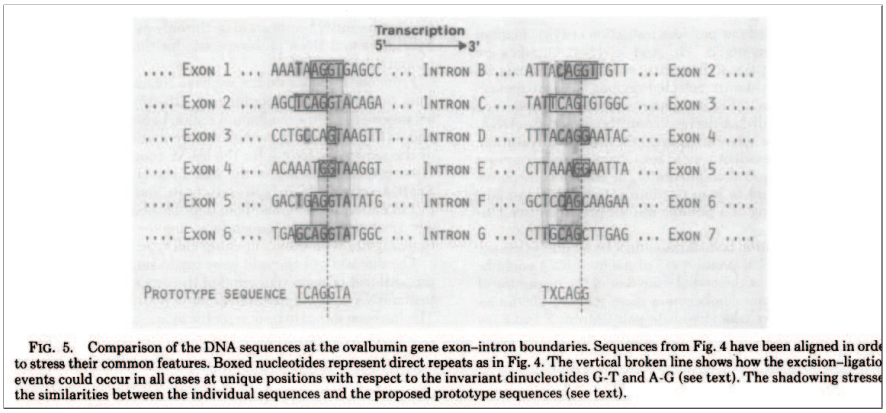


FIG. 5. Comparison of the DNA sequences at the ovalbumin gene exon–intron boundaries. Sequences from Fig. 4 have been aligned in orde to stress their common features. Boxed nucleotides represent direct repeats as in Fig. 4. The vertical broken line shows how the excision–ligatio events could occur in all cases at unique positions with respect to the invariant dinucleotides G-T and A-G (see text). The shadowing stresse the similarities between the individual sequences and the proposed prototype sequences (see text).

Figure 1.2 GT-AG rule: Figure by Breatnach and co-workers (Breathnach et al., 1978): These authors showed that introns could be placed on the DNA so that their first and last dinucleotide would be GT and AG. This finding is nowadays known as the "GT-AG”-rule.

1980). These authors also found another signal, located around the GT and extending around 5nt into the intron. The fact that this sequence showed complementary sequence to U1snRNA, gave the first clues, that part of the donor recognition could be achieved by base-pairing of the pre-mRNA with U1snRNA (Rogers and Wall, 1980; Lerner et al., 1980). Already then, it was suspected that U1snRNA would not be the only small nuclear RNA (snRNA), that could be part of the machinery that would achieve accurate intron removal (Lerner et al., 1980). In vitro experiments subsequently showed that during intron removal, the 5’end of the intron is covalently bound to an adenosine residue upstream of the ppy-tract (Ruskin et al., 1984). Because of the branched nature of the resulting structure, the adenosine residue was termed "branch point” or “branch site”. Aiding to understanding this last sequence motif was the discovery that U2snRNP binds pre-mRNAs in the area of the branch point (Black et al., 1985), which was later shown to be preceded by binding of an auxiliary factor - U2AF

- to the 3' splice site region (Ruskin et al., 1988). U2AF in turn was then shown to contain two distinct subunits, a 65kDA subunit "U2AF65", which binds to the ppy-tract, and a 35kDa subunit "U2AF35" (Zamore and Green, 1989). The role of U2AF35, contacting the AG-dinucleotide of the 3' splice site and stabilizing the U2AF65-RNA interaction, was elucidated only 10 years later (Merendino et al., 1999; Wu et al., 1999; Zorio and Blumenthal, 1999).

Much earlier, the first bioinformatic analysis related to splicing had been undertaken: By 1982, increasing numbers of known introns - though very few by today's standard - in higher eukaryotes had enabled statistical analysis of exon-intron and intron-exon boundary sequences. Such an analysis gave the first frequency based splice site consensus (Mount, 1982). This author determined the frequency of nucleotides at each position around the exon-intron and intron-exon boundary separately. His approach can be seen as the basis for splice site models that are still in use today, such as Markov chains used in Geneid (Parra et al., 2000).

Primary Sequence Elements Are Not Enough: Identification of Splicing Regulatory Sequences

Exonic Elements

While the sequence elements discussed above clearly play an important role in splice site selection, they are not the only elements to do so. Thus Reed and Maniatis (1986) and Mardon et al. (1987) showed that exonic se-

quences also played an important role for splicing outcome. Subsequently, it was demonstrated that such exonic sequences were involved in splice site selection by acting as binding sites for protein factors, for example tra-2 (Hedley and Maniatis, 1991). Generally speaking, SR-proteins are important interactors of these exonic RNA-oligomers (see for example Lavigueur et al. (1993)), that enhance splice site usage. Nowadays these enhancing oligomers are known as "exonic splicing enhancers" or "ESE". Based on the finding that other sequences could also negatively regulate splice site usage, and in analogy to the term ESE, such sequences were termed "exonic splicing silencers" or "ESS" (Amendt et al., 1995), although the protein factors, that interacted with them, remained unknown at the time. Natural candidates were hnRNPs, especially hnRNPA1, as it had been shown to antagonize the effect of SF2, an SR-protein, on splice site choice, both in-vitro (Mayeda and Krainer, 1992) and in-vivo (Cáceres et al., 1994). Indeed, hnRNPA1 is recruited to some ESS-sequences, leading to splicing repression (Del Gatto-Konczak et al., 1999; Zhu et al., 2001). It should be noted that the word "exonic splicing silencer" might be misleading, as they have also been shown to act as suppressors of pseudoexons (Sironi et al., 2004). What is really meant by the word "exonic" is, that the silencing (or enhancing) sequence is located in between the splice site pair that it acts on.

ISE and ISS

The importance of the previously described ESE and ESS is by now, at the time of writing, widely recognized (Wang and Burge, 2008). Yet, other sequence elements located up- and downstream of the exon, have also been shown to have effects on splice site choice. Such elements are now

called "intronic splicing enhancers" (ISE) or "intronic splicing silencers" (ISS), depending on whether they favor or disfavor splice site usage. An interesting twist to the classification into ESS, ESE, ISS and ISE was provided by Yeo et al. (2007), who defined conserved intronic splicing regulatory elements and showed that almost half of the defined RNA-words had previously been published as ESS (Yeo et al., 2007). An example of this is the finding that binding sites for the splicing factor PTB were found to be associated, on the one hand, with PTB-repressed alternative exons when located within or upstream of the regulated exon, however also with PTB-activated alternative exons when located downstream of such an exon (Llorian et al., 2010). In this sense, a PTB-binding site could fulfill three different roles: Those of an upstream ISS, and ESS and a downstream ISE.

Searches for Regulatory Elements

As pointed out earlier, the removal of introns from a pre-mRNAs depends on well known sequence elements such as acceptor, donor, ppy-tract and branchpoint. Yet, the number of suitable arrangements of these elements that are not exons (so called "pseudo exons") outnumber real exons by up to an order of magnitude (Sun and Chasin, 2000). This hints to the importance of ESE and ESS-motifs, that have been discussed above.

A large body of research has characterized ESE and ESS sequences experimentally (see for example Liu et al. (1998)), sometimes already making use of bioinformatic approaches to interpret the determined binding sites. In order to determine the whole set - or at least large parts of it - of ESE and ESS at once, a variety of experimental and often semi-computational

studies (bioinformatic sequence analysis followed by experiments or vice versa) have been utilized. In this way, ESEs have been predicted based on where they are preferential located (Fairbrother et al., 2002; Zhang and Chasin, 2004) or through conservation analysis (Goren et al., 2006). Also sequences to which SR-proteins bind well have been determined and a computational tool locate these in given RNA-sequences (Cartegni et al., 2003) has been made available. Similarly, ESS have been searched for using a mixture of experimental and statistical techniques (Zhang and Chasin, 2004; Wang et al., 2004; Goren et al., 2006). It has been shown (Goren et al., 2006) that the same splicing related oligomer can have effects of different strength (strong or weak effect, also investigated by Graveley et al. (1998)), as well as effects of different quality (splicing enhancing or silencing effect) depending on its position within the exon. Despite this considerable amount of research dedicated to exonic splicing regulatory elements, the ensemble of these elements still remains controversial. As much as 25% of all 4096 RNA-hexamers might be related to splicing, according to single publications (Stadler et al., 2006; Zhang et al., 2008), and the union of all these sets corresponds to an even larger percentage of hexamers. Intronic splicing regulatory sequences have been searched for, based on the idea that they should be conserved in the vicinity of exons (Yeo et al., 2007).

A Combined View

The literature described in the previous sections can be summarized as follows: U1- and U2snRNP, which assemble at the donor and the acceptor respectively, themselves have a certain capacity to recognize the consensus sequence of the two splice sites by RNA-base-pairing (see for example

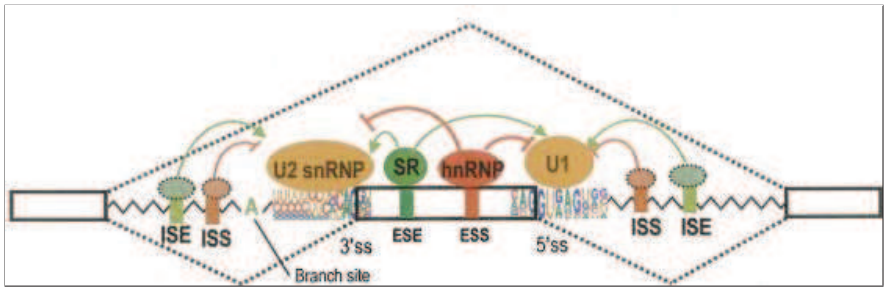


Figure 1.3 Figure by Zefeng Wang and Christopher Burge (Wang and Burge, 2008): Sequence elements in the RNA that influence splicing. The 3' splice site is recognized by the U2snRNP. Similarity of the ppy-tract to the consensus, represented here as a pictogram starting upstream of the exon and extending a bit into the exon, favors this recognition. The donor consensus is also represented by a pictogram. ESE elements located within the exon can be bound by SR-proteins and this binding favors splice site recognition, while binding of hnRNP proteins to ESS hinders it. Favoring or disfavoring elements can also be located in the surrounding introns, in which case they are names "ISE" and "ISS".

Rogers and Wall (1980) in the case of the donor). The sequence of the splice site relates to how easily it is recognized, with similarity to the splice site consensus roughly correlating with easy recognition by base-pairing with the snRNPs. Similarity to the donor and acceptor site can be measured, using a variety of methods. Geneid (Parra et al., 2000), maxEnt (Yeo and Burge, 2004) and idlBNs (Castelo and Guigó, 2004) are examples of these. In addition, splicing can be aided or hindered by the binding of protein factors to multiple sequence elements. Such sequence elements can in turn be located within the exon but also in the surrounding introns. The large body of research devoted to these elements has been reviewed by Zefeng Wang and Christopher Burge (Wang and Burge, 2008). Figure 1.3 is taken from their review and illustrates the sequence elements and their protein partners that influence splicing. The decision whether an exon is included into the mature RNA is thought to be a function of all the elements and binding

events summarized in this section. Thereby, exon recognition can be considered to be controlled by a variety of factors recognizing splice site sequences as well as ESE, ESS, ISE and ISS. The ensemble of these factors controls exon recognition in a combinatorial way (see Smith and Valcárcel (2000) for review). Yeo et al. (2007), as described above, showed that some of these binding site sets partially overlap. In part, this could have been suspected, because pseudoexons, whose splicing is often suppressed by the presence of ESS within them, are located in introns, not necessarily far away from real exons; and these are the regions that Yeo and co-workers investigated to find conserved words.

From a computational point of view, the entirety of the previous elements should provide a code that decides the fate of an exon in a given situation. Thus, knowledge about arrangement of splice sites and auxiliary splicing factor binding sites should facilitate the prediction of splicing. In this way, Wang et al. (2004) used splice site strength, ESE and ESS sequences as well as intronic G-triplets to predict the exon intron structure of cDNA sequences aligned to the genome. These authors created the term "splicing simulation", roughly signifying "prediction of splicing outcome using elements that are used by the splicing machinery". While using ESS, ESE and G-triplets considerably raised simulation accuracy (Wang et al., 2004), an important number of exons could not be predicted in this way. Focusing on tissue specific alternative splicing decisions Barash et al. (2010) showed that such a code could in principle be derived. These authors achieved considerable success in predicting, whether a given exon would be increasingly or decreasingly included in a tissue pair comparison (Barash et al., 2010).

Exon and Intron Definition Concepts

It is clear that an intron is recognized as a unit at the latest when its sequence is removed from the pre-mRNA. The "exon definition" concept (Robberson et al., 1990) however postulates that prior to intron removal, the exon is often recognized as a unit by the splicing machinery. This section reviews why this modus operandi is almost certainly dominant in higher eukaryotes. Vertebrate exons are considerably shorter than vertebrate introns, a statement that is not true for non-vertebrate eukaryotes, such as fungi (see Hawkins (1988) for early evidence). A recent estimate (Zheng et al., 2005) puts the median length of human introns at 1508nts, more than 10fold that of human constitutive exons (120nts). Such observations planted first doubts that the vertebrate intron, although it is the unit that is finally excised from the RNA molecule, might not be the unit that is always recognized initially. In this way, Robberson et al. (1990) showed that the presence of a 5'splice site of an exon affected splicing of the upstream intron. This made the authors suggest that the first unit to be recognized in the pre-mRNA molecule is the exon and called this process exon definition (Robberson et al., 1990). The ideas and observations, that led to the formulation of this concept, have been reviewed by Berget (1995).

One guiding principle of splice site pairing mechanisms is thus, that it is governed by exon- and intron-length. For higher eukaryotes exon definition preceding intron removal would be the most frequent mechanism, while direct intron definition would dominate in lower eukaryotes (Berget, 1995). Examples of experimental data, that are consistent with this model, are the following: When large exons were placed in surroundings with short introns, splicing occurred effectively, supposedly through intron definition. Yet, when placing the same large exons in contexts with large introns exon

skipping was observed, supposedly because neither exon definition nor intron definition could work (Sterner et al., 1996). Moreover, intron definition was shown to cease when introns are longer than between 200 and 250nts (Fox-Walsh et al., 2005). Experimental evidence putting ESE and ESS and exon definition into perspective comes from the FAS receptor. Binding of U1snRNP at the donor of exon 6 facilitates U2AF binding to the acceptor of the same exon. Binding of PTB to an exonic splicing silencer, however, can inhibit the positive effect of U1snRNP binding at the donor on U2AF binding at the acceptor (Izquierdo et al., 2005). Further molecular insights into how exon definition could work have been given by Schneider et al. (2010). These authors purified cross exon complexes formed in vitro and found U4, U5 and U6 to be present in such complexes (Schneider et al., 2010).

In summary one can say, that in most cases in higher eukaryotes, initially, exons are defined by the splicing machinery, through a variety of protein-RNA and RNA-RNA interactions. The intron is then removed in a subsequent step.

1.2 Alternative Splicing

General Comments

So far, we have looked at the elements that are involved in splicing and have scratched the surface of how they affect splicing decisions. We have not yet touched the question, why eukaryotic genes are often organized in a split way, that is interrupted by introns. There are two possible naive answers to this question: First, genes could contain introns, because the intronic RNA sequences are necessary for “something else” than the spliced RNA. Second, modularization, that is separation of “a whole” into “smaller parts”, allows to combine modules (exons) in multiple distinct ways. Here, we will focus on the second answer, that has been demonstrated to affect almost all multi-exon genes (Harrow et al., 2006; Wang et al., 2008) and is known as "alternative splicing”.

Early Evidence & Importance

Just as the discovery of splicing, initial evidence for alternative splicing came from viruses. In the late seventies Berk and Sharp (1978) found that the same viral DNA could lead to two different mRNAs, by choosing different "splice points”, in the terminology of the time. Importantly, these two mRNAs corresponded to two distinct and known proteins (Berk and Sharp, 1978). In today’s terminology this alternative splicing event would be called "alternative donor” or "alternative 5’ splice site”.

The importance of alternative splicing is highlighted, first, by large scale studies, second, by single examples of multiple functional isoforms produced from a single gene and third, by implications for disease. Here, we will have a look at the first two points. Estimates of how many human genes are alternatively spliced are rising each time a new report is published: Modrek and co-workers (Modrek et al., 2001) analyzed EST-data and estimated that at least 42% of all human genes are alternatively spliced. Later Johnson et al. (2003) estimated that "at least 74%" of all human multi-exon gene are alternatively spliced based on exon junction microarrays, and in the framework of the ENCODE project this number was estimated as 86% (Harrow et al., 2006). The most recent estimates vary between 95% (Pan et al., 2008) and 98% and 100% (Wang et al., 2008) using mRNAseq data. This last estimate is satisfying in the sense that it is unlikely to be raised very much in the near future.

Early examples of alternative splicing were connected to coding of distinct proteins by the same gene (e.g. Berk and Sharp (1978)) and multiple cases exist, where this function has been demonstrated in detail. Nilsen and Graveley (2010) list examples of this in fly, worm, human and mouse. Another example is the FAS-receptor, which has been shown to mediate apoptosis (Itoh et al., 1991). From its gene two different spliced mRNAs can be produced, that differ by the inclusion of exon 6. The apoptosis mediating isoform is produced when exon 6 is included and the resulting mRNA encodes the FAS-receptor. Exon 6 exclusion produces an mRNA that is translated into a soluble protein, which neither binds to the membrane nor promotes apoptosis (Cheng et al., 1994). Examples like this justify the idea that alternative splicing can contribute substantially to the diversity of the human proteome (e.g. Nilsen and Graveley (2010)). While this is certainly true, it is worth noting that alternative splicing of a gene can be

functional without coding for multiple proteins. Several lines of evidence support this idea: When alternative splicing introduces a premature stop codon in many transcripts of a gene, leading to degradation of the resulting mRNAs through NMD, this can lead to a decrease in gene expression of the affected gene. Regulation through this coupling of splicing and NMD appears to affect many splicing factors themselves - see Lareau et al. (2007) and also Cartegni et al. (2002) for review. Second, a very similar argument can be made, with reduced mRNA stability taking the place of NMD, in the sense that expression of alternative mRNA isoforms with reduced stability effectively leads to reduced expression (Sureau et al., 2001). Third, differentially spliced mature alternative RNAs can differ in their sub-cellular localization: In this way, Sun et al. (2010) showed that three alternative RNA isoforms of the SF2/ASF gene, which encodes a splicing factor itself, are retained in the nucleus and thus cannot be translated. Fourth, an observation probably connected to the previous three possibilities, alternative splicing can occur in UTRs and this is not infrequent (Sammeth et al., 2008), so that in these cases no effect on the amino-acid sequence of the encoded protein can be evoked. Last but not least, also for non-coding RNAs one can find annotated alternative splicing events in databases. For all these reasons, the importance of alternative splicing goes beyond the encoding of multiple proteins in one gene.

Types of alternative splicing

Alternative splicing events can be extremely complex and involve the alternative selection of multiple exons or splice sites, as illustrated by the

DSCAM gene in *Drosophila melanogaster* (Schmucker et al., 2000). In principle, such complex events, as well as other much less complex events, can be broken down into four different "atomic" events:[1] Alternative 5'ss usage, alternative 3'ss usage, inclusion or skipping of an entire exon and splicing or retention of an entire intron ("intron retention"). These cases are illustrated in figure 1.4 (taken from Nilsen and Graveley (2010)).

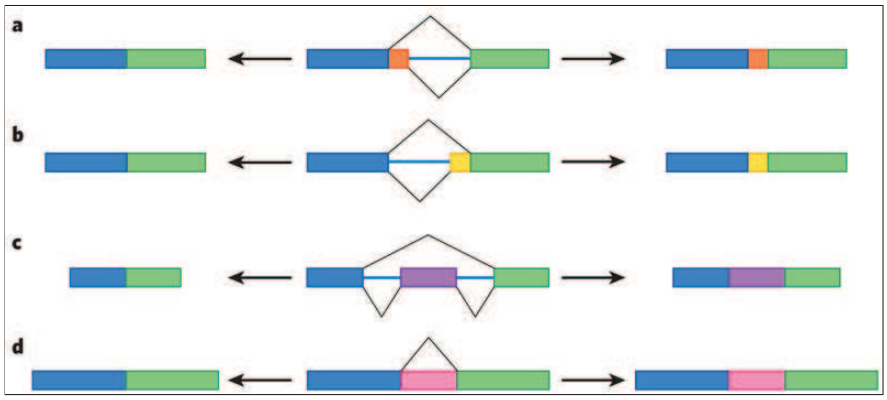


Figure 1.4 Figure by Nilsen and Graveley (Nilsen and Graveley, 2010): Four most basic types of alternative splicing. Production of two different spliced molecules, depending on which 5'splice site (**a**) or 3'splice site (**b**) is chosen. Likewise different spliced molecules can be produced depending on whether an exon is included or not (**c**) or whether an intron is spliced or not (**d**).

In principle any combination of the above four basic types is a possible alternative splicing event. In this way, other types of alternative splicing

[1]Sometimes two more types are referred to as alternative splicing events: A longer first intron in combination with an alternative upstream transcription start site or a longer last intron in combination with an alternative polyA-site. Such cases are certainly "alternative" in the sense that different lariats are formed. However, the spliceosome might be presented with pre-mRNAs that do not contain the same splice-signals in the two cases. In this case, it is rather the transcription machinery that "has made alternative decisions", which make it impossible for the splicing machinery to splice constitutively.

events, such as “mutual exclusive exons” can be viewed as a combination of two differentially included exons, whose inclusions are counter regulated. The frequency with which different types of alternative splicing events occur differs between organisms (Sammeth et al., 2008). Exon skipping, for example, is the most frequent form of alternative splicing in mammals, where intron retention, on the other hand, is rare (Sammeth et al., 2008). This is contrasted by higher rates of intron retention and lower frequencies of exon skipping in worm and fly (Sammeth et al., 2008). In cases where an exon skipping event occurs in the coding part of a gene with both splicing isoforms coding for a protein, one would expect the length of the skipped exon to be a multiple of 3, so that its skipping does not cause a frame-shift. Indeed, it has been shown that coding exons involved in exon-skipping events, which are conserved between human and mouse, tend to be divisible by 3 more often than such exons in UTRs (Magen and Ast, 2005). This, in itself, suggests that many exon skipping events lead to mature RNAs that are translated to protein.

Regulation of alternative splicing

As noted earlier, the decision whether a given exon is in- or excluded in the mature RNA is thought to depend (1) on the binding of splicing auxiliary factors to ESE, ESS, ISE and ISS, (2) on correct assembly of the spliceosome components to acceptor and donor sequences and (3) on the correct interaction of these molecules during exon definition. Therefore, anything that alters any of these three interactions with respect to a reference situation can potentially lead to alternative splice site usage. Two generally different causes of alternative splicing should, nevertheless, be distinguished: Those,

that organisms use in order to regulate alternative splicing in a tissue or time specific manner or in reaction to natural external stimuli and those, that are due to alterations of the genetic material of a cell, either due to mutations, for example in disease or due to experiments. This section reviews examples of tissue specific regulation of alternative splicing.

An interesting case of tissue specific splicing regulation involves two homologous splicing factor proteins: PTB and nPTB — also known as brPTB. The latter is expressed in neuronal cells, whereas the former is not (Markovtsov et al., 2000; Polydorides et al., 2000). Although the two proteins are encoded by different genes, they have strong sequence similarity (Markovtsov et al., 2000). Both proteins differ in RNA binding and splicing repression properties (Markovtsov et al., 2000) and nPTB has been shown to influence the action of other RNA binding proteins such as hnRNPH, KSRP and Nova (Markovtsov et al., 2000; Polydorides et al., 2000). Expression of PTB downregulates nPTB expression and downregulation of PTB induces nPTB expression (Boutz et al., 2007). Thus, one of the two can be expressed, but not both simultaneously, and the switch from one to the other affects splicing decisions of many exons (Boutz et al., 2007). This switch between PTB and nPTB has been connected to differentiation of P19 mouse embryonal carcinoma cells and, supposedly, is one important aspect of such differentiation events (Boutz et al., 2007). The exact modus operandi of PTB in regulation of alternative exon inclusion is currently under debate. Xue et al. (2009) report that PTB binding close to an alternative exon results in skipping of this exon whereas PTB-binding close to the flanking exons would enhance exon inclusion. Llorian et al. (2010), on the other hand, argue, that PTB acts as a repressor, when bound upstream or within a regulated exon, but as an activator when bound downstream of a regulated exon. While in the framework of this introduction, it is not possible to opt for either of these

different concepts, it is worth noting, that the second concept is very much closer to that proposed for alternative splicing regulation by Fox2: Yeo et al. (2009) find that Fox2 bound upstream of a regulated exon represses exon inclusion whereas Fox2 bound downstream enhances it. Also the "RNA map" of Nova (Ule et al., 2006) is conceptually close to this interpretation. While downstream bound Nova related to enhancing activity, exonic bound Nova had silencing effects.

Whereas inclusion or exclusion of alternative exons clearly depends on auxiliary splicing factors, such as the ones mentioned above, alternative exons have a number of intrinsic characteristics, that supposedly predispose them to being alternative. Zheng et al. (2005) have collected many of these features: First, both in human and mouse, they are shorter and their length distribution shows a higher standard deviation than that of constitutive exons. The authors connect this to the concept of exon definition proposed by Robberson et al. (1990). In the same way, alternative exons have acceptors and donors that deviate more from the two consensuses, again both in human and mouse (Zheng et al., 2005). Supposedly both characteristics make their inclusion into mature RNAs more dependent on regulating auxiliary factors. Furthermore, alternative exons tend to be more often divisible by 3 and keep the reading frame of the encoded protein than expected from various background models (Magen and Ast, 2005; Zheng et al., 2005). This characteristic is however certainly a consequence of their coding capacity, rather than a determinant of splicing.

Alternative Splicing & Disease

As pointed out earlier, a plethora of sequence elements, including acceptor, donor, ESE, ESS, ISE and ISS are involved in the splicing process. Based on a set of 238 ESE-hexamers, it has been estimated that 80% of human exons contain three or more ESE (Fairbrother et al., 2002); yet, it has been suspected, that the actual set of ESE is up to five times larger (Stadler et al., 2006; Zhang et al., 2008) than the set by Fairbrother et al. (2002). Thus, there are plenty of sequence elements whose disruption by a mutation can potentially change splicing outcome. In agreement with this, 6 out of 19 synonymous substitutions in exon 12 of the human cystic fibrosis transmembrane conductance regulator (CFTR) gene strongly affected splicing outcome, while most others had mild effects (Pagani et al., 2005). Already in 1992 it was estimated that 15% of all human mutations causing disease, could do so by affecting splicing (Krawczak et al., 1992). More recently, the idea has been put forward that this percentage might actually be close to 60% (López-Bigas et al., 2005). A number of cases where mutations in sequence elements of a gene cause splicing defects and therefore lead to disease are well studied. The SMN protein that is encoded by the SMN1 and SMN2 genes is one of the better known examples (see Cooper et al. (2009) for review).

As noted earlier, RNA sequence elements are mostly functional in splice site selection, when they are bound by spliceosomal or auxiliary splicing factors. Therefore, any change in the functionality of a splicing factor can in principle lead to splicing changes in many target exons of this splicing factor. Indeed, a recently described example, in which different types of mis-regulation of a splicing factor lead to splicing changes of a variety of target exons, is the FOX2 gene (Venables et al., 2009). This publication de-

scribes that FOX2 expression is downregulated in one cancer and its splicing is changed in another cancer type. Using high throughput PCR in healthy and cancerous tissue, the authors show that inclusion levels of one third to one half of all “active” alternative splice forms can discriminate between healthy and cancerous tissue (Venables et al., 2009). Mis-regulation of alternative splicing can therefore be considered a broad contribution to cancer. A "broad contribution” is however not necessarily a causal contribution. That deregulation of some alternative splicing events could be at least partially causal for cancer is supported by the following piece of evidence. It is generally accepted, that under “normal” conditions, cancerous cells should undergo apoptosis but escape this fate (Letai, 2006). In this respect, it is very interesting that inclusion of FAS exon 6 into the mature RNA leads to apoptosis, while skipping does not do so (Cheng et al., 1994). Similar statements, where two mRNA-isoforms seem to make the difference with respect to apoptosis, can be made for the genes Bcl-x and caspase-2 (see David and Manley (2010) for review).

1.3 Co-transcriptionality of splicing

Discovery and Implications

INITIAL EVIDENCE showing that splicing could occur co-transcriptionally came from an experiment by Beyer and Osheim (1988), see also (Neugebauer, 2002; Kornblihtt et al., 2004; Allemand et al., 2008; Moore and Proudfoot, 2009; de Almeida and Carmo-Fonseca, 2008) for review. These authors made use of electron

micrographs for Drosophila genes and showed that intronic sequences are looped out and removed from the pre-mRNA sequence, while it is still attached to the chromatin template. Figure 1.5 is taken from their publication and visualizes this discovery. This finding is important for two - somewhat related - reasons: First, if splicing can take place while both the DNA and the nascent RNA molecule are "connected" to PolII, the latter can become a player in splicing decisions. We will discuss this point in the present section. Second, since splicing can occur in proximity to the chromatin template, chromatin organization is also given a chance to influence splicing, either influencing PolII behavior or by a more direct interaction with splicing factors. These last two points will be treated in the following section 1.4.

The Pol2-CTD

DNA-dependent RNA polymerase II is generally assumed to transcribe all eukaryotic mRNAs (Kornberg, 1999). Since splicing of pre-mRNAs can occur co-transcriptionally (see above), the properties of the RNA PolII become of interest from a splicing point of view. One important property of RNA PolII is the C-terminal domain of the RPB1 subunit. This C-terminal domain, nowadays referred to as the PolII-CTD, contains 52 repeats of the amino acid sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser, which offers multiple phosphorylation sites (Corden et al., 1985). The serine residues are referred to as serine-2, serine-5 and serine-7, due to their position in this repeated amino acid sequence. Especially serine-5 and serine-2 phosphorylation have been associated with transcription: Close to yeast transcription start sites serine-5-phosporylation is dominant, whereas serine-2 phosphorylation is

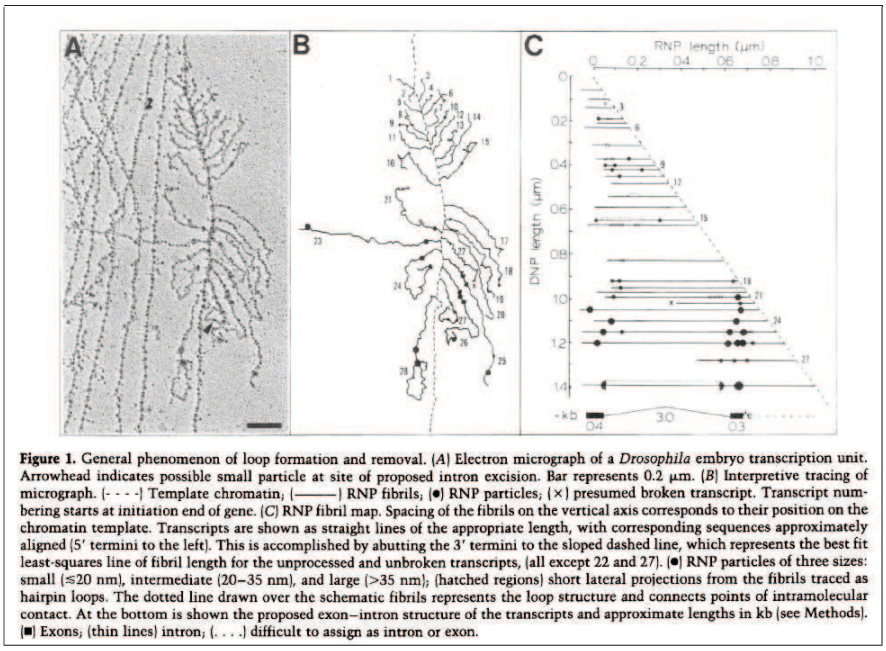


Figure 1. General phenomenon of loop formation and removal. (*A*) Electron micrograph of a *Drosophila* embryo transcription unit. Arrowhead indicates possible small particle at site of proposed intron excision. Bar represents 0.2 μm. (*B*) Interpretive tracing of micrograph. {- - - -} Template chromatin; {———} RNP fibrils; {●} RNP particles; {×} presumed broken transcript. Transcript numbering starts at initiation end of gene. (*C*) RNP fibril map. Spacing of the fibrils on the vertical axis corresponds to their position on the chromatin template. Transcripts are shown as straight lines of the appropriate length, with corresponding sequences approximately aligned (5′ termini to the left). This is accomplished by abutting the 3′ termini to the sloped dashed line, which represents the best fit least-squares line of fibril length for the unprocessed and unbroken transcripts, (all except 22 and 27). {●} RNP particles of three sizes: small (≤20 nm), intermediate (20–35 nm), and large (>35 nm); (hatched regions) short lateral projections from the fibrils traced as hairpin loops. The dotted line drawn over the schematic fibrils represents the loop structure and connects points of intramolecular contact. At the bottom is shown the proposed exon–intron structure of the transcripts and approximate lengths in kb (see Methods). {■} Exons; (thin lines) intron; {. . . .} difficult to assign as intron or exon.

Figure 1.5 Figure by Ann Beyer and Yvonne Osheim (Beyer and Osheim, 1988): Co-transcriptional looping and removal of an intronic sequence. Of special interest for the reader are transcripts 27 and 28. The latter shows the a looped, but unspliced intron, while the former shows a transcript where this intron is already removed.

mainly found in the gene body (Komarnitsky et al., 2000). McCracken et al. (1997) showed that in-vivo splicing, 3'end processing and transcription termination all depend on the CTD-domain of PolII, hinting that these processes did not only occur in physical proximity to transcription, but that there could be a functional connection. Further evidence, linking the CTD and its phosphorylation states to mRNA processing, exists. For example, in mammals the elongation factor Spt6 binds to phosphorylated serine-2 of the CTD. Transcripts produced with a mutant Spt6, that cannot perform this binding, however, show defects in terms of splicing (Yoh et al., 2007), hinting the functional coupling between transcription and splicing.

Coupling of transcription and splicing

The findings described previously indicate the important role of the Pol2-CTD in coupling of transcription and splicing. Much earlier, evidence existed, showing, that transcription and splicing not only could occur simultaneously, but that the dynamics of the former can influence the latter: In 1997 Cramer et al. (1997) showed, that when the fibronectin gene was expressed from different promoters, an internal exon was differentially included. Conversely, elongation rate was also shown to affect exon inclusion decisions (Kadener et al., 2001; de la Mata et al., 2003), with polymerase speed and exon inclusion anti-correlating in this case. Further observations consistent with such a kinetic model exist: First, activation of transcription initiation and elongation through transcriptional activators of the IIb class reduce exon inclusion (Nogues et al., 2002). Second, introduction of RNA-Pol2 pausing sites can affect alternative splicing outcome (Roberts et al., 1998), by delaying the synthesis of regulatory elements. Third, also in yeast, a slower Pol2 favored exon inclusion (Howe et al., 2003). Finally, understanding of coupling between transcription and splicing has been recently advanced by the finding that Pol2 transiently accumulates on acceptors in a splicing dependent way in budding yeast. This co-occurs with changes in the phosphorylation states of serine 5 and serine 2 in the Pol2-CTD (Alexander et al., 2010).

1.4 Intragenic chromatin organization and splicing

[2] A relationship between chromatin structure and splicing had already been speculated on as early as 1991 (Beckmann and Trifonov, 1991), after the observation that the distance between consecutive splice splice junctions follows a periodic pattern, compatible with nucleosome phasing. Moreover, examples of direct chromatin - splicing interactions have been found recently: A subunit of the chromatin remodeling complex SWI/SNF regulates splicing (Batsché et al., 2006) and histone modifications have been shown to be involved in splicing regulation (Schor et al., 2009; Alló et al., 2009; Nogues et al., 2002; Sims et al., 2007) - these findings are described in more detail in other parts of this introduction. The investigation of the relationship between chromatin structure (i.e., at single nucleosome level) and splicing has been confounded, however, by the lack of high resolution nucleosome and chromatin status maps of higher eukaryotic genomes. Only recently, with the advent of genome wide tiling arrays and massively parallel sequencing, has the delineation of such maps become feasible. Thus, Kolasinska-Zwierz et al. (2009) have produced genome wide maps of three histone H3 tail modifications in Caenorhabditis elegans using chromatin immunoprecipitation (ChIP) followed by microarray hybridization. They found that Trimethylation of Lys36 (H3K36me3) is enriched within exons relative to introns, providing the first experimental indication that chromatin structure could indeed be related to the exonic structure of genes. Global genome wide occupancy maps for all nucleosomes, irrespectively of modification status, have recently also become available

[2] The text of this part is in large parts based on (Tilgner and Guigó, 2010)

for human (CD4+ T-cells, Schones et al. (2008)) and worm (Valouev et al., 2008) (whole worm tissue mix). These have been obtained by high-throughput sequencing of DNA from micrococcal nuclease (MNase) digested chromatin preparations. More recently nucleosome maps for two medaka strains (blastulae, Sasaki et al. (2009)) and retained nucleosomes in human sperm (Hammoud et al., 2009) have been produced.

Nucleosome organization on exons

A flurry of papers (Schwartz et al., 2009; Tilgner et al., 2009; Andersson et al., 2009; Nahkuri et al., 2009; Spies et al., 2009; Hon et al., 2009) appeared late in 2009, describing complementary computational analysis of these maps (Schones et al., 2008; Valouev et al., 2008; Sasaki et al., 2009; Hammoud et al., 2009). All these analyzes provide strong evidence that internal exons are enriched in nucleosomes both in human and worm, and uncover a number of features characterizing the relationship between the exonic structure of genes and nucleosome occupancy:

Nucleosome occupancy in exons is independent of transcription. Indeed, while highly expressed genes have lower nucleosome level in and around exons, consistent with nucleosome depletion during RNA-Pol2 passage (Lee et al., 2004), the nucleosome peak is observed in exons both from expressed and silent genes (Schwartz et al., 2009; Tilgner et al., 2009; Andersson et al., 2009; Nahkuri et al., 2009). This is in contrast to the largely expression dependent exonic H3K36me3 peak (Kolasinska-Zwierz et al., 2009).

Nucleosome enrichment is as strong on exons as at TSSs. While nucleosome enrichment on exons was discovered after that at the +1 nucleosome of TSSs in CD4+ T-cells (Schones et al., 2008), it is at least equal in strength (Andersson et al., 2009; Spies et al., 2009).

Exons with weak splice sites have stronger nucleosome occupancy. Moreover, in exons with strong splice sites an extended region of nucleosome occupancy occurs upstream from the acceptor site—a region which is absent in exons with weak splice sites (Tilgner et al., 2009). The patterns are more obvious with acceptor sites (Tilgner et al., 2009; Spies et al., 2009) but are also apparent with donor sites (Spies et al., 2009).

Exons surrounded by longer introns have stronger nucleosome occupancy. Mammalian introns are frequently, but not always, order(s) of magnitude longer than exons and this feature of gene architecture is also related to nucleosome occupancy on exons: exons surrounded by long introns show higher nucleosome occupancy than exons surrounded by short introns (Spies et al., 2009).

Nucleosome occupancy is stronger in longer exons. For exons shorter than 50 bps the nucleosome peak is almost absent (Andersson et al., 2009) and it tends to grow and "move" downstream from the acceptor site towards the center of the exon as exon length increases (Schwartz et al., 2009; Tilgner et al., 2009). This behavior is more evident in exons with weak acceptor sites (Tilgner et al., 2009).

Pseudoexons are depleted of nucleosomes. (Tilgner et al., 2009) Pseudoexons are intronic regions of length similar to that of bona fide exons and flanked by strong splice sites, but with no evidence of inclusion in mature RNA sequences.

Sequence dependent computational predictions of nucleosome positioning recapitulate the nucleosome peaks on exons. (Schwartz et al., 2009; Tilgner et al., 2009) Computational predictions are also able to detect the differences in nucleosome occupancy observed between exons with weak and strong splice sites on the one hand, and exons and pseudoexons on the other hand (Tilgner et al., 2009).

Nucleosome occupancy in exons appears to be conserved across metazoans. (Kolasinska-Zwierz et al., 2009; Valouev et al., 2008; Schwartz et al., 2009; Tilgner et al., 2009; Andersson et al., 2009; Nahkuri et al., 2009) Predicted nucleosome peaks are observed in exons from seven metazoans (Schwartz et al., 2009) (human, mouse, chicken, zebrafish, ciona, fruitfly and worm) suggesting, together with the experimental results in worm (Valouev et al., 2008), that the relation between nucleosomes and splicing is broadly conserved through metazoan evolution.

Splicing regulatory motifs and chromatin structure are related. Intronic splicing regulatory motifs (Yeo et al., 2007; Voelker and Berglund, 2007) are depleted in nucleosomes, while exonic splicing enhancers and silencers (Wang et al., 2004; Fairbrother et al., 2002; Goren et al., 2006), appear not to be (Schwartz et al., 2009).

Nucleosome occupancy on exons is present independently of sequence conservation and GC content. Nucleosome occupancy on exons is present independent of sequence conservation (Nahkuri et al., 2009) and GC content (Schwartz et al., 2009; Andersson et al., 2009; Nahkuri et al., 2009), although the strength of nucleosome occupancy and GC-content are correlated (Schwartz et al., 2009). In fact, sequence composition could largely explain differential nucleosome occupancy between exons and pseudoexons (Spies et al., 2009), but simple GC content is not the only factor (Tilgner et al., 2009).

Taken all together these observations are suggestive for a role of nucleosome occupancy in splicing, a possibility raised by most authors (Figure 1.6) More specifically, the interplay between nucleosome occupancy upstream and downstream from potential acceptor sites would serve as an additional mark for exons: nucleosome depletion upstream from the acceptor site coupled with nucleosome occupancy within the exon would promote the inclusion of exons with weak splice sites, which are commonly assumed in need of additional factors for proper recognition. A similar pattern would facilitate the identification of exons surrounded by long introns—in which decoy pseudoexons are very abundant, confounding proper exon recognition. Conversely, nucleosome depletion downstream from a potential acceptor site coupled with upstream occupancy would prevent the inclusion of pseudoexons with strong splice sites. The molecular mechanisms by means of which nucleosomes mark exons remain to be elucidated, but different non-mutually exclusive hypothesis have been put forward. First, nucleosome positioning could influence transcription kinetics (Schwartz et al., 2009; Tilgner et al., 2009; Spies et al., 2009). Second, nucleosomes could directly or indirectly contribute to specifically

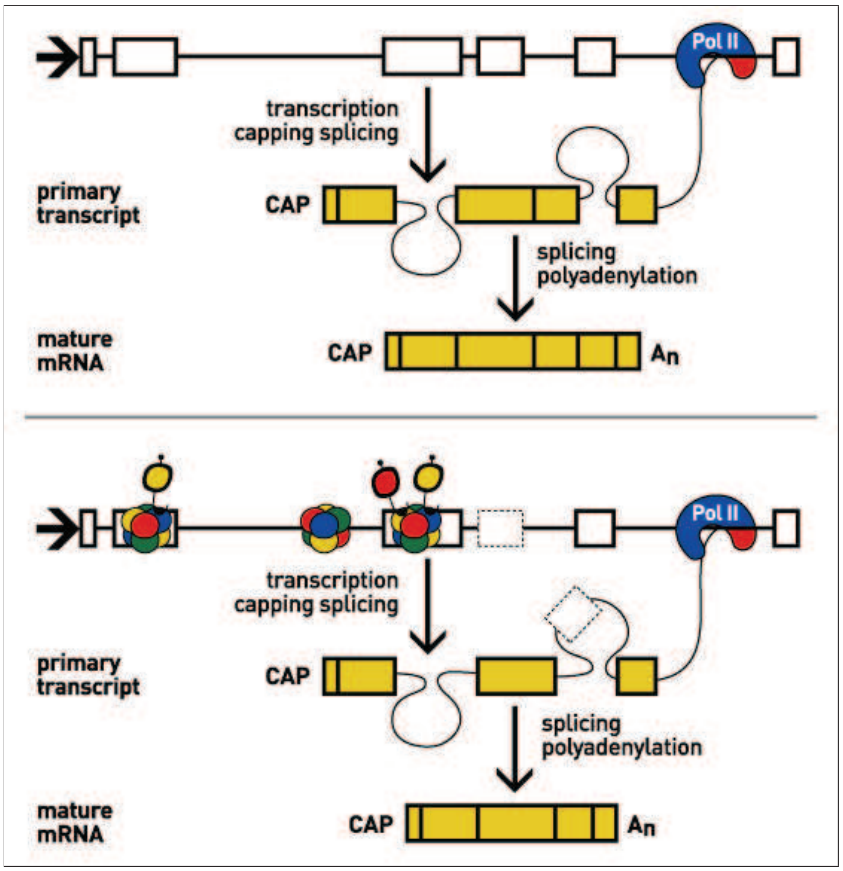


Figure 1.6 Extension of the model by Kornblihtt et al. (2004): Chromatin involvement in co-transcriptional splicing. View of co-transcriptional splicing as proposed and drawn by Kornblihtt et al (Kornblihtt et al., 2004). Splicing is co-transcriptional, allowing transcription kinetics and thereby broad chromatin structure (e.g., open/closed chromatin state) to influence splicing (top). Extension of the top-panel-model, in which stable positioning of possibly single nucleosomes and histone modifications can also influence splicing. In particular nucleosome occupancy within exons will contribute to their inclusion in the mature transcript.

recruit splicing factors during transcription (Schwartz et al., 2009; Tilgner et al., 2009; Andersson et al., 2009; Spies et al., 2009; Hon et al., 2009), a possibility supported by recent findings that some splicing factors can bind to nucleosomes in a histone—tail—modification regulated fashion (Sims et al., 2007; Loomis et al., 2009). Both hypotheses are in-line with a

model under which nucleosomes could co-transcriptionally enhance exon definition (Tilgner et al., 2009; Nahkuri et al., 2009; Spies et al., 2009) - a phenomenon by which splicing factors bound at the flanking splice sites stabilize each other.

Strictly speaking, however, the observations above are only indicative of correlation between nucleosome occupancy and splicing, but they are not conclusive of directionality in the relation or of causation. It could be argued, for instance, that the elevated GC content of exons contributes to position nucleosomes—which have also been postulated to prefer GC-rich regions. If so, preferential positioning of nucleosomes in exons would simply reflect the sequence composition of the latter, but it would not imply that nucleosomes are part of the mechanism for exon selection. While, in absence of further experimental evidence, this hypothesis cannot be completely ruled out, we believe that the characteristic nucleosome occupancy patterns in weak vs. strong splice sites, in exons surrounded by long introns, and in pseudoexons are strongly suggestive of a functional implication of nucleosome positioning in splicing. Also in support of such an implication is the observation that the median length of internal exons is similar to the length of the DNA sequence wrapping around the nucleosome in at least seven investigated metazoans (Schwartz et al., 2009), consistent with nucleosomes playing a role in exon definition. It looks, indeed, more plausible that nucleosome length influences exon length rather than the other way around, if only, because exons constitute only a small fraction of the sequence of the genomes, and genomes without exons or with very few exons nevertheless have nucleosomes. Interestingly, the length of human exons with weak splice sites, in which nucleosomes are positioned more stably, is even closer to the nucleosome length and less variable (Tilgner et al., 2009). A mechanistic connection between nucleosome positioning

and splicing would also provide a complementary explanation for the elevated GC content in exons—usually attributed to protein coding bias. Indeed, we speculate that such elevated GC content may partially result from the need of exons to accommodate GC rich nucleosome sequences contributing to the proper recognition of the exons' splice sites (Tilgner et al., 2009). In support of this hypothesis non-coding exons are also relatively enriched in nucleosomes (Schwartz et al., 2009; Tilgner et al., 2009). Interestingly, they also exhibit elevated relative GC content, which in this case cannot be explained by protein coding bias.

Histone modifications on exons

Particularly intriguing is the role of histone modifications in this mechanism. Indeed, a variety of histone modifications have been reported to exhibit specific patterns on exons (see Table 1). Some of these patterns could simply reflect increased nucleosome occupancy within exons, but others could constitute additional determinants contributing to proper splicing of exons. In this regard, results on H3K36me3 are controversial. Thus, Kolasinska-Zwierz et al. (2009) found enrichment of H3K36me3 to be significantly stronger than nucleosome occupancy in expressed genes in worm. Similarly, Spies and co-workers (Spies et al., 2009), after analysis of data by Barski et al. (2007) found exonic H3K36me3 (and H3K27me2) enrichment to be significantly stronger than nucleosome enrichment in human CD4+ T-cells. In contrast, two other reports analyzing the same human data, found that the nucleosome peak and the H3K36me3 peak are strikingly similar (Schwartz et al., 2009; Tilgner et al., 2009), although an extra stepwise H3K36me3

increase at the acceptor could not be explained by nucleosome occupancy alone (Tilgner et al., 2009; Nahkuri et al., 2009).

Histone	aa	Modification	Non uniform around exons	Different from nucleosome around exons
H3	K4	me1	Nakuri et al.,[32] Spies et al.[33]	Spies et al.[33]
		me2	Spies et al.[33]	Spies et al.[33]
		me3	Spies et al.[33]	
	K9	me1	Nakuri et al.,[32] Spies et al.[33]	Spies et al.[33]
		me2	Spies et al.[33]	Spies et al.[33]
		me3	Hon et al.[34]	Spies et al.[33]
	K27	me1	Andersson et al.,[31] Nakuri et al.,[32] Spies et al.,[33] Hon et al.[34]	Spies et al.[33]
		me2	Andersson et al.,[31] Spies et al.[33]	Spies et al.[33]
		me3	Andersson et al.,[31] Spies et al.[33]	Spies et al.[33]
	K36	me1	Spies et al.[33]	Spies et al.[33]
		me3	Kolasinska-Zwierz et al.,[24] Schwartz et al.,[29] Tilgner et al.,[30] Andersson et al.,[31] Nakuri et al.,[32] Spies et al.,[33] Hon et al.[34]	Kolasinska-Zwierz et al.,[24] Tilgner et al.,[30] Nakuri et al.,[32] Spies et al.[33]
	K79	me1	Schwartz et al.,[29] Andersson et al.,[31] Spies et al.[33]	Spies et al.[33]
		me2	Spies et al.[33]	Spies et al.[33]
		me3	Spies et al.,[33] Hon et al.[34]	Spies et al.[33]
	R2	me1	Spies et al.[33]	Spies et al.[33]
		me2 (as)	Spies et al.[33]	Spies et al.[33]
H4	K20	me1	Schwartz et al.,[29] Spies et al.,[33] Hon et al.[34]	Tilgner et al.,[30] Spies et al.[33]
		me3	Spies et al.[33]	Spies et al.[33]
H2A/H4	R3	H2A + H4R3me2	Spies et al.[33]	Spies et al.[33]
H2B	K5	me1	Schwartz et al.,[29] Andersson et al.,[31] Nakuri et al.,[32] Spies et al.,[33] Hon et al.[34]	

Figure 1.7 Published histone modification characteristics on exons. Histone modifications (columns 1 to 3) that have been reported to be enriched on exons (column 4) and to exhibit a different enrichment on exons than nucleosomes (column 5). Note that not all reports investigated all modifications or investigated difference between histone modification and nucleosome peaks; Hon and co-workers (Hon et al., 2009) above all investigated modification "signatures" (combinations of histone modifications). Aiming for completeness we mention all modifications that are part of an exon related signature.

Taken together, these observations can be interpreted in different and not mutually exclusive ways: Nucleosome enrichment on exons and H3K36me3 (-H3K27me2) marking of exonic nucleosomes could be two separate phenomena, the combination of which leads to the exonic peak obtained from H3K36me3 (-H3K27me2) data sets. However nucleosome positioning on exons could also be a means of enhancing H3K36me3 (-H3K27me2)-marking of exons as compared to introns.

Some modifications, on the other hand, appear to mark exons depending on their position in the transcript. Thus, Hon et al. (2009) find H3K36me3

enrichment on exons towards the 3'end of the gene and enrichment in H2BK5me1 and H4K20me1 towards the 5'end of the gene. These authors also report two chromatin signatures, marking exons with a combination of at least two histone modifications. This is especially exciting, as combinatorics on RNA-motif level have been proposed to influence splicing (Smith and Valcárcel, 2000; Wang and Burge, 2008; Modafferi and Black, 1999; Han et al., 2005). An expansion of this combinatorial logic to multiple histone modifications, could have an impact on splicing specificity, which is to date not sufficiently explained. Chromatin structure and regulated chromatin modifications could thus play a role in alternative splicing regulation. Preliminary data suggests that possibility. Indeed alternative exons have lower peaks of H3K36me3 than surrounding constitutive exons in worm and mouse (Kolasinska-Zwierz et al., 2009), although Spies et al. (2009) could not confirm this finding in humans. Supporting a connection between alternative splicing and histone modifications, human exonic H3K36me3 levels were found to correlate with exon expression and inclusion (Andersson et al., 2009; Hon et al., 2009) as measured by exon arrays in very similar cells (Oberdoerffer et al., 2008). Consistently, constitutive exons were found to exhibit stronger nucleosome peaks than alternative exons (Schwartz et al., 2009).

The relationship between chromatin and splicing uncovered by a number of groups during 2009 was mostly unexpected. The genome wide maps of chromatin structure and chromatin modifications that led to the discovery were not produced with the aim of understanding splicing. The relationship actually escaped the producers of the data, and it took more than a year before it was independently discovered by a number of groups through quite straightforward and, in some cases, serendipitous bioinformatic analyzes. The relevance to our understanding of splicing of the findings triggered by

these chromatin maps augurs that a very active field of research will soon emerge in the intersection of chromatin and splicing. Indeed, in spite of the scarce and premature data, preliminary results already indicate that chromatin information can be used to predict splicing behavior. Thus, inclusion levels of alternative exons, appear to be predictable from their relative nucleosome occupancy (Schwartz et al., 2009; Tilgner et al., 2009). On the other hand using nucleosome positioning as an additional factor in a model for splicing simulation provides modest but significant gains, when compared with a model using splice sites and regulatory sequences only (Spies et al., 2009).

The accumulation of genome wide data on nucleosome positioning and chromatin modifications across multiple cell types and conditions (and species), coupled with data on alternative inclusion of exons across the same conditions - which, thanks to massively parallel sequencing of RNA (RNASeq), can also be easily obtained genome wide - will substantially contribute to the elucidation of the histone modifications and nucleosome remodeling events that may play a role in the splicing of exons specific to a particular cell type or condition. This could lead to the delineation of a combinatorial code of histone modifications for alternative splicing regulation. On the other hand, directed experiments with artificial multi-exonic gene constructs, in which the possibility exists of manipulating positioning of nucleosomes and histone modifications, and to monitor how these manipulations influence splicing of exons, will shed light on the molecular events that mediate the participation of chromatin in splicing. Exciting times lie ahead in the investigation of splicing; the unexpected discovery that chromatin structure may play a role in splicing, may provide a definitive impulse to the understanding of this phenomenon—one of the most puzzling, from the evolutionary standpoint, in the pathway from the DNA to protein se-

quences.

Experimental evidence

Apart from the previously described large scale studies, effects of local chromatin structure changes on alternative splicing have been characterized in more detail on single alternative splicing examples. First, a chromatin remodeler, the SWI/SNF complex was shown to be involved in alternative splicing regulation (Batsché et al., 2006). Second, a role for lysine 4 trimethylation in splicing had been suggested based on the observation that this mark itself and factors that recognize it are important for efficient splicing (Sims et al., 2007). Third, based on mass spectrometry experiments, factors that bind to methylated lysine 9 were shown to include two well known splicing factors, SRp20 and SF2/ASF (Loomis et al., 2009). Furthermore, Schor et al. (2009) showed that neuronal cell depolarization induces chromatin changes in close vicinity of an alternative exon. More exactly lysine 9 acetylation and lysine 36 methylation were increased around an internal alternatively spliced exon of the NCAM gene, whose inclusion was decreased. Similarly Alló et al. (2009) demonstrated that small interfering RNAs (siRNAs) can increase exon inclusion of an alternative exon of the fibronectin gene. This was paralleled by heterochromatization of the chromatin surrounding the alternative exon. The effect was abolished when lysine 9 methylation and histone acetylation was inhibited (Alló et al., 2009). Some of findings can be interpreted as instances of local chromatin organization influencing Pol2-elongation rate, which in turn would influence splicing decisions. This is, however, not the only imaginable way how chromatin structure can influence inclusion levels of alternative exons. The before-

hand mentioned findings by Loomis et al. (2009) already suggested that a more direct interaction is possible. Direct data supporting such an interaction has been presented by Luco et al. (2010), whose findings support the following model: Chromatin organization, in this case mainly levels of lysine 36-trimethylation, interact with the chromatin binding factor MRG-15. MRG-15 then co-transcriptionally recruits the hnRNP protein PTB to the pre-mRNA (Luco et al., 2010). Interestingly, chromatin organization in this respect appears to be most important when PTB binding sites are weak. Therefore, one can view these interactions as an extension of the ESE-ESS logic: Weak splice sites, whose recognition by U2- and U1snRNP is problematic are thought to be especially dependent on the binding of auxiliary splicing factors to ESE or ISE. If this binding is weak itself, chromatin appears to be the next auxiliary layer, at least in the case of PTB. Thus, chromatin structure in all its aspects, transcription dynamics and splicing are related to each other.

1.5 Summary

In summary, one can say that our understanding of the determinants of splicing is still limited. This is exemplified by the fact that no computer program today can summarize all the available knowledge about splicing in order to tell what is an exon and what is not, or which exon will be spliced or not in a given cell type, for example. The determinants of splicing can be divided in three layers:

- Splice sites which are recognized by the spliceosome.
- ESE, ESS, ISE and ISS which can be bound by auxiliary splicing

factors.

- Chromatin organization and transcription dynamics.

Supposedly research in all three areas will further our understanding of splicing. My personal opinion, which might very well be proved wrong, is that the interactions between the last two will contribute most to a global understanding of splicing.

Bibliography

R D Alexander, S A Innocente, J D Barrass, and J D Beggs. Splicing-dependent rna polymerase pausing in yeast. *Mol Cell*, 40(4):582–93, Nov 2010. doi: 10.1016/j.molcel.2010.11.005.

E Allemand, E Batsché, and C Muchardt. Splicing, transcription, and chromatin: a ménage à trois. *Curr Opin Genet Dev*, 18(2):145–51, Apr 2008. doi: 10.1016/j.gde.2008.01.006.

M Alló, V Buggiano, J P Fededa, E Petrillo, I Schor, M de la Mata, E Agirre, M Plass, E Eyras, S A Elela, R Klinck, B Chabot, and A R Kornblihtt. Control of alternative splicing through sirna-mediated transcriptional gene silencing. *Nat Struct Mol Biol*, 16(7):717–24, Jul 2009. doi: 10.1038/nsmb.1620.

B A Amendt, Z H Si, and C M Stoltzfus. Presence of exon splicing silencers within human immunodeficiency virus type 1 tat exon 2 and tat-rev exon 3: evidence for inhibition mediated by cellular factors. *Mol Cell Biol*, 15(11):6480, Nov 1995.

R Andersson, S Enroth, A Rada-Iglesias, C Wadelius, and J Komorowski. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res*, 19(10):1732–41, Oct 2009. doi: 10.1101/gr.092353.109.

Y Barash, J A Calarco, W Gao, Q Pan, X Wang, O Shai, B J Blencowe, and B J Frey. Deciphering the splicing code. *Nature*, 465(7294):53–9, May 2010. doi: 10.1038/nature09000.

A Barski, S Cuddapah, K Cui, T Y Roh, D E Schones, Z Wang, G Wei, I Chepelev, and K Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, May 2007. doi: 10.1016/j.cell.2007.05.009.

E Batsché, M Yaniv, and C Muchardt. The human swi/snf subunit brm is a regulator of alternative splicing. *Nat Struct Mol Biol*, 13(1):22–9, Jan 2006. doi: 10.1038/nsmb1030.

J S Beckmann and E N Trifonov. Splice junctions follow a 205-base ladder. *Proc Natl Acad Sci U S A*, 88(6):2380–3, Mar 1991.

S M Berget. Exon recognition in vertebrate splicing. *J Biol Chem*, 270(6):2411–4, Feb 1995.

S M Berget, C Moore, and P A Sharp. Spliced segments at the 5' terminus of adenovirus 2 late mrna. *Proc Natl Acad Sci U S A*, 74(8):3171–5, Aug 1977.

A J Berk and P A Sharp. Spliced early mrnas of simian virus 40. *Proc Natl Acad Sci U S A*, 75(3):1274–8, Mar 1978.

A L Beyer and Y N Osheim. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev*, 2(6):754–65, Jun 1988.

D L Black, B Chabot, and J A Steitz. U2 as well as u1 small nuclear ribonucleoproteins are involved in premessenger rna splicing. *Cell*, 42(3):737–50, Oct 1985.

P L Boutz, P Stoilov, Q Li, C H Lin, G Chawla, K Ostrow, L Shiue, M Ares, Jr, and D L Black. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev*, 21(13):1636–52, Jul 2007. doi: 10.1101/gad.1558107.

R Breathnach, J L Mandel, and P Chambon. Ovalbumin gene is split in chicken dna. *Nature*, 270(5635):314–9, Nov 1977.

R Breathnach, C Benoist, K O'Hare, F Gannon, and P Chambon. Ovalbumin gene: evidence for a leader sequence in mrna and dna sequences at the exon-intron boundaries. *Proc Natl Acad Sci U S A*, 75(10):4853–7, Oct 1978.

J F Cáceres, S Stamm, D M Helfman, and A R Krainer. Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. *Science*, 265 (5179):1706–9, Sep 1994.

L Cartegni, S L Chew, and A R Krainer. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*, 3(4):285–98, Apr 2002. doi: 10.1038/nrg775.

L Cartegni, J Wang, Z Zhu, M Q Zhang, and A R Krainer. Esefinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res*, 31(13):3568–71, Jul 2003.

R Castelo and R Guigó. Splice site identification by idlbns. *Bioinformatics*, 20 Suppl 1:i69–76, Aug 2004. doi: 10.1093/bioinformatics/bth932.

J F Catterall, B W O'Malley, M A Robertson, R Staden, Y Tanaka, and G G Brownlee. Nucleotide sequence homology at 12 intron–exon junctions in the chick ovalbumin gene. *Nature*, 275(5680):510–3, Oct 1978.

J Cheng, T Zhou, C Liu, J P Shapiro, M J Brauer, M C Kiefer, P J Barr, and J D Mountz. Protection from fas-mediated apoptosis by a soluble form of the fas molecule. *Science*, 263(5154):1759–62, Mar 1994.

L T Chow, R E Gelinas, T R Broker, and R J Roberts. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger rna. *Cell*, 12(1):1–8, Sep 1977.

T A Cooper, L Wan, and G Dreyfuss. Rna and disease. *Cell*, 136(4):777–93, Feb 2009. doi: 10.1016/j.cell.2009.02.011.

J L Corden, D L Cadena, J M Ahearn, Jr, and M E Dahmus. A unique structure at the carboxyl terminus of the largest subunit of eukaryotic rna polymerase ii. *Proc Natl Acad Sci U S A*, 82(23):7934–8, Dec 1985.

P Cramer, C G Pesce, F E Baralle, and A R Kornblihtt. Functional association between promoter structure and transcript alternative splicing. *Proc Natl Acad Sci U S A*, 94(21):11456–60, Oct 1997.

C J David and J L Manley. Alternative pre-mrna splicing regulation in cancer: pathways and programs unhinged. *Genes Dev*, 24(21):2343–64, Nov 2010. doi: 10.1101/gad.1973010.

S F de Almeida and M Carmo-Fonseca. The ctd role in cotranscriptional rna processing and surveillance. *FEBS Lett*, 582(14):1971–6, Jun 2008. doi: 10.1016/j.febslet.2008.04.019.

M de la Mata, C R Alonso, S Kadener, J P Fededa, M Blaustein, F Pelisch, P Cramer, D Bentley, and A R Kornblihtt. A slow rna polymerase ii affects alternative splicing in vivo. *Mol Cell*, 12(2):525–32, Aug 2003.

F Del Gatto-Konczak, M Olive, M C Gesnel, and R Breathnach. hnrnp a1 recruited to an exon in vivo can function as an exon splicing silencer. *Mol Cell Biol*, 19 (1):251–60, Jan 1999.

W G Fairbrother, R F Yeh, P A Sharp, and C B Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–13, Aug 2002. doi: 10.1126/science.1073774.

K L Fox-Walsh, Y Dou, B J Lam, S P Hung, P F Baldi, and K J Hertel. The architecture of pre-mrnas affects mechanisms of splice-site pairing. *Proc Natl Acad Sci U S A*, 102(45):16176–81, Nov 2005. doi: 10.1073/pnas.0508489102.

A Goren, O Ram, M Amit, H Keren, G Lev-Maor, I Vig, T Pupko, and G Ast. Comparative analysis identifies exonic splicing regulatory sequences–the complex definition of enhancers and silencers. *Mol Cell*, 22(6):769–81, Jun 2006. doi: 10.1016/j.molcel.2006.05.008.

B R Graveley, K J Hertel, and T Maniatis. A systematic analysis of the factors that determine the strength of pre-mrna splicing enhancers. *EMBO J*, 17(22): 6747–56, Nov 1998. doi: 10.1093/emboj/17.22.6747.

S S Hammoud, D A Nix, H Zhang, J Purwar, D T Carrell, and B R Cairns. Distinctive chromatin in human sperm packages genes for embryo development. *Nature*, 460(7254):473–8, Jul 2009. doi: 10.1038/nature08162.

K Han, G Yeo, P An, C B Burge, and P J Grabowski. A combinatorial code for splicing silencing: Uagg and gggg motifs. *PLoS Biol*, 3(5):e158, May 2005. doi: 10.1371/journal.pbio.0030158.

J Harrow, F Denoeud, A Frankish, A Reymond, C K Chen, J Chrast, J Lagarde, J G Gilbert, R Storey, D Swarbreck, C Rossier, C Ucla, T Hubbard, S E Antonarakis, and R Guigo. Gencode: producing a reference annotation for encode. *Genome Biol*, 7 Suppl 1:S4.1–9, 2006. doi: 10.1186/gb-2006-7-s1-s4.

J D Hawkins. A survey on intron and exon lengths. *Nucleic Acids Res*, 16(21): 9893–908, Nov 1988.

M L Hedley and T Maniatis. Sex-specific splicing and polyadenylation of dsx pre-mrna requires a sequence that binds specifically to tra-2 protein in vitro. *Cell*, 65(4):579–86, May 1991.

G Hon, W Wang, and B Ren. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol*, 5(11):e1000566, Nov 2009. doi: 10.1371/journal.pcbi.1000566.

K J Howe, C M Kane, and M Ares, Jr. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in saccharomyces cerevisiae. *RNA*, 9(8):993–1006, Aug 2003.

N Itoh, S Yonehara, A Ishii, M Yonehara, S Mizushima, M Sameshima, A Hase, Y Seto, and S Nagata. The polypeptide encoded by the cdna for human cell surface antigen fas can mediate apoptosis. *Cell*, 66(2):233–43, Jul 1991.

J M Izquierdo, N Majós, S Bonnal, C Martínez, R Castelo, R Guigó, D Bilbao, and J Valcárcel. Regulation of fas alternative splicing by antagonistic effects of tia-1 and ptb on exon definition. *Mol Cell*, 19(4):475–84, Aug 2005. doi: 10.1016/j.molcel.2005.06.015.

J M Johnson, J Castle, P Garrett-Engele, Z Kan, P M Loerch, C D Armour, R Santos, E E Schadt, R Stoughton, and D D Shoemaker. Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays. *Science*, 302(5653):2141–4, Dec 2003. doi: 10.1126/science.1090100.

S Kadener, P Cramer, G Nogués, D Cazalla, M de la Mata, J P Fededa, S E Werbajh, A Srebrow, and A R Kornblihtt. Antagonistic effects of t-ag and vp16 reveal a role for rna pol ii elongation on alternative splicing. *EMBO J*, 20(20):5759–68, Oct 2001. doi: 10.1093/emboj/20.20.5759.

P Kolasinska-Zwierz, T Down, I Latorre, T Liu, X S Liu, and J Ahringer. Differential chromatin marking of introns and expressed exons by h3k36me3. *Nat Genet*, 41(3):376–81, Mar 2009. doi: 10.1038/ng.322.

P Komarnitsky, E J Cho, and S Buratowski. Different phosphorylated forms of rna polymerase ii and associated mrna processing factors during transcription. *Genes Dev*, 14(19):2452–60, Oct 2000.

R D Kornberg. Eukaryotic transcriptional control. *Trends Cell Biol*, 9(12):M46–9, Dec 1999.

A R Kornblihtt, M de la Mata, J P Fededa, M J Munoz, and G Nogues. Multiple links between transcription and splicing. *RNA*, 10(10):1489–98, Oct 2004. doi: 10.1261/rna.7100104.

M Krawczak, J Reiss, and D N Cooper. The mutational spectrum of single base-pair substitutions in mrna splice junctions of human genes: causes and consequences. *Hum Genet*, 90(1-2):41–54, 1992.

L F Lareau, A N Brooks, D A Soergel, Q Meng, and S E Brenner. The coupling

of alternative splicing and nonsense-mediated mrna decay. *Adv Exp Med Biol*, 623:190–211, 2007.

A Lavigueur, H La Branche, A R Kornblihtt, and B Chabot. A splicing enhancer in the human fibronectin alternate ed1 exon interacts with sr proteins and stimulates u2 snrnp binding. *Genes Dev*, 7(12A):2405–17, Dec 1993.

C K Lee, Y Shibata, B Rao, B D Strahl, and J D Lieb. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet*, 36(8):900–5, Aug 2004. doi: 10.1038/ng1400.

M R Lerner, J A Boyle, S M Mount, S L Wolin, and J A Steitz. Are snrnps involved in splicing? *Nature*, 283(5743):220–4, Jan 1980.

A Letai. Restoring cancer's death sentence. *Cancer Cell*, 10(5):343–5, Nov 2006. doi: 10.1016/j.ccr.2006.10.014.

H X Liu, M Zhang, and A R Krainer. Identification of functional exonic splicing enhancer motifs recognized by individual sr proteins. *Genes Dev*, 12(13):1998–2012, Jul 1998.

M Llorian, S Schwartz, T A Clark, D Hollander, L Y Tan, R Spellman, A Gordon, A C Schweitzer, P de la Grange, G Ast, and C W Smith. Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by ptb. *Nat Struct Mol Biol*, 17(9):1114–23, Sep 2010. doi: 10.1038/nsmb.1881.

R J Loomis, Y Naoe, J B Parker, V Savic, M R Bozovsky, T Macfarlan, J L Manley, and D Chakravarti. Chromatin binding of srp20 and asf/sf2 and dissociation from mitotic chromosomes is modulated by histone h3 serine 10 phosphorylation. *Mol Cell*, 33(4):450–61, Feb 2009. doi: 10.1016/j.molcel.2009.02.003.

N López-Bigas, B Audit, C Ouzounis, G Parra, and R Guigó. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett*, 579(9):1900–3, Mar 2005. doi: 10.1016/j.febslet.2005.02.047.

R F Luco, Q Pan, K Tominaga, B J Blencowe, O M Pereira-Smith, and T Misteli. Regulation of alternative splicing by histone modifications. *Science*, 327(5968): 996–1000, Feb 2010. doi: 10.1126/science.1184208.

A Magen and G Ast. The importance of being divisible by three in alternative splicing. *Nucleic Acids Res*, 33(17):5574–82, 2005. doi: 10.1093/nar/gki858.

H J Mardon, G Sebastio, and F E Baralle. A role for exon sequences in alternative splicing of the human fibronectin gene. *Nucleic Acids Res*, 15(19):7725–33, Oct 1987.

V Markovtsov, J M Nikolic, J A Goldman, C W Turck, M Y Chou, and D L Black. Cooperative assembly of an hnrnp complex induced by a tissue-specific homolog of polypyrimidine tract binding protein. *Mol Cell Biol*, 20(20):7463–79, Oct 2000.

A Mayeda and A R Krainer. Regulation of alternative pre-mrna splicing by hnrnp a1 and splicing factor sf2. *Cell*, 68(2):365–75, Jan 1992.

S McCracken, N Fong, K Yankulov, S Ballantyne, G Pan, J Greenblatt, S D Patterson, M Wickens, and D L Bentley. The c-terminal domain of rna polymerase ii couples mrna processing to transcription. *Nature*, 385(6614):357–61, Jan 1997. doi: 10.1038/385357a0.

L Merendino, S Guth, D Bilbao, C Martínez, and J Valcárcel. Inhibition of msl-2 splicing by sex-lethal reveals interaction between u2af35 and the 3' splice site ag. *Nature*, 402(6763):838–41, Dec 1999. doi: 10.1038/45602.

E F Modafferi and D L Black. Combinatorial control of a neuron-specific exon. *RNA*, 5(5):687–706, May 1999.

B Modrek, A Resch, C Grasso, and C Lee. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res*, 29(13): 2850–9, Jul 2001.

M J Moore and N J Proudfoot. Pre-mrna processing reaches back to transcription and ahead to translation. *Cell*, 136(4):688–700, Feb 2009. doi: 10.1016/j.cell.2009.02.001.

S M Mount. A catalogue of splice junction sequences. *Nucleic Acids Res*, 10(2):459–72, Jan 1982.

S Nahkuri, R J Taft, and J S Mattick. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. *Cell Cycle*, 8(20):3420–4, Oct 2009.

K M Neugebauer. On the importance of being co-transcriptional. *J Cell Sci*, 115 (Pt 20):3865–71, Oct 2002.

T W Nilsen and B R Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–63, Jan 2010. doi: 10.1038/nature08909.

G Nogues, S Kadener, P Cramer, D Bentley, and A R Kornblihtt. Transcriptional activators differ in their abilities to control alternative splicing. *J Biol Chem*, 277 (45):43110–4, Nov 2002. doi: 10.1074/jbc.M208418200.

S Oberdoerffer, L F Moita, D Neems, R P Freitas, N Hacohen, and A Rao. Regulation of cd45 alternative splicing by heterogeneous ribonucleoprotein, hnrnpll. *Science*, 321(5889):686–91, Aug 2008. doi: 10.1126/science.1157610.

F Pagani, M Raponi, and F E Baralle. Synonymous mutations in cftr exon 12 affect splicing and are not neutral in evolution. *Proc Natl Acad Sci U S A*, 102(18):6368–72, May 2005. doi: 10.1073/pnas.0502288102.

Q Pan, O Shai, L J Lee, B J Frey, and B J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–5, Dec 2008. doi: 10.1038/ng.259.

G. Parra, E. Blanco, and R. Guigó. Geneid in drosophila. *Genome Research*, 10:511–515, 2000.

A D Polydorides, H J Okano, Y Y Yang, G Stefani, and R B Darnell. A brain-enriched polypyrimidine tract-binding protein antagonizes the ability of nova to

regulate neuron-specific alternative splicing. *Proc Natl Acad Sci U S A*, 97(12): 6350–5, Jun 2000. doi: 10.1073/pnas.110128397.

R Reed and T Maniatis. A role for exon sequences and splice-site proximity in splice-site selection. *Cell*, 46(5):681–90, Aug 1986.

B L Robberson, G J Cote, and S M Berget. Exon definition may facilitate splice site selection in rnas with multiple exons. *Mol Cell Biol*, 10(1):84–94, Jan 1990.

G C Roberts, C Gooding, H Y Mak, N J Proudfoot, and C W Smith. Co-transcriptional commitment to alternative splice site selection. *Nucleic Acids Res*, 26(24):5568–72, Dec 1998.

J Rogers and R Wall. A mechanism for rna splicing. *Proc Natl Acad Sci U S A*, 77 (4):1877–9, Apr 1980.

B Ruskin, A R Krainer, T Maniatis, and M R Green. Excision of an intact intron as a novel lariat structure during pre-mrna splicing in vitro. *Cell*, 38(1):317–31, Aug 1984.

B Ruskin, P D Zamore, and M R Green. A factor, u2af, is required for u2 snrnp binding and splicing complex assembly. *Cell*, 52(2):207–19, Jan 1988.

M Sammeth, S Foissac, and R Guigó. A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol*, 4(8):e1000147, 2008. doi: 10.1371/journal.pcbi.1000147.

S Sasaki, C C Mello, A Shimada, Y Nakatani, S Hashimoto, M Ogawa, K Matsushima, S G Gu, M Kasahara, B Ahsan, A Sasaki, T Saito, Y Suzuki, S Sugano, Y Kohara, H Takeda, A Fire, and S Morishita. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science*, 323 (5912):401–4, Jan 2009. doi: 10.1126/science.1163183.

D Schmucker, J C Clemens, H Shu, C A Worby, J Xiao, M Muda, J E Dixon, and S L Zipursky. Drosophila dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6):671–84, Jun 2000.

M Schneider, C L Will, M Anokhina, J Tazi, H Urlaub, and R Lührmann. Exon definition complexes contain the tri-snrnp and can be directly converted into b-like precatalytic splicing complexes. *Mol Cell*, 38(2):223–35, Apr 2010. doi: 10.1016/j.molcel.2010.02.027.

D E Schones, K Cui, S Cuddapah, T Y Roh, A Barski, Z Wang, G Wei, and K Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132 (5):887–98, Mar 2008. doi: 10.1016/j.cell.2008.02.022.

I E Schor, N Rascovan, F Pelisch, M Alló, and A R Kornblihtt. Neuronal cell depolarization induces intragenic chromatin modifications affecting ncam alternative splicing. *Proc Natl Acad Sci U S A*, 106(11):4325–30, Mar 2009. doi: 10.1073/pnas.0810666106.

S Schwartz, E Meshorer, and G Ast. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, 16(9):990–5, Sep 2009. doi: 10.1038/nsmb.1659.

I Seif, G Khoury, and R Dhar. Bkv splice sequences based on analysis of preferred donor and acceptor sites. *Nucleic Acids Res*, 6(10):3387–98, Jul 1979.

R J Sims, 3rd, S Millhouse, C F Chen, B A Lewis, H Erdjument-Bromage, P Tempst, J L Manley, and D Reinberg. Recognition of trimethylated histone h3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mrna splicing. *Mol Cell*, 28(4):665–76, Nov 2007. doi: 10.1016/j.molcel.2007.11.010.

M Sironi, G Menozzi, L Riva, R Cagliani, G P Comi, N Bresolin, R Giorda, and U Pozzoli. Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Res*, 32(5):1783–91, 2004. doi: 10.1093/nar/gkh341.

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

N Spies, C B Nielsen, R A Padgett, and C B Burge. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell*, 36(2):245–54, Oct 2009. doi: 10.1016/j.molcel.2009.10.008.

M B Stadler, N Shomron, G W Yeo, A Schneider, X Xiao, and C B Burge. Inference of splicing regulatory activities by sequence neighborhood analysis. *PLoS Genet*, 2(11):e191, Nov 2006. doi: 10.1371/journal.pgen.0020191.

D A Sterner, T Carlo, and S M Berget. Architectural limits on split genes. *Proc Natl Acad Sci U S A*, 93(26):15081–5, Dec 1996.

H Sun and L A Chasin. Multiple splicing defects in an intronic false exon. *Mol Cell Biol*, 20(17):6414–25, Sep 2000.

S Sun, Z Zhang, R Sinha, R Karni, and A R Krainer. Sf2/asf autoregulation involves multiple layers of post-transcriptional and translational control. *Nat Struct Mol Biol*, 17(3):306–12, Mar 2010. doi: 10.1038/nsmb.1750.

A Sureau, R Gattoni, Y Dooghe, J Stévenin, and J Soret. Sc35 autoregulates its expression by promoting splicing events that destabilize its mrnas. *EMBO J*, 20 (7):1785–96, Apr 2001. doi: 10.1093/emboj/20.7.1785.

H Tilgner and R Guigó. From chromatin to splicing: Rna-processing as a total artwork. *Epigenetics*, 5(3), Apr 2010.

H Tilgner, C Nikolaou, S Althammer, M Sammeth, M Beato, J Valcárcel, and R Guigó. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol*, 16(9):996–1001, Sep 2009. doi: 10.1038/nsmb.1658.

J Ule, G Stefani, A Mele, M Ruggiu, X Wang, B Taneri, T Gaasterland, B J Blencowe, and R B Darnell. An rna map predicting nova-dependent splicing regulation. *Nature*, 444(7119):580–6, Nov 2006. doi: 10.1038/nature05304.

A Valouev, J Ichikawa, T Tonthat, J Stuart, S Ranade, H Peckham, K Zeng, J A Malek, G Costa, K McKernan, A Sidow, A Fire, and S M Johnson. A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res*, 18(7):1051–63, Jul 2008. doi: 10.1101/gr.076463.108.

J P Venables, R Klinck, C Koh, J Gervais-Bird, A Bramard, L Inkel, M Durand, S Couture, U Froehlich, E Lapointe, J F Lucier, P Thibault, C Rancourt,

K Tremblay, P Prinos, B Chabot, and S A Elela. Cancer-associated regulation of alternative splicing. *Nat Struct Mol Biol*, 16(6):670–6, Jun 2009. doi: 10.1038/nsmb.1608.

R B Voelker and J A Berglund. A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. *Genome Res*, 17(7):1023–33, Jul 2007. doi: 10.1101/gr.6017807.

E T Wang, R Sandberg, S Luo, I Khrebtukova, L Zhang, C Mayr, S F Kingsmore, G P Schroth, and C B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6, Nov 2008. doi: 10.1038/nature07509.

Z Wang and C B Burge. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–13, May 2008. doi: 10.1261/rna.876308.

Z Wang, M E Rolish, G Yeo, V Tung, M Mawson, and C B Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–45, Dec 2004. doi: 10.1016/j.cell.2004.11.010.

S Wu, C M Romfo, T W Nilsen, and M R Green. Functional recognition of the 3' splice site ag by the splicing factor u2af35. *Nature*, 402(6763):832–5, Dec 1999. doi: 10.1038/45590.

Y Xue, Y Zhou, T Wu, T Zhu, X Ji, Y S Kwon, C Zhang, G Yeo, D L Black, H Sun, X D Fu, and Y Zhang. 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(6):996–1006, Dec 2009. doi: 10.1016/j.molcel.2009.12.003.

G Yeo and C B Burge. Maximum entropy modeling of short sequence motifs with applications to rna splicing signals. *J Comput Biol*, 11(2-3):377–94, 2004. doi: 10.1089/1066527041410418.

G W Yeo, E L Van Nostrand, E L Nostrand, and T Y Liang. Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet*, 3(5):e85, May 2007. doi: 10.1371/journal.pgen.0030085.

G W Yeo, N G Coufal, T Y Liang, G E Peng, X D Fu, and F H Gage. An rna code for the fox2 splicing regulator revealed by mapping rna-protein interactions in stem cells. *Nat Struct Mol Biol*, 16(2):130–7, Feb 2009. doi: 10.1038/nsmb.1545.

S M Yoh, H Cho, L Pickle, R M Evans, and K A Jones. The spt6 sh2 domain binds ser2-p rnapii to direct iws1-dependent mrna splicing and export. *Genes Dev*, 21 (2):160–74, Jan 2007. doi: 10.1101/gad.1503107.

P D Zamore and M R Green. Identification, purification, and biochemical characterization of u2 small nuclear ribonucleoprotein auxiliary factor. *Proc Natl Acad Sci U S A*, 86(23):9243–7, Dec 1989.

C Zhang, W H Li, A R Krainer, and M Q Zhang. Rna landscape of evolution for optimal exon and intron discrimination. *Proc Natl Acad Sci U S A*, 105(15): 5797–802, Apr 2008. doi: 10.1073/pnas.0801692105.

X H Zhang and L A Chasin. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*, 18(11):1241–50, Jun 2004. doi: 10.1101/gad.1195304.

C L Zheng, X D Fu, and M Gribskov. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA*, 11(12):1777–87, Dec 2005. doi: 10.1261/rna.2660805.

J Zhu, A Mayeda, and A R Krainer. Exon identity established through differential antagonism between exonic splicing silencer-bound hnrnp a1 and enhancer-bound sr proteins. *Mol Cell*, 8(6):1351–61, Dec 2001.

D A Zorio and T Blumenthal. Both subunits of u2af recognize the 3' splice site in caenorhabditis elegans. *Nature*, 402(6763):835–8, Dec 1999. doi: 10.1038/45597.

PART II

Results

Chapter 2

RNA Words and Splicing Simulation

Summary

This chapter describes a small set of ESE and ESS words ("anchors") that guarantee the presence of larger published words. With the exception of one word (GAA), the minimal length of such words is 4. A subset of these words perform equally well than larger published sets of ESE and ESS when used for splicing simulation. Further gains in splicing simulation accuracy by using larger sets of anchors remain however small. For roughly half of these anchors evolutionary conservation of wobble-positions can be shown to concentrate on these words itself.

2.1 Introduction

Combining some commonly used sets of ESE and ESS hexamers (Cartegni et al., 2003; Goren et al., 2006; Stadler et al., 2006; Zhang et al., 2008)[1], a total of 2468 hexamers have been published as having ESE or ESS activity, a number that corresponds to 60.3% of all possible hexamers. Therefore large parts of the human genome are covered by such sequences. It is worthwhile noting that although hexamers are the most commonly used unit when ESE and ESS are defined, larger and smaller sequences have also been published. Due to these numbers of published ESE and ESS, we are faced frequently with situations of the following type: The hexamers GAAGAa, GAAGAc, GAAGAg, GAAGAt are published ESE hexamers. One is therefore tempted to say that the most important part of these hexamers is the pentamer GAAGA (although the 6th nucleotide might have a role in prioritizing the hexamers with respect to each other). These observations motivate the following two questions. First, down to what minimal size can an oligomer sensibly be called ESE or ESS ? Second, is it possible to define a "smaller" set of "high biological importance" ? An approach to answer both questions is based on the following idea: Can we define shorter words that guarantee the presence of at least one published (usually larger) word in its vicinity ? Here we present such an approach and call such sequences "'ESE-anchors" (or "ESS-anchors"), showing that with one exception the minimal size for such anchor oligomers is four nucleotides.

The term splicing simulation was first used by Wang and co-workers (Wang

[1] A reader with good knowledge about splicing might miss two very popular sets (Wang et al., 2004; Fairbrother et al., 2002) in this listing. These two collections are not listed separately, because they are part of another listed publication (Stadler et al., 2006). Adding them would therefore not change the total number of published hexamers, mentioned above.

et al., 2004). Its goal can be loosely defined as "given the pre-mRNA sequence, predict the position of all exons using only the information sources that the spliceosome uses". Splicing simulation is important for two main reasons. First, it allows to quantify how much about splicing we really understand, as a perfect understanding of splicing should allow perfect simulation of splicing. Second it can serve to prioritize ESE (or ESS) oligomers, since the broader and stronger the effect of an ESE, the more should it contribute to splicing simulation accuracy. Conversely if an ESE does not contribute to splicing simulation accuracy, it either is functional in very few cases only or its rules of functionality are very different from the ones implemented in the splicing simulator. Here we iteratively rank ESE- and ESS-anchors according to their contribution to splicing simulation accuracy based on an earlier proposed model (Wang et al., 2004) and show that a relatively small subset of "ESE/S-anchors" performs as well for splicing simulation as larger published sets. We show that one can achieve small gains (approx. 3pp) without overfitting by adding further anchor sequences. Furthermore we show that for some of the ESE tetramer anchors that do contribute to splicing simulation, evolutionary conservation concentrates on them.

2.2 Results

With one exception tetramers appear to be the smallest unit for which the words "ESE" and "ESS" make sense. In order to define minimally-sized ESE words of "high biological importance" we defined trimers, tetramers and pentamers that guarantee the presence of a published ESE

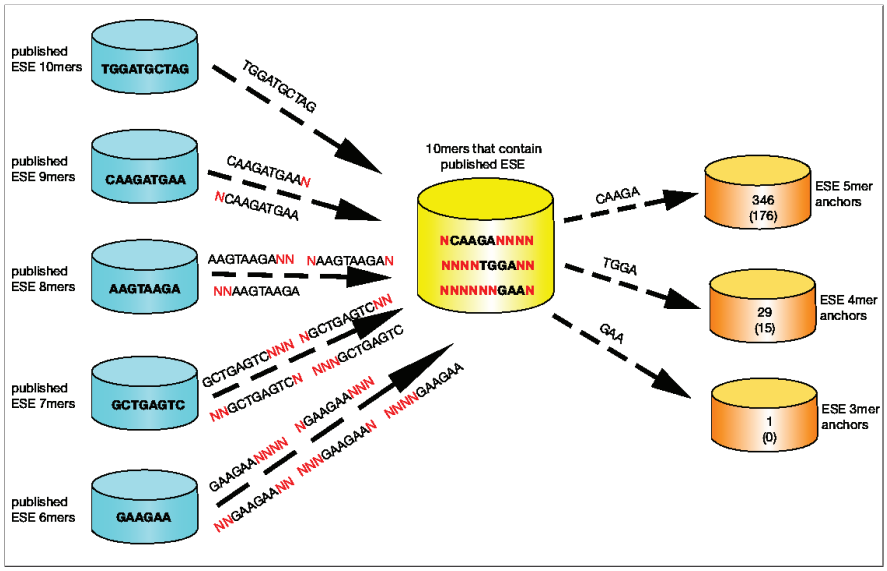


Figure 2.1 Illustration of the definition of ESE anchors of length 3,4 and 5. NCAAGANNNN designates all 10mers that have CAAGA at position 2. We find 1 ESE-trimer- (GAA), 29 ESE-tetramer- and 346 ESE-pentamer-anchors. ESE-hexamer-anchors could be defined, but will include by definition all published ESE hexamers. For ESS we found no ESS-trimer, 16 ESS tetramer and 176 ESS-pentamer-anchors.

in their vicinity. More exactly "ESE-anchor" words w are defined by a fixed position i so that all 10mers having w at position i contained a published ESE (Methods, figure 2.1). The approach is identical for ESS with published ESS as input of course. We found that with the exception of 1 word (GAA) these ESE and ESS-anchors had to be of at least 4nts. For the 29 ESE-tetramer-anchors we found none were identically in the input ESE set, as the minimal size of input ESE words was six nucleotides; for the 15 ESS-tetramer anchors however, 8 had been published as such making their presence in the anchor set trivial.

Tetramer ESE and ESS anchors seem close to optimality for splicing simulation among all 4mers. We implemented an iterative

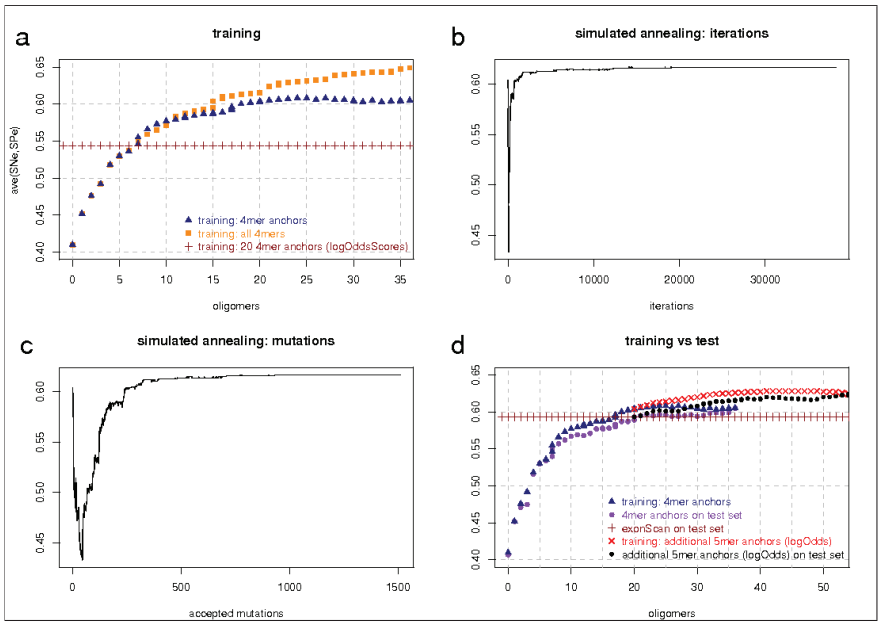


Figure 2.2 iterative exploration of scores. Training with 4mer anchors and training with all 256 4mers. Comparison with logOdds-scores for 19 anchors and downstream GGGs (**a**). Accuracy changes during simulated annealing as a function of iterations (**b**) and as a function of accepted score mutations (**c**). Training and performance on the test set with iteratively explored scores. Comparison with ExonScan performance on test set. Further training and performance evaluation on the test set with 5mer anchors (**d**).

approach (see Methods) to define which anchor sequences and GGGs and which scores would contribute most to raising splicing simulation accuracy. Similarly to Wang et al. (2004), we chose the parameter EW so that SNe and SPe were almost identical, yielding an accuracy of 41% when only splice sites were used on the training set. The first three oligomers (TTTT as an ESS, GGG as a downstream element and AAGA as an acceptor ESE) raised accuracy by 4.5, 2.4 and 1.6pp respectively. When GGG and 19 different 4mer anchors were used, simulation accuracy had passed to 60.4% (SNe =60.7%, SPe=60.1%) on the training set. Accuracy saturated quickly reaching the absolute maximum at 60.8% using GGGs and 23

4mer anchors (figure 2.2a, blue triangles). To address the question of optimality of the chosen 4mer anchors ("anchor parameter set") given this iterative training procedure, we performed the same approach using GGGs and all 256 possible 4mers ("all 4mer parameter set", figure 2.2a, orange triangles). The first 5 oligomers and scores chosen under this scheme (TTTT,GGG,AAGA,TAGG and TGGA) were identical to those in the "anchor parameter set". Only the sixth 4mer (CGGA, a non anchor sequence used as an acceptor ESE) outperformed the chosen anchor (GAAG used as a donor ESE) by 0.1pp (net accuracy gain 0.7pp instead of 0.6pp). Both trained parameter sets achieved similar accuracy (with slight advantages for the "all 4mer parameter set") until 20 oligomers were used. At this point the "all 4mer parameter set" could find 4mers that did further increase accuracy while the "anchor parameter set" could not (figure 2.2a). This approach relied on the same training procedure; therefore it did not prove whether a different training procedure could find scores that yield better performance, escaping local maxima which the iterative approach might be caught in. In order to answer this question rigorously, one would have to simulate splicing using $50^{26} > 10^{44}$ different parameter sets (assuming 50 different scores for each parameter), a number of simulations for which we currently do not have the infrastructure. We attempted to circumvent this problem in two ways. First, fixing the 20 oligomers (corresponding to 26 parameters), we divised a simulated annealing algorithm that mutated the scores of the 26 parameters (Methods). After 38256 iterations during which 1511 score mutations were accepted accuracy had been raised from 60.4% to 61.7%, however no increases in simulation accuracy had been achieved on the last 24135 iterations (or 739 accepted score mutations, figure 2.2b,c). Furthermore one of the obtained mutated parameter sets (after 5299 iterations, accuracy gain of 1pp on the training set) led to an

accuracy loss of 0.4% on the test set. We conclude that it is unlikely that different scores for the same tetramers and of the same granularity would lead to accuracy increases that translate to accuracy increases on the test set.

Explored scores give advantages over pure logOdds-scores. In order to assess whether these scores obtained by exploration during the iterative training presented advantages over pure logOdds-scores, we calculated logOdds-scores for the 20 first oligomers (represented by 26 parameters) in the "anchor parameter set" (Methods). Simulation using these 26 logOdds-scores on the training set gave an accuracy of 54.4% while the "anchor parameter set" with explored scores has achieved 60.4% (figure 2.2a).

Explored scores show some but little overfitting. Subsequently we simulated exons on the test set. First using only splice sites, then using splice sites and the first trained oligomer and score in the "anchor parameter set". Iteratively we repeated this with all increasing subsets of the "anchor parameter set". Although some overfitting (better performance on the training set than on the test set) could be observed (2.2d, blue and purple dots), it generally remained small and decreased as more parameters were added to the model.

20 oligomers perform as well as ExonScan. In order to compare the performance of the "anchor parameter set" to ExonScan (Wang et al., 2004), we submitted the entire test set to the ExonScan-webserver (Wang et al., 2004) and evaluated its accuracy (0.593) when using all its ESE, ESS and GGGs. This performance equaled the performance of the "anchor parame-

ter set" using 20 oligomers (or 26 parameters, figure 2.2d). The "anchor parameter set" however mostly (with the exception of ESS) used less sequence information in terms of the percentage of the transcript being covered (see table 2.1). In order to raise accuracy further we relaunched the training procedure using a 5mer anchor subset (see Methods) on top of GGGs and the 19 4mer anchors. This approach showed that is is possible to gain an additional 3pp (up to an accuracy of 62.4%) on the test set using an additional 34 pentamers (figure 2.2d). However, also this curve saturated quickly, so that further large gains using this approach are unlikely. We explored adding a couple of other 5mer and 6mer sets into the prediction process, but while some could lead a simulation accuracy of up to 68% on the training set, none was more accurate than 63% on the test set (data not shown).

(2.1)

region	19anchors, GGG	exonScan
all	30.6%	35.4
upstr	0.0%	4.4
downstr	4.4%	4.4
accESE	11.6%	23.2
donESE	10.8%	23.2
ESS	15.1%	11.7

Table 2.1 Fraction of transcripts used for splicing simulation. For all considered subsets apart from ESS, and for all subsets together, our approach uses less predictive features than exonScan (Wang et al., 2004).

Correlation between explored scores and pure logOdds-scores. We next investigated the exact relationship between logOdds-scores and explored scores (2.3a,b) and found a spearman correlation of 0.80 ($p < 7.3e-7$) when using the first 20 oligomers (26 parameters) of the "anchor parameter set". A large part of this high correlation appears however to be due to the fact that ESS tend to receive negative scores whereas all other elements tend to receive positive scores in both approaches. Dividing

correlation analysis into ESS and all other elements, led to a non-significant spearman correlation of 0.37 for the former and a spearman correlation of 0.53 ($p < 0.021$) for the latter (2.3b). In summary, it appears that logOdds-scores are a good scoring scheme for many oligomers, but that they do not always capture all the information that an oligomer can provide. Examples where logOdds-scores seem to fail are TAGG or repetitive elements such as TTTT (figure 2.3a).

Evolutionary conservation of anchors used in simulation. If the 19 oligomers of the "anchor parameter set" represent "the most important parts" of larger ESE and ESS sequences, then one would expect, at least for ESE anchors, that conservation of wobble positions focuses on these anchors. In order to check this we made use of a set of roughly 6900 aligned exons between human and mouse (methods) published by Plass and Eyras (2006). We first mapped the 9 ESE anchors we used as donor ESE to human exons. Then we mapped larger ESE sequences (Cartegni et al., 2003; Smith et al., 2006; Stadler et al., 2006; Goren et al., 2006; Zhang et al., 2008), that contain these anchors to the human exonic sequences. We then counted for each anchor conserved and non-conserved wobble positions that fell inside the anchor and outside (but into a larger published ESE). In this way we obtained a 2x2 table for each anchor. For both, 4mer anchors that were used as acceptor ESE and that were used as donor ESE we found that 5 out of 9 showed significantly elevated conservation inside the anchors (controlling for an FDR in the Benjamini Hochberg sense, table 2.2).

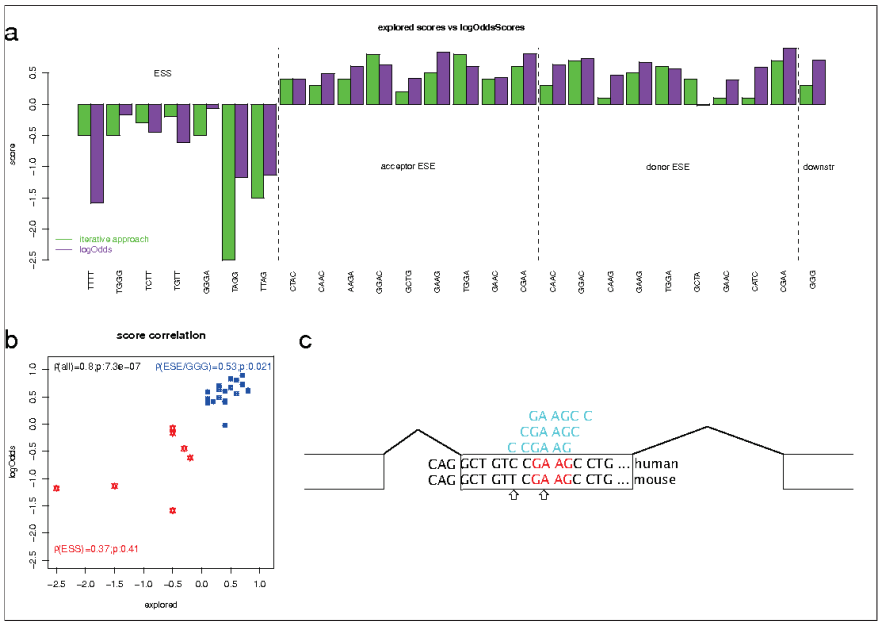


Figure 2.3 Analysis of tetramer anchors used for simulation. Comparison of iteratively explored scores with a-priori trained logOdds-scores (**a**). Correlation analysis of iteratively explored scores and a-priori trained logOdds-scores (**b**).Illustration of conservation analysis. For each ESE anchor used for simulation, conserved and non conserved wobble positions are mapped inside the anchor and outside (but within larger published ESE that contain the anchor). This results in a 2x2 table for each anchor, which is tested using a fisher test(**c**).

(2.2)

accESE	donESE
TGGA	TGGA
CAAC	CAAC
GAAG	GAAG
GCTG	CATC
GAAC	CAAG

Table 2.2 ESE-anchors with concentrated conservation. The 5 out of 9 acceptor ESE -and 5 out of 9 donor ESE- that showed concentrated conservation between human and mouse.

2.3 Discussion

All published ESE and ESS sets were published with good reason. The union of these sets however comprises more than half of all hexamers

and additional sequences of different lengths. Therefore any random sequence will be full of them, with many positions covered by multiple ESE, unless it is highly optimized in order not to contain any ESE or ESS hexamers. Here we have represented multiple ESE sequences by a single signal word ("anchor"), that guarantees the presence of at least one published ESE, although the anchor does not exactly define the position of the published ESE. More exactly we have first defined all 10mers that contain a published ESE and then represented between $1 \times 4^6 = 4096$ and $(10 - 4 + 1) \times 4^6 = 28672$ of these 10mers by a single 4mer. To use 4mers has not been fixed a-priori, but is a result of this approach - with an exception of one trimer (GAA) that also guarantees the presence of published ESE in its surroundings. The fact that tetramers appear to be smallest possible ESS and ESS unit is comforting, as it is the smallest unit of more than codon size that presents a natural separation from the genetic code and the coding capacity of many exons.

In terms of using these anchor sequences for splicing simulation, the importance of this approach does not lie so much in the achieved simulation accuracy, but rather in the ranking of the oligomers on the one hand and the characterization of the elements that are really needed to achieve the level of accuracy of ExonScan. The finding that for a little more than 50% of the ESE anchors used for simulation, conservation between human and mouse concentrates on them supports the idea that these anchors are really oligomers of elevated biological importance. On the other hand however one must ask, "what about the rest"? Here it should be noted that a mutation of an anchor not necessarily destroys the ESE capacity of an oligomer. In this way the anchor GAAC (non significant in the conservation analysis as a donor ESE) is only one mutation away from 3 other anchors (GGAC,CAAC and GAAG), that were also used for simulation.

With respect to simulation accuracy, ExonScan's exceptional performance should be noted. It seems to have captured most of information that is present under the given model. Under this model and test data set we can achieve three percentage points more in simulation accuracy (e.g. ave(SNe,SPe)=62.4% instead of 59.3% using one trimer, 19 tetramer and 34 pentamer anchors - a total of 54 oligomers), depending on how many anchors are used, but we could not achieve the qualitative jump in simulation accuracy (something like from 59% to 80%) that we had hoped for. While we cannot rule out that a subset of scored oligomers (among the exponential number of possible subsets) could lead to important gains in splicing simulation accuracy, we believe that more important advances might be achieved by the following four ideas. First, in the used model, oligomers were scored using the same distance criteria. These distance rules might however be ESE-specific; usage of ESE specific distances has given good results for prediction of inclusion level changes (Barash et al., 2010) and might do well for splicing simulation also. Second, in the present model all pre-mRNAs are treated as if their cellular context (e.g. splicing factor expression levels) were identical. Assuming one were to have expression levels of all splicing factors and their binding sites, one could make the score of an ESE depend on the expression of the associated splicing factor(s). In this way an ESE might have a 0-score in one RNA (expressed and spliced in cell type 1) but a strongly positive score in another RNA (expressed and spliced in cell type 2). Growing evidence shows that epigenetic states can influence splicing. Therefore, in a similar way as for splicing factors, one could also make the score of an ESE (or ESS, GGG, etc.) depend on epigenetic states. This is particularly appealing, as it would introduce co-transcriptional splicing into the splicing simulation world, something that the current model ignores completely. Last, the current

model only scores exons and tries to assemble the transcript structure from all predicted exons - which typically strongly outnumber real exons. Given the large numbers of exon candidates and the small exon size in comparison to the transcript size, one might think of this process as "assembling a large structure from many pieces of poor information" or in other words "knowing the location of one exon is not a lot of knowledge about a transcript". On the other hand if one were able to use information on which exon is spliced to which other exon, one could add intron locations to the simulation process. Given that introns are typically much larger than exons, this could be thought of as "using fewer pieces of rich information" or in other words "knowing the location of a (long) intron is a lot of information about a transcript", as it has the power to exclude many pseudoexons from the simulated structure.

2.4 Methods

Definition of ESE and ESS anchors. Here an approach to do this is described for ESE only. The definition for ESS proceeds analogously. Intuitively, we look for a 4mer for example and a fixed position within a 10mer, so that all 10mers having the 4mer at this fixed position, contain a published ESE. Figure 2.1 gives an illustration of this approach, more formal definitions are given by equations 2.4 to 2.5.

Let o_m and o_n be two oligomers of length m and n with $m > n$ and let furthermore $cont(o_m, o_n)$ denote that o_n is a substring of o_m. Using the

sets of published ESE $ESE_4, ESE_5, ..., ESE_{10}$[2] of oligomers of length 4 to 10, we then define $PUBESE_{10}$ (see equation 2.4), the set of 10mers that contain a published ESE:

$$PUBESE_{10} := \{o_{10} : \exists o \in \bigcup_{i=4}^{10} ESE_i : cont(o_{10}, o)\} \qquad (2.3)$$

Then we define our ESE anchors for a fixed size m ($anchESE_m$). Equation 2.4 shows the formula for the set $anchESE_4$:

[2]The set ESE_4 was actually empty, since no 4mer had been published as ESE in the publications we queried. These ESE sets $ESE_5, ESE_6, ..., ESE_{10}$ contained oligomers published in (Cartegni et al., 2003; Fairbrother et al., 2002; Goren et al., 2006; Smith et al., 2006; Stadler et al., 2006; Zhang and Chasin, 2004; Zhang et al., 2008) and included ESE that had been collected in previous efforts to centralize published ESE (Stamm et al., 2006; Goren and Ast, 2006). These ESE were originally published by Liu et al. (1998); Cavaloc et al. (1999); Heinrichs and Baker (1995); Schaal and Maniatis (1999); Tacke et al. (1998); Modafferi and Black (1999). For ESS the set ESS_4 contained 8 words which by definition end up in the set $anchESS_4$. Published ESS were taken from Stadler et al. (2006); Wang et al. (2004); Zhang and Chasin (2004); Zhang et al. (2008) and Stamm et al. (2006); Goren and Ast (2006). The sequences of the latter trace back to Del Gatto et al. (1996); Burd and Dreyfuss (1994); Caputi and Zahler (2001); Modafferi and Black (1999); Chou et al. (1999); Min et al. (1997); Ashiya and Grabowski (1997); Chan and Black (1995); Chen et al. (1999); DeMaria and Brewer (1996); Ishikawa et al. (1993); Kajita et al. (1995); Jacquenet et al. (2001); Kiledjian and Dreyfuss (1992); Leffers et al. (1995); Matunis et al. (1994); Myer and Steitz (1995); Ostrowski et al. (2001); Swanson and Dreyfuss (1988); Pérez et al. (1997); Brooks and Rigby (2000); Sokolowski et al. (1999); Soltaninassab et al. (1998); Soulard et al. (1993); Spångberg et al. (2000); Takahashi et al. (2000); Lu et al. (1999); Thisted et al. (2001); Sironi et al. (2004); Ule et al. (2006).

$$\begin{aligned}
anchESE_4 &:= \{\text{4mers } o_4 : \forall \text{6mers } o_6 : o_4.o_6 \in PUBESE_{10}\} \\
&\cup \{\text{4mers } o_4 : \forall \text{1mers } o_1 \forall \text{5mers } o_5 : o_1.o_4.o_5 \in PUBESE_{10}\} \\
&\cup \{\text{4mers } o_4 : \forall \text{2mers } o_2 \forall \text{4mers } o_4' : o_2.o_4.o_4' \in PUBESE_{10}\} \\
&\cup \{\text{4mers } o_4 : \forall \text{3mers } o_3 \forall \text{3mers } o_3' : o_3.o_4.o_3' \in PUBESE_{10}\} \\
&\cup \{\text{4mers } o_4 : \forall \text{4mers } o_4' \forall \text{2mers } o_2 : o_4'.o_4.o_2 \in PUBESE_{10}\} \\
&\cup \{\text{4mers } o_4 : \forall \text{5mers } o_5 \forall \text{1mers } o_1 : o_5.o_4.o_1 \in PUBESE_{10}\} \\
&\cup \{\text{4mers } o_4 : \forall \text{6mers } o_6 : o_6.o_4 \in PUBESE_{10}\} \qquad (2.4)
\end{aligned}$$

More generally (but less well readable) for any $i = 1,..,9$ the set $anchESE_i$ is defined by equation 2.5 (where o_0 has to be interpreted as the empty string or simply as "no sequence"):

$$anchESE_i := \bigcup_{j=0}^{10-i} \{\text{i-mers } o_i : \forall o_j, o_{10-i-j} : o_j.o_i.o_{10-i-j} \in PUBESE_{10}\} \qquad (2.5)$$

ESE- and ESS-anchors of 3,4 and 5 were defined using this approach. With the exception of one word (GAA), the shortest anchors were 4mers; however a non negligible number of 4mers were labeled anchors. This suggests that the smallest possible unit for which one can sensibly claim ESE or ESS activity are tetramers. Table 2.6 shows the numbers of ESE- and ESS-anchors for 3-5nts. The same numbers for 6nt anchors are not included in this table, since (as follows straight from the definition) 6mer anchors contain all published 6mers. As previously noted there is a large number of published ESE and ESS of length 6 and increasing this set through the definition of further anchors is unlikely to further our understanding of splicing.

Generally speaking there are always fewer ESS anchors than ESE anchors for a fixed length, a fact that traces back to the larger number of published ESE as compared to published ESS. This is consistent with the observation, that when similar a-priori computational methods were used to define ESE and ESS, the number of ESE was larger (Zhang and Chasin, 2004).

$i =$	$\#anchESE_i$	$\#anchESS_i$
3	1	0
4	29	15
5	346	176

(2.6)

Table 2.3 Number of ESE- and ESS-anchors of length 3,4 and 5. Anchors of length 6 and longer are not shown, because by definition they will contain at least 60% of all possible hexamers.

As noted earlier none of the 29 ESE 4mer anchors was in the input set of ESE_4, however 8 out of the defined 15 ESS 4mer anchors had been previously published and can therefore not be considered novel.

To predict a spliced structure on the given pre-mRNA sequence any splicing simulator has to execute three steps: First all putative splice sites within the pre-mRNA and their strengths have to be found. Then these splice sites have to be combined to putative exons. Finally from the set of all putative exons a final spliced transcript has to be predicted.

Every AG-dinucleotide (GT-dinucleotides resp.) in the pre-mRNA is considered a putative acceptor a (putative donor d resp.) and is given an acceptor score $s(a)$ (donor score $s(d)$ resp.) using splice site models such as for example markov chains, an approach followed by Geneid (Guigó et al., 1992; Parra et al., 2000). The results presented here are however based on a splice site model called maxEnt (Yeo and Burge, 2004), in order to make the comparison with ExonScan, which uses maxEnt (Wang et al., 2004), more exact. Most splice site models make use of sequence starting upstream of a and include up to three downstream nucleotides of a for computation of $s(a)$. Similarly 3 nucleotides upstream of d are also often taken into account to compute $s(d)$. Both of these three-nucleotide stretches are con-

sidered part of the splice site model.

The exon score is additively determined from the scores of the acceptors and donors, all scored enhancers and scored silencers as well as an empirically determined parameter optimizing simulation accuracy. Acceptor and donor enhancers are specified in two sets $aESE$ and $dESE$ (short for "acceptor ESE" and "donor ESE") and silencers in a set ESS ("ESS"). Generally specified enhancers and silencers are not taken into account when they overlap the splice site models. Specified acceptor enhancers are only taken into account when they are closer to the acceptor than the to the donor and within a maximum of 80nts of the acceptor. A similar rule (closer to the donor than to the acceptor and within maximally 80nts of the donor) applies to specified donor enhancers. More formally the exon-scoring-step proceeds as follows: For each acceptor-donor-pair (a, d) that is separated by no more than 250nts and no less than 50 nts a putative exon $e = (a, d)$ is created. Denoting its sequence with $n_1 n_2 ... n_l$ the two limits for acceptor and donor enhancers are defined as $x = min(80, l/2)$ and $y = max(l - 80, l/2) + 1$. The 100 nucleotides upstream and downstream of the exon will be written as $n_{-100} n_{-99} ... n_{-1}$ and $n_{l+1} n_{l+2} ... n_{l+100}$ respectively. Assuming that $aESE$, $dESE$ and ESS contain oligomers of g different lengths $k_1, ..., k_g$ the score of the putative exon $e = (a, d)$ is calculated by the following formula:

$$
\begin{aligned}
score(e) &= SF * (s(a) + s(d)) && (2.7)\\
&+ EW && (2.8)\\
&+ EF * \sum_{i=1}^{g} (c_{k_i}^{ESS} + c_{k_i}^{aESE} + c_{k_i}^{dESE} + c_{k_i}^{Up} + + c_{k_i}^{Down}) && (2.9)
\end{aligned}
$$

where

$$c_{k_i}^{ESS} = \sum_{j=4}^{l-3-k_i+1} s_{ESS}(n_j..n_{j+k_i-1})$$

is the **c**ontribution to $score(e)$ of all elements of ESS of length k_i,

$$c_{k_i}^{aESE} = \sum_{j=4}^{x-k_i+1} s_{aESE}(n_j..n_{j+k_i-1})$$

the **c**ontribution of all elements of aESE of length k_i, similarly

$$c_{k_i}^{dESE} = \sum_{j=y}^{l-3-k_i+1} s_{dESE}(n_j..n_{j+k_i-1})$$

the **c**ontribution of all elements of dESE of length k_i. In the same way **c**ontribution of all elements of Up and Down of length k_i are given by

$$c_{k_i}^{Up} = \sum_{j=-100}^{40-k_i+1} s_{Up}(n_j..n_{j+k_i-1})$$

and

$$c_{k_i}^{Down} = \sum_{j=l+10}^{l+70-k_i+1} s_{Down}(n_j..n_{j+k_i-1})$$

Finally

$$s_{ESS}(n_j..n_{j+k_i-1}) = \begin{cases} \neq 0 & n_j..n_{j+k_i-1} \in ESS \\ 0 & \text{else} \end{cases} \quad (2.10)$$

the silencing score of the k-mer starting at position j in $e = (a, d)$. The terms $s_{aESE}(n_j..n_{j+k_i-1})$, $s_{dESE}(n_j..n_{j+k_i-1})$, $s_{Up}(n_j..n_{j+k_i-1})$ and $s_{Down}(n_j..n_{j+k_i-1})$ are defined in close analogy with the definition of $s_{ESS}(n_j..n_{j+k_i-1})$.

Term (2.7) gives the contribution of the splice sites to the final score of the acceptor-donor-pair, (2.8) is a correctional parameter that has been used earlier (Parra et al., 2000) and which can be interpreted (and calculated) as the log-ratio of the prior probabilities of (a, d) being an exon and (a, d) not being an exon (Parra; Castelo and Guigó, 2004). In practice however it is estimated in order to guarantee that SNe and SPe are similar. Term 2.9 finally gives the contribution of all regulatory elements to the score of the putative exon. The factors SF and EF are parameters allowing to weight splice site contribution and the contribution of regulatory elements differently. We chose however to have $SF = EF = 1$ so that the scores of all elements directly reflect their contribution to the prediction.

Despite this rather complicated formula a pre-calculation scaling linearly with the transcript length, allows to evaluate the score of a putative exon in $O(g)$. Due to the length constraints on predicted exons, scoring all acceptor-donor-pairs is feasible in $O(g * t)$ where t is the transcript length.

Assembling putative exons to gene structures: Given the set of all scored putative exons $\{e_1, ..., e_q\}$, a dynamic programming algorithm, close to the one described in (Guigó, 1998), is employed in order produce the set of exons $\{e_{i_1}, ..., e_{i_m}\}$ that maximizes $\sum_{j=1}^{m} score(e_{i_j})$ among all sets of exons that respect the constraints given in the parameter file. By default an intron is only required to have at least 40nts. A maximal intron length requirement can also imposed, this is however currently set to infinity. As shown in (Guigó, 1998), this algorithm scales linearly with q and again due

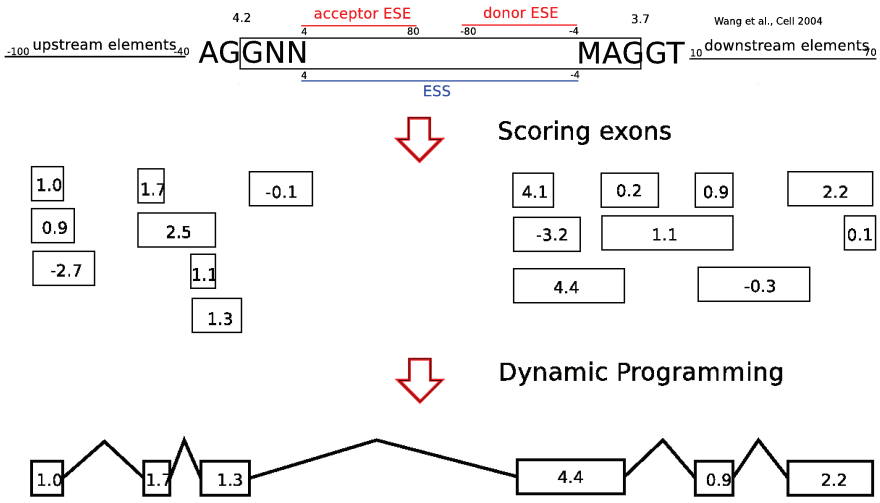


Figure 2.4 Illustration of the simulation process. Splice site scores and scores for ESS, acceptor ESE and donor ESE as well as for up- and downstream elements are additively combined into an exon score. From all exon candidates the predicted exons are chosen by a dynamic programming algorithm.

to the previously mentioned length constraints it does so as well with the transcript length.

Pre $\equiv$ t, A: training Set and annotation
(* initialization *)
$ESS = \emptyset; aESE = \emptyset; dESE = \emptyset; Up = \emptyset; Down = \emptyset$
$s_{ESS} = \emptyset; s_{aESE} = \emptyset; s_{dESE} = \emptyset; s_{Up} = \emptyset; s_{Down} = \emptyset;$
$(SNe, SPe) = SimEva(s_{ESS}, s_{aESE}, s_{dESE}, s_{Up}, s_{Down}, t, A)$
(* iteration *)
while true **do**
(* choose best possible ESS and score *)
if $SNe > SPe$ **then**
$S = (anchESS_4 \setminus ESS) \times \{-5.0, -4.9, \ldots, -0.1\}$
$(k^*, s^*) = argmax_{(k,s)\in S} SimEva(s_{ESS} \cup \{(k,s)\}, s_{aESE}, s_{dESE}, s_{Up}, s_{Down}, t, A)$
$(SNe, SPe) = max_{(k,s)\in S} SimEva(s_{ESS} \cup \{(k,s)\}, s_{aESE}, s_{dESE}, s_{Up}, s_{Down}, t, A)$
$s_{ESS} = s_{ESS} \cup \{(k^*, s^*)\}$
$ESS = ESS \cup \{k^*\}$
(* choose best possible ESE or GGG and score *)
else (* $SNe \leq SPe$ *)
(* acceptor ESE *)
$S = (anchESE_4 \setminus aESE) \times \{0.1, 0.2, \ldots, 5.0\}$
$(k_1^*, s_1^*) = argmax_{(k,s)\in S} SimEva(s_{ESS}, s_{aESE} \cup \{(k,s)\}, s_{dESE}, s_{Up}, s_{Down}, t, A)$
$(SNe_1, SPe_1) = max_{(k,s)\in S} SimEva(s_{ESS}, s_{aESE} \cup \{(k,s)\}, s_{dESE}, s_{Up}, s_{Down}, t, A)$
(* donor ESE *)
$S = (anchESE_4 \setminus dESE) \times \{0.1, 0.2, \ldots, 5.0\}$
$(k_2^*, s_2^*) = argmax_{(k,s)\in S} SimEva(s_{ESS}, s_{aESE}, s_{dESE} \cup \{(k,s)\}, s_{Up}, s_{Down}, t, A)$
$(SNe_2, SPe_2) = max_{(k,s)\in S} SimEva(s_{ESS}, s_{aESE}, s_{dESE} \cup \{(k,s)\}, s_{Up}, s_{Down}, t, A)$
(* upstream GGG *)
$S = (\{GGG\} \setminus Up) \times \{0.1, 0.2, \ldots, 5.0\}$
$(k_3^*, s_3^*) = argmax_{(k,s)\in S} SimEva(s_{ESS}, s_{aESE}, s_{dESE}, s_{Up} \cup \{(k,s)\}, s_{Down}, t, A)$
$(SNe_3, SPe_3) = max_{(k,s)\in S} SimEva(s_{ESS}, s_{aESE}, s_{dESE}, s_{Up} \cup \{(k,s)\}, s_{Down}, t, A)$
(* downstream GGG *)
$S = (\{GGG\} \setminus Down) \times \{0.1, 0.2, \ldots, 5.0\}$
$(k_4^*, s_4^*) = argmax_{(k,s)\in S} SimEva(s_{ESS}, s_{aESE}, s_{dESE}, s_{Up}, s_{Down} \cup \{(k,s)\}, t, A)$
$(SNe_3, SPe_3) = max_{(k,s)\in S} SimEva(s_{ESS}, s_{aESE}, s_{dESE}, s_{Up}, s_{Down} \cup \{(k,s)\}, t, A)$
if $argmax_{i\in\{1,2,3,4\}} ave(SNe_i, SPe_i) == 1$ **then**
$s_{aESE} = s_{aESE} \cup \{(k_1^*, s_1^*)\}$
$aESE = aESE \cup \{k_1^*\}$
if $argmax_{i\in\{1,2,3,4\}} ave(SNe_i, SPe_i) == 2$ **then**
$s_{dESE} = s_{dESE} \cup \{(k_2^*, s_2^*)\}$
$dESE = dESE \cup \{k_2^*\}$
if $argmax_{i\in\{1,2,3,4\}} ave(SNe_i, SPe_i) == 3$ **then**
$s_{Up} = s_{Up} \cup \{(k_3^*, s_3^*)\}$
$Up = Up \cup \{k_3^*\}$
if $argmax_{i\in\{1,2,3,4\}} ave(SNe_i, SPe_i) == 4$ **then**
$s_{Down} = s_{Down} \cup \{(k_4^*, s_4^*)\}$
$Down = Down \cup \{k_4^*\}$
(* update *)
$(SNe, SPe) = SimEva(s_{ESS}, s_{aESE}, s_{dESE}, s_{Up}, s_{Down}, t, A)$

Figure 2.5 Iterative definition of parameters: In each iteration one more scored anchor is added to the simulator's parameters. Anchors and scores are prioritized by their influence on simulation accuracy.

The training procedure: the performance of a splicing simulator depends critically on its parameters, e.g. the splice site models and the sets aESE,dESE and ESS as well as the scores associated to their elements. ExonScan (Wang et al., 2004) uses logOdds-scores trained from appearance frequencies of hexamers in annotated exons and introns. Such scores can be

calculated efficiently and yield good performance. Such an approach does however not answer the following two questions directly:

- which ESE and ESS are the most important ones for a "good prediction" ?
- what is the optimal score for each ESE and ESS ?

In order to answer this question exhaustively, one would need to explore all possible parameter settings of the simulator. Limiting the possible elements of αESE,$dESE$ and ESS to the ESE and ESS anchors (e.g. of size 4) described above. Then αESE,$dESE$ and ESS could be defined in $2^{15} * 2^{29} * 2^{29}$ possible ways. The below calculation illustrates this number as 66553.45 times the earth's age in seconds (assuming earth's age to be roughly 4.5 billion years).

$$\frac{2^{15} * 2^{29} * 2^{29} assignments}{4.5 * 10^9 yr * 365 \frac{days}{year} * 24 \frac{h}{day} * 60 \frac{min}{h} * 60 \frac{s}{min}} = 66553.45 \frac{assignments}{s}$$

Using pre-calculated fixed scores (such as logOdds-scores) and assuming to have the computational power to run and evaluate 66553.45 simulations per second, one could therefore explore all these assignments in roughly the time earth exists. When however for each assignment all scores for the elements of αESE,$dESE$ and ESS have to be optimized much more computational time will be necessary.

In order to approach the problem, this section describes an iterative approach, where one element of the sets αESE,$dESE$ and ESS and its associated score is determined at a time (see 2.5 for the pseudocode of this approach). This is done, by choosing an anchor and an associated score in each iteration: The anchor and the score that raise simulation accuracy most.

Simulation accuracy was here defined as the average of sensitivity and specificity on exon level. ESEs generally raise sensitivity whereas ESSs tend to raise specificity (Wang et al., 2004). Aiming at keeping SNe ($SNe :=$ sensitivity on exon level) and SPe ($SPe :=$ specificity on exon level) close to each other and in order to achieve a natural balance between ESE and ESS, two kinds of iterations were defined. When $SNe < SPe$, the goal is to raise SNe through the incorporation of an element that has positive influence on exon recognition (e.g. an ESE or an intronic GGG). When $SNe \geq SPe$ on the other hand an element with negative influence on exon recognition (e.g. an ESS) is looked for.

Simulated Annealing (Kirkpatrick et al., 1983) is a technique used in optimization problems in order to escape local optima. In each iteration a random score mutation is proposed. Here the optimization problem is being stated as a maximization problem, the objective function being simulation accuracy. If a proposed score mutation leads to increases in simulation accuracy the mutation is accepted. If on the other hand it leads to decreases in simulation accuracy, it is accepted with a certain probability. This probability decreases with the decrease in simulation accuracy and with iteration numbers. The analogy to the annealing process in physiscs is given by the annealing temperature T that decreases with increasing iterations. Here we chose to implement a very simple version, where $T = 1/i$, i being the iteration number.

```
Pre ≡ t, A, ESS, aESE, dESE, Up, Down, s_ESS, s_aESE, s_dESE, s_Up, s_Down: training Set, annotation and
    parameter sets
    (* initialization *)
    (SNe, SPe) = SimEva(s_ESS, s_aESE, s_dESE, s_Up, s_Down, t, A)
    i = 0
    n = |s_ESS| + |s_aESE| + |s_dESE| + |s_Up| + |s_Down|
5:  (* iteration *)
    while true do
        i = i + 1
        (* annealing temperature *)
        T = 1/i
10:     (* random score *)
        s* = rand({0, 0.1, ..., 3.3})
        (* random word *)
        j = rand({1, ..., n})
        (* ESS *)
15:     if j ≤ |s_ESS| then
            (k, s) = rand(s_ESS)
            s* = (−1) × s*
            (SNe*, SPe*) = SimEva((s_ESS \ {(k, s)}) ∪ {(k, s*)}, s_aESE, s_dESE, s_Up, s_Down, t, A)
            acceptProb = min(1, exp( (ave(SNe*, SPe*) − ave(SNe, SPe)) / T ))
20:         r = rand([0, 1])
            if r ≤ acceptProb then
                s_ESS = (s_ESS \ {(k, s)}) ∪ {(k, s*)}
        (* acceptor ESE *)
        if |s_ESS| < j ≤ |s_ESS| + |s_aESE| then
25:         (k, s) = rand(s_aESE)
            (SNe*, SPe*) = SimEva(s_ESS, (s_aESE \ {(k, s)}) ∪ {(k, s*)}, s_dESE, s_Up, s_Down, t, A)
            acceptProb = min(1, exp( (ave(SNe*, SPe*) − ave(SNe, SPe)) / T ))
            r = rand([0, 1])
            if r ≤ acceptProb then
30:             s_aESE = (s_aESE \ {(k, s)}) ∪ {(k, s*)}
        (* donor ESE *)
        if |s_ESS| + |s_aESE| < j ≤ |s_ESS| + |s_aESE| + |s_dESE| then
            (* analogosly as for acceptor ESE *)
        (* upstream GGG *)
35:     if |s_ESS| + |s_aESE| + |s_dESE| < j ≤ |s_ESS| + |s_aESE| + |s_dESE| + |s_Up then
            (* analogosly as for acceptor ESE *)
        (* downstream GGG *)
        if |s_ESS| + |s_aESE| + |s_dESE| + |s_Up| < j ≤ n then
            (* analogosly as for acceptor ESE *)
40:     (* update *)
        (SNe, SPe) = SimEva(s_ESS, s_aESE, s_dESE, s_Up, s_Down, t, A)
```

Figure 2.6 Simulated annealing In each iteration a score mutation is proposed. Score mutation that improve simulation accuracy are always accepted. Score mutations that decrease simulation accuracy are accepted with a given probability that decreases with simulation accuracy losses and over time.

Transcripts: training and test set: In this manuscript we will exploit the exon definition model for U2 introns only in order to simulate splicing of an entire pre-mRNA. Wang et al. (2004) published 1820 transcripts (with 11630 internal exons) for which at the time no evidence for alternative splicing was available. We have discarded all transcripts that did not comply with the following constraints:

1. All internal exons respect minimal (at least 50nts) and maximal exon lengths (at most 250nts) for exon-definition. These limits have also been adopted for EXONSCAN (Wang et al., 2004) for simulation and are slightly more conservative than the 50-300nts mentioned by Berget (1995). 1069 transcripts with 4997 internal exons meet this criterion.

2. All internal exons of the transcripts have a GT-donor and an AG-acceptor. This is fulfilled by 1728 transcripts with 10467 internal exons.

3. No evidence for U12 splicing can be found within the transcript. This has been verified using Geneid-1.3 (Alioto & Guigo unpublished). 1768 transcripts (with 11076 internal exons) do not show any evidence for U12-splicing.

4. All introns of the transcript have at least 40 nts. This has been enforced to guarantee that one can trust the annotated introns. Only one transcript (with six internal exons) did not respect this rule, having an intron with 31nts only.

For two duplicates (two pairs of transcripts with the same sequence and the same annotation) one transcript each has also been discarded. The remaining 1000 transcripts having 4400 internal exons were split into a *training set* and a *test set* of 500 transcripts each. As only internal exons are simulated so far, the sequences of the first and last exon as well as 40 adjacent nucleotides (the minimal allowed length for introns) have been masked.

LogOdds-score calculation: All logOdds-scores were calculated on the training set. For ESS logOdds-scores, the training set pre-mRNAs were separated into internal exons and remaining sequences. From exonic sequences the first and last three nucleotides were removed. The logOdds-score was calculated as the $\log_2$ of the exonic frequency divided by the frequency in the remaining sequence. With this definition 4mer anchor ESS that were used for simulation all received negative scores, as was expected. For acceptor ESE logOdds-scores, we defined for every internal exon the region (see definition of the simulator) where acceptor ESE are taken into account by the simulator. LogOdds-scores were calculated as the $\log_2$-ratio of the frequency of a word in the area vs the frequency in the remaining part of the transcripts. A similar definition applied to logOdds-scores for donor ESE and upstream and downstream GGGs. In this way, 4mer ESE anchors that were used for simulation (with the exception of GCTA) and downstream GGGs received positive scores, again as was expected.

Aligned exons between human and mouse: From the exon alignment set published by Plass and Eyras (2006), we selected those 6887 aligned exons that

- were labeled as constitutive
- were entirely coding
- had a gap-free alignment
- contained no Ns in both the human and the mouse sequence
- had between 50 and 250bps in both human and mouse
- had AG-acceptors and GT donors in human

Additional 5mer anchor training after using downstream GGGs and 19 4mer anchors: We defined the subsets of 5mer anchors that did not contain any of the previously used 4mer anchors for each category (ESS, acceptor ESE and donor ESE). As we had previously observed (data not shown) that 5mers tend to lead to larger overfitting than 4mers, when scores are optimized, we decided to used fixed logOdds-scores rather than exploring scores. In order to reduce the number of parameters only the 10% strongest 5mer anchor ESS and ESE were used. Initially we added all of these 5mers at once, multiplying all ESE by a coefficient m_{ESE} and all ESS by a coefficient m_{ESS}. The two coefficients were optimised so as to achieve the highest possible accuracy. In order to assess which of these oligomers contributed most, we removed them again from the parameter set and then added them one by one, always the one raising accuracy most. All iterative sub-parameter sets were also tested on the test set.

Bibliography

M Ashiya and P J Grabowski. A neuron-specific splicing switch mediated by an array of pre-mrna repressor sites: evidence of a regulatory role for the polypyrimidine tract binding protein and a brain-specific ptb counterpart. *RNA*, 3(9):996–1015, Sep 1997.

Y Barash, J A Calarco, W Gao, Q Pan, X Wang, O Shai, B J Blencowe, and B J Frey. Deciphering the splicing code. *Nature*, 465(7294):53–9, May 2010. doi: 10.1038/nature09000.

S M Berget. Exon recognition in vertebrate splicing. *J Biol Chem*, 270(6):2411–4, Feb 1995.

S A Brooks and W F Rigby. Characterization of the mrna ligands bound by the rna

binding protein hnrnp a2 utilizing a novel in vivo technique. *Nucleic Acids Res*, 28(10):E49, May 2000.

C G Burd and G Dreyfuss. Rna binding specificity of hnrnp a1: significance of hnrnp a1 high-affinity binding sites in pre-mrna splicing. *EMBO J*, 13(5):1197–204, Mar 1994.

M Caputi and A M Zahler. Determination of the rna binding specificity of the heterogeneous nuclear ribonucleoprotein (hnrnp) h/h'/f/2h9 family. *J Biol Chem*, 276(47):43850–9, Nov 2001. doi: 10.1074/jbc.M102861200.

L Cartegni, J Wang, Z Zhu, M Q Zhang, and A R Krainer. Esefinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res*, 31(13):3568–71, Jul 2003.

R Castelo and R Guigó. Splice site identification by idlbns. *Bioinformatics*, 20 Suppl 1:i69–76, Aug 2004. doi: 10.1093/bioinformatics/bth932.

Y Cavaloc, C F Bourgeois, L Kister, and J Stévenin. The splicing factors 9g8 and srp20 transactivate splicing through different and specific enhancers. *RNA*, 5(3): 468–83, Mar 1999.

R C Chan and D L Black. Conserved intron elements repress splicing of a neuron-specific c-src exon in vitro. *Mol Cell Biol*, 15(11):6377–85, Nov 1995.

C D Chen, R Kobayashi, and D M Helfman. Binding of hnrnp h to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. *Genes Dev*, 13(5):593–606, Mar 1999.

M Y Chou, N Rooke, C W Turck, and D L Black. hnrnp h is a component of a splicing enhancer complex that activates a c-src alternative exon in neuronal cells. *Mol Cell Biol*, 19(1):69–77, Jan 1999.

F Del Gatto, M C Gesnel, and R Breathnach. The exon sequence tagg can inhibit splicing. *Nucleic Acids Res*, 24(11):2017–21, Jun 1996.

C T DeMaria and G Brewer. Auf1 binding affinity to a+u-rich elements correlates with rapid mrna degradation. *J Biol Chem*, 271(21):12179–84, May 1996.

W G Fairbrother, R F Yeh, P A Sharp, and C B Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–13, Aug 2002. doi: 10.1126/science.1073774.

A Goren and G Ast. Esrsearch. `http://esrsearch.tau.ac.il/`, 2006.

A Goren, O Ram, M Amit, H Keren, G Lev-Maor, I Vig, T Pupko, and G Ast. Comparative analysis identifies exonic splicing regulatory sequences–the complex definition of enhancers and silencers. *Mol Cell*, 22(6):769–81, Jun 2006. doi: 10.1016/j.molcel.2006.05.008.

R. Guigó. Assembling genes from predicted exons in linear time with dynamic programming. *Journal of Computational Biology*, 5:681–702, 1998.

R. Guigó, S. Knudsen, N. Drake, and T. Smith. Prediction of gene structure. *Journal of Molecular Biology*, 226:141–157, 1992.

V Heinrichs and B S Baker. The *Drosophila* sr protein rbp1 contributes to the regulation of doublesex alternative splicing by recognizing rbp1 rna target sequences. *EMBO J*, 14(16):3987–4000, Aug 1995.

F Ishikawa, M J Matunis, G Dreyfuss, and T R Cech. Nuclear proteins that bind the pre-mrna 3' splice site sequence r(uuag/g) and the human telomeric dna sequence d(ttaggg)n. *Mol Cell Biol*, 13(7):4301–10, Jul 1993.

S Jacquenet, A Méreau, P S Bilodeau, L Damier, C M Stoltzfus, and C Branlant. A second exon splicing silencer within human immunodeficiency virus type 1 tat exon 2 represses splicing of tat mrna and binds protein hnrnp h. *J Biol Chem*, 276(44):40464–75, Nov 2001. doi: 10.1074/jbc.M104070200.

Y Kajita, J Nakayama, M Aizawa, and F Ishikawa. The uuag-specific rna binding protein, heterogeneous nuclear ribonucleoprotein d0. common modular structure and binding properties of the 2xrbd-gly family. *J Biol Chem*, 270(38):22167–75, Sep 1995.

M Kiledjian and G Dreyfuss. Primary structure and binding activity of the hnrnp u protein: binding rna through rgg box. *EMBO J*, 11(7):2655–64, Jul 1992.

S Kirkpatrick, C D Gelatt, Jr, and M P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–80, May 1983. doi: 10.1126/science.220.4598.671.

H Leffers, K Dejgaard, and J E Celis. Characterisation of two major cellular poly(rc)-binding human proteins, each containing three k-homologous (kh) domains. *Eur J Biochem*, 230(2):447–53, Jun 1995.

H X Liu, M Zhang, and A R Krainer. Identification of functional exonic splicing enhancer motifs recognized by individual sr proteins. *Genes Dev*, 12(13):1998–2012, Jul 1998.

X Lu, N A Timchenko, and L T Timchenko. Cardiac elav-type rna-binding protein (etr-3) binds to rna cug repeats expanded in myotonic dystrophy. *Hum Mol Genet*, 8(1):53–60, Jan 1999.

M J Matunis, J Xing, and G Dreyfuss. The hnrnp f protein: unique primary structure, nucleic acid-binding properties, and subcellular localization. *Nucleic Acids Res*, 22(6):1059–67, Mar 1994.

H Min, C W Turck, J M Nikolic, and D L Black. A new regulatory protein, ksrp, mediates exon inclusion through an intronic splicing enhancer. *Genes Dev*, 11 (8):1023–36, Apr 1997.

E F Modafferi and D L Black. Combinatorial control of a neuron-specific exon. *RNA*, 5(5):687–706, May 1999.

V E Myer and J A Steitz. Isolation and characterization of a novel, low abundance hnrnp protein: A0. *RNA*, 1(2):171–82, Apr 1995.

J Ostrowski, Y Kawata, D S Schullery, O N Denisenko, Y Higaki, C K Abrass, and K Bomsztyk. Insulin alters heterogeneous nuclear ribonucleoprotein k protein binding to dna and rna. *Proc Natl Acad Sci U S A*, 98(16):9044–9, Jul 2001. doi: 10.1073/pnas.161284098.

G Parra. PhD thesis.

G. Parra, E. Blanco, and R. Guigó. Geneid in drosophila. *Genome Research*, 10: 511–515, 2000.

I Pérez, J G McAfee, and J G Patton. Multiple rrms contribute to rna binding specificity and affinity for polypyrimidine tract binding protein. *Biochemistry*, 36(39):11881–90, Sep 1997. doi: 10.1021/bi9711745.

M Plass and E Eyras. Differentiated evolutionary rates in alternative exons and the implications for splicing regulation. *BMC Evol Biol*, 6:50, 2006. doi: 10.1186/1471-2148-6-50.

T D Schaal and T Maniatis. Selection and characterization of pre-mrna splicing enhancers: identification of novel sr protein-specific enhancer sequences. *Mol Cell Biol*, 19(3):1705–19, Mar 1999.

M Sironi, G Menozzi, L Riva, R Cagliani, G P Comi, N Bresolin, R Giorda, and U Pozzoli. Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Res*, 32(5):1783–91, 2004. doi: 10.1093/nar/gkh341.

P J Smith, C Zhang, J Wang, S L Chew, M Q Zhang, and A R Krainer. An increased specificity score matrix for the prediction of sf2/asf-specific exonic splicing enhancers. *Hum Mol Genet*, 15(16):2490–508, Aug 2006. doi: 10.1093/hmg/ddl171.

M Sokolowski, H Furneaux, and S Schwartz. The inhibitory activity of the au-rich rna element in the human papillomavirus type 1 late 3' untranslated region correlates with its affinity for the elav-like hur protein. *J Virol*, 73(2):1080–91, Feb 1999.

S R Soltaninassab, J G McAfee, L Shahied-Milam, and W M LeStourgeon. Oligonucleotide binding specificities of the hnrnp c protein tetramer. *Nucleic Acids Res*, 26(14):3410–7, Jul 1998.

M Soulard, V Della Valle, M C Siomi, S Piñol Roma, P Codogno, C Bauvy, M Bellini, J C Lacroix, G Monod, and G Dreyfuss. hnrnp g: sequence and

characterization of a glycosylated rna-binding protein. *Nucleic Acids Res*, 21 (18):4210–7, Sep 1993.

K Spångberg, L Wiklund, and S Schwartz. Hur, a protein implicated in oncogene and growth factor mrna decay, binds to the 3' ends of hepatitis c virus rna of both polarities. *Virology*, 274(2):378–90, Sep 2000. doi: 10.1006/viro.2000.0461.

M B Stadler, N Shomron, G W Yeo, A Schneider, X Xiao, and C B Burge. Inference of splicing regulatory activities by sequence neighborhood analysis. *PLoS Genet*, 2(11):e191, Nov 2006. doi: 10.1371/journal.pgen.0020191.

S Stamm, J J Riethoven, V Le Texier, C Gopalakrishnan, V Kumanduri, Y Tang, N L Barbosa-Morais, and T A Thanaraj. Asd: a bioinformatics resource on alternative splicing. *Nucleic Acids Res*, 34(Database issue):D46–55, Jan 2006. doi: 10.1093/nar/gkj031.

M S Swanson and G Dreyfuss. Classification and purification of proteins of heterogeneous nuclear ribonucleoprotein particles by rna-binding specificities. *Mol Cell Biol*, 8(5):2237–41, May 1988.

R Tacke, M Tohyama, S Ogawa, and J L Manley. Human tra2 proteins are sequence-specific activators of pre-mrna splicing. *Cell*, 93(1):139–48, Apr 1998.

N Takahashi, N Sasagawa, K Suzuki, and S Ishiura. The cug-binding protein binds specifically to ug dinucleotide repeats in a yeast three-hybrid system. *Biochem Biophys Res Commun*, 277(2):518–23, Oct 2000. doi: 10.1006/bbrc.2000.3694.

T Thisted, D L Lyakhov, and S A Liebhaber. Optimized rna targets of two closely related triple kh domain proteins, heterogeneous nuclear ribonucleoprotein k and alphacp-2kl, suggest distinct modes of rna recognition. *J Biol Chem*, 276(20): 17484–96, May 2001. doi: 10.1074/jbc.M010594200.

J Ule, G Stefani, A Mele, M Ruggiu, X Wang, B Taneri, T Gaasterland, B J Blencowe, and R B Darnell. An rna map predicting nova-dependent splicing regulation. *Nature*, 444(7119):580–6, Nov 2006. doi: 10.1038/nature05304.

Z Wang, M E Rolish, G Yeo, V Tung, M Mawson, and C B Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–45, Dec 2004. doi: 10.1016/j.cell.2004.11.010.

G Yeo and C B Burge. Maximum entropy modeling of short sequence motifs with applications to rna splicing signals. *J Comput Biol*, 11(2-3):377–94, 2004. doi: 10.1089/1066527041410418.

C Zhang, W H Li, A R Krainer, and M Q Zhang. Rna landscape of evolution for optimal exon and intron discrimination. *Proc Natl Acad Sci U S A*, 105(15): 5797–802, Apr 2008. doi: 10.1073/pnas.0801692105.

X H Zhang and L A Chasin. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*, 18(11):1241–50, Jun 2004. doi: 10.1101/gad.1195304.

Chapter 3

On the relationship between chromatin and splicing

Summary

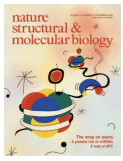
This chapter deals with the relationship between chromatin structure and splicing decisions. The first part is devoted to the analysis of nucleosome and histone-modification maps in human CD4+ T-cells (Barski et al., 2007; Schones et al., 2008) and their behavior on spliced exons. The second part shows analyses of ENCODE data showing that co-transcriptional splicing appears to be the rule rather than the exception in humans. Furthermore chromatin changes associated to alternaive exon inclusion appear to be very widespread.

3.1 Analysing chromatin behavior on exons in human CD4+ T-cells

In this part we analysed how three types of information sources behaved in the genomic vicinity of human exons:

- human nucleosome maps generated by MNaseSeq expriments in CD4+ T-cells (Schones et al., 2008).
- human histone modification maps genrated by MNaseSeq expriments followed by antibody treatment against various histone modifications (Barski et al., 2007).
- human DNA sequences that (dis-)favor nucleosome positioning.

These results were published in the following publication



Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcárcel and Guigó R.
"Nucleosome positioning as a determinant of exon recognition."
Nat Struct Mol Biol., 16: 996–1001, 2009.

Material corresponding to this publication can be accessed at:

- full text and pdf:

  http://www.nature.com/nsmb/journal/v16/n9/full/nsmb.1658.html

- supplemental material:

  http://www.nature.com/nsmb/journal/v16/n9/suppinfo/nsmb.1658_S1.html

Supplemental Information for:

Nucleosome Positioning as a Determinant of Exon Recognition

Hagen Tilgner[1§], Christoforos Nikolaou[1§], Sonja Althammer[1], Michael Sammeth[1], Miguel Beato[1], Juan Valcárcel[1,2], Roderic Guigó[1*]

1 Center for Genomic Regulation, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

2 Institució Catalana de Recerca i Estudis Avançats

§ equal contribution

* Correspondence to:

Roderic Guigó

Bioinformatics and Genomics Program

Center for Genomic Regulation (CRG)

C/ Dr. Aiguader 88

08003 Barcelona, Spain

Telephone number: +34 93 3160110

FAX number: +34 93 3969983

e-mail: roderic.guigo@crg.cat

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1658

3.2 From co-transcriptional splicing to alternative splicing and chromatin changes: An ENCODE view

In this part we analysed RNAseq and ChipSeq data within the ENCODE project to show that

- co-transcriptional splicing appears to the rule in humans.
- alternative exon inclusion is frequently accompagnied by chromatin changes

Genomic analysis of ENCODE data shows the prominence of co-transcriptional splicing in human cell lines and suggests a weak but very widespread role of chromatin organization in alternative splicing.

Hagen Tilgner, David González, João Curado, Sarah Djebali, Angelika Merkel, Rory Johnson, Camilla Iannone, Andrea Tanzer, Tobias Warnecke, the ENCODE consortium, Juan Valcárcel and Roderic Guigó

Centre de Regulació Genòmica, Dr. Aiguader 88, E-08003, Barcelona, Catalonia, Spain.

Introduction

Whether an alternative exon is included or excluded in a mature RNA is considered a matter of combinatorial control, involving splice sites, additional binding sites and the factors that recognize them[1]. Recently evidence accumulated that this combinatorial control might actually involve more factors than previously anticipated. First, splicing can occur co-transcriptionally[2,3] and this mode of intron removal seems to be the most prominent one in yeast[4]. Second, some splicing factors are known to interact with modified histone tails[5] and intragenic histone modifications have been shown experimentally to be involved in alternative splicing decisions on single genes[6,7]. Third, a wave of recent publications in 2009 showed that chromatin structure related to exon-intron structure[8-13] genome wide. Finally lysine 36 trimethylation and lysine 4 methylation states have been linked to the control of alternative exons through PTB[14], thereby offering a mechanistic explanation of how histone modification states can influence splicing outcome. Here, analyzing data, generated within the ENCODE project[15], we show that co-transcriptional splicing appears to be very widespread suggesting that chromatin structure could also play a role in the splicing of a larger number of exons. Using RNAseq data from multiple cell lines we define alternatively skipped exons leading to 400-1000 well defined skipping events per cell type comparison. These events may involve multiple exons. Comparing chromatin changes between multiple cell lines on these alternative exons, we find that chromatin structure appears to vary differently depending on whether alternative exons increase or decrease inclusion from one cell type to

another. Especially H3k9ac and H3K4me1,2 changes appear to be associated to inclusion level changes.

Results

R0. Co-transcriptional splicing appears to be widespread and has a tendency to occur in 5´to 3´order on a genome wide level

We analyzed chromatin-associated long total RNAseq reads mapped to the human genome and to exon-exon junctions in the K562 cell line. For ~30.000 suitable internal exons (Methods) we counted the number of reads mapping to the genome but overlapping the exon-intron junction (which argue for an exon in an RNA molecule that has not yet been spliced), as well as the number of reads linking the exon to another exon of the same gene or linking an upstream exon to a downstream exon (both of which argue for splicing of the exon). Based on these numbers we defined a ratio expressing the fraction of completed splicing of this exon (Methods). This ratio was termed “completed Splicing Index” and abbreviated as coSI (fig1). A very large portion of exons exhibited unspliced reads (that is a coSI value <1, fig 2a). Comparing the exons of two genes that are well studied examples of co-transcriptional splicing (the fibronectin[16-18] and c-src[18] genes) to exons of other genes, we found exons of the fibronectin and c-src genes to show significantly higher coSI values than exons of other genes (fig 2b, left and middle boxplot). Furthermore exons of these two genes also showed higher coSI values than exons of the 4000 longest genes we found in the gencode-v3c annotation (Fig 2b, middle and right boxplot).

Moreover coSI values increased with increasing distance from the polyA/cleavage site (fig 2c), but decreased from exon 3 on progressively with exon order until exon 9 (fig 2d) in genes with at least 10 exons. 94.5% (6397 out of 6771) genes showed at least one exon with at least four spliced and four unspliced reads and in the cytosolic fraction only 51.9% (3777 out of 7281) genes verified this statement. We then repeated the same analysis with RNAseq-data from the long nuclear polyA-, long nuclear polyA+ and long cytosolic polyA+ fractions (fig S1 to fig S3). The characteristics observed in the chromatin fraction (fig 2) were present but less strongly in the nuclear polyA- fraction, supposedly because both fractions contain a large number of transcripts for which transcription has not been completed (Fig2, S1). The nuclear polyA+ fraction contained much lower numbers of unspliced reads (Fig S2), although some polyadenylated transcripts do contain unspliced introns. In the cytosolic fraction for none of the investigated exon subsets the median coSI value descended below 0.997 (comparing to median coSI values as low as 0.698 in the chromatin fraction), indicating that in the cytosolic RNA basically all exons are spliced (fig S3). Note that even in the cytosolic RNA sample, we observe differences in coSI values depending on exon order and distance to the polyA-site. This may originate from small numbers of cells undergoing mitosis among the cells used for fractionation, which would introduce a small fraction of chromatin associated RNA into the cytosolic fraction. Replacing unspliced reads by numbers of CAGE reads overlapping an exon in the coSI calculation for the chromatin fraction ruled out the possibility that unspliced reads mainly originate from novel, not annotated TSS (Methods, fig S4). In fact for this coSI-CAGE value some of the previously described trends were inversed (fig2, S4). Correlation analysis

confirmed that the two replicates of the chromatin fraction were more similar to each other than to any other fraction in terms of coSI values (fig 3a). Furthermore the chromatin fraction showed a stronger correlation with the nuclear polyA- fraction than with the nuclear and cytosolic polyA+ fractions (fig3b-d), supposedly because the former two fractions contain a large number of molecules that have not yet been polyadenylated and undergo splicing.

Considering 3603 "suitable" genes (Methods), we calculated gene-based coSI-values (Methods) in the chromatin, nuclear polyA-, nuclear polyA+ and in the cytosolic polyA+ fraction. Importantly in this analysis every gene contributes equally, although shorter genes tend to have less splicing events. The median gene coSI-value in the chromatin fraction and in the nuclear polyA- fractions were 0.33 and 0.47, while in the nuclear and the cytosolic polyA+ fraction the medians were 0.74 and 0.97, indicating that for at least half of all genes one third (one half respectively) of all splicing events are carried out while the transcription is ongoing (before polyadenylation occurs). Again, we wish to emphasize that the gene based view underestimates completed splicing rates, as genes with few splicing events are counted equally to those with massive splicing.

From these observations we conclude that co-transcriptional splicing (that is splicing while the RNA is still attached to Pol2 and to the chromatin template) is widespread. An alternative interpretation, that we think is unlikely, but that cannot be completely ruled out is based on the idea that the reads we consider as unspliced reads actually originated from novel, unannotated alternative upstream acceptors or alternative downstream donors ("novel outer splice sites"). In this case one would have to assume that such novel splice sites occur

very often in transcripts for which transcription is still continuing, but less for transcripts that do not (yet) have a polyA-tail. Furthermore such transcripts would be only rarely polyadenylated and those that are polyadenylated would not be exported to the nucleus. Moreover the fibronectin and the c-src gene would have higher proportion of these novel "outer" alternative splice sites as compared to all other genes and as compared to the 4000 longest genes. Finally one would have to assume also that the novel "outer" alternative splice sites occur more often as one approaches the polyA-site.

R1. Exons in different coSI bins show different splicing determinants

Splicing determinants are currently assumed to involve three distinct layers of signals in the DNA template: First, acceptor and donor sequences; Second, binding sites for auxiliary splicing factors (ESE, ESS, ISE and ISS) and third, chromatin organization, although it is not yet clear how general a mechanism the latter is. We therefore monitored how these three characteristics behaved in 10 different bins of exons according to coSI values (Methods). Significantly different splice site scores were observed in bin 2 and 9 (figS5 $p < 0.015$, two-sided Wilcoxon rank sum test), although these bins were not the ones showing the strongest differences. Furthermore exons in bins 2 and 9 were different in terms of binding sites for splicing factors (e.g. exonic hnRNPA1 binding sites, Fig S6a, upstream SF2/ASF binding sites, FigS6b, and downstream hnRNPA0 binding sites, FigS6c, to name only a few, Methods). The tendency of coSI values to increase with increasing distance from the polyA-site suggested that exons with different coSI values could also be placed in different chromatin organizations, since chromatin organization has shown to be changing along the body of

expressed genes[19]. Indeed exons in different coSI-bins also showed different values for a variety of chromatin marks and Pol2 occupancies (FigS7, Methods). These observations were significant but in theory compatible with positional effects within a gene. We therefore employed an approach, that dissects the influence of distance to the TSS and polyA-site and the above chromatin characteristics represented by 9 variables (pure Mnase digestion data and Pol2, K36me3, K27me3, K4me1,2,3, K9ac and H4k20me1): Using these 11 variables and focusing on a smaller exon set that was perfectly mappable for 36mers (Methods), a decision tree[20] was built predicting whether an exon belonged to a low coSI class (bins 1-3) or to a high coSI class (bins 8-10). The resulting decision tree (Fig S8) showed that although distance to the polyA site and to the TSS are important variables contributing to the differential chromatin observations on exons, they do not cause this phenomenon entirely. Indeed chromatin associated variables remained significant predictors of the coSI class of an exon.

R2. Sub-cellular expression patterns of spliceosomal RNAs

If splicing occurs mostly co-transcriptionally and therefore in proximity to the chromatin template, one would expect that the splicing machinery would also be visible in proximity to chromatin. We therefore investigated the sub-cellular location of U1-U6, U6atac and U12 snRNAs (Methods) based on small RNAseq datasets in five different sub-cellular locations. UxRNAs were mainly found in the chromatic and in the nucleoplasmic fraction (Fig4a), showed some expression in nucleoli and entire nuclei, but were almost absent in the cytosolic fraction. Ribosomal RNAs however showed a different pattern with the strongest expression in the cytosolic and nucleoplasmic fractions and some expression in the chromatic fraction (Fig4b). Also snoRNA expression showed a different

expression pattern, with strongest expression in nucleoli, but also in the chromatic and nucleoplasmic fractions (Fig 4c). We wish to stress that mapping small RNAseq data is complicated, because many small RNAs are often generated from multiple genomic locations. Comparisons of two different sets of small RNAs might therefore be unsafe; yet comparing the same small RNA set in different fractions should not be problematic, as equal limitations concern all fractions.

R3. A large number of skipped exons in pair-wise cell type comparisons

We made use of a method to call alternatively skipped exons similar to that published by Wang and co-workers[21] (Methods, figS9). We then applied this method to long nuclear polyA+ RNAseq samples, as instances of splicing that have not yet gone through NMD or export filters and thereby represent splicing events as carried out in the nucleus. Initially the K562 and Gm12878 cell lines were used for investigation. Retaining only one exon per gene for each direction of inclusion changes, we found a total of 641 (461 changing towards more inclusion in Gm12878 - referred to as "upregulated exons" - and 180 changing towards more exclusion in Gm12878 – referred to as "downregulated exons" in the following) exons that showed significantly changed inclusion in Gm12878 as compared to K562. Estimation of inclusion level changes (Methods) was contradiction free with the calling of up- and downregulated exons (Fig5a Methods).

The skipped exons thus defined behaved consistently to properties of alternative exons described in the literature[22,23]. First they showed lower splice site strengths compared to exons that were not called alternatively spliced (Figure

5b). Second they tended to be shorter (Fig 5c) and when they were coding exons of a gene where only one skipped exon was found their length was more often divisible by three (Fig 5d). CAGE data suggested that the genes associated to the two sets of exons (changing towards more inclusion and towards less inclusion in Gm12878 as compared to K562) do not show significantly different gene expression changes (Methods, Figure 5e). Similarly the set of upregulated and downregulated exons did not differ significantly in terms of mappability (Fig5f, using GEM-mappability). Based on "splicing rainbow" and ESE-finder splicing factor binding sites we furthermore monitored the occurrence of such binding sites within the exon (Fig 5g), upstream of the exon (fig 5h) and downstream of the exon (fig 5i). For each "stair arrangement", that is a splicing factor for which upregulated exons had a binding site more often than non-AS-exons which in turn had a binding site more often than downregulated exons (or vice-versa), we computed a p-value, controlling for multiple testing in the Bejamini-Hochberg sense (Methods). For K562 and Gm12878 hnRNPA1 and hnRNPG binding sites seemed to make the strongest differences within exons (fig 5g). For exonic SF2/ASF-binding sites both utilized matrices did not agree (fig 5g). Upstream only SRp20 made a difference (Fig5h), while downstream binding sites did not appear to segregate between up- and downregulated exons. Exonic binding of hnRNPs is assumed to contribute to exon skipping. Consistent with a higher fraction of downregulated exons (in Gm12878 as compared to K562) exhibiting exonic hnRNPA1 binding sites (Fig 5g), we found hnRNPA1 to be upregulated in Gm12878 as compared to K562 (cytosolic RPKM of 231 in Gm12878 instead of 193 in K562, representing an upregulation of ~20%). In the opposite way the lower fraction of downregulated exons containing an exonic hnRNPG site (Fig 5g,

row 17) was consistent with a 3fold downregulation of hnRNPG (cytosolic RPKM of 0.0045 in Gm12878 instead of 0.014 in K562). For SRp20 the detected change in cytosolic RPKMs was 3% (100.9 in Gm12878 instead of 97.9 in K562), which we interpret as no change.

In the following we repeated this analysis for the 14=6*5/2 -1 pair-wise cell type comparisons. Table S1-S3 show the numbers and characteristics of skipped exons per cell type comparison. Cell type comparisons with larger numbers of differentially skipped exons also tended to yield more splicing factors whose binding sites segregated between up-and down-regulated exons (Fig S10-S23). For some cell type comparisons few characteristics of alternative exons investigated above (see Fig5b to 5e for the K562-Gm12878 comparison) did not hold (Fig S10 to S23). There were however two more cell type comparisons (Gm12878 vs. HelaS3, figS14 and Helas3 vs. Hepg2, figS18) for which all investigated characteristics held true. Subsequently we defined for every alternative exon and cell type comparison an upstream and a downstream exon that was not labeled as alternative in this cell type comparison.

R4. Changing chromatin signals on differentially included exons:

We then compared changes in chromatin signature on up- and downregulated exons. The counts described above were normalized for sequencing depth and fragment length. Positions were then aligned to the acceptors of alternative exons and the signal was averaged over all exons in each class and normalized by sequencing depth and fragment length (Methods). Finally we computed the differential chromatin signal for each cell type comparison and chromatin type separately, as the difference of the normalized, averaged signal in the two cell

types. In the K562 vs. Gm12878 comparison H3k4me2, H3k9ac, H3k36me3, Pol2 and nucleosomes showed significantly higher changes on upregulated exons than on downregulated exons, (fig 6b for H3k9ac) while the changes were not significantly different on non alternative up-and downstream exons (Methods, fig6a,c for H3k9ac). For H3k36me3 however the change on the upstream and downstream exon appeared also higher for the upregulated exons, graphically speaking, even though it was not significant (data not shown). H3k27me3 showed very different results (fig S24). These results did not appear to stem from cell type specific deletions of the genomic sequence: When combining all reads from all experiments, each exon had at least 8 reads overlapping this exon with at least 15bps in K562 and at least 12 such reads in Gm12878. Table S4 summarizes the results for all histone modifications in the three investigated cell type comparisons. Making use of the two other cell type comparisons (Gm12878 vs. HelaS3 and HelaS3 vs. HepG2, where nucleosome and Pol2 data was not available), we found that only H3k9ac showed similar significant results on the alternative exons (Table S4). For the comparison between HelaS3 and HepG2, significance was however only present when using a Bejamini Hochberg correction, but not when using a Bonferoni correction. In addition, for HelaS3 vs. HepG2 H3k4me3 showed interesting significant results on the alternative exons but also on the upstream exon. For the Gm12878 vs. HelaS3 comparison, H3k4me2/3 both showed significant differences on alternative exons, but not on the up-and downstream exons.

Discussion

Co-transcriptional splicing is known to be widespread in yeast and has been documented in higher eukaryotes for a number of genes, including very widely studied genes, such as the fibronectin and the c-src genes. While we cannot quantify how prominent co-transcriptional splicing exactly is, our results clearly demonstrate that co-transcriptional splicing is prominent enough to leave strong traces in genome wide data sets. Consistently, considerable amounts of spliceosomal RNAs can be found in chromatin. Therefore co-transcriptional intron-removal should be considered the rule rather than the exception, also in higher eukaryotes. Importantly our data does not allow us to see when a splice site is committed to splicing, but rather when it is spliced. Exons we see unspliced might already have been committed to splicing, but their adjacent introns have not yet been removed. Therefore transcription-mediated influences on splicing are probably even larger than what we can see with the data, presented here.

A variety of recent studies have linked chromatin structure to splicing. Co-transcriptionality of splicing is not an absolute prerequisite for a chromatin-splicing connection, because chromatin organization could very well link to splicing commitment co-transcriptionally, while intron-removal might occur only post-transcriptionally. However co-transcriptional intron removal opens the door for chromatin influences on splicing even wider. Building on our observation that co-transcriptional intron removal is very widespread in humans, we decided to investigate chromatin changes that co-occur with differential exon inclusion between human cell types. Using nuclear polyA+ RNAseq data from RNA molecules longer than 200nts, we defined differentially

skipped exons in 15 pair-wise cell type comparisons, leading to between 500 and 1100 cases that are suitable for chromatin investigation in each cell line. When comparing chromatin changes on exons that decreased inclusion and increased inclusion from one cell type to another, H3k9ac showed a significant difference in all three investigated cell type comparisons. For one of these cell type comparisons, however, only an FDR-control showed significance, while a Bonferoni approach did not. Lysine four di-and trimethylation also showed such behavior in one and two cell type comparisons. These results suggest that chromatin influences on alternative splicing decisions are very widespread although weak.

In summary we believe that these results show that chromatin organization and chromatin changes need to be considered, in order to understand splicing and alternative splicing at a large scale.

Methods

RNAseq mapping

RNAseq reads were mapped to the genome and to all ordered exon pairs within the same gene, using RAP (D. González et al, will be described elsewhere), allowing two mismatches. For each exon ex in the gencode v3c annotation four different numbers were counted:

1. JIR(ex): the number of reads mapping to a junction linking ex to another exon of the same gene
2. JER(ex): the number of reads linking an upstream exon of ex to a downstream exon of ex
3. NSR(ex): the number of reads mapping to the genome with at least 4bps inside the exon and at least 4bps outside of the exon
4. EIR(ex): the number of reads mapping to the genome and entirely within the exon ex.

Exon coSI definition

The completed splicing index of the exon ex was then defined as:

coSI(ex)=(0.5*JIR(ex)+JER(ex))/ (0.5*JIR(ex)+0.5*JER(ex) + NSR(ex))

By definition this index is a rational number between 0 and 1, a value of 0 indicating that all the reads mapping to junctions and exon borders of this exon are unspliced. A value of 1 on the other hand indicates that all these reads indicate splicing.

Exon set used for coSI values

Using the gencode v3c and the UCSC annotation (downloaded from the UCSC browser on October 6th 2010) we determined all exons

1. that were internal in all transcripts they appeared in, in both gene annotations

2. that were not overlapped by any non-identical exons in both gene annotations. Identity of exons is defined by their location (chromosome, start, end strand)
3. that were separated by at least 70nts from any other exon
4. that were at least 75bps long and at most 450
5. for which JIR(ex)+JER(ex) + NSR(ex)>=10 in the given RNAseq mapping set (this criterion is specific to whatever RNAseq data set is used)
6. that had an AG acceptor and a GT donor
7. for which all 75mers in a 600bps surrounding around the acceptor were mappable
8. that were part of transcripts annotated as protein coding, NMD, processed transcript, non-coding, retained intron, ambiguous ORF, antisense in the gencode v3c annotation.

For the first replicate of the chromatin fraction this resulted in 30748 exons.

coSI-CAGE analysis

For each exon ex the number of CAGE reads overlapping the exon was counted and termed as cage(ex). This number was used instead of the number NSR(ex) in the coSI calculation, in order to assess whether the observations made in the coSI analysis were due to undiscovered TSS. In order to guarantee that the coSIcage ratio could be defined for all of the above exons, a pseudocount was introduced, leading to the following formula:

coSIcage(ex)=(0.5*JIR(ex)+JER(ex)+1)/ (0.5*JIR(ex)+0.5*JER(ex) + cage(ex)+1)

Exon set used for decision tree analysis

Of the previous 30748 exons we chose those that had fell into the lowest three coSI bins (that is having a coSI value <=0.63146) and those that fell into the highest three coSI bins (that is a coSI value of at least 0.839506). We then excluded all exons whose gene has multiple TSS or polyA-sites in the gencode v3c annotation or that had a 36mer that was not mappable within 450bps of the acceptor. This resulted in 1383 exons with "low" coSI values and 1054 exons with "high" coSI values.

ChipSeq counts for histone modifications, polymerase and nucleosomes

ChipSeq reads for each dataset were extended to the full fragment length (147bps for nucleosomes, 300bps for histone modifications and 225bps for Pol2). For each position in the genome we counted the number of extended reads overlapping this position. We refer to this number of the "count of the position".

ChipSeq levels at acceptors

The "absolute level" of an acceptor is defined as the log2 of the average count in the 147bps after the acceptor in transcription direction. For exons, which had no ChipSeq counts at all in this region, a pseudocount was used, so that the "absolute level" was log2(1/147).

Splice site strength measure

For each exon we used maxEnt[24] in order to calculate an acceptor score and a donor score and represented the "exon strength" by the sum of these two scores.

Small RNA data sets

We determined all genomic loci given in the gencode v3c annotation for the three sets of small RNAs as follows.

1. UxRNAs: all 1392 geneids were retrieved for which the gene name was among U1,U2,U3,U4,U5,U6, U6atac and U12.
2. rRNAs: all 455 geneids were retrieved for which the gene name contained the word “rRNA”. This resulted in 425 geneids for the 5s_rRNA, 6 geneids for the “5_8S_rRNA” and 24 geneids for the “SSU_rRNA_5”
3. snoRNAs: all 697 geneids were retrieved for which the gene type contained the word “snoRNA” but none of the above UxRNAs.

Gene coSI definition

In order to define the completed splicing index of a gene g, we counted for each gene the number of distinct splice sites (“nSPL(g)”). Reads spliced between two of such sites were counted as spliced reads for this gene (“JR(g)”). Similarly we counted for g the number of splice sites that appear as internal exon-ends in the projection of g onto the genome (“nNSP(g)”). Reads overlapping these sites with at least four bps on both sites were counted as unspliced reads of the gene (“NSR(g)”). The coSI for the gene g was then defines as

coSI(g)=(JR(g)/ nSPL(g))/ ((JR(g)/ nSPL(g)) + (NSR(g)/ nNSP(g)))

Again by definition this index is a rational number between 0 and 1, a value of 0 indicating that all the reads mapping to junctions and exon borders of this gene are unspliced. A value of 1 on the other hand indicates that all these reads indicate splicing.

Gene set used for coSI values

Using the gencode v3c we determined all genes

1. that were on chromosome 1-22 and X
2. that had at least 1 splice site as defined above (nSPL(g)>=1) and at least 1 region that can indicate uncompleted splicing (nNSP(g) >=1)
3. for which JR(g)+ NSR(g)>=100 in all four considered fractions
4. whose exons did not overlap the exons of another gene
5. for which all non-overlapping exons were spaced by at least 70bps

This resulted in 3604 genes.

RNAseq mapping and RPKM calculation for small RNAs

For each sub-cellular compartment and gene in the gencode v3c annotation and gene RPKM was calculated using RAP (D. González et al, unpublished) and 36bp-reads. The RAP (RNAseq Analysis Pipeline) will be described elsewhere.

Alternatively skipped exon calling

Again using the gencode v3c and the UCSC annotation mRNAs we determined all exons

4. that were internal in all transcripts they appeared in, in all three gene annotations
5. that were not overlapped by any non-identical exons in the three gene annotations. Identity of exons is defined by their location (chromosome, start, end strand)
6. that were at least 50bps long and at most 450
7. surrounded by AG-GT splice sites

8. located on chr1-22 and X

Then for a given cell type comparison (here we use K562 and Gm12878 as an example) only exons with minimal AS evidence were retained, that it (JIR_{K562}(ex)>=1 AND $JER_{\_Gm12878}$>=1) OR (JER_{K562}(ex)>=1 AND $JIR_{\_Gm12878}$>=1).

For the remaining N=64103 exons a two by two table was constructed containing junction inclusion reads and junction exclusion reads in the two cell types. Two one-sided fisher tests were run and corrected for multiple testing for the N tests in the Benjamini-Hochberg sense, resulting in three disjoint sets of exons:

1. exons that are significantly more included in Gm12878 (which will be referred to as “upregulated”, even though the choice of the direction from K562 to Gm12878 is clearly arbitrary)
2. exons that are significantly less included in Gm12878 (which will be referred to as “downregulated”)
3. exons whose inclusion is not significantly changed between the two cell types (which will be referred to as “non-AS exons” for the sake of conciseness and clarity although “non-significant AS exons” would be more correct.)

We then chose a subset of them for comparison with chromatin changes using the following criteria. For up- and downregulated exons we then zoomed in on the subsets that

i. were at least 600nts away from any annotated TSS or TTS
ii. had between 50 and 450bps

iii. for which the CAGE values of the associated gene did not change more than 10fold between the two cell types

iv. that had at least 1 JIR or at least 1 read mapping entirely to the exon in each cell type.

v. for which at least 75% of all positions in a 900bp window around the acceptor were uniquely mappable.

vi. whose inclusion changed by at least 0.1 or two-fold.

Frequently genes contained more than one alternatively spliced exon thus defined. In order to avoid gene specific characteristics that might be introduced into these sets by genes that contribute many alternative exons (as for example the TTN gene where 212 exons passed the fisher test), we chose in both subsets separately for every gene the exon that whose p-value was most significant.

For non-AS exons a similar procedure was carried out, removing however the "inclusion changed by at least 0.1 or two-fold"-criterion and choosing the exon per gene whose estimated inclusion change was minimal among all non-AS exons of that gene (instead of the exon with the smallest p-value). Fig S9 illustrates this approach.

CAGE gene expression values

For each TSS the number of CAGE-read 5´ends mapping on the same strand and within 50bps were counted and normalized by sequencing depth (in millions of reads). For each gene its expression as estimated by CAGE was averaged over all TSS of this gene and over both considered replicates.

Mappability calculation

Mappability for the hg19 genome was calculated using the GEM-mapper for 36bp and 75bp reads. For each acceptor in the genome, mappability was represented in transcription direction.

Binding site analysis

Published binding sites for RNA binding proteins including hnRNPs[25] and SRproteins [25-27] were collected from previous collections, stemming from a variety of sources. Some "additional RNA binding sites" were taken from http://esrsearch.tau.ac.il/[28] and added to the previous lists if they were not already in one of the previous lists. For each exon we monitored absence and presence of a binding site for a given RNA binding protein in three different regions:

i. within the exon
ii. within the region between 140 and 41bps upstream of the exon
iii. within the region between 11 and 110bps downstream of the exon

References

1. Smith, C.W. & Valcarcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci* **25**, 381-8 (2000).
2. Beyer, A.L. & Osheim, Y.N. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev* **2**, 754-65 (1988).
3. Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J. & Nogues, G. Multiple links between transcription and splicing. *RNA* **10**, 1489-98 (2004).
4. Carrillo Oesterreich, F., Preibisch, S. & Neugebauer, K.M. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Mol Cell* **40**, 571-81.
5. Sims, R.J., 3rd et al. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. *Mol Cell* **28**, 665-76 (2007).

6. Schor, I.E., Rascovan, N., Pelisch, F., Allo, M. & Kornblihtt, A.R. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. *Proc Natl Acad Sci U S A* **106**, 4325-30 (2009).
7. Allo, M. et al. Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nat Struct Mol Biol* **16**, 717-24 (2009).
8. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* **16**, 990-5 (2009).
9. Tilgner, H. et al. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16**, 996-1001 (2009).
10. Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C. & Komorowski, J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res* **19**, 1732-41 (2009).
11. Nahkuri, S., Taft, R.J. & Mattick, J.S. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. *Cell Cycle* **8**, 3420-4 (2009).
12. Spies, N., Nielsen, C.B., Padgett, R.A. & Burge, C.B. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell* **36**, 245-54 (2009).
13. Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol* **5**, e1000566 (2009).
14. Luco, R.F. et al. Regulation of alternative splicing by histone modifications. *Science* **327**, 996-1000.
15. The ENCODE Project: ENCyclopedia Of DNA Elements.
16. Cramer, P., Pesce, C.G., Baralle, F.E. & Kornblihtt, A.R. Functional association between promoter structure and transcript alternative splicing. *Proc Natl Acad Sci U S A* **94**, 11456-60 (1997).
17. Kadener, S. et al. Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing. *EMBO J* **20**, 5759-68 (2001).
18. Pandya-Jones, A. & Black, D.L. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15**, 1896-908 (2009).
19. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-37 (2007).
20. Horthorn, T., Hornik, K. & Zeileis, A. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* **15(3)**, 651-74 (2006).
21. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-6 (2008).
22. Zheng, C.L., Fu, X.D. & Gribskov, M. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA* **11**, 1777-87 (2005).
23. Magen, A. & Ast, G. The importance of being divisible by three in alternative splicing. *Nucleic Acids Res* **33**, 5574-82 (2005).
24. Yeo, G. & Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-94 (2004).
25. Stamm, S. et al. ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res* **34**, D46-55 (2006).

26. Smith, P.J. et al. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet* **15**, 2490-508 (2006).
27. Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q. & Krainer, A.R. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res* **31**, 3568-71 (2003).
28. Goren, A. & Ast, G. ESRsearch. (2006).

Figures

Hagen Tilgner

April 5, 2011

fig1

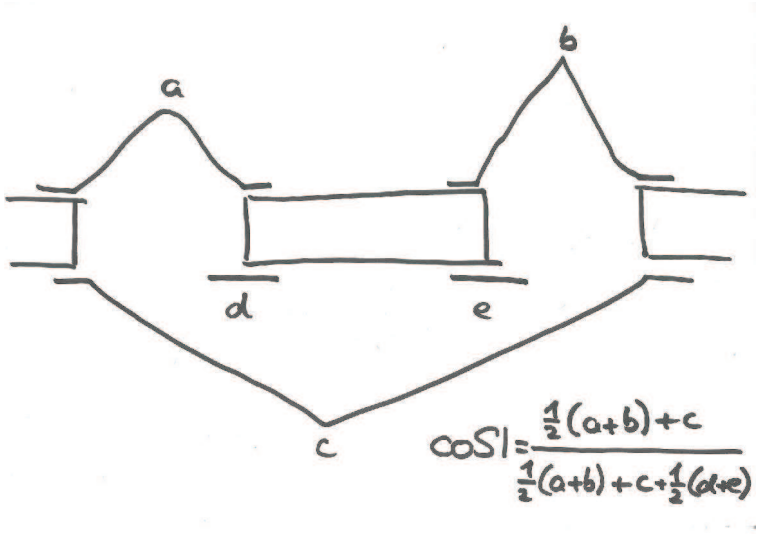


Fig1: coSI

fig2 and S1-S3

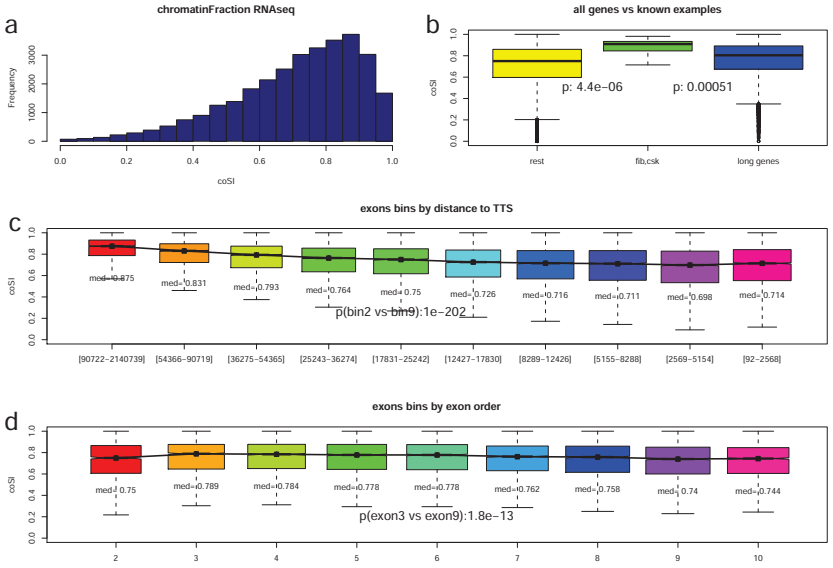
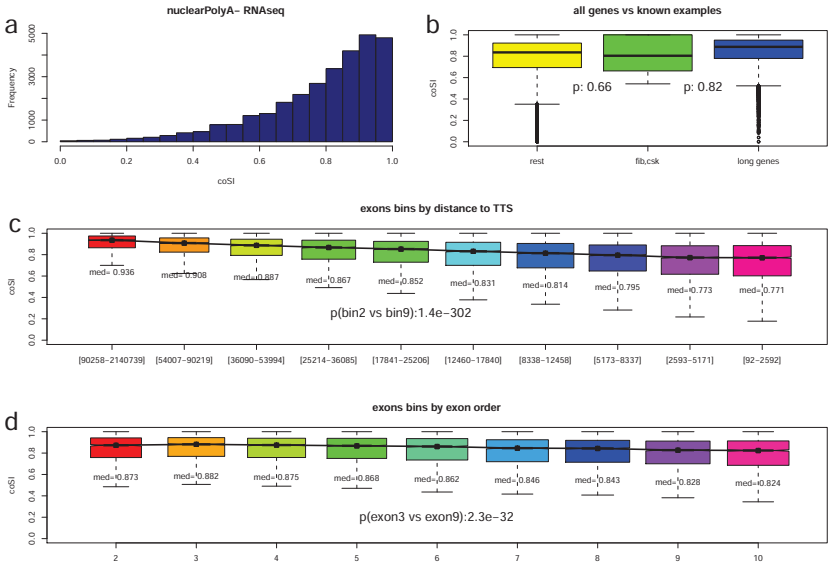
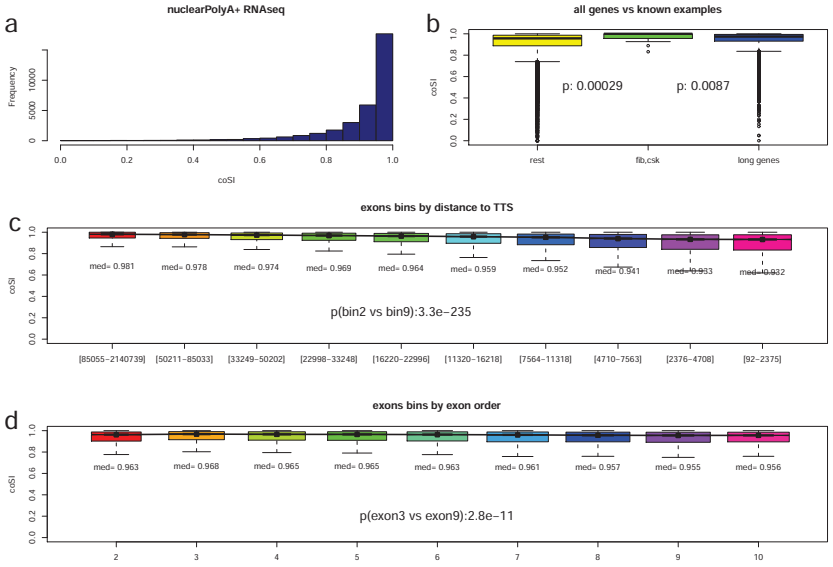


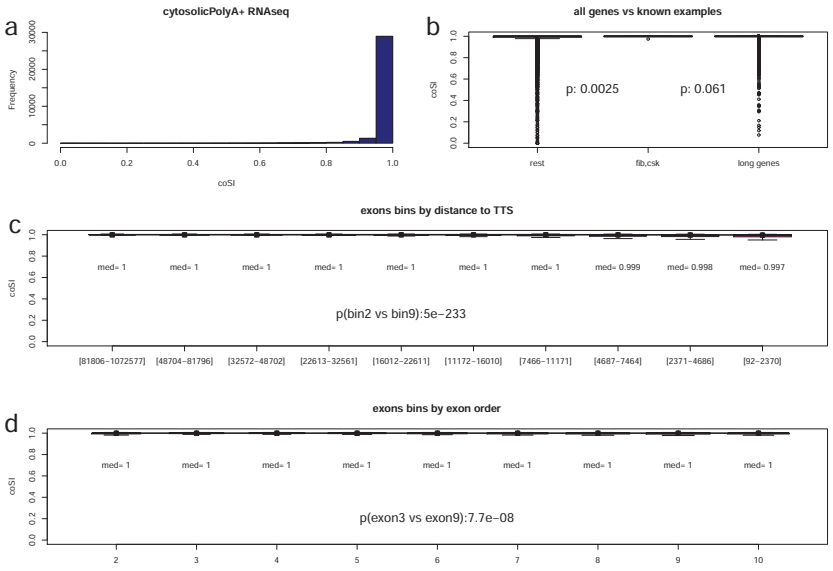
Fig2: spliced and unspliced in the chromatin fraction



FigS1: spliced and unspliced in the nuclear polyA- fraction

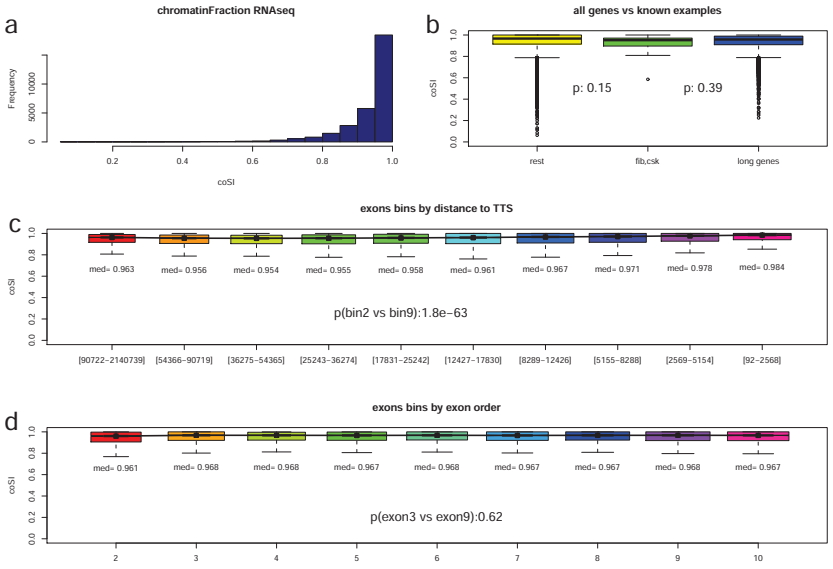


FigS2: spliced and unspliced in the nuclear polyA+ fraction



FigS3: spliced and unspliced in the cytosolic fraction

figS4



FigS4: spliced and CAGE reads in the chromatin fraction

fig3

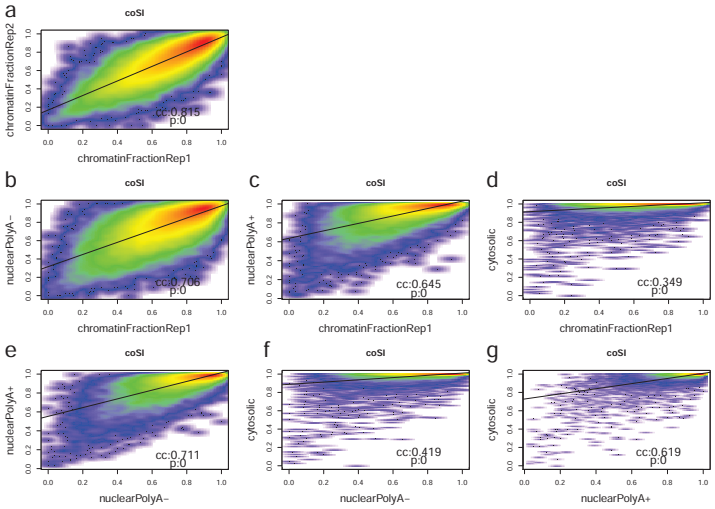
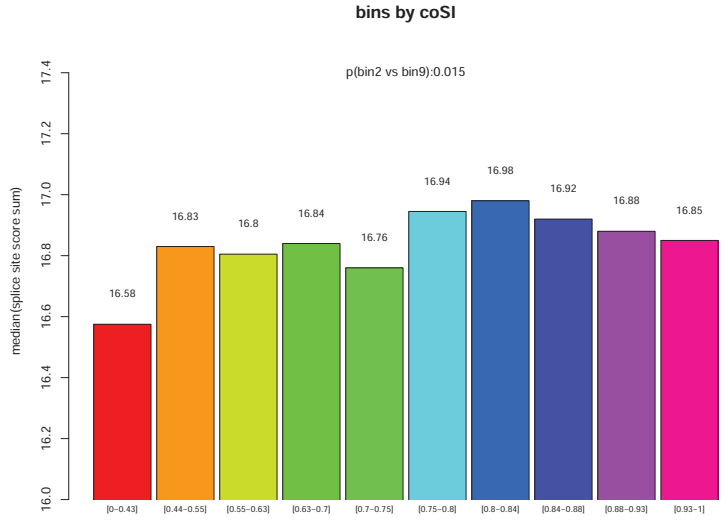
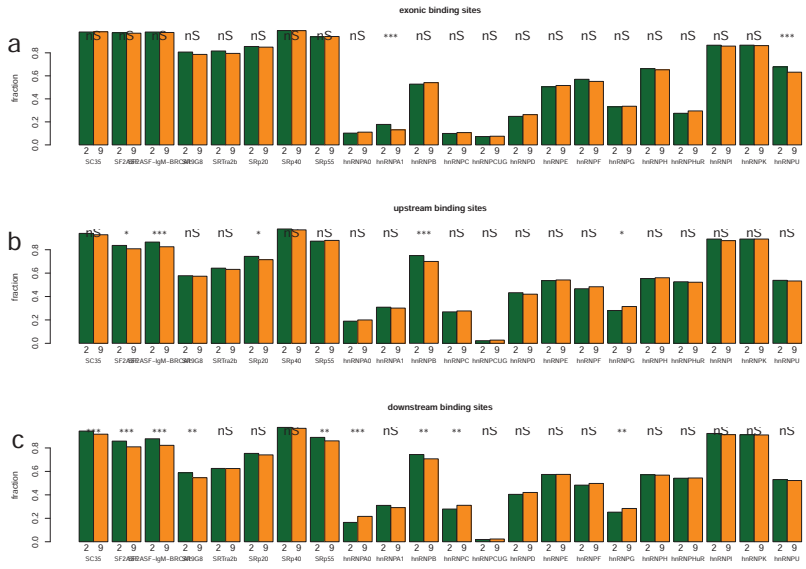


Fig3: coSI exon correlations between replicates and cell fractions

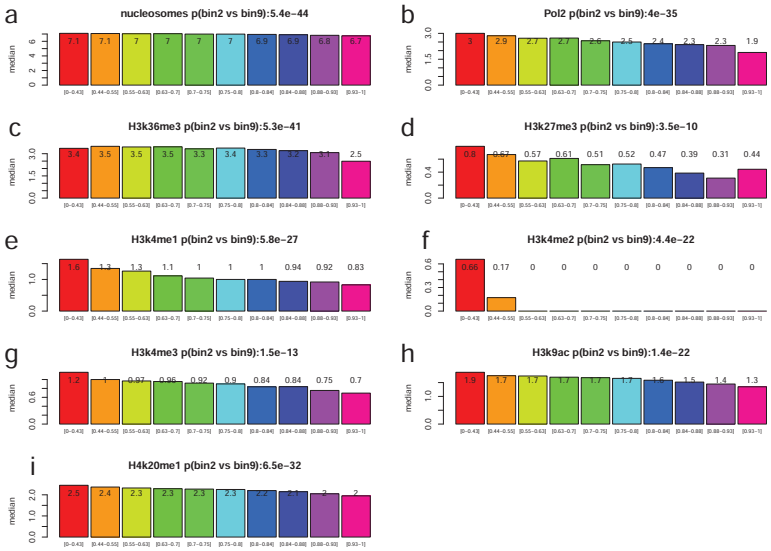
figS5-S8



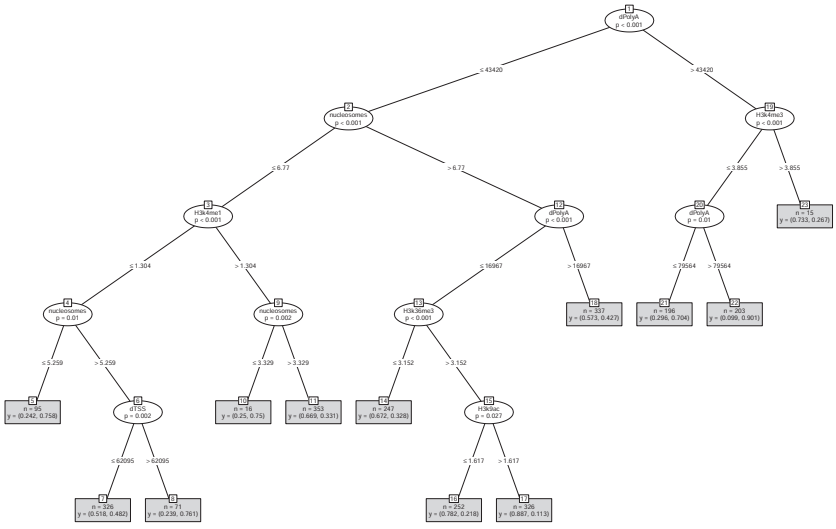
FigS5: splice site strength in different coSI bins



FigS6: bindin site analysis between coSI bins 2 and 9



FigS7: chromatin bins in different coSI bins



FigS8: telling high coSI exons from low coSI exons using chromatin and position within gene

fig 4

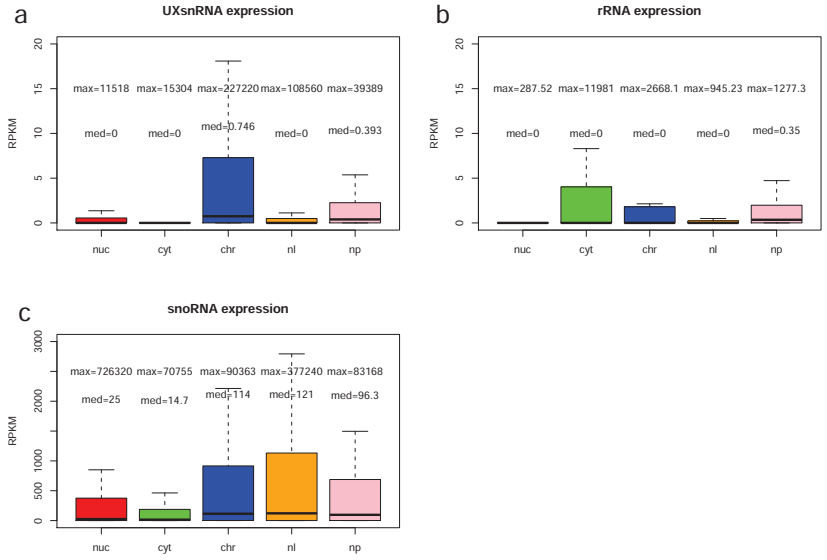
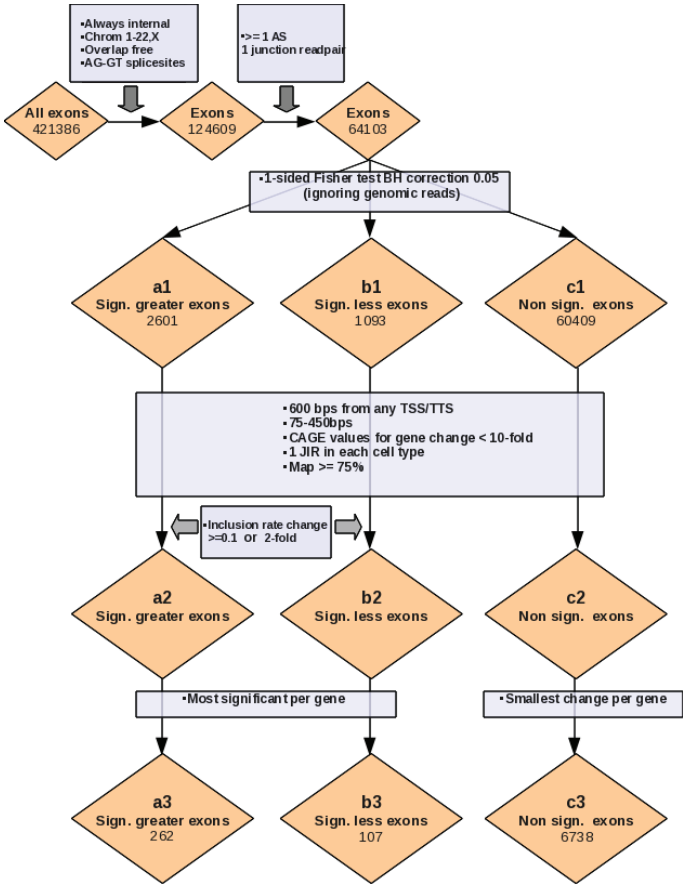


Fig4: sub-cellular expression levels of small RNAs

fig S9


•Always internal
•Chrom 1-22,X
•Overlap free
•AG-GT splicesites
•>= 1 AS
1 junction readpair
All exons
421386
Exons
124609
Exons
64103
•1-sided Fisher test BH correction 0.05
(ignoring genomic reads)
a1
Sign. greater exons
2601
b1
Sign. less exons
1093
c1
Non sign. exons
60409
•600 bps from any TSS/TTS
•75-450bps
•CAGE values for gene change < 10-fold
•1 JIR in each cell type
•Map >= 75%
•Inclusion rate change
>=0.1 or 2-fold
a2
Sign. greater exons
b2
Sign. less exons
c2
Non sign. exons
•Most significant per gene
•Smallest change per gene
a3
Sign. greater exons
262
b3
Sign. less exons
107
c3
Non sign. exons
6738

figS9

fig 5

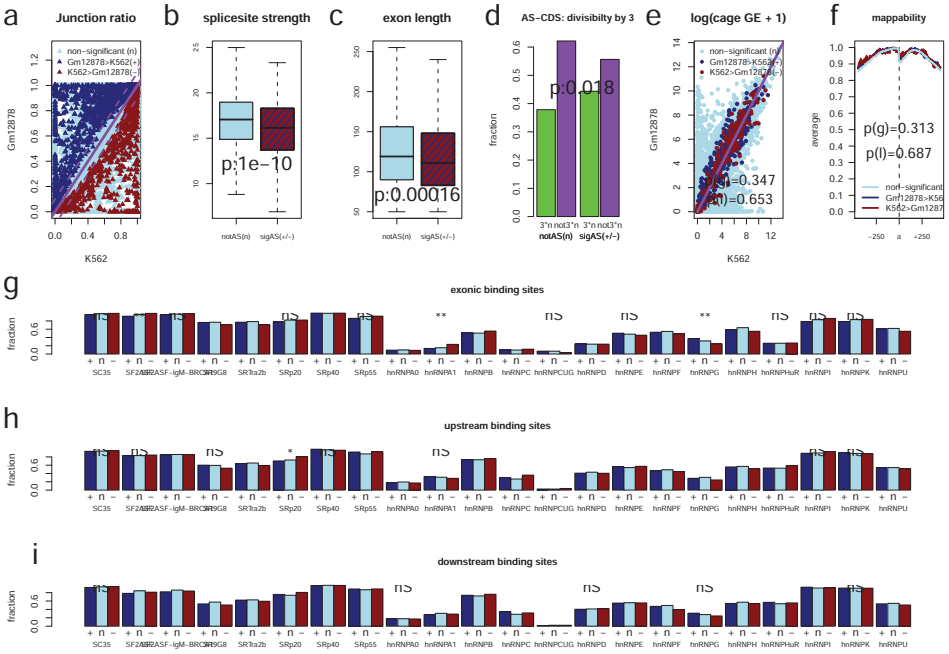


Fig5: calling AS exons: K562 vs Gm12878

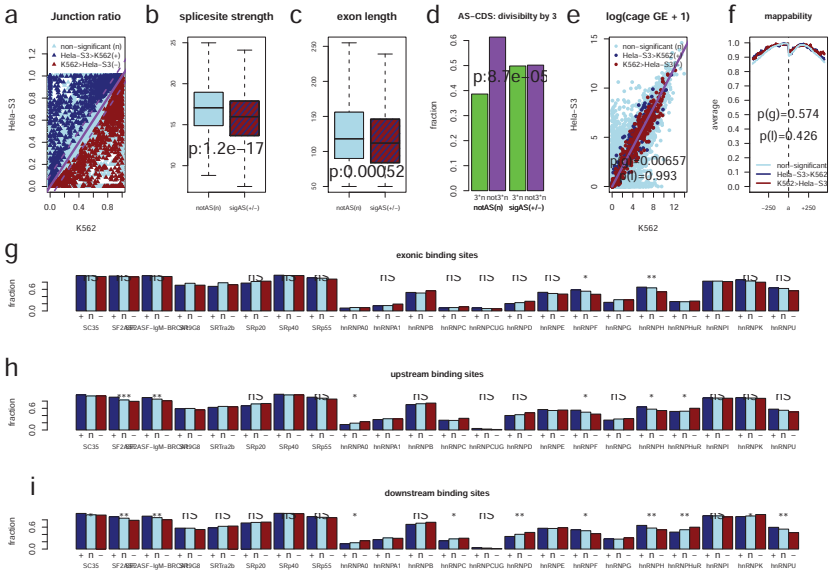
table S1-S3

pair	*up*	*down*
$K562 - Gm12878$	2601	1093
$K562 - H1hesc$	2451	1210
$K562 - HelaS3$	1837	1240
$K562 - HepG2$	2293	1351
$K562 - Huvec$	2661	2376
$Gm12878 - H1hesc$	1747	1698
$Gm12878 - HelaS3$	1935	2299
$Gm12878 - HepG2$	1342	1547
$Gm12878 - Huvec$	1408	2524
$Huvec - H1hesc$	2795	1780
$Huvec - HelaS3$	3186	2302
$Huvec - HepG2$	1855	1171
$HelaS3 - H1hesc$	1790	1385
$HelaS3 - HepG2$	1852	1693
$HepG2 - H1hesc$	1860	1428

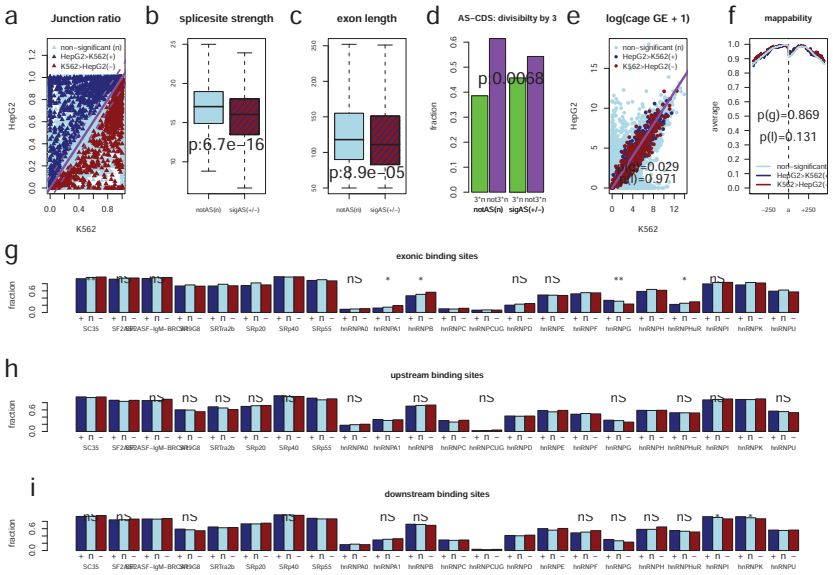
pair	*up*	*down*
$K562 - Gm12878$	461	180
$K562 - H1hesc$	379	332
$K562 - HelaS3$	284	349
$K562 - HepG2$	375	338
$K562 - Huvec$	525	491
$Gm12878 - H1hesc$	267	462
$Gm12878 - HelaS3$	242	604
$Gm12878 - HepG2$	184	381
$Gm12878 - Huvec$	260	498
$Huvec - H1hesc$	471	483
$Huvec - HelaS3$	503	601
$Huvec - HepG2$	293	289
$HelaS3 - H1hesc$	419	348
$HelaS3 - HepG2$	420	365
$HepG2 - H1hesc$	442	393

pair	*strength*	*length*	%3	*GE*	*mappability*
$K562 - Gm12878$	*yes*	*yes*	*yes*	*yes*	*yes*
$K562 - H1hesc$	*yes*	*no*	*yes*	*no*	*no*
$K562 - HelaS3$	*yes*	*yes*	*yes*	*no*	*yes*
$K562 - HepG2$	*yes*	*yes*	*yes*	*no*	*yes*
$K562 - Huvec$	*yes*	*yes*	*no*	*yes*	*yes*
$Gm12878 - H1hesc$	*yes*	*no*	*yes*	*no*	*no*
$Gm12878 - HelaS3$	*yes*	*yes*	*yes*	*yes*	*yes*
$Gm12878 - HepG2$	*yes*	*yes*	*no*	*yes*	*yes*
$Gm12878 - Huvec$	*yes*	*no*	*yes*	*yes*	*yes*
$Huvec - H1hesc$	*yes*	*no*	*yes*	*yes*	*no*
$Huvec - HelaS3$	*yes*	*yes*	*no*	*yes*	*yes*
$Huvec - HepG2$	*yes*	*yes*	*yes*	*no*	*no*
$HelaS3 - H1hesc$	*yes*	*yes*	*yes*	*no*	*yes*
$HelaS3 - HepG2$	*yes*	*yes*	*yes*	*yes*	*yes*
$HepG2 - H1hesc$	*yes*	*yes*	*yes*	*no*	*yes*
total	15/15	11/15	12/15	8/15	11/15
woH1hesc	10/10	9/10	7/10	7/10	9/10

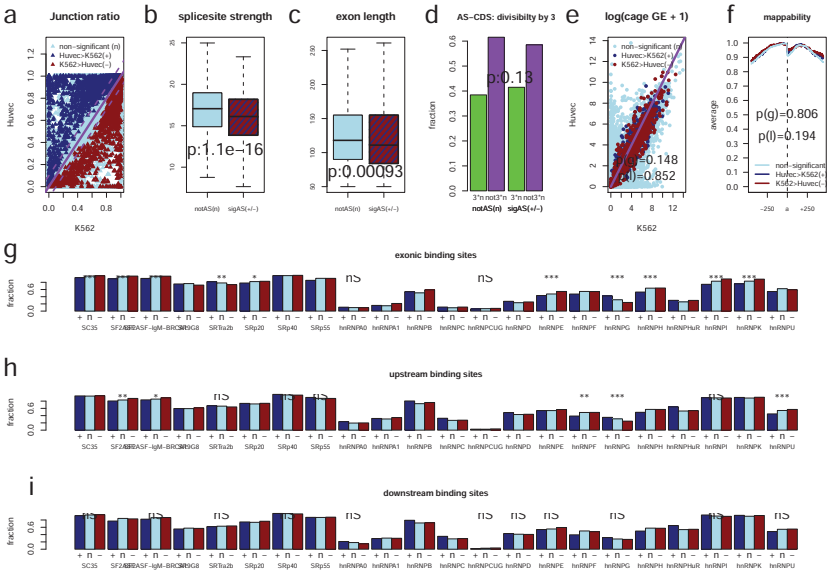
fig S10-S23



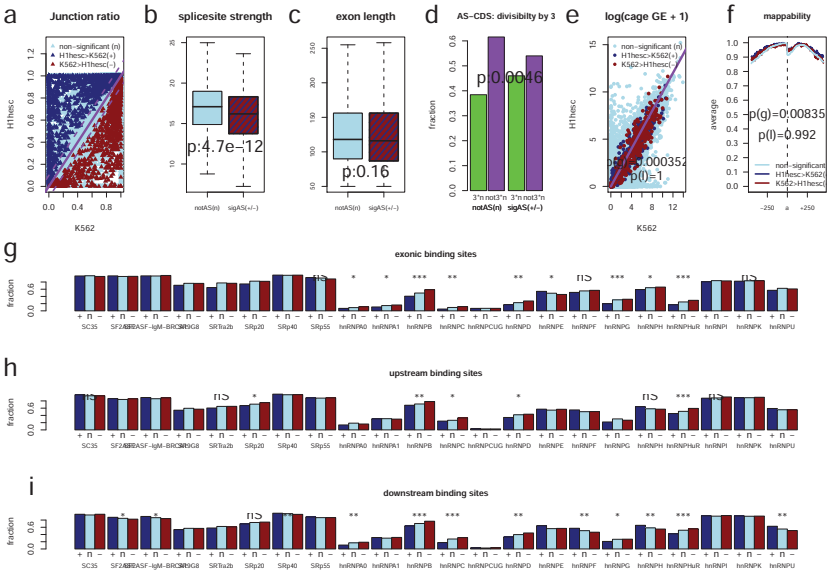
FigS10



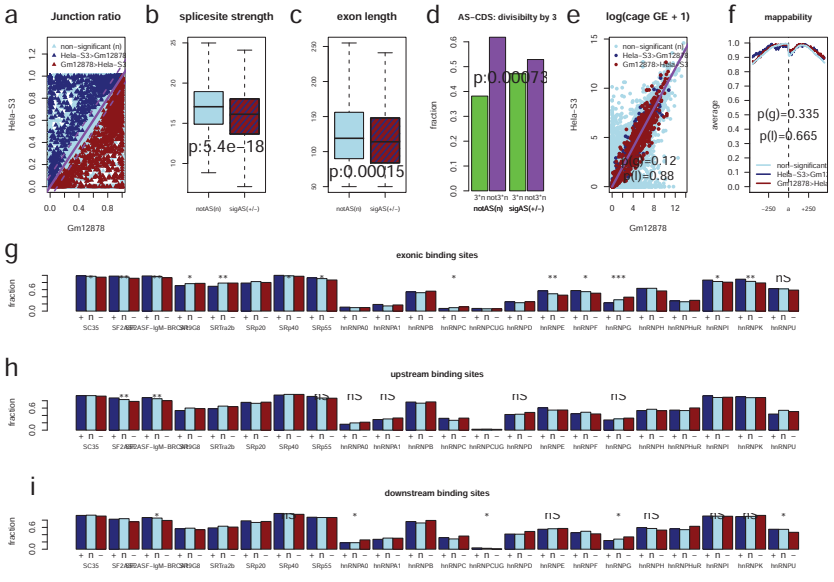
FigS11



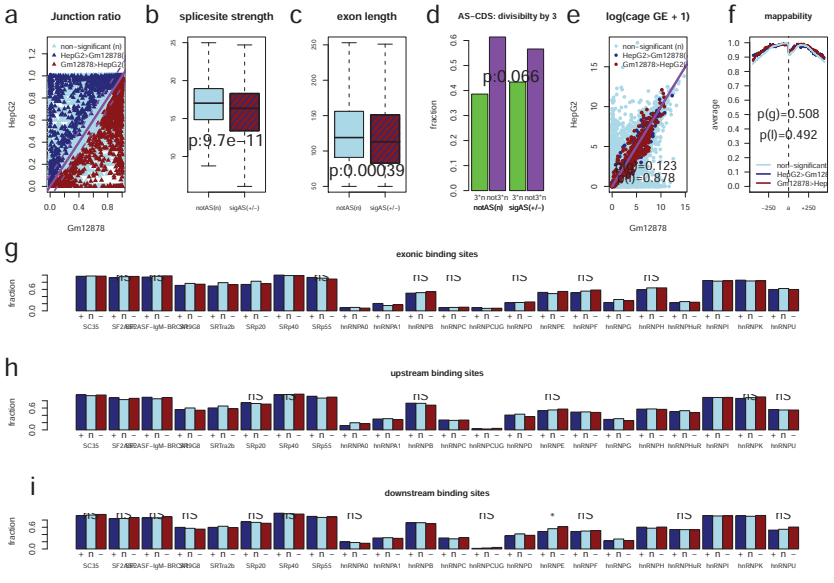
FigS12



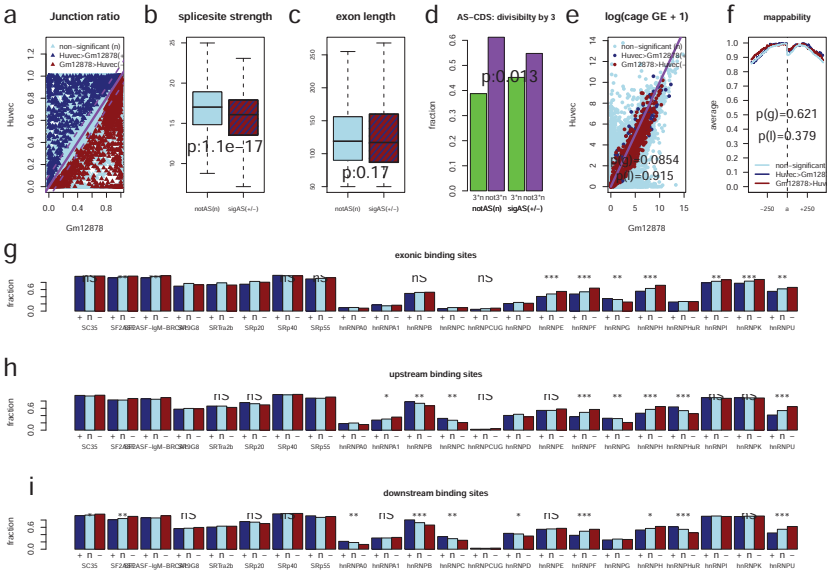
FigS13



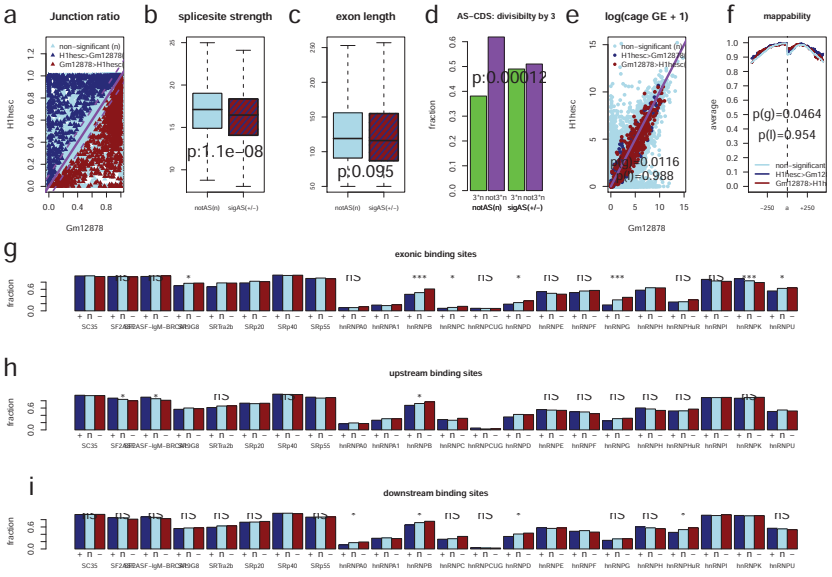
FigS14



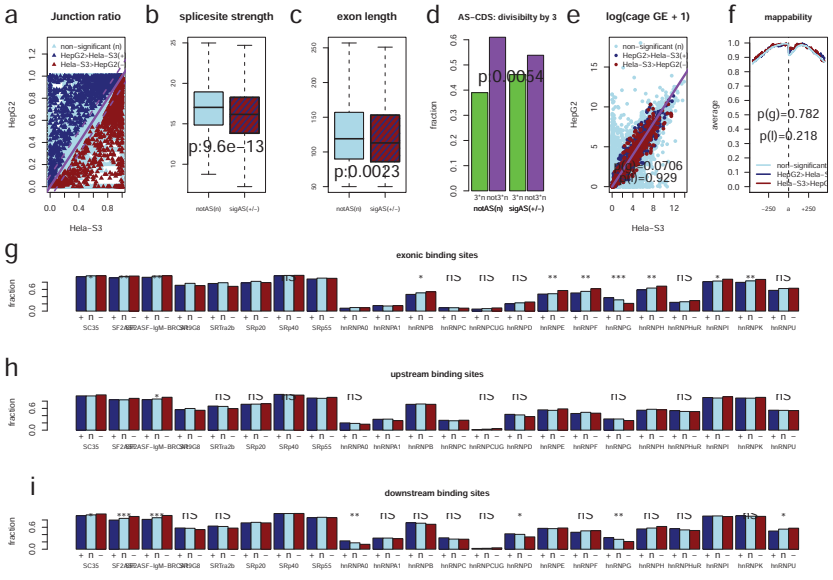
FigS15



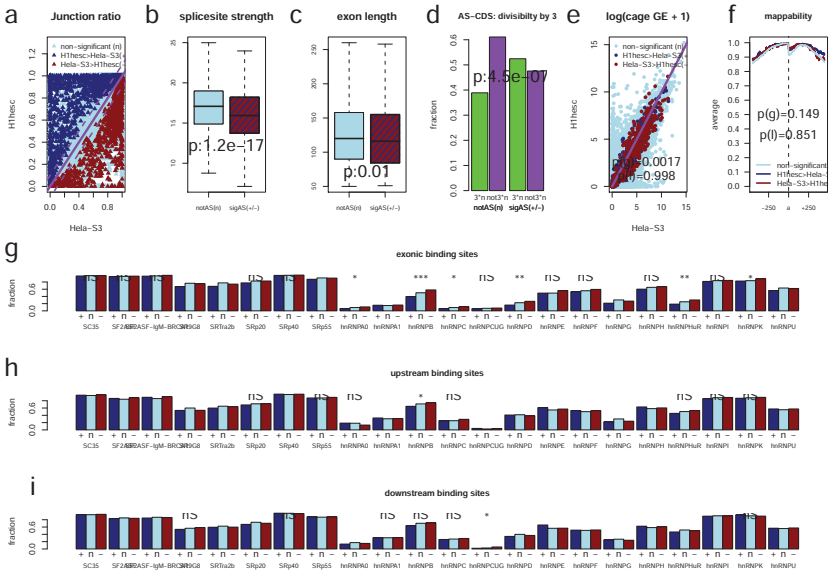
FigS16



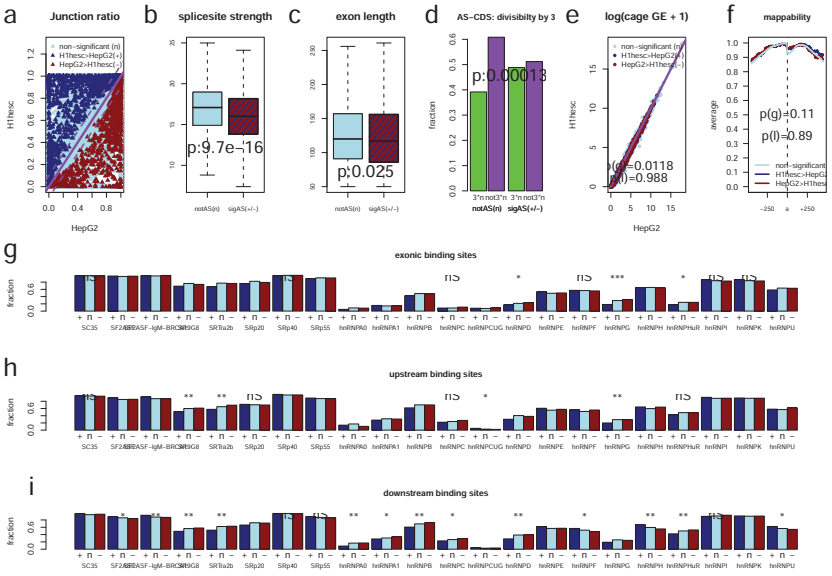
FigS17



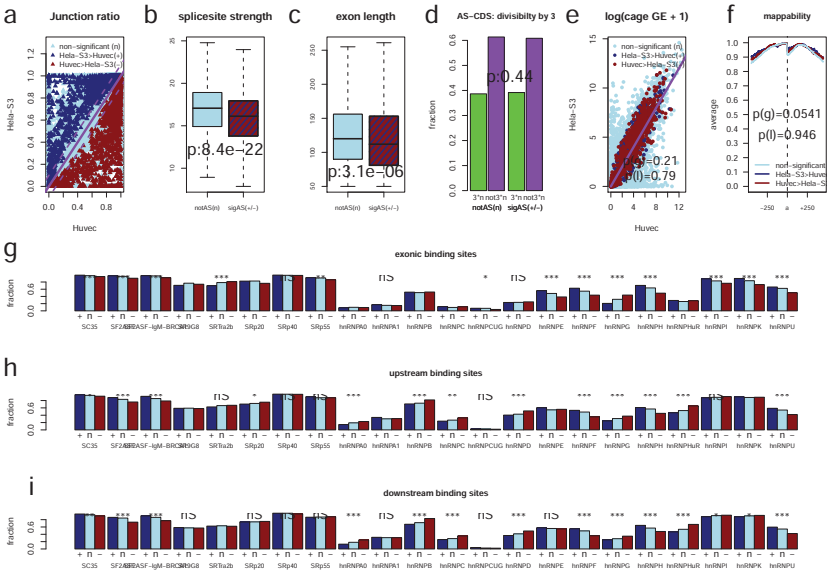
FigS18



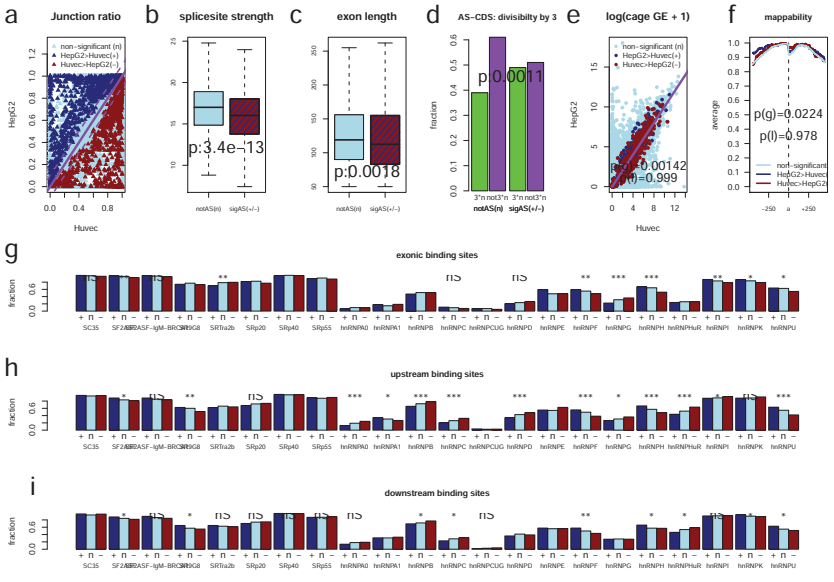
FigS19



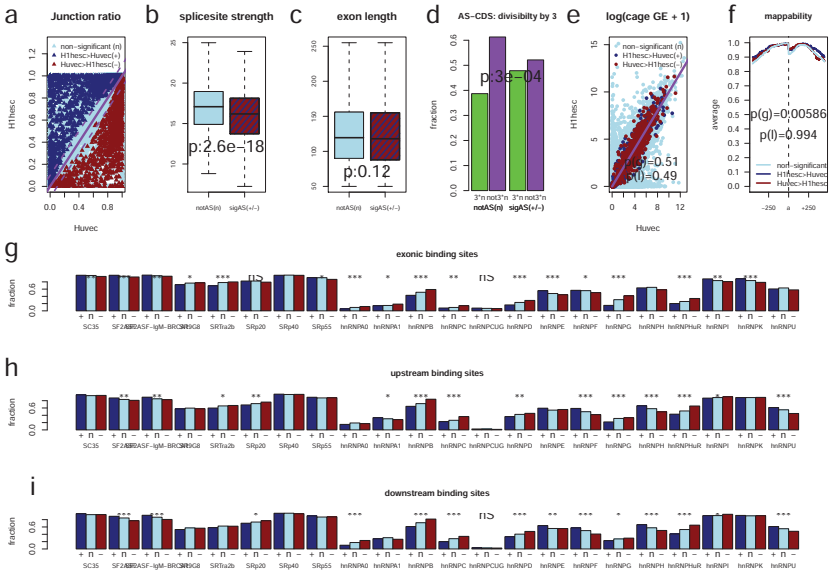
FigS20



FigS21



FigS22



FigS23

fig6

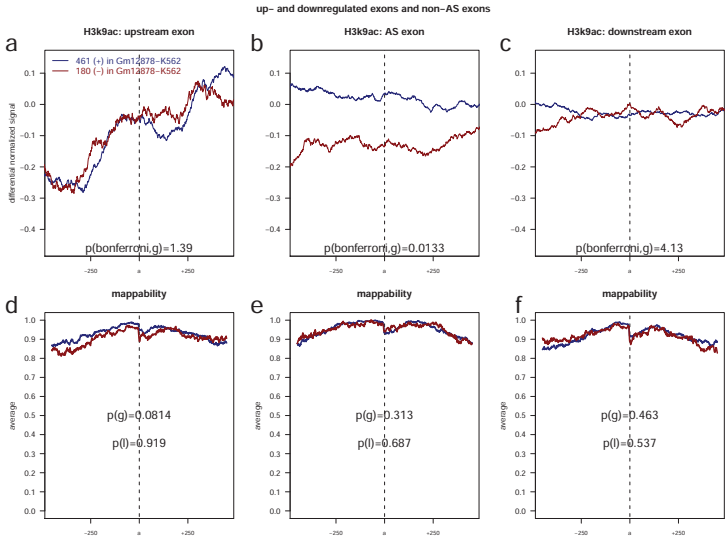
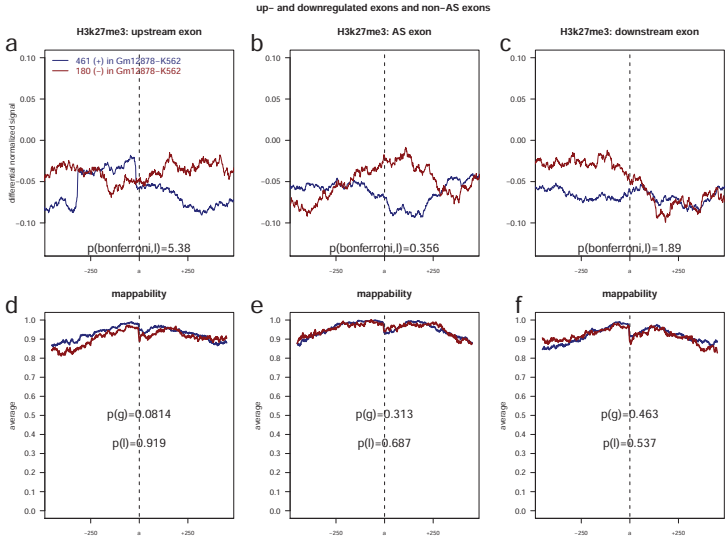


Fig5

figS24



FigS24

table S4: AS vs histone mods

Note: P-values are 1-sided wilcoxon rank sum tests, corrected for multiple testing (Bonferroni or Benjamini Hochberg)
p-vals:
significant after correction for multiple testing
not significant after correction for multiple testing
chromatin:
seems splicing relevant for many exons
does not seem perfectly relevant for splicing of many exons
upExon: upstream non alternative exon within same gene
ASexon: the alternative exon
doExon: downstream non alternative exon within same gene

celltypes	*chrom*	$p(upExon)$	$p(ASexon)$	$p(doExon)$
$K562^{*}$	$k36me3$	no	0.006	no
vs	$k27me3$	no	no	no
$Gm12878$	$k9ac$	no	0.013	no
$(n = 9)$	$k4me1$	0.027	no	no
	$k4me2$	no	0.0026	no
	$k4me3$	no	no	no
	$H4k20me1$	no	no	no
	$MNase$	no	0.03	no
	$Pol2$	no	0.047	no
$Gm12878^{*}$	$k36me3$	no	no	no
vs	$k9ac$	no	0.00001	no
$HelaS3$	$k4me2$	no	0.007	no
$(n = 5)$	$k4me3$	no	0.0067	no
	$H4k20me1$	no	no	0.02
$HelaS3^{**}$	$k27ac$	no	0.04	no
vs	$k36me3$	no	no	no
$HepG2$	$k4me2$	0.047	no	no
$(n = 6)$	$k4me3$	0.005	0.0016	no
	$k9ac$	0.024	0.021	no
	$k20me1$	0.0015	0.002	0.0009

using a Bonferroni (*) or a Benjamini-Hochberg(**) correction

Bibliography

A Barski, S Cuddapah, K Cui, T Y Roh, D E Schones, Z Wang, G Wei, I Chepelev, and K Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, May 2007. doi: 10.1016/j.cell.2007.05.009.

D E Schones, K Cui, S Cuddapah, T Y Roh, A Barski, Z Wang, G Wei, and K Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132 (5):887–98, Mar 2008. doi: 10.1016/j.cell.2008.02.022.

PART III

General Discussion

Chapter 4

General Discussion

Summary

Every chapter and manuscript in this thesis contains a discussion that is devoted to the specifics of the topics that are touched upon. Here I hope to clarify some subjects which were discussed only briefly in the manuscripts due to space limitations. Also I will try to put things into a broader perspective, explaining how one idea led to another ... at the danger that some might judge it as “non-specific blabla”.

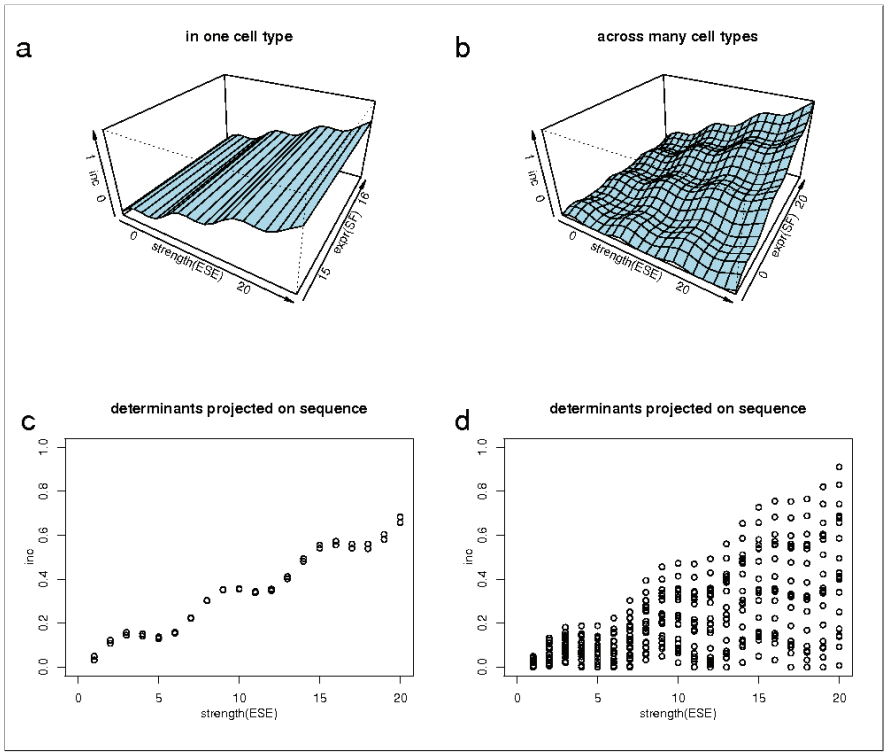


Figure 4.1 Modeling splicing from the RNA sequence: Exon inclusion as a function of splicing factor expression and binding site strength. When considering exons spliced in the same cell type, splicing factor expression is likely fairly constant **(a)**, an assumption rather unlikely when exons spliced in a variety of cell types are considered **(b)**. In the former case, projecting into the plane defined binding site strength and exon inclusion leads to few and small discrepancies **(c)**, in the latter case more and stronger discrepancies appear **(d)**.

4.1 Predictive capacity of splicing simulation

This thesis was carried out with the goal to achieve an understanding of splicing on a large scale. More exactly, the idea was that given the molecular information present at the moment when the splicing process begins to

take place, one should be able to predict the outcome of this process - or the probabilities of the different possible alternative outcomes. In the field of gene prediction Guigó et al. (1992) and Solovyev and Salamov (1994), to name only a few, had laid the groundwork, by developing a variety of methods, in order to predict genes, including exon-intron structure, from sequence features. Wang et al. (2004) had given a pure splicing angle to this, by only using sequence elements that were proved to play a role in splicing. In this way, they combined splice site strength, the absence/presence of ESE, ESS and G-triplets into a model, in order to predict the exon-intron structure of cDNA-alignments to the genome. From a mechanistic point of view, these approaches can be seen as a projection of all the determinants of splicing onto the RNA sequence. In order to illustrate this, I will use a slightly different, although related topic: prediction of exonic inclusion levels. For simplicity of the thought experiment, let us assume that there is exactly one binding site per exon, an ESE, for exactly one splicing factor and that this binding event is the only determinant of splicing for all considered exons. Then, looking at many exons, we expect their inclusion levels to be a function of the ESE strength, that is how easily it is recognized by the splicing factor, and of the expression of this splicing factor. Supposedly, both of these parameters would positively influence exon inclusion (Figure 4.1**a,b**). When considering exons in a given cell type or focusing on constitutive exons only (Figure 4.1**a**), it appears safe to assume that expression of the splicing factor is relatively constant. Therefore when we abstract from splicing factor expression and project the $(\mathtt{ESEstrength}, \mathtt{SFexpression}, \mathtt{inclusion})$ triples orthogonally onto the $(\mathtt{ESEstrength}, \mathtt{inclusion})$ plane, the different resulting inclusion levels are relatively similar for a given ESE strength (Figure 4.1**c**). When, however, many exons are alternative or exon inclusion values are taken

from a variety of cell types (Figure 4.1**b**), splicing factor expression supposedly varies between those cell types. Performing the same projection (or abstraction) then leads to situations where for the same ESE strength value many values of exon inclusion are possible (Figure 4.1**d**) - in this situation a strictly sequence dependent splicing simulator will not be able to predict all inclusion levels correctly. It is in the former situation (Figure 4.1**a,c**) that splicing simulation as carried out by Wang et al. (2004) and by ourselves supposedly works best. From a 2005 perspective the approach to work only on genes with splicing but without alternative splicing seemed sensible, so that all exons could be considered as constitutive. In the light of more recent findings (Harrow et al., 2006; Wang et al., 2008), it seems, however, unlikely that a large number of such genes really exists. With the advent of high throughput-sequencing technologies on the other hand, in this case RNAseq technologies, cell type specific gene annotations should be produced in the near future. I suspect that it should then be possible to find a large number of genes, for which one spliceform is clearly dominant within the cell type. I believe that using the presented approach (see chapter 2) could then be useful, in order to explore the limits of entirely sequence dependent splicing simulation.

By definition the employed approach (see chapter 2) can only hope to achieve as much accuracy as information is contained in the sequence. I had hoped, that a considerable amount of information remained to be discovered under the given additive model (see chapter 2). This model is simple in nature as it uses equal distance constraints for all regulatory sequences of the same kind (e.g. ESE). More exactly this means that the ESE GAAG receives the same score when it is located 5nts or 80nts downstream of the acceptor, but is not taken into account when being located more downstream. The same is true for the ESE TGGA. Since

it is very much possible that different ESE (or ESE, ISE or ISS) are subject to different distance constraints, one could therefore define different distance constraints for each ESE word. Furthermore, one could also have the score of an ESE vary depending on its position in the exon. Both approaches theoretically seem promising, but a considerable amount of effort would have to be invested, in order to estimate all these distance constraints specifically for each ESE. Not considering these complications, the obtained results suggest, but do not prove beyond doubt, that large gains are not to be expected using the employed model (see chapter 2) with the utilized aligned exon-intron structures. These aligned cDNAs stem from a variety of tissues or cell lines - therefore, as discussed above, working with alignments specific for one cell type could help. It is worth noting that sequence elements can predict tissue specific behavior (Barash et al., 2010). This abstraction from splicing factor expression level is, however, only possible, when we know the cell type, in which an inclusion level for a given exon is observed. An alternative, yet probably quite labor-some approach, could be to determine the splicing factor expression levels in all tissues, as well as, the tissue of origin of an exon-intron structure and to incorporate both into the simulation process.

As pointed out in the discussion of the splicing simulation chapter (see chapter 2), I believe that larger progress could be made for mammalian gene structures when information about which exons are spliced together is used - provided, an underlying mechanism exists. This thought follows almost directly from the fact that mammalian introns are up to orders of magnitudes larger than mammalian exons (Zheng et al., 2005). Mammalian internal exons are frequently shorter than 100nts and thereby represent only a smaller fraction of the entire transcript. Thus, knowledge about the position of an exon tends to bear very little information about the

exon-intron structure of a transcript. Contrarily, introns can be much larger; hence knowledge about the position of an intron can be a lot of knowledge about the exon-intron structure of a transcript. Therefore, given one exon, information allowing to predict which exon it should be spliced to would be most welcome.

A biologically intriguing limitation of the utilized model (see chapter 2) is that the concept of co-transcriptional splicing is not considered at all, although it is a subject that is increasingly researched. A good overview of this field is given in Neugebauer (2002); Kornblihtt et al. (2004); Allemand et al. (2008); Moore and Proudfoot (2009); de Almeida and Carmo-Fonseca (2008). The same statement is true for chromatin influences on splicing. Incorporating both into the splicing simulation process could provide considerable advances. These observations are one further motivation for exploring intragenic chromatin structure, which will be discussed in the following.

4.2 Chromatin behavior on exons

Exploring chromatin organization on exons

When the splicing process occurs co-transcriptionally, Pol2 is by definition still transcribing the DNA. This opens the door for chromatin-splicing interactions in at least the following two ways: First, chromatin is well known to be linked to Pol2-dynamics (see Struhl (1998) for an overview of earlier evidence) which in turn relates to splicing decisions (de la Mata et al., 2003;

Howe et al., 2003) - a model referred to as the "kinetic model". Second, such splicing events then occur spatially close to the DNA-chromatin template. An interaction between chromatin and splicing factors not passing through Pol2 therefore seemed imaginable. In fact such an interaction between lysine 36 methylation states and the splicing factor PTB via the intermediate of MRG-15 has been demonstrated recently (Luco et al., 2010), a finding compatible with the "recruitment model". Importantly both models are not necessarily exclusive. In fact, it seems very much imaginable that kinetics influence recruitment or vice versa.

From a 2008 perspective, it seemed natural to investigate first the most basic elements of chromatin organization - the nucleosome, basically for four reasons: First, in the literature a relationship between nucleosome characteristics and exon-intron structure of genes had already been claimed (Beckmann and Trifonov, 1991). Second, as pointed out before, nucleosomes are the most basic aspect of chromatin organization, with modifications of histone tails being the next layer. Third, *symCurv*, a software to localize nucleosomes in the genome, was being developed in the laboratory by Christoforos Nikoloau and Sonja Althammer, who described their findings in Nikolaou et al. (2010). Besides many observations concerning TSS and transcription, *symCurv* also suggested that exonic DNA favored nucleosome occupancy. Fourth and finally, a genome wide map of nucleosome positioning had just been made available (Schones et al., 2008).

The basic strategy to investigate this finding was fairly simple in nature and its guiding principle are two comparisons of two sets each. If a certain property is to have a positive influence on splicing, one would assume that it is more prevalent in exons than in non-exonic sequences. Natural candidates for non-exonic sequences are of course introns themselves and introns have been used in this context (see for example Schwartz et al. (2009)). We,

however, focused on rather special parts of introns, so-called pseudoexons that must be distinguished from genuine exons by the splicing machinery. Although pseudoexons are only parts of introns, they are nevertheless abundant (Sun and Chasin, 2000). Following this logic, we first compared exonic nucleosome signals to pseudoexonic nucleosome signals. The second comparison is based on the idea that certain exons have weaker splicing signals and are therefore assumed to require an additional layer of elements aiding to the specificity of splicing (Fairbrother et al., 2002). We therefore compared exons that have very weak splice sites to those that have very strong splice sites. These two comparisons are in fact similar to those employed by Fairbrother et al. (2002) to define ESE.

Importantly histone modifications, above all H3K36me3, have been shown to peak on exons (Kolasinska-Zwierz et al., 2009). Quantitatively, however, in our hands a large portion of this peak is already present in the underlying nucleosome peak. Yet, when normalizing for the nucleosome peak, a stepwise increase at the acceptor was observed that extended at least 350bps into the downstream intron (Tilgner et al., 2009), an observation also confirmed by Nahkuri et al. (2009). Another modification, H4K20me1 showed a totally flat average curve around internal exons (Tilgner et al., 2009), in contrast to all other modifications (data by Barski et al. (2007)) that we investigated. This was unexpected, as any histone modification, that is randomly distributed on nucleosomes, should show a peak reminiscent of the nucleosome peak on exons. Since H4K20me1 did not show any peak on exons, there are two possible explanations: Either there is basically no H4K20me1 around exons and the reads giving the constant but non zero signal for this modification around exons are simply artifacts. The second more appealing explanation would be the existence of a mechanism that prevents H4K20me1-marking of exonic nucleosomes with a possible influence

on splicing. At the time of writing it is unfortunately not possible to distinguish between these two possibilities.

An interpretation of the MNase peak on exons is that nucleosomes would have an influence on splicing, for example by modulating Pol2-elongation rate or by serving as the primary template allowing to have well localized histone modifications. While the nucleosome peak on exons has been observed by a variety of groups besides us (Schwartz et al., 2009; Andersson et al., 2009; Nahkuri et al., 2009; Spies et al., 2009; Hon et al., 2009; Cuddapah et al., 2011), none of them (including us), so far, has been able to show a direct influence of the position of the nucleosome on alternative splicing. This could be due to the difficulty of manipulating nucleosome positioning. As of today I am not aware of any experimental way to manipulate nucleosome positions. It is, however, possible that nucleosome positions have no effect on alternative splicing. There are a couple of interpretations compatible with this idea:

- Nucleosomes could have no influence on splicing at all. In fact this possibility has been raised by Spies et al. (2009), although the same authors also say that nucleosome positions aid to splicing simulation and in this way behave like splicing defining elements. The argument that nucleosome positioning could be circumstantial basically postulates that different exon categories have different amounts of RNA sequence elements, such as ESE and ESS, that allow them to be spliced efficiently. These different sequence compositions would lead to differences in nucleosome positioning, which would then be totally circumstantial. It is clear that GC content plays an important role for nucleosome occupancy (Schwartz et al., 2009; Tilgner et al., 2009). Yet, in our hands GC-content of exons relative to the surrounding intronic

sequence does not fully account for the differences in MNase signal between exons and pseudoexons. The same is true for the difference between exons with weak and strong splice sites. Also, the signal on exons only cannot account for it, as exemplified by first and last exons, where GC content and nucleosome signal show opposite trends. GC content is, however, only one measure of sequence composition and Spies et al. (2009) argue that higher order sequence composition measures could account for nucleosome occupancy observations on exons. However, similar observations could also be interpreted in other ways. We have for example suggested that the higher GC content of exons could stem at least in part from the necessity of exonic DNA to harbor nucleosomes (Tilgner and Guigó, 2010). The higher GC content of exons is sometimes explained by their coding-potential; however, we found both higher GC content with respect to the surrounding introns and an MNase peak in exons from non-coding RNAs as well, suggesting that nucleosome occupancy is not a consequence of the exons' coding capacity.

- positioned nucleosomes could favor exon inclusion of the exon they are positioned on, but never move. This would suggest that nucleosome occupancy cannot contribute to differential alternative splicing. An observation by Schwartz et al. (2009) could be interpreted as supporting this idea: Alternative exons have lower nucleosome occupancy than constitutive exons. This lower occupancy would predispose them to being alternative, but not directly influence inclusion levels in a given tissue or cell type.

Frequency of co-transcriptional splicing in humans

As described in chapter 3.2, co-transcriptional splicing is not a strict requirement for chromatin to influence splicing decisions. Yet, it does offer a wealth of opportunities for chromatin to influence splicing, simply because of the spatial proximity of chromatin and RNA during transcription and because chromatin can influence transcription dynamics, which in turn can influence inclusion levels (de la Mata et al., 2003; Howe et al., 2003). As a step towards understanding chromatin influences on splicing, we decided to assess the frequency of co-transcriptional splicing in the framework of the ENCODE project. We did so by introducing the "completed splicing index" (coSI) for each exon. The observations made in RNAseq data from the chromatin fraction, as well as, in the nuclear polyA- fraction, but only very slightly in the nuclear polyA+ and the cytosolic fractions, support the idea that co-transcriptional splicing is wide-spread in humans. Co-transcriptional splicing seems to be the rule when exons are far from the polyA-site and a gradient of lower rates of completed splicing are observed towards the 3'end. Thus, splicing generally tends to proceed in a 5' to 3' direction, nicely fitting with a "first come first served rule" although exceptions are not excluded. Depending on how much splicing was completed exons showed differences in splice site strength, binding sites for SRproteins and hnRNAPs, as well as, in chromatin organization. Chromatin structure and Pol occupancy clearly also change along the gene body (Barski et al., 2007), as do coSI values. Therefore exons that are spliced early and late are "embedded" in different chromatin surroundings, simply because of their position within the gene, an observation interesting in itself. The same line of thinking could, however, also suggest that the correlations between coSI values and chromatin are completely circumstantial. Decision tree analy-

sis supported the idea that chromatin organization does contain predictive capacity for separating exons into high and low coSI classes - predictive capacity that is not entirely contained in the position within the gene, represented by the distance to polyA-site and TSS. The future will probably bring more insight in this direction, as data monitoring nascent RNA will be increasingly available. What is clear, is that this data presents yet another piece of evidence that on a genomic level chromatin structure and splicing should be viewed as connected - through the link of co-transcriptional splicing.

Monitoring changes in chromatin organization on alternative exons

The above observations point to a wide spread existence of co-transcriptional splicing. This motivated the analysis of chromatin organization changes on differentially included exons between multiple tissues. As a first step, it was necessary to determine alternatively skipped exons, when considering a pair of RNAseq experiments from two different cell types (or tissues). The method we used is based on the Fisher-test, an approach that was previously used by Wang et al. (2008). Since RNA sequencing is a fairly recent technique and determining alternative exons from it even more so, we invested a lot of energy into making sure that these alternative exons behave like alternative exons. Predominantly all investigated characteristics behaved as expected, when we employed this method, in order to determine alternatively spliced exons in 15 pairwise cell type comparisons. We found between 500 and 1100 alternatively skipped exons, suitable for investigating chromatin, per cell type comparison.

In order to investigate chromatin changes on alternative exons, a simple approach would have been to look at histone modification levels on an exon in both cell types. Unfortunately some histone modifications showed stronger peaks on all exons in one cell type as compared to another. The causes for this remain obscure, but it could be either due to genomic differences or to differences in experiments between the two cell types. The strength of the peak change varied between different exon-sets, allowing to find significant differences between differentially up- and downregulated exons using a Wilcoxon rank sum test. A variety of histone modifications showed such significant characteristics in one cell type comparison (for example K562 vs Gm12878). Only H3K9ac showed up in all three cell type comparisons, though, while H3K4me3 showed up in two cell type comparisons. These histone modifications therefore represented the best candidates, that would make a difference on a genomic level in terms of influences on splicing. Importantly, this does not imply that each exon showing a higher H3K9ac-peak will necessarily be more highly included. Rather, it means that there is a significant majority of such exons, that will do so.
Of course, it was of interest for us to exploit the predictive capacity of histone modification changes for alternative exon inclusion. Preliminary results, with some contribution of mine, but mainly by João Curado, point to significant but weak predictive capacity of H3K9ac as a predictor of alternative splicing. This statement was true, when

- qualitatively predicting, whether an exon will be more highly included or more lowly included in the Gm12878 cell line than in the K562 cell line, using logistic regression models.

- quantitatively predicting how much an exon will change inclusion levels between the K562 and the Gm12878 cell line, using linear regres-

sion models.

In both cases other chromatin variables showed more significant behavior and it was possible to find other variable-combinations, but not involving H3K9ac. It is important to note, that a linear or logistic regression differs from a Wilcoxon rank sum test: The former two use the exact values of the *predictor variables* - in our case histone modifications, while the latter uses only the ranking of this variables.

The most problematic step, when building regression models, is the transformation from raw signal to a *predictor variable*. That is, for each exon the chromatin behavior needs to be represented as one variables (or some few variables). We performed this by looking at the change of H3K9ac (and separately for other chromatin data) in a 300bp window around the acceptor, simply because graphical analysis suggested, that this was a good region to be looked at for most histone modifications. For H4K20me1 this was not the case. It is therefore possible that by changing the way of calculating *predictor variables*, one would get more significant results. This could be the case especially for H4K20me1. For future modeling projects investigating the transformation step might be crucial, although I do not see a straightforward way to do this well.

4.3 Defining all determinants of splicing

When discussing what are the molecular determinants of splicing outcome, one should consider three fundamentally different ways to define “determi-

nants of splicing".

First, any molecule whose introduction into the nucleus affects (a) splicing decision(s) could be defined as a determinant of splicing. In this way, many molecules would qualify as determinants of splicing. For example, introducing a large number of RNAs with a binding site for an SR-protein would probably alter the splicing of other pre-mRNAs in which the SR-protein interferes, simply because a certain number of SR-proteins would be bound by the introduced RNAs and less SR-proteins would be available for splicing of the original pre-mRNAs. In order to model splicing as a function of determinants according to this very broad definition, it would probably be necessary to model chemical reactions in the nucleus. Assuming that one could determine the concentrations of all molecules in the nucleus and sensibly describe all chemical reactions, this could be done following the lines of stochastic simulation, as for example described by Gillespie (1977).

Second, "determinants of splicing" could also be defined as any molecule, for which a functional interaction with the splicing process and an effect on splicing outcome can be shown. This definition is somewhat tighter buy would include molecules introduced into the cell during experiments. Following this definition, intragenic histone modifications (Schor et al., 2009; Alló et al., 2009; Luco et al., 2010) and siRNAs (Alló et al., 2009) can be considered the latest proved layers of splicing determinants. I personally think that it is very likely that nucleosome occupancy also belongs to this group, if only because the nucleosome fundamentally influences Pol2-dynamics (Hodges et al., 2009). Therefore, removing a nucleosome from a DNA template would most likely change transcription dynamics profoundly and therefore influence splicing.

Third, "determinants of splicing" could also be defined as "endogenous ways the cell or organism uses to regulate splicing", which is much more

restrictive in nature. Again, histone modifications appear to be "splicing determinants" according to this definition, as for example shown by Luco et al. (2010). For siRNAs this definition applies fairly likely, since endogenous siRNAs likely regulate gene expression (see Alló et al. (2009) and references therein). For nucleosome occupancy in itself this is much more uncertain, as it is not clear how the cell would "(re-)move" a nucleosome from/on the DNA template in intragenic regions.

Again, from a computational perspective, the hope is that knowledge about these epigenetic factors can contribute to elucidating exon-intron structures in a given context. More specifically, epigenetic surroundings could make an oligomer have ESE activity, while the same oligomer might not have such an activity in different surroundings. This fits nicely with the model by Luco et al. (2010), whereby lysine 36 methylation states would make the difference when PTB binding sites are weak, and could solve the riddle why so many hexamers have been published as ESE or ESS. Supposedly, this would be a step forward on the way to a computational method which really simulates the splicing process, in the sense that it takes as input all configurations of all "determinants of splicing" and reliably predicts the spliced RNA molecules. I hope that this thesis is a step towards this goal.

4.4 Outlook

Before thinking about the changes that will occur in the future, it might be helpful to look at the changes that have occurred recently. With respect to this thesis, the most important change has been the availability of high throughput sequencing data. While initially, I put all my efforts into simu-

lating or modeling splicing, the second part was made up of data analysis, for the simple reason that there was a lot of interesting data to be analyzed. In the fields I have touched, the importance of high throughput sequencing has been highlighted by the the initial round of MNase-seq (Barski et al., 2007; Schones et al., 2008) and RNA-seq (Wang et al., 2008). Right now, within the ENCODE project (ENCODE-Consortium), this is taken a step further by performing RNAseq in a variety of sub-cellular compartments and in different cell types. We are seeing transcripts localized to specific compartments, so that the "atomic" spatial unit of bioinformatics is not anymore "the cell", but "the cytosolic fraction" or "the nucleoplasm". From a personal point of view, it is this kind of data that confers meaning to words like "nucleoplasm" to me. Probably three years ago, I would have needed a minute to grasp the idea of "sub-cellular localization", while it is now part of my everyday vocabulary. In this way, it looks like bioinformatic and experimental biology show some convergence.

It does not take a lot of vision to predict that more sequencing will be carried out, across sub-cellular compartments and across time points, in order to describe the spatial transcript-distribution and its dynamics in time. From a splicing point of view, I find the idea of analyzing RNA, while it is being transcribed and spliced especially appealing. Similarly, I believe that the dynamics of chromatin will be described in much more detail in the years to come. Interestingly, already with the scarce data of limited resolution, judging by the standard five years from now, the combinations of histone modifications that are meaningful have been started to be studied genome wide (Ernst and Kellis, 2010).

At the time of writing, ChIPseq experiments, using sonication, frequently make use of fragments of 200-400bp. This means that the exact same biological situation can lead to quite different read-representations. For in-

stance, a nucleosome, carrying a lysine 36 trimethylation mark could lead to a read at starting right at the nucleosome, but also to one starting 200bp upstream. Yet, the latter read could also very well indicate a lysine 36 trimethylation on the nucleosome, right before our first nucleosome. This is especially annoying when comparing multiple cell lines. Although the current data allows tremendous insights, I believe that we will greatly profit from advances in resolution. Similarly, RNAseq technologies for splicing investigation currently lack resolution, in the sense that a read only provides information about one or, at best, two exons - meaning one junction. Ideally, we would sequence the entire transcript at once, so that one read alone would provide TSS, all splicing events and polyA site for the transcript. Also measurement of expression would then be, what it is supposed to be: The number of observed molecules. Nevertheless, even increasing the read length 10fold, which seems very much feasible, would already provide a wealth of opportunities, as one read would also provide information about a larger number of exons and junctions.

Assuming that the next years will describe with sufficient accuracy, where and when RNA molecules are localized, how they interact and so on ... what will we still be missing ? In my mind, the strongest limitation to the kind of analysis I have been performing, is that high throughput sequencing can only investigate nucleic acid chains. Although mass spectrometry has provided crucial insights into the word of proteins, I am not aware of any large-scale data set describing proteins with similar detail as sequencing technologies have (and will) describe(d) the DNA and RNA world. I do not know, how this could be achieved, but assuming that someday it will be achieved, I think that future “omics”-generations will probably mention three crucial steps, that opened possibilities in their field:

- the sequencing of “the human genome” (Lander et al., 2001; Venter et al., 2001)

- high throughput sequencing (see for example Mortazavi et al. (2008) for review)

- an equivalent of the latter for proteins.

Of course, the cell contains more molecules than just DNA, RNAs and proteins. While it is not straight-forward to name and quantify all molecules in the cell, high-throughput methods will be providing crucial information on the way towards this goal. In the end, the task will be to grasp all, or at least the most important, reactions between them. Many of these molecules - water, for instance - are present in very high quantities, others, like DNA, in very low numbers. Therefore a stochastic formulation of chemical reactions is likely to provide good insights. The theoretical background for this is available (Gillespie, 1977). With respect to splicing, it is worth noting, that co-transcriptional splicing by definition occurs during transcription on a single chromatin template. Furthermore, also by definition, each RNA molecule that is still being transcribed differs in sequence from any other molecule undergoing transcription on the same gene. Therefore, a lot of the involved molecules exist exactly once in a given cell. These considerations lead to the idea, that, also for splicing, a stochastic formulation - or simulation - is adequate and might provide great insights.

Bibliography

E Allemand, E Batsché, and C Muchardt. Splicing, transcription, and chromatin: a ménage à trois. *Curr Opin Genet Dev*, 18(2):145–51, Apr 2008. doi: 10.1016/j.gde.2008.01.006.

M Alló, V Buggiano, J P Fededa, E Petrillo, I Schor, M de la Mata, E Agirre, M Plass, E Eyras, S A Elela, R Klinck, B Chabot, and A R Kornblihtt. Control of alternative splicing through sirna-mediated transcriptional gene silencing. *Nat Struct Mol Biol*, 16(7):717–24, Jul 2009. doi: 10.1038/nsmb.1620.

R Andersson, S Enroth, A Rada-Iglesias, C Wadelius, and J Komorowski. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res*, 19(10):1732–41, Oct 2009. doi: 10.1101/gr.092353.109.

Y Barash, J A Calarco, W Gao, Q Pan, X Wang, O Shai, B J Blencowe, and B J Frey. Deciphering the splicing code. *Nature*, 465(7294):53–9, May 2010. doi: 10.1038/nature09000.

A Barski, S Cuddapah, K Cui, T Y Roh, D E Schones, Z Wang, G Wei, I Chepelev, and K Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, May 2007. doi: 10.1016/j.cell.2007.05.009.

J S Beckmann and E N Trifonov. Splice junctions follow a 205-base ladder. *Proc Natl Acad Sci U S A*, 88(6):2380–3, Mar 1991.

S Cuddapah, D E Schones, K Cui, T Y Roh, A Barski, G Wei, M Rochman, M Bustin, and K Zhao. Genomic profiling of hmgn1 reveals an association with chromatin at regulatory regions. *Mol Cell Biol*, 31(4):700–9, Feb 2011. doi: 10.1128/MCB.00740-10.

S F de Almeida and M Carmo-Fonseca. The ctd role in cotranscriptional rna processing and surveillance. *FEBS Lett*, 582(14):1971–6, Jun 2008. doi: 10.1016/j.febslet.2008.04.019.

M de la Mata, C R Alonso, S Kadener, J P Fededa, M Blaustein, F Pelisch, P Cramer, D Bentley, and A R Kornblihtt. A slow rna polymerase ii affects alternative splicing in vivo. *Mol Cell*, 12(2):525–32, Aug 2003.

The ENCODE-Consortium. The encode project: Encyclopedia of dna elements. `http://www.genome.gov/10005107`.

J Ernst and M Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, 28(8):817–25, Aug 2010. doi: 10.1038/nbt.1662.

W G Fairbrother, R F Yeh, P A Sharp, and C B Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–13, Aug 2002. doi: 10.1126/science.1073774.

D T Gillespie. Exact stochastic simulation of coupled chemical-reactions. *JOURNAL OF PHYSICAL CHEMISTRY*, 81(25):2340–61, 1977.

R. Guigó, S. Knudsen, N. Drake, and T. Smith. Prediction of gene structure. *Journal of Molecular Biology*, 226:141–157, 1992.

J Harrow, F Denoeud, A Frankish, A Reymond, C K Chen, J Chrast, J Lagarde, J G Gilbert, R Storey, D Swarbreck, C Rossier, C Ucla, T Hubbard, S E Antonarakis, and R Guigo. Gencode: producing a reference annotation for encode. *Genome Biol*, 7 Suppl 1:S4.1–9, 2006. doi: 10.1186/gb-2006-7-s1-s4.

C Hodges, L Bintu, L Lubkowska, M Kashlev, and C Bustamante. Nucleosomal fluctuations govern the transcription dynamics of rna polymerase ii. *Science*, 325(5940):626–8, Jul 2009. doi: 10.1126/science.1172926.

G Hon, W Wang, and B Ren. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol*, 5(11):e1000566, Nov 2009. doi: 10.1371/journal.pcbi.1000566.

K J Howe, C M Kane, and M Ares, Jr. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in saccharomyces cerevisiae. *RNA*, 9(8):993–1006, Aug 2003.

P Kolasinska-Zwierz, T Down, I Latorre, T Liu, X S Liu, and J Ahringer. Differential chromatin marking of introns and expressed exons by h3k36me3. *Nat Genet*, 41(3):376–81, Mar 2009. doi: 10.1038/ng.322.

A R Kornblihtt, M de la Mata, J P Fededa, M J Munoz, and G Nogues. Multiple links between transcription and splicing. *RNA*, 10(10):1489–98, Oct 2004. doi: 10.1261/rna.7100104.

E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, and others (International Human Genome Sequencing Consortium, IHGSC). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

R F Luco, Q Pan, K Tominaga, B J Blencowe, O M Pereira-Smith, and T Misteli. Regulation of alternative splicing by histone modifications. *Science*, 327(5968): 996–1000, Feb 2010. doi: 10.1126/science.1184208.

M J Moore and N J Proudfoot. Pre-mrna processing reaches back to transcription and ahead to translation. *Cell*, 136(4):688–700, Feb 2009. doi: 10.1016/j.cell.2009.02.001.

A Mortazavi, B A Williams, K McCue, L Schaeffer, and B Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5(7):621–8, Jul 2008. doi: 10.1038/nmeth.1226.

S Nahkuri, R J Taft, and J S Mattick. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. *Cell Cycle*, 8(20):3420–4, Oct 2009.

K M Neugebauer. On the importance of being co-transcriptional. *J Cell Sci*, 115 (Pt 20):3865–71, Oct 2002.

C Nikolaou, S Althammer, M Beato, and R Guigó. Structural constraints revealed in consistent nucleosome positions in the genome of s. cerevisiae. *Epigenetics Chromatin*, 3(1):20, 2010. doi: 10.1186/1756-8935-3-20.

D E Schones, K Cui, S Cuddapah, T Y Roh, A Barski, Z Wang, G Wei, and K Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132 (5):887–98, Mar 2008. doi: 10.1016/j.cell.2008.02.022.

I E Schor, N Rascovan, F Pelisch, M Alló, and A R Kornblihtt. Neuronal cell depolarization induces intragenic chromatin modifications affecting ncam alternative splicing. *Proc Natl Acad Sci U S A*, 106(11):4325–30, Mar 2009. doi: 10.1073/pnas.0810666106.

S Schwartz, E Meshorer, and G Ast. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, 16(9):990–5, Sep 2009. doi: 10.1038/nsmb.1659.

V.V. Solovyev and A.A. Salamov. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Research*, 22:5156–5163, 1994.

N Spies, C B Nielsen, R A Padgett, and C B Burge. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell*, 36(2):245–54, Oct 2009. doi: 10.1016/j.molcel.2009.10.008.

K Struhl. Histone acetylation and transcriptional regulatory mechanisms. *Genes Dev*, 12(5):599–606, Mar 1998.

H Sun and L A Chasin. Multiple splicing defects in an intronic false exon. *Mol Cell Biol*, 20(17):6414–25, Sep 2000.

H Tilgner and R Guigó. From chromatin to splicing: Rna-processing as a total artwork. *Epigenetics*, 5(3), Apr 2010.

H Tilgner, C Nikolaou, S Althammer, M Sammeth, M Beato, J Valcárcel, and R Guigó. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol*, 16(9):996–1001, Sep 2009. doi: 10.1038/nsmb.1658.

J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

E T Wang, R Sandberg, S Luo, I Khrebtukova, L Zhang, C Mayr, S F Kingsmore, G P Schroth, and C B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6, Nov 2008. doi: 10.1038/nature07509.

Z Wang, M E Rolish, G Yeo, V Tung, M Mawson, and C B Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–45, Dec 2004. doi: 10.1016/j.cell.2004.11.010.

C L Zheng, X D Fu, and M Gribskov. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA*, 11(12):1777–87, Dec 2005. doi: 10.1261/rna.2660805.

Conclusions

- With one exception, 4mers appear to be smallest unit for which ESE or ESS activity can be claimed. 19 such 4mers and G-triplets perform as well as much larger published sets of ESE and ESS for splicing simulation. Small further gains in splicing simulation accuracy can be achieved, but large gains are not likely with the employed model.

- nucleosomes are well positioned on internal exons. This correlates with aspects of exon-intron architecture such as the splice site strength. Weak-splice-site-exons tend to have exon-lengths more close to that of a nucleosome than exons with strong splice sites.

- co-transcriptional splicing appears to be very wide-spread in humans and tends to occur in 5' to 3' direction as judged from a genome wide data set. Consistently spliceosomal RNAs were found in the chromatin fraction.

- in 15 pairwise cell type comparisons we could define a large number of differentially skipped exons. Exon inclusion changes between two cell types tend to co-occur with chromatin changes. H3K9ac and H3K4me1,2 are the most promising candidates for a wide-spread influence on splicing.

PART IV

Appendices

Abbreviations

A : adenosine

aa : amino acid

ave : average

BH : Benjamini-Hochberg

C : cytidine

CAGE : cap analysis gene expression

ChIPseq : chromatin immunoprecipitation (ChIP) followed by sequencing

chr : chromatin OR chromosome

coSI : completed splicing index

cyt : cytosol

DNA : deoxyribonucleic acid

ENCODE : ENCyclopedia Of DNA Elements

ESE : exonic splicing enhancer

ESS : exonic splicing silencer

EST : expressed sequence tag

G : guanosine

hnRNP : heterogeneous nuclear ribonucleoprotein

ISE : intronic splicing enhancer

ISS : intronic splicing silencer

kDa : kilo dalton

nl : nucleolus

NMD : nonsense-mediated decay

np : nucleoplasm

nt : nucleotides

nuc : nucleus

pp : percentage point(s)

ppy-tract : polypyrimidine tract

RNA : ribonucleic acid

RNAseq : RNA sequencing

SRp : protein of the serine/arginine-rich protein family

T : thymidine

U : uridine

UTR : untranslated region

Titles in the GBL Dissertation Series

2001-02 Moisés Burset.

Estudi computacional de l'especificació dels llocs d'splicing.

[Computational analysis of the splice sites definition.]

Departament de Genètica, Universitat de Barcelona.

2001-04 Sergi Castellano.

Towards the characterization of the eukaryotic selenoproteome: a computational approach.

Departament de Ciències Experimentals i de la Salut

Universitat Pompeu Fabra.

2002-04 Genís Parra.

Computational identification of genes: "ab initio" and comparative approaches.

Departament de Ciències Experimentals i de la Salut

Universitat Pompeu Fabra.

2001-05 Josep F. Abril.

Comparative Analysis of Eukaryotic Gene Sequence Features.

Departament de Ciències Experimentals i de la Salut

Universitat Pompeu Fabra.

2001-06 Enrique Blanco.

Meta-Alignment of Biological Sequences.

Departament de Llenguatges i Sistemes Informàtics

Universitat Politècnica de Catalunya.

2002-04 Charles Chapple.

Finding a needle in a haystack: The eukaryotic Selenoproteome

Departament de Ciències Experimentals i de la Salut

Universitat Pompeu Fabra.

2006-11 Hagen Tilgner.

Modelling Splicing.

Departament de Bioinformatica i Genòmica

Universitat Pompeu Fabra.