Bioinformatics analysis of intron retention events associated with the minor spliceosome

Ali Oghabian

Institute of Biotechnology

Helsinki Institute of Life Sciences (HiLife)

Faculty of Biological and Environmental Sciences

Doctoral Programme in Integrative Life Science

University of Helsinki

ACADEMIC DISSERTATION

To be presented for public examination with
the permission of the Faculty of Biological and Environmental Sciences
in Lecture hall 2402 (Telkänpönttö) in Biocenter 3, Viikinkaari 1
on November 27th 2018 at 12 o'clock noon.

Helsinki 2018

Supervisors

Docent Mikko Frilander
Institute of Biotechnology
University of Helsinki
Helsinki, Finland

Associate professor Dario Greco
Institute of Biosciences and Medical Technologies
University of Tampere
Tampere, Finland

Thesis advisory committee

Docent Petri Auvinen
Institute of Biotechnology
University of Helsinki
Helsinki, Finland

Professor Jukka Corander
Department of Mathematics and Statistics
University of Helsinki
Helsinki, Finland

Reviewers

Professor Tero Aittokallio
Department of Mathematics and Statistics
University of Turku
Turku, Finland

Professor Garry Wong
Department of Health Sciences
University of Macau
Macau, China

Opponent

Professor Rickard Sandberg
Department of Cell and Molecular Biology
Karolinska Institutet
Stockholm Sweden

Custos

Professor Liisa Holm
Organismal and Evolutionary Biology Research Programme
Institute of Biotechnology
University of Helsinki
Helsinki, Finland

ISBN 978-951-51-4698-4 (paperback)
ISBN 978-951-51-4699-1 (PDF)
ISSN 2342-3161 (print)
ISSN 2342-317X (online)
http://ethesis.helsinki.fi

Cover layout by Anita Tienhaara

Juvenes Print - Suomen Yliopistopaino Oy, Helsinki, Finland 2018

“ We're forever teetering on the brink of the unknowable,
and trying to understand what can't be understood. ”

Isaac Asimov, Caves of steel

Table of Contents

List of original publications

I. Oghabian, A., Greco, D., Frilander, M.J., (2018). IntEREst: intron-exon retention estimator. *BMC Bioinformatics*, 19, 130.

II. Niemelä, E.H.*, Oghabian, A.*, Staals, R.H., Greco, D., Pruijn, G.J., Frilander, M.J., (2014). Global analysis of the nuclear processing of transcripts with unspliced U12-type introns by the exosome. *Nucleic Acids Res*, 42, 7358–7369.

III. Argente, J. , Flores, R., Gutiérrez-Arumí, A., Verma, B., Martos-Moreno, G.Á., Cuscó, I., Oghabian, A., Chowen, J.A., Frilander, M.J., Pérez-Jurado, L.A., (2014). Defective minor spliceosome mRNA processing results in isolated familial growth hormone deficiency. *EMBO Mol Med*, 6, 299–306.

*- equal contribution

I. AO developed the software, ran all the analyses, and together with other authors wrote the manuscript. The analysis include mapping and normalization of RNAseq data, annotation of U2/U12-type introns, measuring intron retention (IR) levels, differential IR analysis and benchmarking and comparison analysis.

II. AO ran all the bioinformatics analyses, and together with other authors wrote the manuscript. The analysis include mapping and normalization of RNAseq data, annotation of U2/U12-type introns, measuring IR levels and differential IR analysis.

III. AO ran the bioinformatics analyses together with other authors. The analysis include mapping and normalization of RNAseq data, isoform reconstruction, annotation of U2/U12-type introns, measuring IR levels and differential IR analysis.

Abbreviations

20K/25K/31K/35K/65K	U11/U12 small nuclear ribonucleoprotein 20/25/35/65 KDa protein
AS	alternative splicing
bam	binary sequence alignment/map
bp	base pair
C. bigsae	Caenorhabditis briggsae
C. elegans	Caenorhabditis elegans
ChIP-seq	chromatin immunoprecipitation with massively parallel DNA sequencing
D. melanogaster	Drosophila melanogaster
DEG	differentially expressed genes/isoforms
deltaPSI	difference in percentage spliced in
DIS3L1/2	DIS3 like 3'-5' exoribonuclease 1/2
DNA	deoxyribonucleic acid
DRB	5,6-dichlorobenzimidazole 1-β-D-ribofuranoside
edgeR	empirical analysis of digital gene expression data in R
EOCA	early-onset cerebellar ataxia
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
FPKM	fragments per kilobase of transcript per million mapped reads
GFP	green fluorescent proteins
GLM	generalized linear model
GO	gene ontology

GTF	Gene transfer format
IGHD	isolated growth hormone deficiency
IGV	integrative genomics viewer
IntEREst	intron-exon retention estimator
IR	intron retention
ISE	intronic splicing enhancer
KD	knockdown
limma	linear models for microarray data
LWS	Lowry Wood syndrome
MDS	myelodysplastic syndrome
miRNA	microRNA
MISO	mixture of isoforms probabilistic model for RNAseq
pre-mRNA	precursor mRNA
PSI	percentage spliced in
PSSM	position-specific scoring matrix
PTC	premature termination codons
PWM	position weight matrices
qPCR	quantitative PCR
RefSeq	NCBI reference sequence database
RFMN	Roifman syndrome
RNA	ribonucleic acid
RNAseq	RNA sequencing
rRNA	ribosomal RNA
RT-qPCR	quantitative reverse transcription PCR
SF	splicing factor

siRNA	small interfering RNA
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
SOLiD	sequencing by oligonucleotide ligation and detection
SR	serine/arginine-rich (protein)
ss	splice site
TALS	Taybi-Linder syndrome
TMM	trimmed mean of m values
TPM	Transcripts Per Million
tRNA	transfer RNA
48K	U11/U12 small nuclear ribonucleoprotein 48 KDa protein
U11-48K	U11/U12 small nuclear ribonucleoprotein 48 KDa protein

Summary

In the Eukaryotes, DNA sequences in genes are often interrupted by non-coding sequences called introns. These sequences are removed from the transcripts via a process known as splicing either while the genes are being transcribed (co-transcriptionally) or after transcription (post-transcriptionally). In higher eukaryotes two separate pre-mRNA splicing machineries have been described: the U12-dependent spliceosome which is responsible for splicing of approximately 700-800 unique introns (known as the U12-type introns), and the U2-dependent spliceosome responsible for splicing all other introns (known as the U2-type introns). The two intron types show divergent sequence elements in their 5' splice site and branch point sequences. In addition, earlier reports have indicated that U12-type introns are spliced with a slower rate comparing to the U2-type introns, suggesting that the splicing of U12-type introns is rate-limiting to the expression of the U12-type intron containing genes. This slower splicing is manifested as unspliced or retained U12-type introns in the otherwise fully processed mRNA products.

In this work I developed a novel computational tool called the intron-exon retention estimator (IntEREst) which allows accurate detection, quantification and differential analysis of the intron retention levels from RNAseq data. Additional features of IntEREst include a tool for identification of U12-type introns, and a number of tools to compare the retention levels of user-defined subclasses of introns across several samples. An already published RNAseq dataset (available under accession GSE63816 in NCBI Gene expression Omnibus database) from patients and control subjects of myelodysplastic syndrome (MDS) was used to assess the functionality by benchmarking IntEREst. This dataset included RNAseq data from MDS patients featuring mutations in the *ZRSRS2* gene that functions in the recognition of U12-type introns, and from control subjects that were either healthy or MDS patients without *ZRSR2* mutations. Additionally, I used a Maize dataset consisting of samples with mutated and wild-type *RGH3* gene, which is an ortholog of human *ZRSR2*. My results indicate that IntEREst is a reliable tool for analyzing intron retention events from RNAseq data producing comparable or better results than the other similar methods.

I used IntEREst to globally compare the retention of the U12-type introns to that of U2-type introns. I found that U12-type introns show on average a 2-fold higher retention levels compared to that of U2-type introns both in human and plant cells. This result recapitulates the findings from earlier studies using a small set of selected genes and generalizes the increased intron retention of U12-type introns to a genome-wide scale. Furthermore, the results of this work provide evidence that transcripts containing unspliced U12-type introns are degraded in the nucleus by the nuclear exosome. Together, these results support the hypothesis that U12-type introns are globally spliced less efficiently than the U2-type introns and can thus regulate the rate of mature mRNA formation with the genes containing U12-type introns. Additionally, intron retention analysis of human/plant cells containing mutations in the U12-dependent spliceosome showed that such defects lead to a further increase in the levels of unspliced U12-type introns.

In conclusion, this thesis extends current knowledge concerning the significance of the correct splicing of U12-type introns and the consequences of their abnormal splicing. Furthermore, it describes a combination of available tools together with a novel software tool (*i.e.* IntEREst) that can be used to measure and compare the efficiency and accuracy of RNA splicing across multiple samples. We show that these tools can reveal valuable information about the molecular mechanisms involved in various conditions, *e.g.* diseases caused by defective spliceosome.

1. Introduction

1.1. Spliceosomal introns and pre-mRNA splicing

Most nuclear genes in higher eukaryotes feature non-coding sequences called introns, which are absent from the final mRNAs and therefore are not coded to proteins (Gilbert, 1978; Berget *et al.*, 1977; Chow *et al.*, 1977; Evans *et al.*, 1977). These sequences are removed by a ribonucleoprotein complex called the spliceosome, hence they bear the name "spliceosomal introns". They are completely absent from prokaryotes and are present in varied numbers (from hundreds to hundreds of thousands) across eukaryotes. For example in humans more than 200,000 spliceosomal introns have been reported (Sakharkar *et al.*, 2004). The presence of these introns in many eukaryotes, together with the conservation of their positions across various lineages of eukaryotes suggests the existence of introns and a nascent spliceosome in the last common ancestor (*i.e.* cenancestor) of eukaryotes (Roy and Gilbert, 2006, 2005; Rogozin *et al.*, 2003). It is thought that the cenancestor of eukaryotes probably featured a significant number of introns and although some modern eukaryotes show evidence of intron gain, *e.g.* 81 gained introns in *C. elegans* and 41 in *C. briggsae* (Coghlan and Wolfe, 2004), throughout evolution introns had a tendency to be lost rather than new ones being inserted (Roy and Gilbert, 2005, 2006). This is not unexpected since introns carry several burdens to the genome: cell needs to make extensive investments to both replicate and transcribe the intronic sequences that will be excised and degraded nearly immediately after their synthesis (Beyer *et al.*, 1981; Beyer and Osheim, 1988). Mutations on the intron splice sites or any defective spliceosome causing mutation may lead to failure to detect and/or remove introns from transcripts (discussed in detail in section 1.2). Moreover, excised introns in the nucleus must be degraded.

However, several benefits have also been ascribed for introns: emergence of new genes as a result of recombination within introns and 'exon-shuffling' (Gilbert, 1978; reviewed by Patthy, 1999); improved fitness by promoting intragenic recombination (Gilbert, 1978; Comeron and Kreitman, 2000); regulation of transcription by housing transcription enhancers and suppressors (Rossi and de Crombrugghe, 1987; Stergachis *et al.*, 2013); increased flexibility for genes to code multiple products as a result of alternative splicing (described in detail in section 1.2); enhanced gene expression (Gruss *et al.*, 1979; Callis *et al.*, 1987); control of chromatin assembly and mRNA transport (Luo and Reed, 1999; Valencia *et al.*, 2008; Schwartz *et al.*, 2009; Spies *et al.*, 2009); enhance mRNA quality control through nonsense-mediated mRNA decay (*i.e.* NMD); NMD regulation through exon junction complexes (Le Hir *et al.*, 2001); and lastly, generation of stable RNA products, *e.g.* small nucleolar RNAs (snoRNA) and microRNAs (miRNA)(Smith and Steitz, 1998; Dieci *et al.*, 2009; Ambros *et al.*, 2003). In a few cases the intron-coded RNAs are the stable products emerging from the primary transcript while the fully spliced mRNA is degraded (Tycowski *et al.*, 1996; Moore, 1996). Notably although these benefits were probably effective in driving the introns to long-term fixation, they were influential only after the introns were already established. However, there have been attempts to explain the origin of introns (with their many burdens on the genome) using other non-adaptive reasons as the basis for such arguments. One such theory claims that mildly deleterious insertions such as those that led to the origin of introns may be tolerated if the effective population is sufficiently small; *i.e.* conditions that probably the cenancestor of eukaryotes lived under (Lynch, 2002).

1.1.1. Did introns emerge late or early during evolution?

The widespread presence of introns in various eukaryote lineages and their scarcity in prokaryotes has led to heated debates and disagreements over the time of their origins. Specifically, there are two main theories of the origins of introns: the "introns early" theory postulates that the early ancestral genomes were mainly built up of short exons separated by introns which, through exon-shuffling, facilitated the construction of complex gene structures of the present day. This theory further postulates that the introns were present in the genome of the common ancestor of eukaryotes and prokaryotes but they were lost during the evolution of prokaryotes, possibly due to selection favoring maximal growth and streamlined genomes, resulting in modern intron-less prokaryotes (Darnell, 1978; Doolittle, 1978; Gilbert, 1987). In contrast, the "introns late" theory argues that the genome of the ancestors of eukaryotes and prokaryotes lacked introns and resembled those of modern prokaryotes. Consequently, according to this theory the introns were later inserted into the already existing intron-less genes (Cavalier-Smith, 1991; Palmer and Logsdon, 1991).

Although completely devoid of spliceosomal introns, it is known that prokaryotes include other non-spliceosomal types of introns, *i.e.* group I, group II self-splicing introns and tRNA introns (Belfort *et al.*, 1995). The tRNA introns are found in tRNA genes of eukaryotes and archaea. They are usually located at their canonical position (*i.e.* upstream the anticodon) and their removal from the RNA is dependent on protein components. In contrast, group I and II are self-splicing introns. They are present in mRNA, tRNA and rRNA of bacteria, chloroplasts, mitochondria, viruses and lower eukaryotes, but are absent from nuclear genomes of the multicellular eukaryotes. However, the group II introns have been described in the chloroplasts and mitochondria of several multicellular eukaryotes (*e.g.* plants and fungi). Extensive evidence suggests that the present group II self-splicing introns and spliceosomal introns share a common evolutionary ancestry. Specifically, in addition to a similar catalytic mechanism between group II self-splicing introns and spliceosomal intron excision, sequence and structural data support the hypothesis that snRNA components of the nuclear spliceosomes have a strong resemblance to individual RNA domains in the group II self-splicing introns. Conversely, the excision of spliceosomal introns is known to be catalyzed by snRNA components of the spliceosome (Valadkhan and Manley, 2001; Shukla and Padgett, 2002; Yan *et al.*, 2015; Papasaikas and Valcarcel, 2016). Thus it is likely that spliceosomal snRNAs and presumably also at least some spliceosomal introns have originated from group II introns that either were transferred from other sources (*e.g.* the primordial mitochondrion or a bacterium), or had prior residence in the nucleus of eukaryotes (Cavalier-Smith, 1991; Lynch, 2007; Zimmerly and Semper, 2015). Recent developments, *e.g.* the discovery of presence of self-splicing introns in genomes of some prokaryotes and the ancestral relationship of group II introns to the spliceosomal snRNAs, render the introns late theory (at least in its extreme form mentioned above) to be highly unlikely. Note that hereafter for the sake of simplicity, I will use the term ‘intron’ without clarifying the specific group, to refer to ‘spliceosomal introns’.

1.1.2. U2- and U12-type introns

There are two known types of spliceosomal introns, namely U2- and U12-type introns. Mammalian genomes contain ~ 700-800 U12-type as compared with the ~ 200 000 U2-type introns. Hence the U12-type introns are also known as the ‘minor’ introns and the U2-type introns as the “major” introns (Turunen *et al.*, 2013). U12-type introns are present in the nuclear genomes of diverse eukaryotic lineages. They are present in the genome of vertebrates (at least

mammals, birds, fish, and amphibians), several insect species (*e.g. Drosophila melanogaster* and silkworm *i.e. Bombyx mori*), in Cnidarians (*e.g.* jellyfish), various plants (*e.g. Arabidopsis thaliana* and maize *i.e. Zea mays*) and even in a number of protists (Russell *et al.*, 2006); and they are absent from several species such as the yeast *Saccharomyces cerevisiae*, nematode *Caenorhabditis elegans*, and most eukaryotic microbes (*e.g.* algae, protists and fungai) (Burge *et al.*, 1998; Alioto, 2007). The two classes of introns feature distinct consensus sequences on their 5′ and 3′ splice sites (ss) and the branch point sequence. Two subclasses of U12-type introns have been described: GT-AG subtype with GT and AG dinucletides at their intron 5′ and 3′ termini, and a less frequent AT-AC subtype with AT and AC dinucleotides at the respective termini (Dietrich *et al.*, 1997; Wu and Krainer, 1997; Burge *et al.*, 1998). Usually genes are either completely devoid of U12-type introns or include only a single U12-type intron together with one or several U2-type introns. Exceptions do, however, exist; *e.g.* the human genes *SPTSSA* and *CCDC56* each feature solely a U12-type intron, and several genes feature more than one U12-type intron, *e.g. DERL2* has two U12-type and four U2-type introns. The genes that contain U12-type introns are mainly associated with ‘information processing’; on the other hand genes related to basic energy metabolism, fatty acid and phospholipid biosynthesis pathways lack these introns (Burge *et al.*, 1998).

Besides the rarity of U12-type introns compared to the U2-type, several other differences between the two intron types have also been noted: first, U12-type introns feature a more conserved 5′ss and branch point site compared to the U2-type introns (figure 1) (Levine and Durbin, 2001; Lopez and Seraphin, 1999). Furthermore, the distances of the branch points to the 3′ss are more restricted in the U12-type introns, *i.e.* generally in the range of [-21,-8] nucleotides upstream of the 3′ss. Additionally, U12-type introns show no or very weak polypyrimidine tract downstream of their branch site and upstream of their 3′ss compared to in the U2-type introns. These conserved features are essential for proper U12-type intron recognition and in fact, mutations in the 5′ss of the U12-type introns can convert them to U2-type introns/5′ss (Dietrich *et al.*, 1997; Burge *et al.*, 1998), Also, U12-type branch point sequences located at positions beyond the [-20,-10] window relatively to 3′ss have been shown to lead to activation of cryptic U2-type splice sites (Dietrich *et al.*, 2001). The highly conserved sequences at the U12-type 5′ss and branch sites/3′ss suggest that they are more inclined to be correctly recognized even in the absence of external splicing factors which typically assist splice site recognition with the U2-type introns (1.2) (Abril *et al.*, 2005; Sheth *et al.*, 2006). The average size of the two introns are not drastically different, *i.e.* ~3600 bp for the U12-type introns *vs* ~4130 for the U2-type. However, in contrast to U2-type introns (where their size peaks around 90 bp) U12 type introns with sizes smaller than 100 bp are extremely rare (Levine and Durbin, 2001).

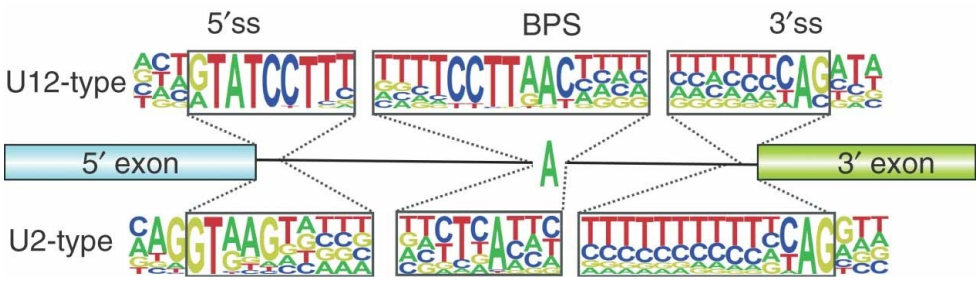

Figure 1. Consensus sequences of 5′ ss, 3′ ss and branch points of U12 and U2 type introns. Adapted from (Turunen, Niemelä, *et al.*, 2013)

1.1.3. U2- vs U12-type spliceosome

As mentioned previously introns are recognized and removed by a ribonucleoprotein complex called the spliceosome. The U2-dependent spliceosome (also known as major spliceosome) consists of 5 small nuclear RNAs (snRNAs), which, together with protein factors make up small nuclear ribonucleoproteins (snRNPs), *i.e.* U1, U2, U4, U5, and U6 (Will and Lührmann, 2011; Matera and Wang, 2014). As shown in figure 2, initially U1 snRNP recognizes the 5′ ss of the U2-type intron while a non-snRNP protein factor SF1 (described in 1.2) detects the branch point (Mount *et al.*, 1983; Zhuang and Weiner, 1986; Liu *et al.*, 2001). Subsequently, a U2AF heterodimer consisting of 65 kDa and 35 and kDa subunits bind to polypyrimidine tract and 3′ ss, respectively, and form the E complex (Zamore and Green, 1989; Zorio and Blumenthal, 1999). The A complex is formed by U2 being recruited by U2AF1 to replace SF1 at the branch point (Ruskin *et al.*, 1988; Valcarcel *et al.*, 1996). Subsequently, the U2 snRNA forms base pairing with the branch site which causes the branch site adenosine to bulge out. The adenosine 2’ hydroxyl can function in the later catalytic steps (Wu and Manley, 1989; Query *et al.*, 1994). Later, a tri-snRNP formed by U4, U5 and U6 associate to the spliceosome which leads to a dissociation of the U1 snRNP, thus forming the B complex (Konarska and Sharp, 1987). After a large number of rearrangements in RNA-RNA interactions and in protein composition (Will and Lührmann, 2011), the spliceosome reaches catalytically active conformation (B*) and catalyzes the first step of splicing (Wu and Manley, 1989; Query *et al.*, 1994). Splicing factors associate to assist spliceosome undergo further conformational changes that leads to forming C complex. Eventually the complex is disassembled and the lariat intron is released. The disassembled spliceosome components are recycled for use in later splicing reactions (Tsai *et al.*, 2005).

In comparison, the U12-dependent spliceosome (minor spliceosome) includes four specific snRNAs, namely U11, U12, U4atac and U6atac; however, it also shares the U5 snRNA with the U2-dependent spliceosome. The U12-dependent spliceosome includes seven specific protein components which all are located in the U11/U12 di-snRNP and not found in the U2-dependent spliceosome (table 1). The assembly of the U12-dependent spliceosome is overall similar to that of major spliceosome except for the differences in the initial intron recognition stage (Patel and Steitz, 2003) (see figure 2). Initially, the U11 and U12 snRNPs (as a preformed U11/U12 di-snRNP) bind simultaneously to the 5′ ss and the branch point sequence of the U12-type introns (Frilander and Steitz, 1999). As opposed to U1 snRNP, U11 neither form base-pairing interactions beyond the exon-intron boundary nor with the first 3 bases of the U12-type 5′ ss (Hall and Padgett, 1994; Kolossova and Padgett, 1997). Instead these nucleotides are recognized by the 48K protein (Turunen *et al.*, 2008). As mentioned previously, U12-type introns do not have a defined polypyrimidine tract and hence the U2AF65/35 heterodimer does not function in the recognition of U12-type introns. Rather, this recognition is more dependent on RNA/RNA interactions at the branch point sequence (Brock *et al.*, 2008). After U12 binding, the branch site adenosine is bulged out similarly as in the major spliceosome (Tarn and Steitz, 1996). Binding ZRSR2 (URP) protein (and possibly ZRSR1 (Horiuchi *et al.*, 2018)) is needed for the 3′ ss recognition and A complex formation (Shen *et al.*, 2010). Next, the U4atac/U6atac.U5 tri-snRNP enters the nascent spliceosome and similarly as in the major spliceosome this leads to a dissociation of U11 snRNP from the 5′ ss and pairing formation of U6atac/5′ ss interaction. Subsequent RNA-RNA rearrangements lead to unwinding of U4atac/U6atac duplex, dissociation of U4atac from the complex, and base pairings between U6atac and U12 (B complex). These lead to catalytic core formation (Tarn and Steitz, 1996; Frilander and Steitz, 2001) and juxtaposition of the 5′ ss and the branch point adenosine for the first catalytic step. Following the

exon ligation the lariat intron is released. Similar to the U2-dependent spliceosome, the spliceosome components are disassembled and recycled (Damianov *et al.*, 2004).

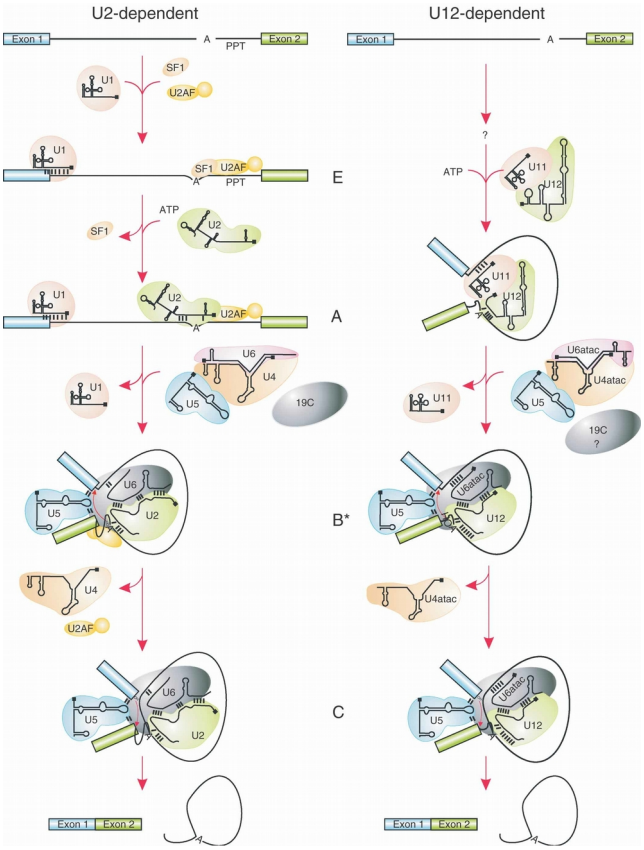

Figure 2. U2-dependent (on the left) and U12-dependent (on the right) spliceosome assembly, including labels that show the formation of E, A, B* and C complexes. Adapted from (Turunen, Niemelä, *et al.*, 2013)

1.1.4. The origins of the major and minor spliceosomes and introns

Since their discovery, various theories have been proposed regarding the origin of the U12-dependent spliceosome and its evolution in relation to that of U2-dependent spliceosome. A prevalent theory is the so-called fission/fusion model which postulates that both spliceosomes are descendants of the same ancient spliceosome which initially evolved independently (presumably in separate organismal lineages) thus leading to fission of the original spliceosome to two separate lineages. A subsequent fusion of these organism led to their coexistence in a single cell/genome (Burge *et al.*, 1998). The three main points in support of the fission/fusion model are: i) the presence of U12-type introns in higher proportions in the U12 genes (*i.e.* U12-type intron containing genes) compared to their expected proportions if they were randomly distributed, ii) the relatively over-representation of nonhomologous U12-type introns in paralogous genes, and iii) the association of the U12-type introns with information processing functions (Burge *et al.*, 1998). The main criticism against this model is that for the fission-fusion to be possible it needs to fulfill two opposing conditions: the dissimilarities between the two lineages at the isolation stage would have had to be both high enough to allow the retention of their separate functions upon reuniting, and low enough to allow the two hybridized lineages to be reproductively compatible; *i.e.* conditions that were unlikely to coexist. Moreover, the number

of introns in the host genomes (in isolation stage) would have had to be effectively small enough to allow all of them to be spliced (Lynch and Richardson, 2002; Lynch, 2007). Moreover, evidence exists that the spliceosomal machinery is evolutionary relatively stable; *e.g.* despite the billion years of separation yeasts can splice mammalian introns (Trachtulec and Forejt, 1999; Kunze *et al.*, 2000).

Protein (HUGO name)	12S U1	17S U2	18S U11/ U12	Function
Sm proteins: B/B', D1, D2, D3, E, F, and G	✓	✓	✓	snRNP core components (Will and Lührmann, 2011)
U1 A (SNRPA)	✓			Structural; RNA binding (Pomeranz Krummel *et al.*, 2009)
U1 C (SNRPC)	✓			5′ss recognition (Pomeranz Krummel *et al.*, 2009)
U1 70K (SNRNP70)	✓			Structural; SR protein interactions(Pomeranz Krummel *et al.*, 2009; Cho *et al.*, 2011)
U2A′ (SNRPA1)		✓		Structural; RNA-binding
U2B″ (SNRPB2)		✓		Structural; RNA-binding
SF3a complex		✓		BPS binding (Gozani *et al.*, 1996)
SF3b complex		✓	✓	BPS binding (Gozani *et al.*, 1996; Will and Lührmann, 2011)
20K (ZMAT5)			✓	Unknown; homology to U1C (Will *et al.*, 2004)
25K (SNRNP25)			✓	Unknown (Will *et al.*, 2004)
31K (ZCRB1)			✓	Unknown; RNA-binding (Kim *et al.*, 2010)
35K (SNRNP35)			✓	SR protein interactions, homology to U1-70K (Will *et al.*, 1999; Lorkovic *et al.*, 2004, 2005)
48K (SNRNP48)			✓	5′ ss recognition (Turunen *et al.*, 2008)
59K (PDCD7)			✓	Structural, binds 48K and 65K (Benecke *et al.*, 2005; Turunen *et al.*, 2008)
65K (RNPC3)			✓	Structural, binds U12 snRNA (Benecke *et al.*, 2005; Norppa *et al.*, 2018)
ZRSR2 (URP)			✓	3′ ss recognition (Shen *et al.*, 2010)
ZRSR1			✓	Intron-less active pseudo gene of ZRSR2 with similar function (Horiuchi *et al.*, 2018)
hPrp43 (DHX15)		✓	✓	
Y Box-1 (YBX1)			✓	

Table 1. Comparison of the protein composition of the major spliceosome U1 and U2 snRNPs and the minor spliceosome U11/U12 di-snRNPs, which both function in the initial intron recognition. Adapted from (Turunen, Niemelä, *et al.*, 2013). Note that the ZRSR2/Urp and Y Box-1 are present in the Major spliceosome but absent from U1 or U2 snRNPs.

An alternative theory proposes that origins of the two type of introns (and spliceosomes) are two organellar group II introns that invaded the primordial nuclear genome (Lynch and Richardson, 2002). Accordingly, the similarities between the two splicing machinery are due to convergence, and the sparseness of the U12-type introns in comparison to the U2-type was later reached (or exacerbated) upon loss or conversion of the U12-type introns to the U2-type. As mentioned in section 1.1.2 the 5′ ss and branchpoint consensus sequences of the U12-type introns are more conserved compared to that of U2-type and mutations in their sequences have shown to convert the introns. In case of complete loss of U12-type introns, similar to in *C. Elegans*, the minor spliceosome would eventually be degenerated preventing any future recolonization of the U12-type introns (Bartschat and Samuelsson, 2010).

1.2. Alternative splicing

When the human genome was first sequenced in early 2000s, initial analysis suggested that the number of protein coding genes in human were about 20,000 to 25,000 (International Human Genome Consortium, 2004). This raised the question of how a limited number of genes can code for a great variety of known phenotypes and functions? In response, alternative splicing (AS) was suggested as a process by which genes can code for many phenotypes (International Human Genome Consortium, 2001). AS is referred to as various combinations of splicing of introns in pre-mRNAs that can lead to different combinations of exons being included in the mRNAs. It has recently been discovered that most mammalian genes (*e.g.* > 95% of human genes) can produce multiple mRNAs through this process (Wang *et al.*, 2008; Pan *et al.*, 2008; Merkin *et al.*, 2012) and many other multicellular species have also been shown to feature AS (Stolc *et al.*, 2004). Moreover, contrary to the assumption that they lack alternative splicing entirely, single cell organisms have recently been discovered to feature a few genes that undergo AS (McGuire *et al.*, 2008) thus pushing the estimated date for the origin of alternative splicing to the pre-multicellular eukaryotes era (Irimia, Rukov, *et al.*, 2007). Various types of AS exist including exon skipping, exon inclusion, alternative 5′ and 3′ ss activation, mutually exclusive exons and intron retention (figure 3). Of these, cassette exon (*i.e.* exon skipping or inclusion) is the most reported (and most studied) AS event whereas intron retention is the least. The effect of AS on protein production can be immense. The potential isoforms emerging from a single gene can be hundreds or thousands (Nilsen and Graveley, 2010) and it can have profound biological consequences, *e.g.* in *Drosophila* various alternative isoforms of the gene *fru* (fruitless) determine the sex (Demir and Dickson, 2005). Alternative splicing is not common in the U12 genes. Of the few cases of alternative 5′ss, 3′ ss and exon skipping that has been reported, either their significance is not known or at least it is unknown whether the events are regulated or resulted from splicing errors (Zhu and Brendel, 2003; Chang *et al.*, 2007; Lin *et al.*, 2010). Additionally, there are a few examples where the major and minor introns/spliceosomes collaborate in mutually exclusive fashion. They allow the introns to be spliced either by the major or minor spliceosome in a tissue specific manner, *e.g. JNK2* gene in mouse (Chang *et al.*, 2007), or *prospero* gene in *D. melanogaster* (Scamborova *et al.*, 2004; Borah *et al.*, 2009).

The regulation of AS (or in general splicing) is regulated by various sequence elements within pre-mRNA collectively known as splicing regulated elements. These elements are located in both intronic and exonic locations and they provide binding sites for splicing factor (SF) proteins that function as either enhancers or inhibitors of intron/exon definition (reviewed by Wang and Burge, 2008). Predominantly common in the mammals, exon definition involves cross-exon interactions of splicing factors that bind near the 3′ ss upstream and 5′ ss downstream of the exon (Robberson *et al.*, 1990). Intron definition, however, involves cross-intron interactions of SFs

binding on the intron and near to its 5′ and 3′ ss (reviewed by Berget, 1995). It is common in plants, fungi and invertebrates. Two of the most studied classes of splicing regulatory elements are exonic splicing enhancers (ESE) and exonic splicing silencers (ESS) which are typically regulated by SR and hnRNP protein families, respectively. An examples of how SFs can regulate alternative splicing is when an exon is silenced due to inhibition, in addition to the possibility that the exon can be excluded from the mRNA (or skipped) alternative exons may find opportunity to be included yielding mutually exclusive exons, alternative 3′ ss or alternative 5′ ss activation. Intron retention regulation is however more complex and is dependent on a combination of splicing factors and dedicated mRNA export factors (Reed and Cheng, 2005; Sakabe and De Souza, 2007; Wang and Burge, 2008). Interestingly, splicing factors are also known to be gene or location dependent, meaning that their function may vary or they may become inactive if their gene or location is changed. As an example some SR proteins promote splicing when bound to exonic sites but inhibit splicing when bound to the intronic sites (Kanopka *et al.*, 1996; Ibrahim *et al.*, 2005). The distance of the SR protein ESEs from the introns are also shown to affect their enhancing efficiency (Graveley *et al.*, 1998). One interesting observation related to the differences between regulation of splicing of U12 and U2-type introns is that due to the more restricted and conserved splice sites of the U12-type introns (1.1.2), the U12-type introns may be less dependent on external splicing factors (SFs) to undergo accurate splicing in comparison to U2-type introns. This may also contribute to the fact that the U12-type introns are rarely alternatively spliced.

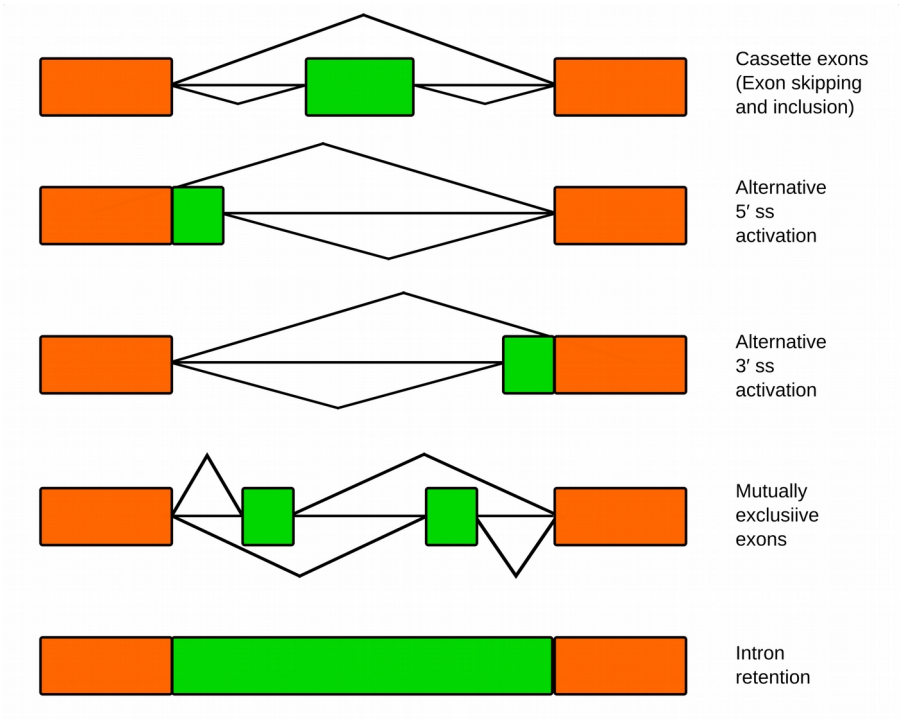

Figure 3. The main patterns of alternative splicing. The common exons are indicated with orange boxes. The exons involved in alternative splicing (the ones differing between isoforms) are shown with green boxes. The alternative splicing events are arranged above or beneath the horizontal lines, *i.e.* the reference introns.

Splicing factors have been described as a selective force driving the evolution of the spliceosome. Some have even argued that the evolution of these proteins built the selective pressure under which constitutively spliced exons adapted to be able to splice alternatively (Ast, 2004; Busch and Hertel, 2012). As an example, a binding of SR protein in the vicinity of an exon (which is solely spliced constitutively) may cause negative selection to that exon and the surrounding splice sites, ultimately leading to weaker surrounding splice sites (due to accumulation of mutations over time) and alternative splicing adaptivity of the exon. Others however argue that accumulation of alternative splice sites near the introns (due to mutations and weak negative selection) and maybe also weak intronic splice sites in the early spliceosomal introns (Irimia, Penny, *et al.*, 2007) can promote the rise of alternative splicing (Keren *et al.*, 2010; Rogozin *et al.*, 2012; Koonin *et al.*, 2013). Whether the SR proteins played a significant role or not, in support of the idea that weak splice sites might be involved with the alternative splicing evolution, it has been reported that alternatively spliced exons indeed feature weaker splice sites compared to constitutively spliced exons (Stamm *et al.*, 1994; Carmel *et al.*, 2004; Sorek *et al.*, 2004).

Features other than SFs also seem to be involved in the regulation of alternative splicing. In fact, various studies have attempted to discover various RNA features (noted as regulatory splicing codes) that are not only associated with the abundance of spliced isoforms but also can be used to predict the occurrence of various splicing events (Wang and Burge, 2008; Hallegger *et al.*, 2010; Barash *et al.*, 2010). Note that similar to SFs, *splicing code* is also not specific to alternative splicing regulation, and is also involved in regulation of constitutive splicing (Black, 2003; Fu, 2004) and may also determine whether an exon is alternatively or constitutively spliced (Barash *et al.*, 2010). One particular study examined 1014 RNA features embedded in exons and their flanking introns and analyzed their effect on the probability of exon inclusion, skipping or no-change (*i.e.* neither inclusion nor skipping) in the transcripts across several samples (Barash *et al.*, 2010). The features were grouped into 4 classes: previously reported or *known* motifs *e.g.* exoninc/intronic enhancers and silencers (*i.e.* 6-8 nucleotides long); novel motifs (*i.e.* 5-7 nucleotides long); short motifs (*i.e.* 1-3 nucleotides long); and features that describe the transcript structure. In order to properly address the tissue specificity of the alternative splicing events and the regulation of SFs the studied samples were also collected from 27 different mouse tissues. Various combination of features were evaluated based on the amount of information for which they accounted with regards to tissue-dependent splicing. Eventually, a combination of ~ 200 of the most significant RNA features were chosen to assemble a *splicing code* that can be used to predict inclusion/skipping of the exons. Interestingly, this assembled splicing code includes features from all 4 classes mentioned above, including features that are related to the structure of the transcripts *e.g.* intron/exon lengths, secondary structures and whether exon inclusion/exclusion introduces premature termination codons (PTC). Furthermore, excluding the RNA structural features from the *splicing code* leads to less accurate predictions, confirming their importance in regulation of alternative splicing (Barash *et al.*, 2010) .

1.2.1. Intron retention and intron detention

As mentioned above, intron retention (IR) is the least reported alternative splicing event in mammals, whereas in lower metazoans, fungi and protozoa it is the most common (Ner Gaon *et al.*, 2004; Keren *et al.*, 2010; McGuire *et al.*, 2008). Weak splice sites, small intron size, high GC content and increase in density of few intron splicing enhancers *i.e.* ISEs such as GGG and a number of ESS regulators are reported to be associated with intron retention hence IR is usually attributed to poor or mis-splicing of the introns due to the weak splice sites or lack of exon/intron

definition (Galante *et al.*, 2004; Sakabe and De Souza, 2007; Amit *et al.*, 2012). Furthermore, the inefficient or slower splicing of introns such as conditions described for U12-dependent spliceosome has been reported to result in the retention of the introns (Patel *et al.*, 2002; Pessa *et al.*, 2006). The retention of introns usually results in transcripts with large exons that include the retained introns and their flanking exons. Due to the inclusion of introns, mRNAs usually include one or more PTCs, which, if located farther than 50-55 nucleotides of an exon-exon junction it can trigger the degradation of transcripts by non-sense mediated mRNA decay (NMD) (Popp and Maquat, 2013). Preventing the production of faulty and potentially harmful proteins, other IR transcripts are degraded by the exosome nuclear degradation (described in details in 1.2.3), or they may be stable in the cell hence, also bearing the name “detained introns” (Boutz *et al.*, 2015).

Although IR has mostly been described as hazardous to the cells and destined for degradation, recently various regulatory functions have been discovered for a number of these transcripts. In plants the majority of the IR transcripts are not susceptible to NMD despite the fact that most feature PTCs (Kalyna *et al.*, 2012). Some IR transcripts may be post-transcriptionally spliced to increase the rate of protein production if needed. Examples of such regulatory functions of IRs has been seen both in plants *e.g. Marsilea vestita* (*i.e.* the hairy water clover) and animals *e.g. Nematostella vectensis* (*i.e.* the starlet sea anemone) (Boothby *et al.*, 2013; Moran *et al.*, 2008). In Drosophila the expression of the X chromosome gene *msl-2* (*i.e.* male-specific lethal 2) is regulated by producing IR isoforms in females and spliced isoforms in the males (Zhou *et al.*, 1995; Bashaw and Baker, 1995). Using this strategy the X-linked gene can produce functional proteins in males only, *i.e.* dosage compensation. Furthermore, in animals IR is reported to be associated with detection of RNA by retrotransposons and also exclusion of exons (with flanking retained introns regulated by hnRNPLL) (Buckley *et al.*, 2011; Cho *et al.*, 2014).

Finally, a number of IR mRNAs has also shown signs of protein production. One example is the production of a novel but truncated Cyclin D1b isoform as a consequence of retention of intron 4, which is upregulated in prostate and esophageal cancers and promote cellular transformation activity as opposed to other isoforms of Cyclin D1 (Solomon *et al.*, 2003; Comstock *et al.*, 2009). Another interesting case is IR of calcineurin gene (studied in mice) that produces CnAβ1 isoform with improved functions comparing to the canonical isoform (Felkin *et al.*, 2011). CnAβ1 plays a crucial role in myoblast proliferation and regeneration stimulation, and it is also reported to improve cardiac function (Felkin *et al.*, 2011; Padrón-Barthe *et al.*, 2018).

1.2.2. Cryptic splicing

In addition to the annotated splice sites in the genome of the eukaryotes, many other weaker and less active (or inactive) splice sites exist in genes that under some circumstances may become more active or even become the dominantly used splice site. Rather than being completely inactive in normal conditions, it is suggested that many of these cryptic splice sites are always active but at very low levels; hence, normally they are not easily detected (Kapustin *et al.*, 2011). It has also been shown that cryptic splice sites situated within the exonic regions of a gene in a species may be spliced under normal conditions in an orthologous gene in another species (Kapustin *et al.*, 2011). Defective intron recognition caused by spliceosomal mutations can activate the cryptic splice sites upon mis-splicing of the original site. These conditions are usually seen in splicing diseases (Tazi *et al.*, 2009; Verma *et al.*, 2018). Cryptic splice sites often cause translational frame shifts which in turn may cause PTCs and as a result lead to degradation of the mRNA by NMD.

In order to detect the cryptic splice sites in a sample, one strict and conservative approach is to consider all active splice sites that are absent form gene annotation databases such as RefSeq (Pruitt *et al.*, 2007) or Ensembl (Zerbino *et al.*, 2018). However due to the fact that many cryptic splice sites may be active in low levels under the normal conditions, one should also take into account the Ensembl/RefSeq annotated splice sites that are detected in low levels under normal conditions but are detected in high levels under the study conditions (*e.g.* splicing disease).

1.2.3. RNA degradation

Defective splicing of the introns or processing of mRNAs can lead to production of abnormal mRNAs. Subsequently, these aberrant mRNAs may lead to production of dysfunctional or harmful proteins. Various systems exist in the cell to degrade these mRNAs. RNA degradation may occur in cytoplasm or in the nucleus. The enzymes that are known to be involved are either endonucleases or exonucleases that cut RNAs internally, or from the 5′/3′ end respectively. Furthermore, the degradation proceeds either from 5′ to the 3′ of the mRNA or from 3′ to 5′ (reviewed by Houseley and Tollervey, 2009).

XRN2 exonuclease in mammals (*i.e. Rat1p* in yeasts) is an example of an enzyme that degrades in 5′→ 3′ direction in the nucleus. Only mRNAs that lack a 5′ cap are targeted by this exonuclease. Therefore most pre-mRNA transcripts that are capped with 7meG and bound by cap-binding complex are immune to this type of degradation. However, if the cap is removed by DCP2 decapping enzyme, for example during pol-II mediated transcription pausing, the resulting cap-less mRNA is degraded and transcription is prematurely terminated by the XRN2 activity. More generally, XRN2 has a central role in transcription termination. In a widely-accepted model for transcription terminaton (so called 'torpedo model') transcription proceeds past the polyadenylation site of the mRNA. Cleavage of the primary transcript by the polyadenylation machinery creates an unprotected 5' end for transcripts extending downstream of the cleavage site. This serves an entry point for the Xrn2 exonuclease, which start to chase the transcribing pol-II. When XRN2 reaches the transcribing pol-II the polymerase is detached from the DNA template, leading to termination of transcription (Luo *et al.*, 2006; Brannan *et al.*, 2012). The XRN2 exonuclease is also reported to co-transcriptionally degrade the abnormally spliced transcripts in mammals (Davidson *et al.*, 2012).

Nonsense mediated mRNA decay (NMD) is another mRNA degradation process that is known to occur in the cytoplasm. This pathway targets mRNAs that include PTCs. PTCs can result from, for example, nonsense or indel mutations, leading to frame shifts, cryptic splice site activation or intron retention (Leeds *et al.*, 1991; Lykke-Andersen *et al.*, 2000). If the PTC is located prior to an exon junction complex and farther than 50-55 nucleotides from the complex, during translation the NMD pathway will be activated (Zhang *et al.*, 1998; Le Hir *et al.*, 2000). NMD is also involved in regulation of alternative splicing. Many RNAs that are subjected to alternative splicing may be detected and degraded by this degradation pathway (Lewis *et al.*, 2003; Baek and Green, 2005). Many splicing factors (*e.g.* SR proteins) are also regulated in similar fashion, through a combination of alternative splicing and NMD (Lareau *et al.*, 2007). Over-expression and protein-production for several genes (*e.g. PTB*) are also reported to be compensated by including more PTCs in their mRNA products, making them susceptible to NMD degradation (Wollerton *et al.*, 2004). Furthermore, U1 and U11 snRNPs were both discovered to regulate their proteins using alternative splicing coupled with NMD (Rösel-Hillgärtner *et al.*, 2013; Verbeeren *et al.*, 2010). In the U1 regulation, activation of any of the three downstream 5′ ss in intron 7 of U1-70K activates a downstream 3′ ss. Regulated by U1C this leads to inclusion of a

PTC containing exon in the final mRNA products. Eventually these PTC containing mRNAs are detected and degraded by NMD which results in lack of functional U1-70K mRNAs and decreased protein levels. Due to requirement of both U1C and 70K for correct functioning of U1 snRNP, the lack of 70K leads to decrease in detection of cryptic 5′ ss, eventually driving the AS balance back towards producing more functional 70K mRNAs (Rösel-Hillgärtner *et al.*, 2013). Similarly the cellular levels of U11 snRNP are regulated through a tandem repeat of U12-dependent 5′ ss known as the USSE that exist in the fourth intron of *U11-48K* gene (and in the 3'UTR of *U11/U12 65K*). The U11-48K is known to be involved in the recognition of the 5′ ss of U12-type introns (Turunen *et al.*, 2008). Binding of U11 to USSE in 48K mRNA activates an upstream U2-dependent 3′ ss, causing frame-shift and PTC in the mRNA products (in comparison to the correctly spliced mRNA). This leads to detection and degradation of the alternatively spliced mRNAs by NMD and thus decreased levels of the 48K mRNA and protein. The lack of 48K leads to decreased recognition of USSE, driving the AS balance to production of more correctly spliced 48K mRNAs (Verbeeren *et al.*, 2010).

The *exosome* is a multi-protein complex active in the nucleolus, nucleus and cytoplasm of eukaryotes. It is constructed of 9 core subunits in the shape of a barrel. Six of the subunits are arranged to form a ring (*i.e.* Rrp41, Rrp45, Rrp46, Rrp43, Mtr3 and Rrp42) whereas 3 other proteins (*i.e.* Csl4, Rrp4 and Rrp40) form a cap on the side of the ring (Makino *et al.*, 2013). The catalytic/ribonuclease activities are carried out by subunits Rrp6 (*i.e.* PM/SCL-100 in mammalian system) and Dis3 (*i.e.* Rrp44) (Liu *et al.*, 2006; Dziembowski *et al.*, 2007). Furthermore, two paralogs of Dis3 are known in humans *i.e.* Dis3L1 which is associated with the exosome core, and Dis3L2 which is not associated (Staals *et al.*, 2010; Tomecki *et al.*, 2010; Astuti *et al.*, 2012). The two proteins Dis3L1 and RRP6 are exoRNases and degrade RNA in 3′→ 5′ direction, while Dis3 degrades RNA from 3′ to 5′ in both exonucleolytic and endonucleolytic manner (Lebreton *et al.*, 2008). Contrary to Dis3 which is mainly nuclear, Dis3L1/L2 are active in the cytoplasm (Tomecki *et al.*, 2010; Malecki *et al.*, 2013). Exosome is also involved in various other pathways such as maturation of several ncRNAs (*e.g.* rRNA, snRNA and snoRNAs) and mRNA quality control through regulation of NMD, non-stop decay (*i.e.* RNAs lacking termination codon) and no-go decay (*i.e.* ribosomal stalling during translation) (reviewed by Chlebowski *et al.*, 2013).

Finally, one of the strategies for RNA quality control and prevention of coding of the truncated mRNAs is detention of these mRNAs in the nucleus (reviewed by Schmid and Jensen, 2010; discussed in more detail in 1.2.1).

1.2.4. U12-type introns and diseases

Several diseases have been discovered that were caused by mutations in U12-dependent spliceosome, namely IGHD, MDS, EOCA, MOPD1/TALS, RFMN and LWS (table 2). All of these diseases are known to feature increased levels of the retention of U12-type introns; while other splicing defects, *e.g.* cryptic 5′ ss or 3′ ss activation, have been recognized in some of these diseases. Despite these reported abnormal splicings, correctly spliced mRNAs are also detected in all patients, hence the disease causing mutations are described as hypomorphic. These mutations lead to decreased levels (rather than the absence) of correctly spliced mRNA products (reviewed by Verma *et al.*, 2018).

IGHD is characterized by pituitary hypoplasia and a consequent lack of growth hormone. The form of this disease associated with minor spliceosome is caused by recessive mutations in the

RNPC3, i.e. a gene that codes for U11/U12-65K protein. The disease-causing mutations are compound heterozygous. In particular, they include one missense mutation *i.e.* P474T together with one of the two nonsense mutations R502X or R205X. The P474T and R502X mutations are located on the C-terminal of the RNA recognition motif (RRM) of 65K protein while the R205X is in its proline-rich region, situated between the two RRMs (figure 4A). The 65K protein is known to be involved in the U11/U12 di-snRNP formation (1.1.3). Another disease that has been associated with defective U12-dependent spliceosome protein component is MDS, *i.e.* a disease characterized by inefficient hematopoiesis and acute myeloid leukemia in extreme cases. Somatic mutations in various genes coding for splicing factors *e.g.U2AF1, ZRSR2, SRSF2 and SF3B1* have been associated with MDS (reviewed by Inoue *et al.*, 2016; Saez *et al.*, 2017). Usually mutations in only one of the mentioned splicing factors has been detected in each patient; however, this mutation usually occurs together with mutations in other genes that code for epigenetic factors and cell signaling/transcriptional regulators (Mian *et al.*, 2013). Here I focus on the MDS disease that is caused by mutations in *ZRSR2*/*Urp* which code for a protein involved in the recognition of 3′ss of U12-type introns (Madan *et al.*, 2015). These mutations include a wide-spread nonsense, missense, frameshift and splice site mutation which are scattered across the gene (figure 4D). Both increase of retention of U12-type introns and activation of cryptic U2-dependent splice sites has been reported in the MDS patients that were diagnosed with mutation in *ZRSR2* (Madan *et al.*, 2015). Both MDS and IGHD will be discussed in further details later (4.4).

The only presently known disease to be caused by mutations in the U12 snRNA gene (*RNU12*) is EOCA. It is characterized by defects in muscle coordination due to cerebellar hypoplasia and degeneration. Many patients also suffer from hypotonia in infancy, learning difficulties, seizures (febrile or complex partial), delayed motor skills. A 84C > T (homozygous) mutation in *RNU12* has been detected in a group of studied EOCA patients (figure 4B; Elsaid *et al.*, 2017). However, mechanistic details on how this mutation disrupts the function of the minor spliceosome is not exactly known at the moment. Both increased retention of U12-type introns and upregulation of U12 was detected in patient cells; however, no activation of cryptic splice sites were reported.

Finally, MOPD1/TALS, RFMN and LWS are 3 diseases known to be caused by defects in *RNU4ATAC* (figure 4C). As mentioned previously, U4atac snRNA joins U6atac and U5 to form a U4atac/U6atac.U5 tri-snRNP which enters the nascent spliceosome after the 5′ ss and 3′ ss are detected by the U11/U12 di-snRNP (1.1.3). The first out of the 3 mentioned diseases associated with defective *RNU4ATAC* (MOPD1/TALS) is a rare autosomal recessive disorder; meaning that the disease is developed only if both copies of the *RNU4ATAC* gene are mutated (He *et al.*, 2011; Edery *et al.*, 2011). In its extreme form, it is characterized by postnatal growth retardation, developmental defects, and microcephaly. The patients usually do not survive beyond 3 years after birth. Most mutations that cause the severe forms of MOPD1/TALS are located in the 5′ stemloop of U4atac snRNA (figure 4C). This region contains binding sites of several proteins, including 15.5 K (Snu13) and PRPF31, both of which are crucial for the correct functioning of U5 (*i.e.* forming U4atac/U6atac.U5 tri-snRNP) (Makarova *et al.*, 2002). The second disease (RFMN) is characterized by defects in growth, facial dysmorphia, cognitive delays and antibody deficiency. All reported cases of RFMN featured compound heterozygous mutations in *RNU4ATAC*, while the most severe cases of the disease shared one mutation with MOPD1/TALS (figure 4C). Other mutations associated with this disease are mainly located on stem II of U4atac/U6atac snRNA complex, with one case featuring mutation in the sm-site (figure 4C). Another autosomal recessive disease associated with defective *RNU4ATAC* is LWS. Mutations in

5′ and 3′ stemloops of the U4atac snRNA are shared with MOPD1/TALS, together with a mutation in the stem II of the U4atac/U6atac are associated with LWS. LWS is characterized with epiphyseal dysplasia and microcephaly; however, the symptoms are mild in comparison to TALS/MOPD1. Note that increase of retention of U12-type introns have been reported in MOPD1/TALS and RFMN patients but currently not in the LWS patients (He *et al.*, 2011; Edery *et al.*, 2011; Merico *et al.*, 2015). This may be due to lack of transcriptome analysis (Farach *et al.*, 2018). However judging on the many *RNU4ATAC* mutations in common between the 2 diseases (MOPD1/TALS and RFMN) and LWS (figure 4C) it would not be unexpected if similar to the two, LWS also features an increase of retention of U12-type introns. The wide range of splicing defects observed in diseases caused by defective U12-dependent spliceosome highlight the importance of suitable tools to be able to detect these splicing abnormalities, which is discussed in the following.

Defective component	Abbreviation	Disease name
U11/U12-65K (RNPC3)	IGHD	Isolated Growth Hormone Deficiency (III)
ZRSR2	MDS	Myelodysplastic Syndrome (Madan *et al.*, 2015)
RNU12	EOCA	Early-onset Cerebellar ataxia (Elsaid *et al.*, 2017)
RNU4ATAC	MOPD1/ TALS	Microcephalic osteodysplastic primordial dwarfism/ Taybi-Linder Syndrome (Edery *et al.*, 2011; He *et al.*, 2011)
RNU4ATAC	RFMN	Roifman syndrome (Merico *et al.*, 2015)
RNU4ATAC	LWS	Lowry Wood syndrome (Farach *et al.*, 2018)

Table 2. Human diseases caused by defective minor spliceosome (adapted from Verma *et al.*, 2018).

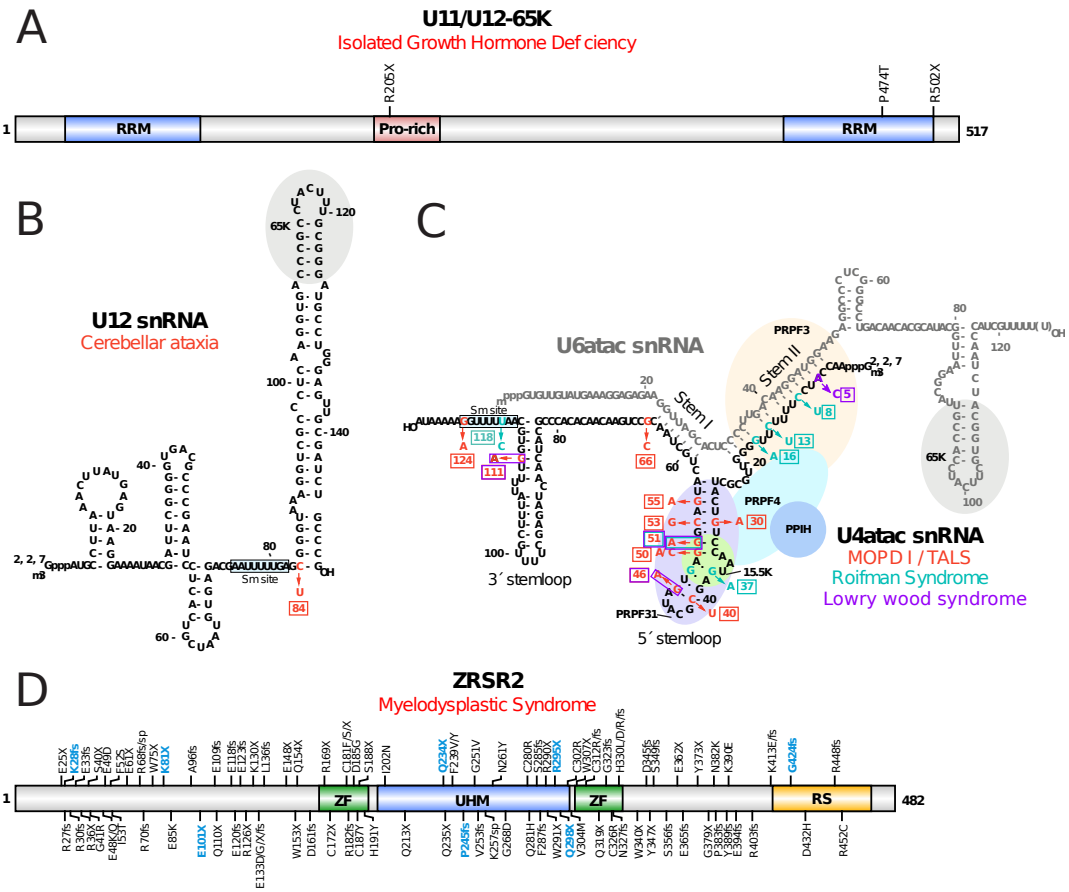

Figure 4. Mutations in various snRNA and protein components of U12-Dependent spliceosome that are associated with known diseases (adapted from Verma *et al.*, 2018), including the recently reported Lowry wood syndrome (Farach *et al.*, 2018). A) U11/U12-65K protein and the location of IGHD associated mutations. B) secondary structure of U12 snRNA and the location of the EOCA associated mutation. C) Secondary structure of U4atac snRNA and the location of MOPDI/TALS, RFMN and LWS associated mutations. D) ZRSR2 protein and the location of MDS associated mutations. Note that the 51G > A mutation in U4atac (C) has been identified in each of the three diseases TALS, RFMN and LWS whereas the 46G > A mutation is present in the two diseases MOPD1/TALS and LWS, and the 5A > C mutation is specific to the LWS disease only.

1.3. Bioinformatics methods for analyzing pre-mRNA splicing

Several bioinformatics methods have been described in details that are extremely useful to carry out various genome-wide splicing/transcriptome analysis such as: identification, detection and quantification of various expressed isoforms in the studied samples; detection of differentially expressed genes/isoforms (DEG) across several samples; IR abundance level estimation in a sample; discovery of significantly differentially retained introns across several samples; and discovery of U12- and U2-type introns within genes. Comparison of various RNAseq platforms and software that carry out mapping of the sequencing reads to the genome/mRNAs, however, is beyond the scope of this thesis (for a comprehensive review see Engström *et al.*, 2013; Li *et al.*, 2014; Su *et al.*, 2014).

1.3.1. Genome-wide transcriptome analysis using RNAseq data

Genes in many species, including a number of known single cell organisms undergo alternative splicing (1.2). Furthermore, in mammalian genes it has been discovered that the majority of the genes (*i.e.* >95% in human) undergo alternative splicing (Pan *et al.*, 2008; Merkin *et al.*, 2012). The occurrence of these events are regulated by splicing factors, many of which exhibit tissue/cell/developmental-stage specific activities that results in different alternative splicing events and gene expression abundance across various samples (Yeo *et al.*, 2004; Castle *et al.*, 2008; Grosso *et al.*, 2008). The generality of alternative splicing highlights the importance of detection of different isoforms and quantification of the isoform levels. Such information can be helpful to understand the biological pathways and mechanisms that are associated with various phenotypes such as progression of a disease (Costa *et al.*, 2013). Several analysis methods are of interest when studying the transcriptome; namely identification of the expressed isoforms including novel isoforms that previously were not discovered; detection of genes and isoforms, *i.e.* whether they are expressed or not; quantification of isoforms; and differential expression of the isoforms (Angelini *et al.*, 2014). Many of the existing software tools tackle these challenges independently, making determination of a best performing method complicated since it requires one to consider all different combinations of the individual tools and their various parameter settings.

1.3.2. Identification,detection and quantification of expressed isoforms using RNAseq data

One of the main questions in RNAseq data analysis is the identity of the isoforms and their relative expression levels in the studied sample. To address this, a number of recent studies have focused on the available software that carry out three main analyses, *i.e.* identification, detection and quantification of all expressed isoforms. These studies also attempted to evaluate and compare the methods that carry out these analysis (Steijger *et al.*, 2013; Angelini *et al.*, 2014). One study (Angelini *et al.*, 2014) evaluated several of these software, including the popular Cufflinks (Trapnell *et al.*, 2010) which is part of the Tuxedo analysis pipeline (Trapnell *et al.*, 2012), by employing several simulated data constructed of either paired-end or single reads of various lengths. Their analysis were carried out under three modes: quantifying isoforms strictly and solely from a predefined annotated set of isoforms; incorporating predefined annotated isoforms to build novel isoforms and then quantifying the new set of isoforms which also includes the novel isoforms; and finally building all isoforms *de novo* while considering no prior knowledge of the possible isoforms and then quantifying the built isoforms.

The available isoform detection/quantification software were overall more successful when they took into account the known isoform annotations either during mapping of the reads to the genome or when detecting the expressed isoforms. This success was however, specific to detection of the high and moderately expressed isoforms since none of the studied methods were able to accurately detect the low expressed isoforms (Angelini *et al.*, 2014). Furthermore, despite the fact that increasing the read length and sequencing depth can improve the results, neither completely resolves the detection problem with the low expressed isoforms. In fact, there seems to be a saturation point for the read depth that once reached, ceases to improve the accuracy but rather leads to discovering more false positives (Angelini *et al.*, 2014).

Apart from detection of the isoforms, discovery of their true expression abundance is even more difficult and prone to errors (Steijger *et al.*, 2013; Angelini *et al.*, 2014). To complicate the

situation even more, alternative splicing complexity of higher eukaryotes such as humans imposes further limitations on the performance of the isoform detection and quantification software (Steijger *et al.*, 2013). To overcome this limitation it has been suggested that incorporating biological replicates may be a solution to improve the accuracy of detection and quantification of the expressed isoforms (Vardhanabhuti *et al.*, 2013; Steijger *et al.*, 2013). Studies show that there remains space for further improvements both on the technical level *e.g.* adaptation of sequencing platforms that use long reads, and in the analysis methods *e.g.* incorporating biological replicates to improve detection and quantification of the isoforms (Steijger *et al.*, 2013; Angelini *et al.*, 2014).

1.3.3. Discovery of differentially expressed genes using RNAseq data

In biological studies, apart from detecting and quantifying the expressed isoforms in the samples, detecting genes/isoforms that are differentially expressed across several studied samples are also of interest. In this regard several recent studies have attempted to evaluate the various available differential expression analysis tools and compare them to one another (Soneson and Delorenzi, 2013; Rapaport *et al.*, 2013; Tang *et al.*, 2015). Here I will describe a brief summary of their findings.

First, several studies have demonstrated that due to the different normalization and algorithms used to detect the differentially expressed genes (DEG), different methods/algorithms (e.g. IntEREst and DESeq) report substantially different number of DEGs (Rapaport *et al.*, 2013; Tang *et al.*, 2015). Reassuringly, the popular methods that are frequently used for DEG discovery in RNAseq data *i.e.* edgeR and DESeq overall perform very well, especially under the conditions where biological replication is limited *i.e.* < 3 (Robinson *et al.*, 2010; Anders and Huber, 2010; Love *et al.*, 2014; Soneson and Delorenzi, 2013). However, Limma that was originally developed for microarray analysis (but were recently adjusted to analyze RNASeq data) perform as well as edgeR or DESeq, especially for data with large numbers of biological replicates (Ritchie *et al.*, 2015). This is surprising since Limma assumes that the data is normally distributed, which may be reasonable for microarray data but does not apply to RNAseq data (Di *et al.*, 2011). For DEG discovery in a more complicated setting *e.g.* among a larger number of groups of samples, edgeR and DESeq outperform Limma (Tang *et al.*, 2015).

In RNAseq data, the read counts assigned to individual genes are described to follow a Poisson distribution across the technical replicates of a sample, while they follow an over-dispersed Poisson such as negative binomial (NB) across biological replicates (Marioni *et al.*, 2008; Di *et al.*, 2011). This over-dispersion arises from the fact that the variance of the expression of genes across multiple samples is larger than their mean expression (Robinson and Smyth, 2007; Di *et al.*, 2011). To model for the dispersion, methods such as edgeR and DESeq assume dependency of the variance v to the mean μ by considering a $v = \mu + \alpha \mu^2$ relationship where α is the dispersion factor that can be estimated from the data. However, judging on the surprisingly good performance of Limma in detecting DEGs in RNAseq data, it is suggested that for a large enough data set with a large number of genes (and more than 3 biological replicates) accurate modeling of the variance may be more important than the discreteness in the data (Rapaport *et al.*, 2013). The normalization used by various methods can also heavily influence the final discovered DEG results (reviewed and benchmarked by Dillies *et al.*, 2013). For instance, edgeR uses trimmed mean of *m* values (TMM) where it assumes that most genes are not differentially expressed across the samples. It initially filters the genes that were not expressed across any of the samples and later also removes the genes showing the most extreme (high/low) expressions

in each sample. Subsequently, edgeR chooses one sample as the reference and scale all the other samples to the reference by measuring the weighted mean of log ratios of the expression values. The scaled values are eventually centered to 1 (Robinson and Oshlack, 2010). DESeq also assumes that most genes are not differentially expressed across the samples. The normalization in DESeq is carried out according to scaling factors measured for each sample. Each scaling factor is measured by the median of log ratio of expression levels of the genes to the geometric mean of the expression of the genes across all samples, while excluding all genes that are not expressed in one or more samples (Anders and Huber, 2010).

Finally, evaluations of DEG analysis software using simulated data suggests that methods show improved performance if the following conditions apply: in comparisons with two groups of samples, one group features symmetric changes of gene expression levels (*i.e.* both up and down-regulated genes rather than solely up-regulated genes); and in comparisons of several groups (*e.g.* G1, G2 and G3) every group features significantly differentially expressed genes compared to the other group with an overall rising/decreasing order *e.g.* G1<G2<G3 or G2<G1<G3 (Soneson and Delorenzi, 2013; Tang *et al.*, 2015). It is notable that expanding the comparison of 2 groups of samples to 3 in the evaluation of DEG discovery software better outlines the differences in performance of these tools and explains the popularity of RNAseq specific methods *e.g.* DESeq and edgeR over the methods originally developed for microarray analysis e.g. Limma (Tang *et al.*, 2015). This shows that in order to better understand the advantages and disadvantages of various DEG discovery methods they should also be evaluated under more complex settings and design experiments that consists of more than a comparison of 2 groups of samples; *i.e.* a criteria that future comparison studies may take into account.

1.3.4. Estimation of IR abundance and differential IR analysis

Conditions that can lead to poor splicing or mis-splicing of the introns *e.g.* weak splice sites, small introns, high GC content and *etc.* together with slower/inefficient splicing of the introns can lead to their retention (1.2.1). Transcripts that contain retained introns are usually destined to degradation, regulatory functions and in some cases protein production (1.2.1). Therefore, discovery of transcripts with retained introns in a sample or identifying those featuring significantly differential intron retention (IR) across several studied samples are of potential interest. In fact, many previously described isoform reconstruction and detection methods (1.3.3) can be applied to detect intron-containing isoforms. Subsequently, these built isoforms can also be used to detect the transcripts with significantly differential IRs across several samples by running various well known DEG discovery methods *e.g.* DESeq, edgeR, Cuffdiff (1.3.4). While it is already a challenging task to accurately detect and quantify various isoforms in a sample (1.3.3), detecting/quantifying isoforms with retained introns has proven to be even more difficult (Steijger *et al.*, 2013; Vanichkina *et al.*, 2018). This is attributed to several particular characteristics of introns and IR transcripts: introns are normally much larger in length compared to exons; hence, they require much more sequencing reads to fully cover their length (*i.e.* conditions usually not met by biological studies performing RNAseq); introns feature more repetitive sequence elements *i.e.* regions to which a small number of reads may be mapped accurately (by the mapping software) due to uncertainties caused by multi-mapping reads; some introns feature non-coding RNAs (*e.g.* miRNAs or snoRNAs) which, if expressed, may influence the read-counts of those introns; the detection of IR isoforms requires excessive sequencing depth if they are less expressed compared to their spliced counterparts *i.e.* isoforms lacking introns entirely (Vanichkina *et al.*, 2018).

To resolve the shortcomings of the available isoform analysis software in detecting, quantifying and differential expression analysis of IR isoforms, recently a number of dedicated software tools have been developed that specifically analyze these isoforms, *e.g.* IRFinder and KMA (Middleton *et al.*, 2017; Pimentel *et al.*, 2016). As for instance, the IRFinder measures the *intronic abundance* (*i.e.* the number of reads that map to the intron) and *exonic abundance* (*i.e.* the sum of the intron-spanning reads that partially mapping to the upstream and downstream exons). It measures the IR ratio for each intron, *i.e.* relative intronic abundance over the sum of intronic abundance and exonic abundance. Furthermore, it excludes intronic regions that overlap the reference features (provided by users as a GTF file) and introns with radically high and low IR ratio. The differentially retained introns are discovered by analyzing IR ratio changes using other methods such as DESeq2 (Love *et al.*, 2014). The existing methods however do not completely resolve all the issues with detecting the IR isoforms that we mentioned above. This leaves room for development of methods for improved IR detection, quantification and differential analysis.

1.3.5. Computational detection of U2 and U12-type introns

Two classes of introns have been discovered in the eukaryotes *i.e.* denoted either as U12- or U2-type introns (1.1.2). Furthermore, two subtypes of U12-type introns have been described to date, the GT-AG variant (with initial GT and terminal AG di-nucleotides) and the AT-AC variant. Both GT-AG and AT-AC terminal dinucleotides are however found in U2-type introns as well (Dietrich *et al.*, 1997); hence, additional sequence information other than the initial and terminal di-nucleotides should be taken to account in order to be able to distinguish between U12- and U2-type introns. Traditionally, consensus sequences were built and applied to detect the two intron types. In fact, this lead to the realization that the splice site patterns for the U12-type introns were more constrained than those of the U2-type (Dietrich *et al.*, 1997; Sharp and Burge, 1997). Upon discovery of an increasing number of U12-type introns across the eukaryotic lineages the usage of position weight matrices (PWM), also known as position-specific scoring matrix (PSSM), for detecting U12-type introns became more popular (Staden, 1984; Burge *et al.*, 1998; Levine and Durbin, 2001; Zhu and Brendel, 2003). PWM is a two dimensional array with one dimension corresponding to the nucleotides (A, C, G and T) and the other dimension corresponding to positions (*e.g.* within splice sites of the introns); each value in the array represent a score describing the probability of finding a nucleotide (or amino acid) at a position. PWM was first introduced and applied to detect ribosome binding sites which define the translational initiation sites in the genes (Stormo *et al.*, 1982). Using PWMs over consensus sequences is advantageous as it allows incorporation of information related to the probability of observing different variants in each position (Staden, 1984). In later studies, splicing donor, branch-point and acceptor PWMs were constructed by measuring logarithm (in base 2) of probability of each nucleotide occurring at a position in the corresponding U12/U2-dependent splice site (box1, formula 3; Staden, 1984; Levine and Durbin, 2001). To estimate these probability values, the number of each nucleotide in each position across several aligned sequences (*i.e.* splice sites of a group of known U2/U12-type introns) were counted and then divided to the number of sequences (box 1, formula 1 and 2). Based on the PWMs match scores were generated for donor/branch-point/acceptor sites of the U2 and U12-type introns (box 1, formula 4). This includes checking the nucleotides within each splice site individually and looking up the $\log_2$-probability of observing the nucleotide at the corresponding position from the PWMs, and eventually summing the extracted $\log_2$-probability values (box 1, formula 4). Eventually, threshold cutoffs for donor/branch-point/acceptor match scores were applied to classify each intron as either U2 or U12-type (Staden, 1984; Levine and Durbin, 2001; Alioto,

2007). As an alternative to measuring these 3 match scores separately (corresponding to the donor, branch-point and acceptor sites), other studies have measured a pair of log odds ratios to detect whether an intronic sequence is U2 or U12-type. These values correspond to the relative probability that the sequence matches the 5′ ss (or respectively 3′ ss) of U12-type introns over the probability that the same sequence matches the 5′ ss (or respectively 3′ ss) of U2-type introns (Burge *et al.*, 1998; Zhu and Brendel, 2003). Furthermore, several other studies have converted the PWMs to log-odds matrices before they used them for scoring the splice sites of the introns. The conversion is carried out by taking the logarithm (base 2) of the relative probabilities at each position of the PWM against a background probability, e.g. uniform probability of 0.25 for each nucleotide (Wasserman and Sandelin, 2004; Sheth *et al.*, 2006; Madan *et al.*, 2015; Merico *et al.*, 2015). Note that when generating the PWM, in order to avoid overfitting and complications caused by calculating the logarithm of zero many studies replace zero with a low pseudo-count value in the PWM when a nucleotide is completely absent from a position in the studied sequences (Burge *et al.*, 1998; Wasserman and Sandelin, 2004; Sheth *et al.*, 2006). Various methods adapted by the mentioned studies, and also the completion of the genome annotations have yielded reports stating different number of U12-type introns for the same species, *e.g.* human. A relatively recent study that adapted in comparison a less strict and more inclusive method of detecting U12-type introns discovered 744 unique U12-type introns within the human genome, *i.e.* GRCh37 (hg19) assembly (Merico *et al.*, 2015). This study focused on the 5′ ss and the branch point nucleotide sequences of the introns only to build PWMs and discover novel U12-type introns. They also excluded the initial di-nucleotides of the 5′ ss since these bases are highly conserved in U12-type introns (*i.e.* either GT or AT).

Detecting the U12-type introns allows comparing various features of these introns to the more common U2-type introns. Moreover, it is the first step in studying the mechanistic differences in the splicing and the regulation of the transcripts that include U12-type introns to those that lack them.

Presuming that $f(b,i)$ is the frequency of observing nucleotide *b* (*i.e.* either A,C,G or T) at position *i* across *N* aligned sequences of size *k*, the probability of observing nucleotide *b* at position *i* of the aligned sequences is $p(b,i)$ and is calculated by the following formula:

Regular probability calculation: $p(b,i)=\frac{f(b,i)}{N}$. (1)

For an efficient computation, often the *logarithm* of the probability is used. To avoid obtaining *null* values (logarithm of zero), the $p(b,i)$ formula can be corrected using pseudo counts.

Corrected probability: $p(b,i)=\frac{f(b,i)+s(b)}{N+\sum_{b'\in\{A,C,G,T\}} s(b')}$. (2)

* $s()$ is the pseudo count function.

**Assuming that the length of the input sequences is *k*, $i=\{1,2,3,\ldots,k\}$.

Each element of the PWM represents a weight score for observing nucleotide *b* at position *i* and is shown here with $W(b,i)$.

PWM construction: $W(b,i)=\log_2\frac{p(b,i)}{p(b)}$. (3)

* $p(b)$ is the background probability of base *b*, usually replaced with 0.25 .

To score sequence $S=s_1s_2s_3...s_k$ according to the PWM (*i.e.* $W(b,i)$ matrix) formula 4 can be used.

Evaluation of a given sequence: $E(S)=\sum_{i=1}^{k} W(s_i,i)$. (4)

The information content in position *i* (in bits) is shown here with B_i and can be calculated as following:

Information content estimation: $B_i=2+\sum_b p(b,i)\log_2 p(b,i)$.

Box 1. Formulas used for constructing PWMs, evaluating sequences based on PWMs and estimating the information content of each position of the aligned sequences *e.g.* splice sites (Wasserman and Sandelin, 2004).

2. Aims of the study

In response to the growing interest to study intron retention (IR) events and attempting to overcome various technical difficulties that are known in detecting these events, this thesis initially presents Intron-Exon Retention Estimator *i.e.* IntEREst (I). IntEREst is an R/Bioconductor package that can be used for detection, quantification and differential analysis of IR levels across various samples (I). The software also supports tools to detect the U12- and U2-type introns based on their splice-site sequences. Furthermore, it supports comparison of retention levels of various subclasses of introns (*e.g.* U12-type introns vs U2-type). Taking advantage of these 2 options the retention levels of the U12-type introns are compared to that of the U2-type introns across 8 MDS patients with mutations in their *ZRSR2* genes and 8 controls (i.e. 4 healthy individuals and 4 patients lacking the *ZRSR2* mutation). We hypothesize higher retention of U12-type introns compared to U2-type in all samples due to the less efficient U12-dependent splicing (in comparison to U2-dependent), and exacerbation of the IR of U12-type introns (compared to that of U2-type introns) as a consequence of defective splicing of U12-type introns caused by *ZRSR2* mutations. The same results were expected to be acquired when analyzing a Maize data set consisting of 6 samples (3 roots and 3 shoots) featuring mutations in *RGH3* (*i.e.* an ortholog of human *ZRSR2*) and 6 samples that lack this mutation: An approximately 2 fold higher retention of U12-type introns compared to U2-type in all samples and the retention increase of U12-type introns for samples with defective RGH3.

In the second study, using an early version of the IntEREst software the long-time speculation that the U12-type introns are spliced less efficiently than the U2-type introns was investigated by comparing the genome-wide retention levels of the U12-type introns to that of U2-type introns (II) . Both nuclear and cytoplasmic extracts from HeLa cells were studied, each under 3 conditions : knockdown of RRP41 subunit of the exosome, knockdown of DIS3 subunit of the exosome and control. We expect genome-wide increased IR for U12-type introns compared to U2-type introns, with a more prominent effect to be observed in the nuclear samples with KD of exosome subunits. We also hypothesize that transcripts containing unspliced U12-type introns are preferentially detected and degraded by the exosome in the nucleus.

Finally, various splicing abnormalities including activation of cryptic U2-dependent splice-sites and increased retention of U12-type introns are seen in the cells from three sisters that were diagnosed with isolated growth hormone deficiency (IGHD) and pituitary hypoplasia caused by defective U11/U12-65K (III). The molecular basis of the disease are studied thoroughly in relation to the splicing abnormalities that are caused by defective 65K, a protein that is known to be involved in U11/U12 di-snRNP formation. These discovered mechanisms show that a defect in minor spliceosome component can lead to tissue specific consequences, through leading to abnormal splicing patterns in the genes containing U12-type introns.

3. Materials and methods

Method	Publication		
	I	II	III
*BA: Differential alternative splicing	√	√	√
BA: Differential expression gene		√	√
BA: Intron retention (detection, quantification and statistical differential analysis)	√	√	√
BA: RNAseq data analysis (*e.g.* read alignment and normalization)	√	√	√
BA: U12-type intron annotation	√	√	√
Cell fractionation		√	
Cell lines and cell culture		√	
Northern blot		√	√
Protein over-expression in mammals		√	
Quantitative PCR		√	
RNA sequencing		√	√
RNAi knock down		√	
RT-PCR		√	√
Site-directed mutagenesis		√	
Plasmid transfection		√	
Western blot		√	√

Table 3: Methods used in the publications. *BA stands for bioinformatics analysis.

In addition to the methods mentioned in tables 3, the following statistical tests have also been applied in this work :

Ordered alternative hypothesis tests are used to infer whether an order exist among two or more groups of IR values (I and II). The methods commonly used for theses tests are Kruskal Wallis and Jonckeere Terpstra (Kruskal and Wallis, 1952; Jonckheere, 1954). Both of these methods are non-parametric and as null hypothesis presume the medians of the compared groups of values to be equal. The alternative hypothesis for which the statistical test is used is that an increasing/ decreasing order exists among the compared groups of values. The latter method (i.e. Jonckheere Terpstra) is used in the publications (I and II) of this thesis, since contrary to the former (i.e. Kruskal Wallis) it can consider *a priori* order across the groups of IR values and tests whether the analyzed groups of IR values follow this particular order.

Enrichment analysis tests are commonly used to discover functions/pathways that may be associated with gene lists extracted from differential gene expression or other bioinformatics analyses. A simple and commonly used approach is to apply statistical methods such as Fisher's exact test, Hypergeometric distribution, Binomial probability and *etc.* to obtain p-values that indicate the probability that equal/more number of genes within the list may hit the feature randomly (II; Reviewed by Huang *et al.*, 2009). The same statistical methods can also be used to test whether two different subsets of genes (e.g. genes with highly retained introns in two separate samples) significantly overlap (II).

Differential expression analysis tools detect the genes, transcripts or exons that are significantly differentially expressed across several samples (Rapaport *et al.*, 2013; Tang *et al.*, 2015). These methods apply various normalization (1.3.3) and statistical tests to detect the significantly differentially expressed genes. The two commonly used methods edgeR and DESeq apply either Fisher’s exact test or Likelihood ratio test to estimate p-values. In contrast, the DESeq2 method supports Wald and Likelihood ratio tests, while DEXSeq applies the Likelihood-ratio test to analyze differential exon usage (Robinson *et al.*, 2010; Anders and Huber, 2010; Love *et al.*, 2014; Anders *et al.*, 2012; Reyes *et al.*, 2013). The p-values are based on the null hypothesis that when comparing controls to the treatment samples, differences observed in the expression of the genes are disregarding of the treatment and follow a Negative Binomial distribution (as expected) across the samples. The Wald test supported by DESeq2 (I), Likelihood ratio test of edgeR (I), Fisher’s exact test of edgeR (II) and the Likelihood ratio test of DEXSeq (I) have been used in the publications of this thesis.

4. Results and discussions

4.1. Development of an intron retention (IR) analysis tool

IR is the least studied alternative splicing event in mammals (1.2; 1.2.1). It is mainly caused by conditions such as weak splice sites, small size introns and is known to play a crucial role in RNA processing of plants, fungi and protozoa (1.2.1). Moreover, defects in spliceosome machinery can result in mis-splicing of the introns and their retention in the mRNA products, *i.e.* one of the main features of human spliceosomal diseases such as MOPD1/TALS (1.2.4). For these reasons the usefulness of a suitable dedicated intron retention analysis tool is imperative. In this regard, due to the fact that specialized computational tools to study IR events have only recently become available (1.3.4; 4.6), previous analysis of IR were performed using either custom developed software pipelines (Madan *et al.*, 2015) or publicly available general transcriptome assembly/abundance measurement software such as Cufflinks (Trapnell *et al.*, 2010). IR analysis has proven to be a difficult task especially using conventional isoform detection software (1.3.4). In this respect, a custom software pipeline called IntEREst (Intron-Exon Retention Estimator) was developed using R programming language to enable accurate detection of IR and comparison of IR levels across several samples (I). IntEREst also provides tools to detect U12-type and U2-type introns within the genome. This allows comparing of the retention levels of U12-type introns to that of U2-type introns at a genome-wide scale.

During the process of developing IntEREst, I initially formulated a custom pipeline to measure the retention levels of U12-type introns and compared them to U2-type introns at genome-wide scale using R programming language (II). In this initial version of IntEREst, the IR differential analysis were solely based on the reads that mapped to introns or intron-exon junctions. The reason for this was that we were initially using older RNAseq technologies that provided short sequence reads; *e.g.* SOLiD4 sequencing with 50 bps on positive and 35 bps on the negative strand (II). Furthermore, virtually no reads were mapped to the exon-exon junctions upon running the exon-exon junction mapping analysis software of SoLiD Bioscope. Therefore, the pipeline used only the intron mapping reads to determine the IR levels, estimate the FPKM normalized read counts and run comparison analysis to extract the introns with the most IR increase/decrease. For the differential IR analysis the edgeR package of R/Bioconductor was applied (Robinson *et al.*, 2010). Further improvements on the IR analysis pipeline resulted to the development of the present R-Bioconductor package of IntEREst (I).

The present version of IntEREst uses, in addition to intron mapping reads, also the exon-exon junction or intron-spanning reads (based on configurations set by the user) to detect the differentially retained introns. The read counts of the introns/exons are then used (together with sample annotation information) to detect the significantly higher or lower retained introns using one of the multiple IR differential analysis modules supported by IntEREst (4.5). The package also incorporates and recognizes standard R/Bioconductor objects such as *Summarized-Experiment*. IntEREst includes two read summarization functions that read mapped sequence-reads from a binary sequence alignment/map file (with *.bam* extension) progressively and enumerate the reads mapping to the introns and exons: *interest()* capable of running in parallel on several computing cores and *interest.sequential()* that runs only sequentially on a single core. In the latest release of IntEREst (v1.4.1) these two functions can be run in one or several modes to

count the intron mapping, exon-exon junction mapping and/or intron spanning reads. Several plotting tools are also supported by IntEREst. IntEREst requires a reference file that includes intron and exon coordinates. Either of Ensembl/RefSeq can be used to build the reference. Note that unless requested otherwise by the user (through the adjustable parameter settings) all overlapping exonic regions are collapsed to form large exonic regions in the reference and the remaining regions (with no overlapping exons) are considered as the intronic regions.

Despite the relatively high performance of IntEREst (4.5 and 4.6), there are several general challenges that are shared with all RNAseq analyses. Poor RNAseq and mapping quality can affect the results and eventually lead to discovery of false significant IRs. Genomic DNA contamination in RNAseq libraries may also cause spurious increases in the levels of the detected intron-mapping reads. Techniques to control for this however exist. When visualizing the binary sequence alignment/map files (with .bam extension) using software such as IGV (Thorvaldsdóttir *et al.*, 2013), if a high number of reads are mapped to the intergenic regions this indicates contamination of mRNAs with genomic DNA. A low percentage mapping of the reads to the genes is also another sign that the RNAseq libraries are contaminated with genomic DNA. Another source of noise in IR analysis is the presence of ribosomal RNAs (rRNAs) in the sample. Due to the high expression of rRNAs (*i.e.* up to 90% of the total RNA), they may bias the statistical differential IR analysis in IntEREst (especially if exon mapping reads are being analyzed). Therefor limiting the expression of rRNA using kits such as RiboMinus, filtering rRNA reads using SortMeRNA (Kopylova *et al.*, 2012), and excluding the ribosomal RNA genes from the reference can prevent biases that may be caused by these highly expressed RNAs. Filtering genes with low levels of intron mapping reads also improves the accuracy of detecting significant IRs (I, additional file 1). Read duplicates produced at the PCR duplication step of the sequencing may bias the results of the differential IR analysis; however, using a simple removal tool that filters all the duplicate reads (i.e. reads with exactly identical sequences that map to the same place in the genome) is not recommended since many of these reads may occur naturally; especially if they map to highly expressed genes (Bansal, 2017). Regions within the introns (or genes) that feature DNA repeats, or small RNAs (e.g. snoRNAs and miRNAs) can also affect the read counts of the introns. If their coordinates are provided, IntEREst can excluded them from the IR analysis. However, if the coordinates of intron coded RNAs or regions with many aligned PCR duplicate reads are not known, read-peak finding tools such as *macs* (usually used for ChIP-seq analysis) can be used to discover these regions (Zhang *et al.*, 2008). Recently, we received requests by several users to add two additional features to IntEREst (v1.4.1). One popular request was a downstream filtering function that removes the unreliably detected IRs, e.g. IRs that were detected based on low number of mapped reads. Additionally, implementation of an alternative IR normalization that is more suitable for cross-sample comparisons was requested. Although the FPKM normalization currently used by IntEREst corrects the IR levels for length of the introns and the library size of the sequenced samples, they do not however sum to a constant value within a sample. In contrast, scaled IR levels that sum to a constant value within individual samples (e.g. Transcripts Per Million/TPM), facilitate a more accurate cross-samples IR comparison. To address these issues, in future updates two functions will be added to IntEREst: An IR filtering function together with a function to scale the normalized levels such that the scaled IR levels in each sample would sum to a constant value (i.e. 1,000,000).

The latest release of the packages can be accessed through https://bioconductor.org/packages/release/bioc/html/IntEREst.html, moreover the latest develop version is available in https://www.bioconductor.org/packages/devel/bioc/html/IntEREst.html.

4.2. Global retention of U12-type introns

Earlier studies (Patel *et al.*, 2002; Pessa *et al.*, 2006) have provided experimental evidence that at least some U12-type introns show elevated IR levels compared to the U2-type introns. To address if this is true for most or all of the U12-type introns we used genome-wide analysis to compare the IR of U12-type introns to that of U2-type in the same genes (I; II and III). In agreement with the earlier reports that focus on a few specific U12-type introns, in two of our studies (*e.g.* I and II) already the FPKM normalized read count values (I, formula 1) showed a global 2-fold higher retention of U12-type introns compared to the U2-type introns (I, figure 2a and b; II, figure 2C and S4).

In study II, a preliminary version of IntEREst together with MISO (Katz *et al.*, 2010) was used to analyze SOLiD sequencing data derived from three individual Hep-2 cell pools (II). The Hep-2 samples were part of an experiment to investigate the nuclear processing of the U12-type introns. Consequently, the Hep-2 cells were treated with siRNAs targeting RRP41 (*i.e.* a core component of the exosome), DIS3 (*i.e.* a catalytic subunit of the exosome) and GFP (green fluorescent protein used as a control) (II, figure S3). Additionally, nuclei were separated from cytoplasm to increase the relative abundance of unspliced intron-containing pre-mRNAs compared to the spliced mRNAs. The nuclear fractions showed low levels of cytoplasmic contamination (<10%) and similarly, almost no nuclear contamination in cytoplasmic fractions were detected (II, figure 2B). Both nuclear and cytoplasmic samples were sequenced using SOLiD, followed by mapping with SoLiD Bioscope (II, table 1), after which the data was analyzed with IntEREst using a reference containing 544 annotated U12-type introns (that was built from RefSeq). A global 2-fold higher level of retained U12-type introns were observed compared to U2-type, either when analyzing reads that map to intron-exon junctions or when analyzing reads that fully map to the introns (II, figure 2C and S4). IR results based on intron-exon junction mapping reads together with the validation of the most stabled introns (see below) ruled out the possibility that the observed increase in IR levels may had been influenced by the stabilization of excised intron lariats (II, figure S4). Similarly, a possibility that stable intron-coded RNAs (miRNAs and snoRNAs) may influence the IR values was ruled out as no annotated intron-coded ncRNAs such as snoRNAs or miRNAs were detected within U12-type introns.

Subsequently, the evidence of increased IR of U12-type introns compared to U2-type has been seen in a great variety of samples, including bone marrow mononuclear cells (I), nuclear fractions from Hep-2 cells (II) and mononuclear blood cells (III). The effect was also observed in 6 studied maize control samples consisting of 3 roots and 3 shoots of maize (I, figure S7A and B), indicating that the higher retention of U12-type introns compared to U2-type is not limited to mammals but extends to plants. Thus, while our results are in line with the earlier studies which claim that at least in a few individual genes U12-type introns are spliced less efficiently compared to the U2-type introns (Patel *et al.*, 2002; Pessa *et al.*, 2006), we can now generalize it to all U12-type introns in different organisms.

Contrary to RNAs with retained U2-type introns (1.2.1), post-transcriptional splicing of the RNAs with retained U12-type introns has not been reported. Furthermore, the nuclear export of RNAs with retained U12-type introns has neither been seen in our studies (II, figure S2C and D), nor been reported by others (Friend *et al.*, 2008). However, there is accumulating evidence that the efficiency of splicing of U12-type introns may be regulated. One possible pathway for

regulation is a negative feedback loop that is described to regulate the levels of U11-48K and U11/U12-65K proteins in the complex that carries out the initial recognition of the U12-type intron (Verbeeren *et al.*, 2010; Turunen, Verma, *et al.*, 2013) and which has recently shown to be regulated during neuronal differentiation (Verbeeren *et al.*, 2017). The other possibility is the p38MAPK signaling pathway, which is known to regulate the stability of U6atac snRNA involved in the catalytic core of U12-dependent spliceosome. In this case, the stabilization of U6atac snRNA has been shown to correlate with an increase in mRNA levels of the genes containing U12-type introns (Younis *et al.*, 2013).

4.3. Nuclear exosome degrades the transcripts containing U12-type introns

The overall retention of U12-type introns suggest that mRNA containing unspliced U12-type introns are retained in the nucleus and subsequently degraded. To identify the pathways involved, we hypothesized that the nuclear exosome (1.2.3) would be involved in the decay of the mRNAs containing unspliced U12-type introns. The hypothesis was based on a preliminary knock-down survey of various RNA degradation pathways on U12-type intron retention (II, figure 1). To test this possibility on a genome-wide scale we knocked down the key RRP41 and DIS3 subunits of the exosome and used RNAseq to investigate the possible stabilization of U12-type intron signals. Consistently, with knock-down levels of 83% for DIS3 and 81% for RRP41 (II, figure 2A), we observed 119 U12-type introns that were differentially retained (either higher or lower) in either of the exosome KD samples (DIS3/RRP41 KD) as compared to control (GFP KD) knock-down. The majority of the discovered U12 introns with high IR (*i.e.* 71 out of 119) were stabilized in the RRP41 KD samples rather than the DIS3 KD samples (II, figure 3A, B and C). The stabilized introns (in RRP41 KD) not only featured higher retention levels than their flanking U2-type introns (II, figure 2C and 3D), but their retention levels were on average 2-fold higher than those in the control samples (II, figure 3C). The genes with stabilized introns included PSMC4, KIFAP and IFT80 (II, table S1) that were previously reported to feature U12-type introns that splice less efficiently than their U2-type introns (Pessa *et al.*, 2006; Singh and Padgett, 2009). The number of significantly differentially retained U12-type introns in DIS3 KD samples were less compared to RRP41 KD and they included more destabilized introns than stabilized. However, there were significant overlaps with the differentially retained U12 type introns extracted from the two exosome KD samples ($p = 1.74E{-}10$ for significantly higher retained U12 type introns and $p = 0.0122$ for significantly less retained, using hypergeometric test). When comparing the log fold-changes of the IR of U12-type introns (from control to the RRP41KD sample) to that of U2-type, it was also realized that their fold-change values were significantly higher than that of their immediate upstream/downstream U2-type introns or that of randomly sampled U2-type introns ($p < 2.2E{-}16$, using Jonckheere trend test; see II, figure 3F). The same significant increase was however not seen in the DIS3 KD (II, figure 3E).

To validate the results, RT-qPCR was performed using two sets of primers: a pair to estimated the IR levels of U12-type introns and an additional pair to estimate the splicing levels of the U2-type introns (II, figure 4A and B). In detail, a reverse strand primer on the U12-type introns and a forward strand primer on the upstream exon flanking the U12-type introns were used for the former; and reverse and forward strand primers *i.e.* both located on the junction of the exons flanking the U2-type introns were used for the latter. These RT-qPCR results confirmed the exacerbation of IR upon RRP41 KD in 7 out of 8 genes that were detected using IntEREst. Note

that for the bioinformatics analysis in this study (II), the exact test function provided by IntEREst (and adapted from edgeR) was used (considering a $p < 0.05$) and due to the short length of the read sequences used here (50+35 bps) the analysis were limited to only take into account reads that map to the introns (and did not incorporate any intron-spanning or exon-exon junction mapping reads).

Previous reports have indicated that the exosome can regulate the processing and transcript termination of snRNAs *e.g.* U4 and U6 snRNAs (van Hoof *et al.*, 2000; Schneider *et al.*, 2012). To rule out the possibility that our exosome knockdowns could affect the IR by influencing the levels of spliceosomal snRNAs, we carried out Northern blot analysis of several U2/U12-dependent snRNAs. We did not observe any changes in their expression levels upon exosome knockdown (II figure S2A and D). Hence (since no variation was observed), the changes in IR levels of the introns cannot be attributed to variations in the levels of snRNAs that are responsible for their splicing.

In summary, the results indicate that the U12 type introns are preferentially stabilized upon the knock-down of exosome subunits. The stronger effect of RRP41 in comparison to DIS3 may be due to co-depletion of other exosome subunits such as RRP6 and other core subunits of exosome in response to depletion of RRP41, *i.e.* a phenomenon that has also been seen in other studies (Kammler *et al.*, 2008). Our results described here, together with the finding that conversion of U12-type introns to U2-type in U12-type intron containing genes has shown to upregulate the protein production (Patel *et al.*, 2002), support the hypothesis that U12-type introns are rate-limiting factors that control the levels of the fully spliced mRNAs of the U12-type intron containing genes. Through this mechanism they may also limit the protein production of the U12-type intron containing genes.

The original hypothesis related to increased IR of U12-type introns postulated, based on *in vitro* splicing experiments, that splicing of U12-type introns is slower than U2-type introns (Patel *et al.*, 2002). To test this possibility we compared the pre-mRNA decay kinetics of selected U12-type introns to those of U2-type introns located in the same genes while the control and exosome KD (Rrp41 KD and DIS3 KD) cells were treated with DRB (Singh and Padgett, 2009). This led to inhibition of CDK7 phosphorylation, which inhibits transcription initiation without influencing either the elongation of the transcripts or the processing of their pre-mRNAs.

Our results (II, figure 4a-c) not only confirmed the higher IR of U12-type introns but also provided evidence that contradicted the original Patel et al. (2002) hypothesis. Specifically, the levels of U12-type introns leveled off but never reached zero; not even after 300 minutes (II, figure 4a-c). This was especially noticeable in the samples in which RRP41 was knocked down (II, figure 4a-c). Given that *slow* splicing entails that over time IR levels should reduce to zero, these results led us to hypothesize that the higher retention of U12-type introns to that of U2-type may be the result of *inefficient* rather than slow splicing (Niemelä and Frilander, 2014). Slow splicing would also suggest that U12-type introns are spliced predominantly post-transcriptionally, but the present evidence has reported only co-transcriptional splicing and it is not yet known whether these introns can undergo post-transcriptional splicing as well. Our observations suggest that a window of opportunity exists in which the U12-type introns can be spliced, but out of that window the introns will either be detained in the nucleus or degraded by the exosome. This limited window of opportunity may be caused by the exclusively co-transcriptional splicing of the U12-type introns, or for other reasons that are currently unknown.

In the future, the *inefficient vs slower* splicing arguments may be settled by accurately estimating the splicing rates of the U12-type introns (*e.g.* using *in vivo* single molecule microscopy) and comparing the measurements to those of the U2-type introns (Martin *et al.*, 2013).

4.4. Consequences of minor spliceosome mutations

In the study of a family with four daughters, three of which were affected by IGHD and pituitary hypoplasia (caused by defective U11/U12-65K protein), in addition to elevated levels of retention of U12-type introns (mentioned in subsection 4.2), several novel cryptic alternative splicing events were detected (III, figure 3G and S1-10). To name a few, expression of defective SPCS2/SPCS3 (*i.e.* associated with peptide hormone metabolism and gene expression) and ARPC5L (*i.e.* associated with actin binding GO category) were detected in the patient cells. Actually these mutations are the possible cause of the observed somatotroph-restricted dysfunction. These abnormal alternative splicing events were themselves the consequences of loss of function of U11/U12 di-snRNP due to its destabilization in the patient cells (caused by defective 65K). The loss of the integrity of the U11/U12 di-snRNP in the patient cells was validated by native gel analysis of nuclear extracts derived from IGHD patient cells and controls (III, figure 3A, lanes 3 and 4), in pull-down experiments using whole cell extracts (III, figure 3B and C) and glycerol gradient analyses (III, figure S11b) of the whole cell extracts. Furthermore, U12 snRNP with reduced mobility was detected in native gel analysis (III, figure 3B and C). Western blot results also showed significant decrease of 65K protein levels in the IGHD patient cells (III, figure 3D) which is consistent with a subsequent work (Norppa *et al.*, 2018) showing that one of the IGHD alleles leads to formation of a premature STOP codon and allele-specific degradation by NMD pathway. Similarly, the other mutation leads to reduced binding to U12 snRNA and likely degradation of the U11/U12-65K protein not bound to U12 snRNA (Norppa *et al.*, 2018). Interestingly, we also observed an upregulation of U4atac snRNA in the patient cells. Similar upregulation can also be observed in the MDS patient dataset (unpublished) suggesting that it may be a compensatory effect for the decreased levels of U11/U12 di-snRNP, but the mechanism of its upregulation is not known at the present time.

In another study (I) we reanalyzed, using IntEREst, the RNAseq data from 12 myelodysplastic syndrome (MDS) patients and 4 healthy individuals (Madan *et al.*, 2015). Of the MDS patients, 8 featured mutations in the gene *ZRSR2* that codes for the ZRSR2/Urp protein responsible for recognition of 3′ ss of U12 type introns (see 1.1.3). The outcomes showed both an elevation of unspliced U12-type intron signal in *ZRSR2* mutated samples (hereafter called ZRSR2mut) in comparison to the U2-type introns and also activation of cryptic U2-type splice sites near the U12-type introns (I, figure 1a). After introns with low read counts were filtered and the DESeq2 (Love *et al.*, 2014) -based function of IntEREst (hereafter called IntEREst-DESeq2) was applied, using an adjusted p-value cutoff ($p_{adj} < 0.01$), 1521 introns were detected to be significantly retained in ZRSR2mut compared to the controls. These controls include the 4 MDS patients that lack the *ZRSR2* mutation (hereafter called ZRSR2wt) and the 4 healthy individuals (hereafter called HEALTHY). The significantly more retained introns in ZRSR2mut included 269 U12-type introns *i.e.* 52.7% of the studied U12-type introns. In contrast, 1252 U2-type introns accounting for only ~0.54% of the studied U2-type introns showed increased IR (I, figure 1a and b). Moreover, while no U12-type introns were significantly less retained in the ZRSR2mut, 89 U2-type introns (~0.03%) showed a significant decrease (I, figure 1a and b).

Increased IR of U12-type introns compared to the U2-type was observed in all samples (I, figure 2a and b) and is consistent with slower/less efficient splicing of the U12-type introns (see 1.2.1 and section 4.4). However, with ZRSR2mut samples this effect was more prominent (I, figure 2a and b). The log2 fold-change of FPKM values were also higher for the U12-type introns (~1.5) than that of U2-type (~ 0.0). This is an indication of IR exacerbation of U12-type introns in ZRSR2mut samples (I, figure 2c and d). Jonckheere Trend test with 10000 permutations also confirmed that the log fold-changes of U12-type introns are significantly higher than that of U2-type introns ($p = 0.0001$). In-line with these results, the deltaPSI ($\Delta\Psi$) measurements were almost twice higher for the U12-type introns ($\Delta\Psi = 1\%$) compared to the U2-type ($\Delta\Psi = 0.6\%$). The $\Delta\Psi$/deltaPsi measurement represents the changes in Ψ/PSI (Percentage Spliced In) when comparing two conditions.

A similar effect was observed in a maize data set investigating the effect of mutations in *RGH3* gene (*i.e.* an ortholog of human *ZRSR2*). This dataset consisted of 12 samples and showed both a ≥2-fold higher retention of U12-type introns compared to U2-type and the exacerbation of IR of U12-type introns as a consequence of *RGH3* mutations similarly as seen with human *ZRSR2* mutations (I, figure S7 A-D). The retention levels of ~46% of the studied U12-type introns were significantly increased in RGH3mut samples (comparing to RGH3wt). In contrast, only ~0.46% of the U2-type introns showed increase in their IR (I, figure S7 A-D).

In addition to increased IR levels of U12-type introns, additional splicing defects featuring the U2-type introns located next to the U12-type introns have been observed in several studies. In our study (III) RNAseq analyses and subsequent RT-PCR validations (III, figure 3G and S1-10) indicated exon skipping and activation of cryptic alternative U2-type introns in U12-type intron containing genes. Similar effects have also been reported with MDS dataset (Madan *et al.*, 2015), and in a knockout study investigating the functions of *ZRSR1* (a paralog of *ZRSR2*, see 1.1.3) in mouse (Horiuchi *et al.*, 2018). In these cases, the possible mechanism affecting the splicing of U2-type introns is disruption of exon-definition interactions between the neighboring U12-dependent and U2-dependent spliceosome. The subsequent outcome of such events is either detention of the incorrectly processed mRNAs in the nucleus followed by their degradation, or in case of cryptic splice site activation or exon skipping events formation of altered proteins or mRNA decay via the NMD pathway (Verma *et al.*, 2018).

One of the unexplained observations is the diverse phenotypic consequences seen with the hereditary disease-causing mutations in the human U12-dependent spliceosome component. They range from very severe i.e. significant developmental defects in multiple organs leading to early death as seen with MOPD1/TALS, to relatively benign diseases i.e. pituitary hypoplasia and associated dwarfism seen with IGHD that can be successfully treated with growth hormone injections (Martos-Moreno *et al.*, 2018). Surprisingly, RNAseq data analysis of IGHD patients show several splicing defects including activation of cryptic alternative U2-type introns in U12-type intron containing genes and nuclear detention of transcripts with retained U12-type introns. One possible explanation for the observed mild effects in IGHD in comparison to those observed in MOPD1/TALS is that the cryptic splicing events in IGHD and other diseases may be eliminated by the NMD pathway. However, in contrast, accumulation of the unspliced U12-type introns that are observed in MOPD1/TALS patient cells may disrupt the actual functioning of the U12-dependent spliceosome or other nuclear functions in the cell (Niemelä and Frilander, 2014; Verma *et al.*, 2018).

4.5. Comparison of the various analysis modules supported by IntEREst

Detecting the genes/isoforms that are differentially expressed across various samples have been the main focus of many studies. To carry out these analysis a number of computational methods have been developed (1.3.3), namely the two widely popular R packages edgeR and DESeq (Robinson *et al.*, 2010; Anders and Huber, 2010). At the same time, due to the various limitations imposed by characteristics shared by many introns *e.g.* their large length and presence of repetitive DNA sequences (see 1.3.4), accurate detection and differential analysis of isoforms containing retained introns have been challenging. In this respect, IntEREst attempts to carry out intron retention (IR) detection and cross-sample differential test by carrying out all its analysis at the intron/exon level. Due to the success and the relatively high performance of edgeR and DESeq (1.3.3), IntEREst adapts these methods (together with DEXSeq) to perform statistical differential analysis. However, since these methods adapt different normalization and differential analysis algorithms, despite their relatively high performance in detecting the differentially expressed isoforms, they typically produce somewhat different results (1.3.3). This motivated us to allow incorporation of any of the three DESeq2 (Love *et al.*, 2014), edgeR (Robinson *et al.*, 2010) and DEXSeq (Anders *et al.*, 2012; Reyes *et al.*, 2013) methods (referred to as IntEREst-DESeq2, IntEREst-edgeR and IntEREst-DEXSeq) to run differential IR analysis. Here, we compare the results that were obtained by running IntEREst-DESeq2, IntEREst-edgeR and IntEREst-DEXSeq on the MDS data (Madan *et al.*, 2015). The results are not restricted to U12-type introns but takes into account all differentially retained introns (including U12- or U2-type). Note that for IntEREst-edgeR, the GLM function of edgeR was used.

The results indicated that although there were a number of significantly higher/lower retained introns that were discovered by each differential intron retention analysis function specifically, the majority of the discovered significantly high and low retained introns (in ZRSR2mut compared to controls) were shared between IntEREst-DESeq2 and IntEREst-edgeR (I, figure 3a and b). However, IntEREst-edgeR returned more significantly less retained introns than the IntEREst-DESeq2, whereas the latter found more significantly higher retained introns (I, figure 3a and b). Interestingly, the IRs discovered by IntEREst-DESeq2 and missed by IntEREst-edgeR mainly displayed weaker foldchange values compared to those discovered by both (I, figure S3).

Despite the high overlap (both significantly more and less retained introns) between the IntEREst-DESeq2 and IntEREst-DEXSeq results, the latter discovered many cases that were not detected by IntEREst-DESeq2. Many of the IRs specific to IntEREst-DEXSeq were, however, proved to be false positives upon manual inspection. The higher and lower retained introns discovered by the IntEREst-DEXSeq actually featured a more symmetric distribution compared to those discovered by IntEREst-DESeq2 (I, figure S4). They were also featured from the same genes twice more frequently than the higher and less retained introns discovered by IntEREst-DESeq2. Overall, the fact that IntEREst supports various settings of read summarization and various significant differential test functions allows the users to choose and apply their preferred method based on their data and biological questions.

4.6. Benchmarking IR analysis tools including IntEREst

As mentioned previously various published methods exist to perform intron retention analysis (see subsection 1.3.4). We evaluated several of the existing methods that can be used to extract

IR events, namely IRFinder (Middleton *et al.*, 2017), MISO (Katz *et al.*, 2010), rMATS (Shen *et al.*, 2014) and SUPPA (Alamancos *et al.*, 2015; Trincado *et al.*, 2018) and we compared their features to that of IntEREst. Some of these methods analyze variations of all alternative splicing events across the samples (*e.g.* MISO, SUPPA and rMATS), whereas others analyze intron retention specifically (e.g. IRFinder and KMA). Methods also differ based on whether their analysis can run in parallel on multiple computing cores, their supported sample size comparison, their support of complicated experiment designs to run sample comparison accordingly, their support of Ψ or PSI (*i.e.* Percentage Spliced In) values, their provided statistical tests to extract significantly retained introns, their support of IR comparison of subgroups of introns, or whether they annotate novel IRs *i.e.* absent from the used transcripts reference (table 4).

All methods but SUPPA and KMA can be run on multiple computing cores in parallel. Apart from MISO that supports only a one-to-one sample comparison, all other methods allow comparisons of multiple samples to multiple samples, however, rMATS and SUPPA do not allow one-to-one sample comparison (*i.e.* running without biological replication). Since IntEREst provides various advanced functions that are based on commonly used R packages for differential gene expression analysis (e.g. DESeq2 and edgeR), unlike the other methods it provides the possibility to define complicated experiment designs to perform IR differential analysis. This allows the running differential intron retention analysis while taking into account sample information such as age, sex *etc*. All methods support Ψ values calculation. SUPPA (version 1) was the only method that lacked statistical tests to discover differential IRs, however in its newest update this functionality has been added (Trincado *et al.*, 2018). IntEREst is the only method that provides tools to compare between retention levels of various subgroups of introns (*e.g.* U12-type introns vs U2-type, or other user-defined groups). Moreover, IntEREst and rMATS are the two methods that can discover novel IR events not present in their used transcript references. Other methods that lack this capability (*e.g.* IRFinder and MISO) are dependent on transcriptome assembly tools *e.g.* Cufflinks (Trapnell *et al.*, 2012) to detect novel significantly differential IRs.

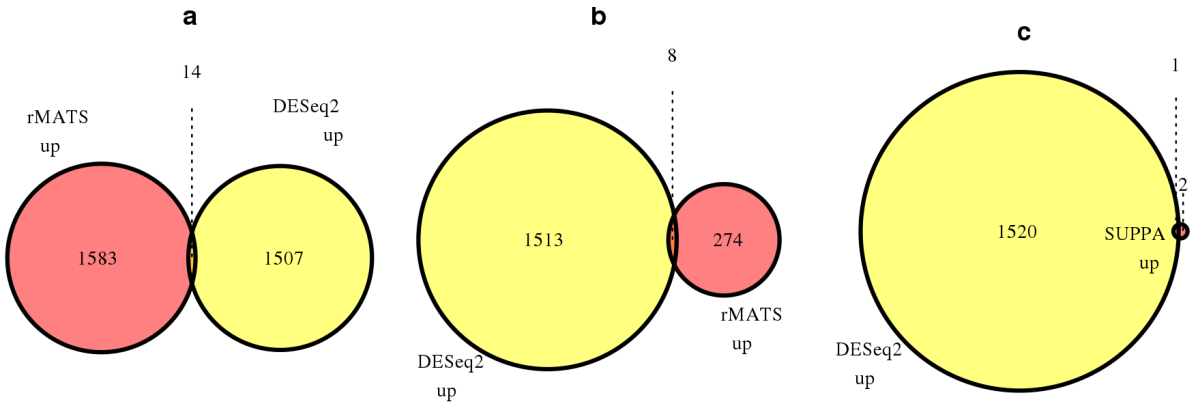

Figure 5. Venn diagrams comparing the significantly retained introns (ZRSR2mut *vs* controls) that were discovered by IntEREst to those discovered by (a) rMATS with ΔΨ > 0.01% (b) rMATS with ΔΨ > 20% (c) SUPPA (unpublished figure).

Method	Parallel run	Sample size compare	Defining experiment design	Supports Ψ values	Significant differential IR	Intron subgroups comparison	Annotated / novel IR
IntEREst	√	N-N (N>0)	√	√	√	√	Annotated + Novel IR events
IRFinder	√	N-N (N>0)	✗	√	√	✗	Annotated IR events
KMA	✗	N-N (N>0)	✗	√	√	✗	Annotated IR events
MISO	√	1-1	✗	√	√	✗	Annotated IR events
rMATS	√	N-N (N>1)	✗	√	√	✗	Annotated + Novel IR events
SUPPA	✗	N-N (N>1)	✗	√	√	✗	Annotated IR events

Table 4. Various tools applicable for discovering significantly retained introns and performing differential IR analysis and their main features (unpublished).

Running IRFinder, 250 significantly increased IRs in ZRSR2mut samples (compared to controls) were discovered most of which (*i.e.* 235) were also discovered by IntEREst-DESeq2 (I, figure 3e). IntEREst-DESeq2 was also more sensitive to lower retained introns compared to IRFinder, discovering more introns with significant increase/decrease of IR (when comparing ZRSR2mut to the controls) despite their low mapped read counts (I, figure S5). When compared to the significant IRs discovered by the original MDS study (Madan *et al.*, 2015) except for a few (*i.e.* 37), all of the significantly increased IRs were also discovered by IntEREst. Most of the IRs missed by IntEREst featured very low IR log fold-changes in our analysis (I, figure S6). Due to lack of the possibility to use biological replicates, MISO was not applicable for analyzing the MDS data. Running rMATS with default parameters (*i.e. cstat* or cutoff splicing difference of 0.0001 or 0.01%) resulted in differential IRs drastically different from those extracted by IntEREst (figure 5a and b). Upon manual inspection, we realized that most of these introns were retained at high levels in both ZRSR2mut and control samples and the increase in the IR levels were not clear when comparing their retention in the ZRSR2mut, to that in the controls. In line with these observations, the log fold-change measurements of retention levels of the introns discovered by rMATS were relatively low regardless of whether the splicing difference cutoff (or ΔΨ) parameter was set to 0.01% (figure 6a) or if it was set as high as 20% (figure 6b). Increasing the rMATS ΔΨ cutoff parameter to 20% yielded results that featured less log fold-change IRs compared to when it was set to 0.01% (figure 6a and b). Running rMATS with the default parameter settings returned only 4 significantly less retained introns, all of which were discarded when the ΔΨ cutoff parameter was raised to 20%. Running SUPPA together with the *diffsplice* tool (Alamancos *et al.*, 2015; Trincado *et al.*, 2018) yielded only 3 significantly increased IRs, 1 of which was also discovered by IntEREst (figure 5c); the other 2 IRs were missed by IntEREst due to the introns/genes being absent from the reference used by IntEREst. We were not able to successfully install/run KMA on the MDs data to obtain significant IRs.

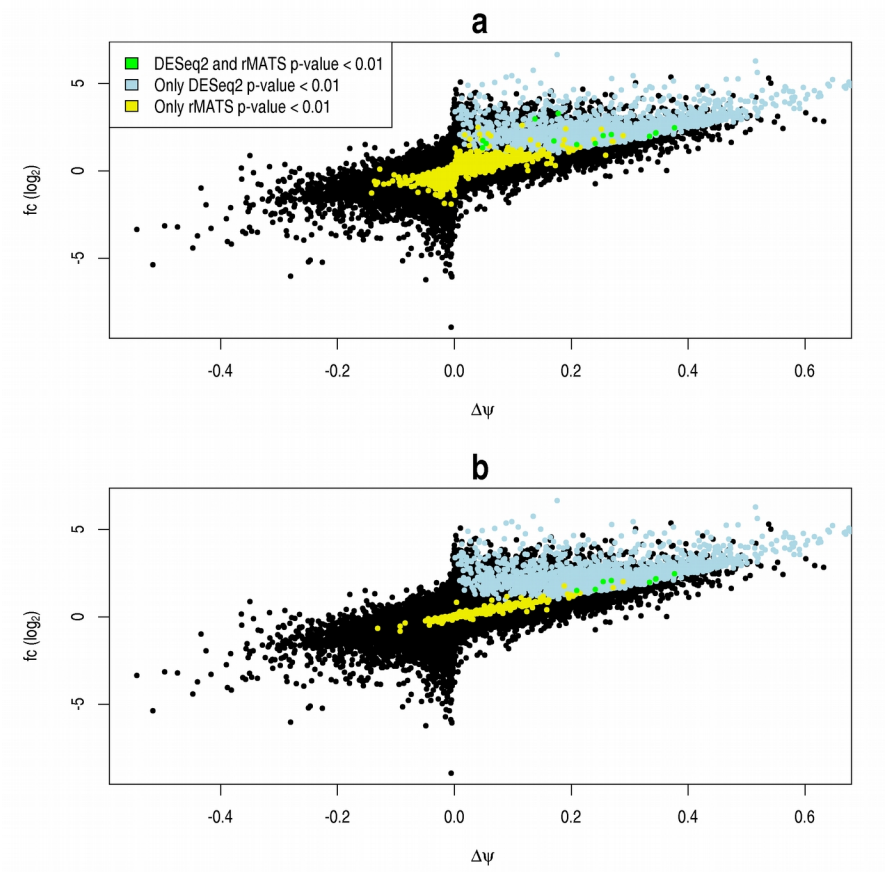

Figure 6. Scatter plot showing the distribution of log2 fold-change relative to the ΔΨ values (when comparing ZRSR2mut to controls). The significant IRs discovered by IntEREst-DESeq2, rMATS or both have been labeled with different colours as depicted in the legend. The results of rMATS for two parameter settings are shown: (a) rMATS run with the default parameter settings, i.e. ΔΨ > 0.01% and (b) rMATS run with ΔΨ > 20% (unpublished figure).

As mentioned above, methods such as SUPPA and rMATS detect significant IRs by measuring the ΔΨ levels. They also implement statistical methods to infer the probability of reaching the observed ΔΨ values by chance (*i.e.* the significance of the observed ΔΨ). In contrast, IntEREst is mainly based on raw read-counts of the introns/exons. Furthermore, IntEREst analyzes the changes of the number of intron mapping reads relative to the changes of the number of intron spanning reads. Admittedly, solely focusing on the fold change of the read levels does not completely reflect the changes in the fraction of various expressed isoforms of a gene. However, we have noticed that a number of isoforms, especially those with retained U12-type introns may be low expressed and feature Ψ measurements as low as 10% or less (I, Additional file 2). Detecting the possible ~2 fold increase of these isoforms in the test samples compared to that in controls are interesting even though this yields to a ΔΨ increase of 10%; which is usually neglected in other studies. This is particularly relevant for spliceosomal disease cases, where the defects are almost always hypomorphic and can lead to mild effects with individual introns, but nevertheless cause diseases upon influencing a large number of splicing events. In other words, by analyzing the genome-wide fold change of the IRs we can detect radical changes of IR levels

despite their low expressions. As a consequence however, a number of significantly differential IRs may also be incorrectly detected based on their extremely low average and high variance of intron read counts across the studied samples. For this reason IntEREst also supports Ψ (and ΔΨ) values measurement tool, the results of which can be combined with traditional fold change measurements. We have found out that sorting the significant IRs by their cross-sample average Ψ values and filtering those with very low mean Ψ values can be helpful to focus on the most reliable detected IRs.

4.7. The effect of expanding biological replicates and the depth of sequencing libraries

Rerunning the IR detection analysis multiple times, while each time allowing either various number of samples or force the overall read counts (*i.e.* library depths) to different limits *i.e.* 5-50 million, confirmed that including more biological replicates and reads both yield more detected significant IRs (II, figure 4a, b and c). However with the library size increase, the rise in the number of discovered significant IRs was not linear and the slope decreased after reaching ~35 million mapped reads (II, figure 4c). This trend was not seen for increase of the number of biological replicates, indicating that there remains room for increasing the number of biological replicates (beyond 16) and thus detect more significant IR results (II, figure 4a and b). This is probably due to the fact that the number of biological replicates in our MDS study was limited (*i.e.* 8 test samples vs 8 controls). However, a study that carried out differential gene expression analysis on a data with full set of 42 biological replicates claimed that at least 6 biological replicates are needed to be able to detect DEGs reliably. Furthermore they recommend that the number of biological replicates should be expanded to 12 if the goal is to detect DEGs with low fold-changes as well as the most extreme cases (Schurch *et al.*, 2016). Given that IntEREst uses the similar tools that were developed for differential gene expression analysis (e.g. DESeq2 and edgeR), we conclude that the same guidelines are applicable for accurate differential IR analysis using IntEREst.

5. Concluding remarks

Analyzing the retention of introns reveals information regarding the efficiency of splicing in various conditions. However, detecting and analyzing these retained introns using conventional isoform analysis software can be difficult due to several characteristics that are shared by most spliceosomal introns, *e.g.* they are large in length and they carry repetitive DNA sequences. I present in I a software tool called intron-exon retention estimator (IntEREst) that can be used to accurately detect, quantify and perform differential analysis of the intron retention in several samples, using RNAseq data from these samples. IntEREst also supports comparison of retention levels of various subclasses of introns (*e.g.* U12-type introns vs U2-type). To test IntEREst in I, I also reanalyzed data constructed from patients with myelodysplastic syndrome (MDS) and healthy individuals, confirming that mutations in the *ZRSR2* gene (responsible for detection of 3′ss of the U12-type introns) exacerbate the nuclear detention of the U12-type intron containing RNAs.

Normally U12-type introns are spliced less efficiently and retained ~2 fold higher than the U2-type introns, and therefore they are thought to be rate-limiting to the expression of transcripts that contain these introns. This was not previously studied at a genome-wide scale to our knowledge. The fate of the mRNAs that contain unspliced U12-type introns was not defined earlier. Using IntEREst, in II we show that there is a genome-wide higher (~2-fold) retention of U12-type introns compared to U2-type introns. We also show that the RNAs that contain unspliced U12-type introns are detected and degraded by the nuclear exosome. Furthermore, our results provide evidence that instead of being spliced slowly, the splicing of U12-type introns is in fact less efficient than that of U2-type introns.

Using other isoform analysis tools (e.g. Cufflinks), in III additional splicing abnormalities including activation of cryptic U2-dependent splice-sites and elevated retention of U2-type introns were also seen in patients diagnosed with isolated growth hormone deficiency (IGHD) caused by mutations in *RNPC3*, *i.e.* the gene coding for U11/U12-65K protein (III). Such defects may have an impact on the severity of the diseases caused by minor spliceosome dysfunction.

Overall, this thesis demonstrates that applying isoform analysis software together with a dedicated intron retention analysis software tool such as IntEREst provides a comprehensive understanding of the biological pathways involved in diseases associated with defective spliceosome function. However, similarly as with any software, the development does not stop at the first version. IntEREst currently (v1.4.1) lacks a comprehensive tool to filter the unreliable results (e.g. those based on low mapped-read counts). Moreover, the scaled (FPKM) IR levels do not sum to constant values within the samples. In future updates, both of these shortcomings would be resolved by adding a downstream IR filtering function together with an additional function that scales the normalized values, thereby they will sum to a constant value (*e.g.* 1 million) within the samples providing higher accuracy in cross-sample IR comparisons.

6. Acknowledgements

This study was carried out at the Institute of Biotechnology in the University of Helsinki. I would like to thank the institute for providing outstanding infrastructure and excellent research facilities. I would also like to thank the directors of the Institute during the time in which I carried out this research, Tomi Mäkelä, Howard Jacobs and Olli Silvennoinen. Additionally, I would like to thank the Doctoral programme in Integrative Life Science (ILS) and the Viikki Doctoral Programme in Molecular Biosciences (VGSB) for partially funding my doctoral studies.

I would like to express my immense gratitude to my supervisor, Mikko Frilander, who has supported and mentored me throughout my studies. He has always been and still is a great source of inspiration to me and I learnt a great deal from him. I specifically appreciated his critical thinking and reasoning, patience, understanding, optimism, and openness towards criticism. The Frilander lab (RNA splicing laboratory) conducts high quality research and is an inspiring and joyful place to do research. These qualities have been brought about by great current and previous lab members: Maureen, Antto, Mariia, Lilli, Bhupendra, Jens, Jouni, Elina, Janne and Heli. Despite our background differences, mine being computer engineering and theirs mainly molecular and biochemically oriented, interaction with the lab members has always felt easy. I felt they always took time to understand the computational analysis results which I presented and patiently explained the biology/biochemistry experiments and processes to me. I specially thank Elina with whom I collaborated on an exciting project which led to a publication of one of the papers in this thesis (II). I would also like to especially thank Antto for translating the abstract of my thesis to Finnish, Bhupendra for sharing many glasses of excellent Indian chai while discussing about life and science, Maureen whom with I had espresso often in the lab and discussed about life, and Jens who has been and still is a great friend and who introduced me to the great joys of playing board games.

I am deeply grateful to my co-supervisor, Dario Grecco, for his assistance and encouragements throughout my studies. Many times in my research, his helpful guides and hints made complicated procedures, such as biological software development and publishing, seem easy.

I would like to thank Professor Rickard Sandberg for accepting to be the opponent and Professor Liisa Holm for accepting to be the custos at my PhD defense. I look forward to interesting discussions at the defense session. I would additionally like to thank Professor Garry Wong and Professor Tero Aittokallio for their extremely helpful and prompt reviews of my thesis. Their comments have significantly improved the legibility and quality of my thesis. Assistance and comments that I received from Professor Juha Partanen and the members of the advisory committee of my PhD thesis, Docent Petri Auvinen and Professor Jukka Corander, were extremely helpful in keeping me on track towards completing my studies; I am extremely thankful for their help and support.

I would like to specially thank my friends Reza, Neda, Mohsen, Leile, Abdollah, Milad, Mehdi, Shirin and Sina for sharing great moments with me and my family, outside the research and science environment. Special thanks to the members of the book club which I attend every now and then, especially Fahimeh, Hassan and Farid for organizing interesting events and recommending fantastic books for reading and discussion. Many thanks to my friends at the former International group of the Kameraseura photography club, especially Javier, Rasa, Eugene, Don, Olivier, Tua and Rich for our fun activities during the time that I was a club member.

I owe a lot to my parents, Mohammad Ali and Fatemeh for bringing me up and their significant role in making me who I am today. Their love and support has been non-stop, extremely heartwarming and encouraging, especially during the time that I have been living and studying in Finland. I thank my sisters Zeynab and Zoha for their constant care and love. I also thank my in laws, especially Ali and Maryam for their support and for accepting and welcoming me in their family.

Finally, this work as it is, would have not have existed without the great patience, understanding and love of my wife, Homa Ehsan, to whom I owe a great deal. Dear Homa, thank you for your love and support during these years and also for your understanding on multiple occasions that I was preoccupied with studying or preparing this thesis. I also thank my recently turned 1 year old son, Iliya, who has brought great joy and excitement to our lives since he joined our family.

Helsinki, October 2018

Ali

References

Abril,J.F. *et al.* (2005) Comparison of splice sites in mammals and chicken. *Genome Res*, **15**, 111–119.

Alamancos,G.P. *et al.* (2015) Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA*, **21**, 1521–1531.

Alioto,T.S. (2007) U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res*, **35**, D110–115.

Ambros,V. *et al.* (2003) MicroRNAs and other tiny endogenous RNAs in C. elegans. *Curr. Biol.*, **13**, 807–818.

Amit,M. *et al.* (2012) Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep*, **1**, 543–556.

Anders,S. *et al.* (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res*, **22**, 2008–2017.

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Angelini,C. *et al.* (2014) Computational approaches for isoform detection and estimation: good and bad news. *BMC Bioinformatics*, **15**, 135.

Ast,G. (2004) How did alternative splicing evolve? *Nat Rev Genet*, **5**, 773–782.

Astuti,D. *et al.* (2012) Germline mutations in DIS3L2 cause the Perlman syndrome of overgrowth and Wilms tumor susceptibility. *Nat Genet*, **44**, 277–284.

Baek,D. and Green,P. (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc Natl Acad Sci USA*, **102**, 12813–12818.

Bansal,V. (2017) A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. *BMC Bioinformatics*, **18**, 43.

Barash,Y. *et al.* (2010) Deciphering the splicing code. *Nature*, **465**, 53–59.

Bartschat,S. and Samuelsson,T. (2010) U12 type introns were lost at multiple occasions during evolution. *BMC Genomics*, **11**, 106.

Bashaw,G.J. and Baker,B.S. (1995) The msl-2 dosage compensation gene of Drosophila encodes a putative DNA-binding protein whose expression is sex specifically regulated by Sex-lethal. *Development*, **121**, 3245–3258.

Belfort,M. *et al.* (1995) Prokaryotic introns and inteins: a panoply of form and function. *J Bacteriol*, **177**, 3897–3903.

Benecke,H. *et al.* (2005) The U11/U12 snRNP 65K protein acts as a molecular bridge, binding the U12 snRNA and U11-59K protein. *EMBO J*, **24**, 3057–3069.

Berget,S.M. (1995) Exon recognition in vertebrate splicing. *J Biol Chem*, **270**, 2411–2414.

Berget,S.M. *et al.* (1977) Spliced segments at the 5’ terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA*, **74**, 3171–3175.

Beyer,A.L. *et al.* (1981) Correlation of hnRNP structure and nascent transcript cleavage. *Cell*, **26**, 155–165.

Beyer,A.L. and Osheim,Y.N. (1988) Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev*, **2**, 754–765.

Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.

Boothby,T.C. *et al.* (2013) Removal of retained introns regulates translation in the rapidly developing gametophyte of Marsilea vestita. *Dev Cell*, **24**, 517–529.

Borah,S. *et al.* (2009) Drosophila hnRNP A1 homologs Hrp36/Hrp38 enhance U2-type versus U12-type splicing to regulate alternative splicing of the prospero twintron. *Proc Natl Acad Sci USA*, **106**, 2577–2582.

Boutz,P.L. *et al.* (2015) Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev*, **29**, 63–80.

Brannan,K. *et al.* (2012) mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol Cell*, **46**, 311–324.

Brock,J.E. *et al.* (2008) Mutational analysis of the U12-dependent branch site consensus sequence. *RNA*, **14**, 2430–2439.

Buckley,P.T. *et al.* (2011) Cytoplasmic intron sequence-retaining transcripts can be dendritically targeted via ID element retrotransposons. *Neuron*, **69**, 877–884.

Burge,C.B. *et al.* (1998) Evolutionary fates and origins of U12-type introns. *Mol Cell*, **2**, 773–785.

Busch,A. and Hertel,K.J. (2012) Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip. Rev. RNA*, **3**, 1–12.

Callis,J. *et al.* (1987) Introns increase gene expression in cultured maize cells. *Genes Dev*, **1**, 1183–1200.

Carmel,I. *et al.* (2004) Comparative analysis detects dependencies among the 5’ splice-site positions. *RNA*, **10**, 828–840.

Castle,J.C. *et al.* (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat Genet*, **40**, 1416–1425.

Cavalier-Smith,T. (1991) Intron phylogeny: a new hypothesis. *Trends Genet*, **7**, 145–148.

Chang,W.C. *et al.* (2007) Alternative splicing and bioinformatic analysis of human U12-type introns. *Nucleic Acids Res*, **35**, 1833–1841.

Chlebowski,A. *et al.* (2013) RNA decay machines: the exosome. *Biochim Biophys Acta*, **1829**, 552–560.

Cho,S. *et al.* (2011) Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proc. Natl. Acad. Sci.*, **108**, 8233–8238.

Cho,V. *et al.* (2014) The RNA-binding protein hnRNPLL induces a T cell alternative splicing program delineated by differential intron retention in polyadenylated RNA. *Genome Biol.*, **15**, R26.

Chow,L.T. *et al.* (1977) An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA. *Cell*, **12**, 1–8.

Coghlan,A. and Wolfe,K.H. (2004) Origins of recently gained introns in Caenorhabditis. *Proc Natl Acad Sci USA*, **101**, 11362–11367.

Comeron,J.M. and Kreitman,M. (2000) The correlation between intron length and recombination in drosophila. Dynamic equilibrium between mutational and selective forces. *Genetics*, **156**, 1175–1190.

Comstock,C.E. *et al.* (2009) Cyclin D1 splice variants: polymorphism, risk, and isoform-specific regulation in prostate cancer. *Clin Cancer Res*, **15**, 5338–5349.

Costa,V. *et al.* (2013) RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur J Hum Genet*, **21**, 134–142.

Damianov,A. *et al.* (2004) Recycling of the U12-type spliceosome requires p110, a component of the U6atac snRNP. *Mol Cell Biol*, **24**, 1700–1708.

Darnell,J.E. (1978) Implications of RNA-RNA splicing in evolution of eukaryotic cells. *Science*, **202**, 1257–1260.
Davidson,L. *et al.* (2012) Co-transcriptional degradation of aberrant pre-mRNA by Xrn2. *EMBO J*, **31**, 2566–2578.
Demir,E. and Dickson,B.J. (2005) fruitless splicing specifies male courtship behavior in Drosophila. *Cell*, **121**, 785–794.
Di,Y. *et al.* (2011) The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.*, **10**.
Dieci,G. *et al.* (2009) Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics*, **94**, 83–88.
Dietrich,R.C. *et al.* (2001) Role of the 3' splice site in U12-dependent intron splicing. *Mol Cell Biol*, **21**, 1942–1952.
Dietrich,R.C. *et al.* (1997) Terminal intron dinucleotide sequences do not distinguish between U2-and U12-dependent introns. *Mol. Cell*, **1**, 151–160.
Dillies,M.A. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinforma.*, **14**, 671–683.
Doolittle,W.F. (1978) Genes in pieces: were they ever together? *Nature*, **272**, 581.
Dziembowski,A. *et al.* (2007) A single subunit, Dis3, is essentially responsible for yeast exosome core activity. *Nat Struct Mol Biol*, **14**, 15–22.
Edery,P. *et al.* (2011) Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA. *Science*, **332**, 240–243.
Elsaid,M.F. *et al.* (2017) Mutation in noncoding RNA RNU12 causes early onset cerebellar ataxia. *Ann Neurol*, **81**, 68–78.
Engström,P.G. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, **10**, 1185.
Evans,R.M. *et al.* (1977) The initiation sites for RNA transcription in Ad2 DNA. *Cell*, **12**, 733–739.
Farach,L.S. *et al.* (2018) The expanding phenotype of RNU4ATAC pathogenic variants to Lowry Wood syndrome. *Am J Med Genet A*, **176**, 465–469.
Felkin,L.E. *et al.* (2011) Calcineurin Splicing Variant Calcineurin Aβ1 Improves Cardiac Function After Myocardial Infarction Without Inducing HypertrophyClinical Perspective. *Circulation*, **123**, 2838–2847.
Friend,K. *et al.* (2008) Minor-class splicing occurs in the nucleus of the Xenopus oocyte. *RNA*, **14**, 1459–1462.
Frilander,M.J. and Steitz,J.A. (2001) Dynamic exchanges of RNA interactions leading to catalytic core formation in the U12-dependent spliceosome. *Mol Cell*, **7**, 217–226.
Frilander,M.J. and Steitz,J.A. (1999) Initial recognition of U12-dependent introns requires both U11/5' splice-site and U12/branchpoint interactions. *Genes Dev*, **13**, 851–863.
Fu,X.-D. (2004) Towards a splicing code. *Cell*, **119**, 736–738.
Galante,P.A. *et al.* (2004) Detection and evaluation of intron retention events in the human transcriptome. *RNA*, **10**, 757–765.
Gilbert,W. (1987) The exon theory of genes. *Cold Spring Harb Symp Quant Biol*, **52**, 901–905.
Gilbert,W. (1978) Why genes in pieces? *Nature*, **271**, 501.
Gozani,O. *et al.* (1996) Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. *Genes Dev*, **10**, 233–243.
Graveley,B.R. *et al.* (1998) A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. *EMBO J*, **17**, 6747–6756.

Grosso,A.R. *et al.* (2008) Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Res*, **36**, 4823–4832.
Gruss,P. *et al.* (1979) Splicing as a requirement for biogenesis of functional 16S mRNA of simian virus 40. *Proc Natl Acad Sci USA*, **76**, 4317–4321.
Hall,S.L. and Padgett,R.A. (1994) Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J. Mol. Biol.*, **239**, 357–65.
Hallegger,M. *et al.* (2010) Alternative splicing: global insights. *FEBS J.*, **277**, 856–866.
He,H. *et al.* (2011) Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. *Science*, **332**, 238–240.
Horiuchi,K. *et al.* (2018) Impaired Spermatogenesis, Muscle, and Erythrocyte Function in U12 Intron Splicing-Defective Zrsr1 Mutant Mice. *Cell Rep*, **23**, 143–155.
Houseley,J. and Tollervey,D. (2009) The many pathways of RNA degradation. *Cell*, **136**, 763–776.
Huang,D.W. *et al.* (2008) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
Ibrahim,E.C. *et al.* (2005) Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proc Natl Acad Sci USA*, **102**, 5002–5007.
Inoue,D. *et al.* (2016) Spliceosomal gene mutations in myelodysplasia: molecular links to clonal abnormalities of hematopoiesis. *Genes Dev*, **30**, 989–1001.
International Human Genome Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860.
International Human Genome Consortium,N. authors (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
Irimia,M., Penny,D., *et al.* (2007) Coevolution of genomic intron number and splice sites. *Trends Genet*, **23**, 321–325.
Irimia,M., Rukov,J.L., *et al.* (2007) Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC Evol Biol*, **7**, 188.
Jonckheere,A.R. (1954) A distribution-free k-sample test against ordered alternatives. *Biometrika*, **41**, 133–145.
Kalyna,M. *et al.* (2012) Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res*, **40**, 2454–2469.
Kammler,S. *et al.* (2008) The RNA exosome component hRrp6 is a target for 5-fluorouracil in human cells. *Mol Cancer Res*, **6**, 990–995.
Kanopka,A. *et al.* (1996) Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature*, **381**, 535–538.
Kapustin,Y. *et al.* (2011) Cryptic splice sites and split genes. *Nucleic Acids Res*, **39**, 5837–5844.
Katz,Y. *et al.* (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*, **7**, 1009–1015.
Keren,H. *et al.* (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*, **11**, 345–355.
Kim,W.Y. *et al.* (2010) The Arabidopsis U12-type spliceosomal protein U11/U12-31K is involved in U12 intron splicing via RNA chaperone activity and affects plant development. *Plant Cell*, **22**, 3951–3962.
Kolossova,I. and Padgett,R.A. (1997) U11 snRNA interacts in vivo with the 5'splice site of U12-dependent (AU-AC) pre-mRNA introns. *Rna*, **3**, 227.
Konarska,M.M. and Sharp,P.A. (1987) Interactions between small nuclear ribonucleoprotein particles in formation of spliceosomes. *Cell*, **49**, 763–774.

Koonin,E.V. *et al.* (2013) Whence genes in pieces: reconstruction of the exon–intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. *Wiley Interdiscip. Rev. RNA*, **4**, 93–105.

Kopylova,E. *et al.* (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28**, 3211–3217.

Kruskal,W.H. and Wallis,W.A. (1952) Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.*, **47**, 583–621.

Kunze,B. *et al.* (2000) Transcription and proper splicing of a mammalian gene in yeast. *Gene*, **246**, 93–102.

Lareau,L.F. *et al.* (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, **446**, 926–929.

Le Hir,H. *et al.* (2001) The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J*, **20**, 4987–4997.

Le Hir,H. *et al.* (2000) The spliceosome deposits multiple proteins 20–24 nucleotides upstream of mRNA exon–exon junctions. *EMBO J.*, **19**, 6860–6869.

Lebreton,A. *et al.* (2008) Endonucleolytic RNA cleavage by a eukaryotic exosome. *Nature*, **456**, 993–996.

Leeds,P. *et al.* (1991) The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon. *Genes Dev*, **5**, 2303–2314.

Levine,A. and Durbin,R. (2001) A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res*, **29**, 4006–4013.

Lewis,B.P. *et al.* (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci USA*, **100**, 189–192.

Li,S. *et al.* (2014) Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.*, **32**, 915.

Lin,C.-F. *et al.* (2010) Evolutionary dynamics of U12-type spliceosomal introns. *BMC Evol. Biol.*, **10**, 47.

Liu,Q. *et al.* (2006) Reconstitution, activities, and structure of the eukaryotic RNA exosome. *Cell*, **127**, 1223–1237.

Liu,Z. *et al.* (2001) Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science*, **294**, 1098–1102.

Lopez,P.J. and Seraphin,B. (1999) Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition. *RNA*, **5**, 1135–1137.

Lorkovic,Z.J. *et al.* (2005) Evolutionary conservation of minor U12-type spliceosome between plants and humans. *RNA*, **11**, 1095–1107.

Lorkovic,Z.J. *et al.* (2004) Interactions of Arabidopsis RS domain containing cyclophilins with SR proteins and U1 and U11 small nuclear ribonucleoprotein-specific proteins suggest their involvement in pre-mRNA Splicing. *J Biol Chem*, **279**, 33890–33898.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.

Luo,M.J. and Reed,R. (1999) Splicing is required for rapid and efficient mRNA export in metazoans. *Proc Natl Acad Sci USA*, **96**, 14937–14942.

Luo,W. *et al.* (2006) The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: implications for a unified allosteric-torpedo model. *Genes Dev*, **20**, 954–965.

Lykke-Andersen,J. *et al.* (2000) Human Upf proteins target an mRNA for nonsense-mediated decay when bound downstream of a termination codon. *Cell*, **103**, 1121–1131.

Lynch,M. (2007) Chapter 9: Genes in peaces. In, *The origins of genome architecture*. Sinauer Associates Sunderland (MA).
Lynch,M. (2002) Intron evolution as a population-genetic process. *Proc Natl Acad Sci USA*, **99**, 6118–6123.
Lynch,M. and Richardson,A.O. (2002) The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.*, **12**, 701–710.
Madan,V. *et al.* (2015) Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. *Nat Commun*, **6**, 6042.
Makarova,O.V. *et al.* (2002) Protein 61K, encoded by a gene (PRPF31) linked to autosomal dominant retinitis pigmentosa, is required for U4/U6*U5 tri-snRNP formation and pre-mRNA splicing. *EMBO J*, **21**, 1148–1157.
Makino,D.L. *et al.* (2013) Crystal structure of an RNA-bound 11-subunit eukaryotic exosome complex. *Nature*, **495**, 70–75.
Malecki,M. *et al.* (2013) The exoribonuclease Dis3L2 defines a novel eukaryotic RNA degradation pathway. *EMBO J*, **32**, 1842–1854.
Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, **18**, 1509–1517.
Martin,R.M. *et al.* (2013) Live-cell visualization of pre-mRNA splicing with single-molecule sensitivity. *Cell Rep*, **4**, 1144–1155.
Martos-Moreno,G.Á. *et al.* (2018) Response to growth hormone in patients with RNPC3 mutations. *EMBO Mol. Med.*, e9143.
Matera,A.G. and Wang,Z. (2014) A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.*, **15**, 108.
McGuire,A.M. *et al.* (2008) Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol*, **9**, R50.
Merico,D. *et al.* (2015) Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. *Nat Commun*, **6**, 8718.
Merkin,J. *et al.* (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**, 1593–1599.
Mian,S.A. *et al.* (2013) Spliceosome mutations exhibit specific associations with epigenetic modifiers and proto-oncogenes mutated in myelodysplastic syndrome. *Haematologica*, **98**, 1058–1066.
Middleton,R. *et al.* (2017) IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol*, **18**, 51.
Moore,M.J. (1996) Gene expression. When the junk isn't junk. *Nature*, **379**, 402–403.
Moran,Y. *et al.* (2008) Intron retention as a posttranscriptional regulatory mechanism of neurotoxin expression at early life stages of the starlet anemone Nematostella vectensis. *J Mol Biol*, **380**, 437–443.
Mount,S.M. *et al.* (1983) The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. *Cell*, **33**, 509–518.
Ner Gaon,H. *et al.* (2004) Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant J.*, **39**, 877–885.
Niemelä,E.H. and Frilander,M.J. (2014) Regulation of gene expression through inefficient splicing of U12-type introns. *RNA Biol*, **11**, 1325–1329.
Nilsen,T.W. and Graveley,B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457–463.

Norppa,A.J. *et al.* (2018) Mutations in the U11/U12-65K protein associated with isolated growth hormone deficiency lead to structural destabilization and impaired binding of U12 snRNA. *RNA*, **24**, 396–409.

Padrón-Barthe,L. *et al.* (2018) Activation of Serine One-Carbon Metabolism by Calcineurin Aβ1 Reduces Myocardial Hypertrophy and Improves Ventricular Function. *J. Am. Coll. Cardiol.*, **71**, 654–667.

Palmer,J.D. and Logsdon,J.M. (1991) The recent origins of introns. *Curr Opin Genet Dev*, **1**, 470–477.

Pan,Q. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, **40**, 1413–1415.

Papasaikas,P. and Valcarcel,J. (2016) The Spliceosome: The Ultimate RNA Chaperone and Sculptor. *Trends Biochem Sci*, **41**, 33–45.

Patel,A.A. *et al.* (2002) The splicing of U12-type introns can be a rate-limiting step in gene expression. *EMBO J*, **21**, 3804–3815.

Patel,A.A. and Steitz,J.A. (2003) Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol*, **4**, 960–970.

Patthy,L. (1999) Genome evolution and the evolution of exon-shuffling–a review. *Gene*, **238**, 103–114.

Pessa,H.K. *et al.* (2006) The abundance of the spliceosomal snRNPs is not limiting the splicing of U12-type introns. *Rna*, **12**, 1883–1892.

Pimentel,H. *et al.* (2016) A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res*, **44**, 838–851.

Pomeranz Krummel,D.A. *et al.* (2009) Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature*, **458**, 475–480.

Popp,M.W. and Maquat,L.E. (2013) Organizing principles of mammalian nonsense-mediated mRNA decay. *Annu Rev Genet*, **47**, 139–165.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**, D61–65.

Query,C.C. *et al.* (1994) Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model. *Genes Dev*, **8**, 587–597.

Rapaport,F. *et al.* (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, 3158.

Reed,R. and Cheng,H. (2005) TREX, SR proteins and export of mRNA. *Curr. Opin. Cell Biol.*, **17**, 269–273.

Reyes,A. *et al.* (2013) Drift and conservation of differential exon usage across tissues in primate species. *Proc Natl Acad Sci USA*, **110**, 15377–15382.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, **43**, e47.

Robberson,B.L. *et al.* (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol Cell Biol*, **10**, 84–94.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, **11**, R25.

Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.

Rogozin,I.B. *et al.* (2012) Origin and evolution of spliceosomal introns. *Biol Direct*, **7**, 11.

Rogozin,I.B. *et al.* (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*, **13**, 1512–1517.

Rösel-Hillgärtner,T.D. *et al.* (2013) A novel intra-U1 snRNP cross-regulation mechanism: alternative splicing switch links U1C and U1-70K expression. *PLoS Genet.*, **9**, e1003856.

Rossi,P. and de Crombrugghe,B. (1987) Identification of a cell-specific transcriptional enhancer in the first intron of the mouse alpha 2 (type I) collagen gene. *Proc Natl Acad Sci USA*, **84**, 5590–5594.

Roy,S.W. and Gilbert,W. (2005) Complex early genes. *Proc Natl Acad Sci USA*, **102**, 1986–1991.

Roy,S.W. and Gilbert,W. (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet*, **7**, 211–221.

Ruskin,B. *et al.* (1988) A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. *Cell*, **52**, 207–219.

Russell,A.G. *et al.* (2006) An early evolutionary origin for the minor spliceosome. *Nature*, **443**, 863–866.

Saez,B. *et al.* (2017) Splicing factor gene mutations in hematologic malignancies. *Blood*, **129**, 1260–1269.

Sakabe,N.J. and De Souza,S.J. (2007) Sequence features responsible for intron retention in human. *BMC Genomics*, **8**, 59.

Sakharkar,M.K. *et al.* (2004) Distributions of exons and introns in the human genome. *Silico Biol Gedrukt*, **4**, 387–393.

Scamborova,P. *et al.* (2004) An intronic enhancer regulates splicing of the twintron of Drosophila melanogaster prospero pre-mRNA by two different spliceosomes. *Mol Cell Biol*, **24**, 1855–1869.

Schmid,M. and Jensen,T.H. (2010) Nuclear quality control of RNA polymerase II transcripts. *Wiley Interdiscip Rev RNA*, **1**, 474–485.

Schneider,C. *et al.* (2012) Transcriptome-wide analysis of exosome targets. *Mol Cell*, **48**, 422–433.

Schurch,N.J. *et al.* (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, **22**, 839–851.

Schwartz,S. *et al.* (2009) Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, **16**, 990–995.

Sharp,P.A. and Burge,C.B. (1997) Classification of introns: U2-type or U12-type. *Cell*, **91**, 875–879.

Shen,H. *et al.* (2010) The U2AF35-related protein Urp contacts the 3′ splice site to promote U12-type intron splicing and the second step of U2-type intron splicing. *Genes Dev.*, **24**, 2389–2394.

Shen,S. *et al.* (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci USA*, **111**, E5593–5601.

Sheth,N. *et al.* (2006) Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.*, **34**, 3955–3967.

Shukla,G.C. and Padgett,R.A. (2002) A catalytically active group II intron domain 5 can function in the U12-dependent spliceosome. *Mol Cell*, **9**, 1145–1150.

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

Smith,C.M. and Steitz,J.A. (1998) Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol Cell Biol*, **18**, 6897–6909.

Solomon,D.A. *et al.* (2003) Cyclin D1 splice variants. Differential effects on localization, RB phosphorylation, and cellular transformation. *J Biol Chem*, **278**, 30339–30347.
Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 91.
Sorek,R. *et al.* (2004) Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. *Mol Cell*, **14**, 221–231.
Spies,N. *et al.* (2009) Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell*, **36**, 245–254.
Staals,R.H. *et al.* (2010) Dis3-like 1: a novel exoribonuclease associated with the human exosome. *EMBO J*, **29**, 2358–2367.
Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences.
Stamm,S. *et al.* (1994) A sequence compilation and comparison of exons that are alternatively spliced in neurons. *Nucleic Acids Res.*, **22**, 1515–1526.
Steijger,T. *et al.* (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, **10**, 1177.
Stergachis,A.B. *et al.* (2013) Exonic transcription factor binding directs codon choice and affects protein evolution. *Science*, **342**, 1367–1372.
Stolc,V. *et al.* (2004) A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science*, **306**, 655–660.
Stormo,G.D. *et al.* (1982) Use of the 'Perceptron'algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.*, **10**, 2997–3011.
Su,Z. *et al.* (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.*, **32**, 903.
Tang,M. *et al.* (2015) Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. *BMC Bioinformatics*, **16**, 361.
Tarn,W.Y. and Steitz,J.A. (1996) A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell*, **84**, 801–811.
Tazi,J. *et al.* (2009) Alternative splicing and disease. *Biochim Biophys Acta*, **1792**, 14–26.
Thorvaldsdóttir,H. *et al.* (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
Tomecki,R. *et al.* (2010) The human core exosome interacts with differentially localized processive RNases: hDIS3 and hDIS3L. *EMBO J*, **29**, 2342–2357.
Trachtulec,Z. and Forejt,J. (1999) Transcription and RNA processing of mammalian genes in Saccharomyces cerevisiae. *Nucleic Acids Res*, **27**, 526–531.
Trapnell,C. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, **7**, 562–578.
Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511–515.
Trincado,J.L. *et al.* (2018) SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.*, **19**, 40.
Tsai,R.T. *et al.* (2005) Spliceosome disassembly catalyzed by Prp43 and its associated components Ntr1 and Ntr2. *Genes Dev*, **19**, 2991–3003.
Turunen,J.J., Verma,B., *et al.* (2013) HnRNPH1/H2, U1 snRNP, and U11 snRNP cooperate to regulate the stability of the U11-48K pre-mRNA. *RNA*, **19**, 380–389.
Turunen,J.J., Niemelä,E.H., *et al.* (2013) The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA*, **4**, 61–76.

Turunen,J.J. *et al.* (2008) The U11-48K protein contacts the 5′ splice site of U12-type introns and the U11-59K protein. *Mol. Cell. Biol.*, **28**, 3548–3560.
Tycowski,K.T. *et al.* (1996) A mammalian gene with introns instead of exons generating stable RNA products. *Nature*, **379**, 464–466.
Valadkhan,S. and Manley,J.L. (2001) Splicing-related catalysis by protein-free snRNAs. *Nature*, **413**, 701–707.
Valcarcel,J. *et al.* (1996) Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected]. *Science*, **273**, 1706–1709.
Valencia,P. *et al.* (2008) Splicing promotes rapid and efficient mRNA export in mammalian cells. *Proc Natl Acad Sci USA*, **105**, 3386–3391.
van Hoof,A. *et al.* (2000) Yeast exosome mutants accumulate 3'-extended polyadenylated forms of U4 small nuclear RNA and small nucleolar RNAs. *Mol Cell Biol*, **20**, 441–452.
Vanichkina,D.P. *et al.* (2018) Challenges in defining the role of intron retention in normal biology and disease. *Semin Cell Dev Biol*, **75**, 40–49.
Vardhanabhuti,S. *et al.* (2013) A Hierarchical Bayesian Model for Estimating and Inferring Differential Isoform Expression for Multi-Sample RNA-Seq Data. *Stat Biosci*, **5**, 119–137.
Verbeeren,J. *et al.* (2017) Alternative exon definition events control the choice between nuclear retention and cytoplasmic export of U11/U12-65K mRNA. *PLoS Genet.*, **13**, e1006824.
Verbeeren,J. *et al.* (2010) An ancient mechanism for splicing control: U11 snRNP as an activator of alternative splicing. *Mol Cell*, **37**, 821–833.
Verma,B. *et al.* (2018) Minor spliceosome and disease. *Semin Cell Dev Biol*, **79**, 103–112.
Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
Wang,Z. and Burge,C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna*, **14**, 802–813.
Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276.
Will,C.L. *et al.* (1999) Identification of both shared and distinct proteins in the major and minor spliceosomes. *Science*, **284**, 2003–2005.
Will,C.L. *et al.* (2004) The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. *RNA*, **10**, 929–941.
Will,C.L. and Lührmann,R. (2011) Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.*, **3**, a003707.
Wollerton,M.C. *et al.* (2004) Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol Cell*, **13**, 91–100.
Wu,J. and Manley,J.L. (1989) Mammalian pre-mRNA branch site selection by U2 snRNP involves base pairing. *Genes Dev*, **3**, 1553–1561.
Wu,Q. and Krainer,A.R. (1997) Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs. *RNA*, **3**, 586–601.
Yan,C. *et al.* (2015) Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science*, **349**, 1182–1191.
Yeo,G. *et al.* (2004) Variation in alternative splicing across human tissues. *Genome Biol*, **5**, R74.
Younis,I. *et al.* (2013) Minor introns are embedded molecular switches regulated by highly unstable U6atac snRNA. *Elife*, **2**, e00780.
Zamore,P.D. and Green,M.R. (1989) Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. *Proc Natl Acad Sci USA*, **86**, 9243–9247.

Zerbino,D.R. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res*, **46**, D754–D761.
Zhang,J. *et al.* (1998) Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. *RNA*, **4**, 801–815.
Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, **9**, R137.
Zhou,S. *et al.* (1995) Male-specific lethal 2, a dosage compensation gene of Drosophila, undergoes sex-specific regulation and encodes a protein with a RING finger and a metallothionein-like cysteine cluster. *EMBO J*, **14**, 2884–2895.
Zhu,W. and Brendel,V. (2003) Identification, characterization and molecular phylogeny of U12-dependent introns in the Arabidopsis thaliana genome. *Nucleic Acids Res*, **31**, 4561–4572.
Zhuang,Y. and Weiner,A.M. (1986) A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell*, **46**, 827–835.
Zimmerly,S. and Semper,C. (2015) Evolution of group II introns. *Mob DNA*, **6**, 7.
Zorio,D.A. and Blumenthal,T. (1999) Both subunits of U2AF recognize the 3' splice site in Caenorhabditis elegans. *Nature*, **402**, 835–838.